

IESM 2017

7th International Conference on
Industrial Engineering and Systems Management

Saarbrücken, Germany, October 11 – 13, 2017

CONFERENCE PROCEEDINGS

Editors

Thomas Bousonville*

Teresa Melo*

Nidhal Rezg**

François Vernadat**

*Saarland University of Applied Sciences, Saarbrücken, Germany

**University of Lorraine, Metz, France

wirtschafts
wissenschaften
htw saar

UNIVERSITÉ
DE LORRAINE

IESM 2017

INDUSTRY AND LOGISTICS 4.0:

TRENDS, BENEFITS AND CHALLENGES FOR INDUSTRIAL GROWTH

www.iesm17.org

Scientific Partners and Sponsors

Preface

Dear Colleagues and Friends,

We are pleased to welcome you to IESM 2017 – the 7th edition of the International Conference on Industrial Engineering and Systems Management. This year's conference edition is a Franco-German co-organization by Saarland University of Applied Sciences (htw saar) and University of Lorraine, extending and deepening the long-term cooperation between the two institutions. It is hosted by htw saar and held at the campus of its Business School.

The challenges industrial engineers are faced in the age unfolding digitalization are embraced by the theme of the 2017 edition of IESM: ***Industry and Logistics 4.0: Trends, Benefits and Challenges for Industrial Growth***. Indeed, by 2020, over 20 billion devices will be communicating, representing an economic added value of $1.9 trillion, according to Gartner. The next industrial revolution, known as Industry 4.0 or Industry of the Future, is underway. Its main objective is to increase global competitiveness through individualization of products and production while maintaining low production costs. The establishment of "smart factories" and "smart services" will offer greater flexibility in operations and more efficient allocation of resources, paving the way for agile manufacturing and scalable logistics solutions. Networks will become increasingly integrated allowing real-time information exchange. Thus, the complexity of systems is becoming more and more important and requires more sophisticated market offers. Industry 4.0 can be both a threat and an opportunity for industries. In fact, traditional businesses are faced with new challenges that will change the rules of productivity and competition. Interdisciplinary thinking is essential, both in the technical and social fields, with a continuous process of learning and collaboration, and a significant impact on professional roles.

Over the course of three days, the conference provides ample opportunities to present and discuss the latest developments from theoretical results and algorithmic progress to application-oriented cases. The conference includes three keynote presentations from academia and industry, illustrating the imminent relevance of the subject for both parties. Overall, 110 peer-reviewed contributions are presented at the conference across 30 sessions. Covering the broad spectrum of the conference theme, they are documented in the present proceedings. In total, around 150 participants from 25 countries attend the conference.

We would like to thank all the people who helped to make IESM 2017 an interesting, inspiring and fruitful event: the members of the scientific, program and organizing committees, the reviewers, session chairs, local organizers, keynote speakers and last but not least, all the participants who presented their ideas and results. Special thanks are also due to our scientific partners and sponsors: fitt gGmbH (Germany), GDR RO 3002 - CNRS (France), I4e2 (Belgium), lgipm (France), ORBIS AG (Germany), State Chancellery of Saarland (Germany), and University of Valenciennes and Hainaut-Cambresis (France).

Saarbrücken and Metz, October 2017

Thomas Bousonville, Teresa Melo,
Nidhal Rezg, François Vernadat

CONFERENCE HISTORY

IESM is a well-established international conference in Industrial Engineering. It has been held every two years since its creation in 2005 when it replaced and extended the scope of the former IEPM (Industrial Engineering and Production Management) conference launched in 1993.

2005	Marrakech, Morocco
2007	Beijing, China
2009	Montréal, Canada
2011	Metz, France
2013	Rabat, Morocco
2015	Seville, Spain

CONFERENCE VENUE



The "Hochschule für Technik und Wirtschaft des Saarlandes" (**htw saar**), is a multi-campus university in Saarbrücken, the capital of the federal state of Saarland, Germany.

Currently, htw saar has some 6000 students and a staff of over 300. The university offers more than 40 fulltime degree programs (Bachelor and Master programs), all accredited by national or international accreditation agencies, as well as a growing number of part-time postgraduate professional education programs in the four schools: Business, Architecture & Civil Engineering, Engineering and Social Sciences.

htw saar also can take credit for being a pioneer in shaping fully integrated German-French degree programs in conjunction with the University of Lorraine, France. In 1978, long before ERASMUS and Bologna, the two universities founded the unique German-French Higher education institution, called ISFATES in French and DFHI in German, which today offers twelve German-French double-degree-programs.

Studies at htw saar are practical-oriented with an excellent combination of hands-on skills along with a strong theoretical knowledge base. Research is fundamental and great importance is attached to applied research and development. Research is conducted with industry partners in Germany and abroad and htw saar belongs to the top ranked Universities of Applied Sciences in Germany for R&D.

IESM 2017 takes place at Campus Rotenbühl, the home of the **Business School** of htw saar. The Business School offers degree programs in various business study areas and industrial engineering.

COMMITTEES

General Chairs

Thomas Bousonville, Saarland University of Applied Sciences, Germany
Teresa Melo, Saarland University of Applied Sciences, Germany
Nidhal Rezg, University of Lorraine, France

Advisory Board

François Vernadat, University of Lorraine, France (Chair)
Abdelhakim Artiba, University of Valenciennes, France
Alexandre Dolgui, IMT Atlantique, France
Yannick Frein, Chairman of GDR-MACS, France
Andrew Kusiak, The University of Iowa, USA
Samir Lamouri, Arts et métiers ParisTech, France
Stephen Newman, University of Bath, UK
Henri Pierreval, SIGMA, University of Clermont-Ferrand, France
Alain Quilliot, Chairman of GDR-RO, France
Katsuhiko Takahashi, Hiroshima University, Japan
MengChu Zhou, New Jersey Institute of Technology, USA

International Scientific Committee

Alexandre Dolgui, IMT Atlantique, France (President)
Le Thi Hoai An, University of Lorraine, France (Vice-President)

Members

Aghezzaf El H. (Belgium),
Ait-Kadi D. (Canada),
Allahverdi A. (Kuwait),
Allaouri H. (France),
Araújo M. (Portugal),
Artiba A. (France),
Bakhtadze N. (Russia),
Baptiste P. (Canada),
Baumgärtel H. (Germany),
Ben-Daya M. (Saudi Arabia),
Benyoucef L. (France),
Bidanda B. (USA),
Billaut J.-C. (France),
Borne P. (France),
Bousonville, T. (Germany),
Bourrières J.-P. (France),
Camarinha-Matos L. (Portugal),
Carlier J. (France),
Castagliola P. (France),
Chan F. (Hong Kong),
Cao L. (Australia),
Cao T. (Vietnam),
Che A. (China),
Chelbi A. (Tunisia),
Childe S. (UK),
Chu C. (France),
Chu F. (France),
Comesaña J. (Spain),
Cormier G. (Canada),
Corominas A. (Spain),
Crainic T. G. (Canada),
Dahane M. (France),
Dauzère-Péres S. (France),
Dehombreux P. (Belgium),
Disney S. (UK),
Drexl M. (Germany),
Duvivier D. (France),
Eremeev A. (Russia),
Framinan J. (Spain),
Frein Y. (France),
Gao L. (China),
Gaudreault J. (Canada),
Gharbi A. (Canada),
Glock C. (Germany),
Goncalvez G. (France),

Grabot B. (France),
Grosse E. (Germany),
Grubbström R. W. (Sweden),
Guinet A. (France),
Günther H.-O. (Germany),
Guschinsky N. (Belarus),
He D. D. (USA),
Hnaien F. (France),
Ho T. B. (Japan),
Hong T.-P (Taiwan),
Kopfer H. (Germany),
Korne, T. (Germany),
Kotov V. (Belarus),
Kovalyov M. (Belarus),
Kusiak A. (USA),
Lamouri S. (France),
Laporte G. (Canada),
Lazarev A. (Russia),
Leisten R. (Germany),
Levin G. (Belarus),
MacCarthy B. (UK),
Macchi M. (Italy),
Melo, T. (Germany),
Meskens N. (Belgium),
Minner S. (Germany),
Moench L. (Germany),
Moukrim A. (France),
Mula J. (Spain),
Nagano M. (Brazil),
Nagi R. (USA),
Najid N. (France),
Nassehi A. (UK),
Nickel S. (Germany),
Nielsen P. (Denmark),
Nourelfath M. (Canada),
Ortiz Baz A. (Spain),
Ouali M.-A. (Canada),
Oztemel E. (Turkey),
Pellerin R. (Canada),
Penz B. (France),
Perez-Gonzalez P. (Spain),
Pierreval H. (France),
Poler R. (Spain),
Póvoa A. (Portugal),
Quilliot A. (France),
Ramudhin A. (UK),
Rezg N. (France),
Riane F. (Morocco),
Roux O. H. (France),
Sauer N. (France),
Ivanov, D. (Germany),
Jaber M. Y. (Canada),
Jamali A. (Morocco),
Jiang Z. (China),
Khatab A. (France),
Sauvey C. (France),
Schutz J. (France),
Selamat A. (Malaysia),
Selle, S. (Germany),
Siarry P. (France),
Talbi E.-G. (France),
Tang L. (China),
Thomas A. (France),
Trentesaux D. (France),
Vernadat F. (France),
Voss S. (Germany),
Wang N. (China),
Wang S. (China),
Werner F. (Germany),
Wu D. (China),
Xie X. (France),
Yalaoui F. (France),
Yang S. (China),
Yu Y. (China).

Student Competition Committee

Anis Chelbi, University of Tunis, Tunisia (Chair)
Thomas Bousonville, Saarland University of Applied Sciences, Germany
François Vernadat, University of Lorraine, France

Organizing Committee

Thomas Bousonville, Saarland University of Applied Sciences, Germany
Sophie Hennequin, University of Lorraine, France
Teresa Melo, Saarland University of Applied Sciences, Germany
Christophe Sauvey, University of Lorraine, France
Jérémie Schutz, University of Lorraine, France
Anne Steinhaus, Saarland University of Applied Sciences, Germany

PROGRAM OVERVIEW

TUESDAY, OCTOBER 10					
17:30 – 21:00	Registration and Get Together				
WEDNESDAY, OCTOBER 11					
08:00 – 09:00	Registration				
09:00 – 09:30	Opening Session (room A-E-08)				
09:30 – 10:30	*Keynote 1:* Dimitris Kiritsis, École Polytechnique Fédérale de Lausanne, Switzerland **"Trends on Enabling IT Technologies for Industrial Information Systems and Applications"**				
10:30 – 11:00	Coffee Break				
11:00 – 13:00	*Parallel Sessions*				
	WD-M1 **Assembly Line Planning** (room A-E-03)	WD-M2 **Supplier Selection & Evaluation** (room A-1-04)	WD-M3 **Machine Scheduling I** (room A-1-05)	WD-M4 **Manufacturing System Design** (room A-1-15)	WD-M5 **Supply Chain Coordination & Risk Management** (room A-1-17)
13:00 – 14:00	Lunch				
14:00 – 16:00	*Parallel Sessions*				
	WD-A1 **Vehicle Routing I** (room A-E-03)	WD-A2 **Maintenance Planning I** (room A-1-04)	WD-A3 **Machine Scheduling II** (room A-1-05)	WD-A4 **Supply Chain Network Planning** (room A-1-15)	
16:00 – 16:30	Coffee Break				
16:30 – 18:00	*Parallel Sessions*				
	WD-L1 **Project Management I** (room A-E-03)	WD-L2 **Maintenance Planning II** (room A-1-04)			
18:00 – 18:30	55^{th} Anniversary, International Journal of Production Research				(room A-E-08)
19:00 – 20:30	Guided Tour of Saarbrücken (meeting point: Town Hall, Rathausplatz 1)				
THURSDAY, OCTOBER 12					
08:30 – 10:30	*Keynote 2:* Thies Uwe Trapp, Robert Bosch GmbH, Germany (room A-E-08) **"Industry 4.0: From Theory to Practice - The Bosch Experience"**				
	Keynote 3: Thierry Collet, Dassault Systèmes, France **"Preparing Academia for the Industry of the Future"**				
10:30 – 11:00	Coffee Break				
11:00 – 13:00	*Parallel Sessions*				
	TH-M1 **Vehicle Routing II** (room A-E-03)	TH-M2 **ICT for Manufacturing & Services I** (room A-1-04)	TH-M3 **Machine Scheduling II** (room A-1-05)	TH-M4 **Automatic Storage & Retrieval Systems** (room A-1-15)	TH-M5 **Production & Design Methods** (room A-1-17)
13:00 – 14:00	Lunch				
14:00 – 16:00	*Parallel Sessions*				
	TH-A1 **Inventory Management** (room A-E-03)	TH-A2 **ICT for Manufacturing & Services II** (room A-1-04)	TH-A3 **Machine Scheduling IV** (room A-1-05)	TH-A4 **Lean Supply Chains** (room A-1-15)	TH-A5 **Supply Chain Management** (room A-1-17)
16:00 – 16:30	Coffee Break				
16:30 – 18:00	*Parallel Sessions*				
	TH-L1 **Project Management II** (room A-E-03)	TH-L2 **Maintenance Planning III** (room A-1-04)	TH-L3 **Forecasting I** (room A-1-05)		
19:00 – 00:00	Gala Dinner - Congress Hall Saarbrücken (Hafenstraße 12)				
FRIDAY, OCTOBER 13					
08:30 – 10:30	*Parallel Sessions*				
	FR-M1 **Production Planning** (room A-E-03)	FR-M2 **ICT for Manufacturing & Services III** (room A-1-04)	FR-M3 **Planning and Scheduling in Air & Maritime Transportation** (room A-1-05)	FR-M4 **Operating Room Scheduling & Health Care Services** (room A-1-15)	
10:30 – 11:00	Coffee Break				
11:00 – 12:30	*Parallel Sessions*				
	FR-L1 **Forecasting II** (room A-E-03)	FR-L2 **ICT for Manufacturing & Services IV** (room A-1-04)			
12:40 – 13:00	Closing Session (room A-E-08)				
13:00 – 14:30	Light Lunch				

KEYNOTE SPEAKERS

Dimitris Kiritsis, École Polytechnique Fédérale de Lausanne, Switzerland

Trends on enabling IT technologies for industrial information systems and applications



Prof. Dr. Dimitris Kiritsis is Faculty Member at the Institute of Mechanical Engineering of the School of Engineering of the École Polytechnique Fédérale de Lausanne, Switzerland, where he is leading a research group on ICT for Sustainable Manufacturing.

He served also as Guest Professor at the IMS Center of the University of Cincinnati, and Invited Professor at the University of Technology of Compiègne, the University of Technology of Belfort-Montbéliard and at ParisTech ENSAM Paris.

Prof. Kiritsis is actively involved in EU research programs in the area of Factories of the Future and Enabling ICT for Sustainable Manufacturing. He has more than 190 publications.

Since September 2013 Prof. Kiritsis is Chair of the Working Group 5.7 Advanced Production Management Systems of the International Federation for Information Processing (IFIP). He is also member of the Advisory Group of the European Council on Leadership on Enabling Industrial Technologies – AG LEIT-NMBP. Furthermore, Prof. Kiritsis is founding fellow member of the International Society for Engineering Asset Management (ISEAM) and of various international scientific communities in his area of interests including EFFRA (European Factories of the Future Research Association).

Thies Uwe Trapp, Robert Bosch GmbH, Germany

Industry 4.0: From Theory to Practice - The Bosch Experience



After completing his studies and receiving his PhD from the Technical University Clausthal, Germany, Dr.-Ing. Thies Uwe Trapp began his professional career at Robert Bosch GmbH, where he has held several positions in process engineering and manufacturing at the division Automotive Electronics in Salzgitter, Germany, and in Suzhou, China. Dr.-Ing. Trapp is currently Director for Connected Industry at Boschs diesel plant in Homburg (Saar), Germany. Prior to this role, he was responsible for process engineering and tool technology at the same plant.

Thierry Collet (Dassault Systèmes, France)

Preparing Academia for the Industry of the Future



Thierry Collet graduated from the ESSEC Business School and has occupied many senior management positions in prestigious IT companies. He started his career as a Consultant then became a sales representative and finally a Sales Manager in various IT companies.

In 1988 he joined ORACLE where he served during 19 years various senior sales management positions in France, was the EMEA Field Marketing Support Senior Director based in California and lastly ran the Western Europe & Middle East Business Operations for 8 years.

Thierry Collet has joined Dassault Systèmes in 2007 as the VP Global Marketing, encompassing all marketing responsibilities for all Dassault Systèmes Brands until September 2011. For almost the last 7 years, he has been the worldwide VP for Global Academia and Education. As such, he is in charge to define, provide, market and implement end-to-end Lifelong Learning solutions based upon the whole range of Dassault Systèmes products helping individuals, customers and academic users to learn using innovative methods and obtain the highest possible employability in their respective domains.

Thierry Collet who is an avid reader and a music critic is also a regular lecturer on Social Media in B2B and a demanded speaker at numerous international conferences.

Table of Contents

Assembly Line Planning

Bounds and bound sets for the bi-objective simple assembly line balancing problem 2
Audrey Cerqueus and Xavier Delorme

Modeling kitting operations in a semi-robotic environment 8
Mohamed El Amine Boudella, Evren Sahin and Yves Dallery

Impact of smoothening and synchronization of components flow on staging delay in assembly systems 14
Shankar Saravana Kumar, Chandrasekharan Rajendran and Rainer Leisten

A MILP approach for robust transfer line design 20
Aleksandr Pirogov, Evgeny Gurevsky, André Rossi and Alexandre Dolgui

Automatic Storage and Retrieval Systems

Design of an automated storage space for reticles in a semiconductor manufacturing facility: A simulation-based approach 26
Ali Ben Salem, Claude Yugma, Emmanuel Troncet and Jacques Pinaton

Analytical models for a joint posture and fatigue analysis in order picking 32
Martina Calzavara, Fabio Sgarbossa, Eric Grosse and Christoph Glock

A three-step heuristic for bi-directional flow-rack AS/RS with duration-of-stay storage policy 38
Zhuxi Chen and Jatinder Gupta

Throughput analysis of random storage systems: Closest eligible location pays 44
Anja Heßler and Christoph Schwindt

Forecasting

Shelf space allocation during promotion periods 51
Mahmoud El-Banna and Ahmed Alabadleh

Obsolescence forecasting strategy of technological components - A comparative study of algorithms 57
Yosra Grichi, Yvan Beauregard, Thien-My Dao and Imen Zaabar

Hybrid forecasting model for sales promotion: Application in the modern retail supply chain 63
Slim Harbi, Mohamed Bahroun and Rihab Mechmech

Components' demand analysis in assemble-to-order systems: An analysis method 69
Mohammed Hichame Benbitour, Evren Sahin and Yves Dallery

The impact of macroeconomic leading indicators for tactical sales forecasting on SKU inventory management 75
Yves R. Sagaert, Stijn De Vuyst, Nikolaos Kourentzes and El-Houssaine Aghezzaf

Information and Communication Technologies for Manufacturing and Services

Smart city projects and improvement of public service supply chains 81
Gautier Aubourg, Jacques Lamothe, François Galasso and Bernard Grabot

Concept and possible application of an automated framework to influence production dispatch based on supply chain events 87
Dennis Bauer, Florian Maier, Thomas Bauernhansl, Bernd Waschneck, Thomas Ponsignon, Daniel Gürster, Bernhard Oberegger, Andreas Felsberger and Gerald Reiner

Some elements regarding the relevance of using a fuzzy performance measurement framework for industrial visual management 93
Lamia Berrah, Laurent Foulloy and Vincent Clivillé

Molding 4.0 - The economics of an injection molding as-a-service business model 99
Alessandro Charalambis, Marco Tonetti and Guido Tosello

Economic analysis of additive manufacturing integration in injection molding process chain ..105
Alessandro Charalambis, Laoucine Kerbache, Guido Tosello, David Bue Pedersen and Michael Mischkot

3D printing technology: Are we ready for the next step? 111
Prodromos Chatzoglou and Vasiliki Michailidou

The software and hardware complex for automatic feed control of steam generators: Modeling and optimization 117
Liudmila Denisova and Vitalii Meshcheryakov

A user-centric process management for system and software engineering projects 123
Mojtaba Hajmoosaei, Hanh Nhi Tran and Christian Percebois

Developement of an industrie 4.0 maturity index for small and medium sized companies 129
Sebastian Häberer, Lina Lau and Fabian Behrendt

Survey of the renovation sector in France: A necessary step towards implementing BIM in microenterprises 135
Laurent Joblot, Dominique Deneux, Thomas Paviot and Samir Lamouri

Industry 4.0 and the SME: A technology-focused review of the empirical literature 141
Alexandre Moeuf, Samir Lamouri, Robert Pellerin, Romain Eburdy and Simon Tamayo

Quantification of Industry 4.0 manufacturing complexity based on information theory 150
Dimitris Mourtzis, Ekaterini Vlachou and Sophia Fotia

IoT for All: Architectural design of an extensible and lightweight IoT analytics platform 156
Stephan Müller, Patrick Wiener, Adrian Bürger and Jens Nimis

Improving visibility in the supply chain with digital technologies 162
Karine Samuel

Inventory Management

Analysis of end-of-quarter inventory in hybrid production environments using simulation ... 169
Lawrence Al-Fandi and Faisal Aqlan

EOQ inventory model for disassembly systems 174
Matthieu Godichaud and Lionel Amodeo

Blood products inventory routing problem with transshipment and substitution 180
Imane Hssini, Nadine Meskens and Fouad Riane

Inventory deployment with uncertainty on production and lead-times 186
Jean Respen and Nicolas Zufferey

Lean Supply Chains

A new metric to evaluate the supply chain agility 192
Selmen Boubaker, Ziad Jemaî, Evren Sahin and Yves Dallery

Sustaining lean through lean experts' roles 198
Florian Magnani, Ali Siadat, Emmanuel Caillaud and Olivier Gaudichau

A leverage potential of chilled distribution center operations 204
Jirapat Wanitwattanakosol, Saran Jangnilwong and Sirikorn Santirojanakul

Machine Scheduling

Scheduling job shops with operators in a free assignment-changing mode 210
Imène Benkalai, Pierre Baptiste and Djamal Rebaine

Existing results and new developments for the single processor scheduling problem with time restrictions 216
Oliver Braun, Rachid Benmansour and Said Hanafi

Unrelated parallel machines with sequence dependent setup times and machine eligibility 220
Fernando Carrasco, Paz Perez-Gonzalez, Jose M. Framinan and Victor Fernandez-Viagas

Resolution of a job-shop problem with precedence constraints between jobs (JSPC) 226
Matthieu Gondran, Philippe Lacomme, Enzo Mura and Nikolay Tchernev

Resolution of a job-shop with routing considering a quality of service for customer: JSSP with routing 232
Matthieu Gondran, Philippe Lacomme and Nikolay Tchernev

Multitasking scheduling with energy cost 238
Ming Liu, Junkai He, Feifeng Zheng and Chengbin Chu

Bi-objective optimization of a reentrant flow shop scheduling with exact time lag considering energy cost 244
Ming Liu, Xin Liu, Feifeng Zheng and Feng Chu

Scheduling reentrant jobs on parallel machines with a remote server to minimize makespan, revisited 250
Ming Liu, Yang Sui, Feifeng Zheng and Chengbin Chu

A prelude to anarchic manufacturing systems 256
Aydin Nassehi and Andrew Ma

Heuristics for hybrid flowshop problem solving with non-identical parallel machines 262
Adriana Maria Pacheco Osorio, Christophe Sauvey, Nathalie Sauer and Trabelsi Wajdi

Exact solution of the single machine total tardiness problem: The power of memorization ... 268
Lei Shang, T'Kindt Vincent and Federico Della Croce

New objective functions based on jobs positions for single machine scheduling with deadlines 273
Thanh Thuy Tien Ta, Kivin Ringard and Jean-Charles Billaut

Robust parallel machine scheduling with job splitting under uncertain processing times 279
Naji Widad, Marie-Laure Espinouse and Van-Dat Cung

Analysis of integrated scheduling and intermediate transportation systems in production lines with variability-related measures 285
Miriam Zacharias, Rainer Leisten and Chandrasekharan Rajendran

A batching and scheduling optimisation for a cutting work-center: Acta-Mobilier case study . 291
Emmanuel Zimmermann, Hind Bril El Haouzi, Philippe Thomas, André Thomas and Mélanie Noyel

Machining elements sequence optimization of thin-walled and multi-cavity aircraft structural parts based on knowledge-based rules and genetic algorithm 297
Dan Zhang and Tao Wei

Maintenance Planning

Optimal joint production, maintenance and quality control for a single machine with random failures using Matlab/Simulink optimization-simulation 304
Abdessamad Ait El Cadi, Ali Gharbi and Abdelhakim Artiba

Opportunity study for the adoption of an extended warranty 311
Zied Boubaker, Soumaya Bouguerra, Anis Chelbi and Nidhal Rezg

A joint selective maintenance and multiple repair-person assignment problem 317
Claver Diallo, Abdelhakim Khatab, Uday Venkatadri and El-Houssaine Aghezzaf

A study of variability induced by events dependency in microelectronic production 323
Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse and Philippe Vialletelle

Prioritization of failure modes of fire and rescue vessels with application case study 329
Mohammad Haidar, El Awady Attia and Ahmed El-Assal

Prediction of complex failures' effects in offshore wind turbines using fuzzy cognitive maps 335
Afshin Jamshidi, Daoud Aït-Kadi and Amar Ramudhin

Optimizing the integrated economic production quantity for a stochastically deteriorating production system under condition-based maintenance 340
Abdelhakim Khatab, Claver Diallo, El-Houssaine Aghezzaf and Uday Venkatadri

Integrated maintenance/production planning plan for a multiple-product manufacturing system with variable maintenance and changeover costs considerations 346
Lahcen Mifdal, Zied Hajej, Sofiene Dellagi and Ilias Majdouline

Integrated preventive maintenance and inspection planning in a deteriorating serial multi-stage manufacturing system 352
Mohammad Rezaei-Malek, Reza Tavakkoli-Moghaddam, Ali Siadat and Jean-Yves Dantan

Manufacturing System Design

Dual strategy for the evaluation of the structures of series production 359
Matthias Heinicke and Ulf Bergmann

Toward a generic method to design reconfigurable manufacturing systems 364
Khaled Lameche, Najib Najid, Pierre Castagna and Khalid Kouiss

Modelling flexible manufacturing systems with power constraints and machine switch on/off . 370
Damien Lamy, Sylverin Kemmoé T., Nikolay Tchernev and Pierre Fenies

In search of an optimal tooling capacity for the vulcanising planning decision problem 376
Federico Trigos and Eduardo M. López

Operating Room Scheduling and Health Care Services

A heuristic for an operating room scheduling problem with durations' uncertainty 382
Mohamed Amine Abdeljaouad, Zied Bahroun, Nour El Houda Saadani and Belhassen Zouari

A mathematical model and a heuristic for an operating room scheduling problem with setup times 388
Mohamed Amine Abdeljaouad, Nour El Houda Saadani, Zied Bahroun and Belhassen Zouari

Planning surgical activities considering post-surgical capacity 394
Valérie Dorval, Abdessamad Ait El Cadi, Daoud Aït-Kadi and Abdelhakim Artiba

A ConWIP approach to manage an integrated surgical suite: the case of the ophthalmology surgical suite in a Tunisian hospital 399
Safa Chabouh, Sondes Hammami, Hanen Bouchriha and Eric Marcon

Multidimensional assignment of services to people with disabilities 405
Anderson Fabian Mosquera Varela, Lorraine Trilling and Thibaud Monteiro

Planning and Scheduling in Air and Maritime Transportation

Aircraft maintenance capacity planning: A decision support framework 412
Duarte Dinis and Ana Barbosa-Póvoa

A two-stage stochastic programming for rescheduling on a single yard crane with stochastic release time of retrieval tasks 418
Ming Liu, Xiaoyi Man, Feifeng Zheng and Chengbin Chu

Fuel-oriented aircraft optimization for ground operations and transfer performance 424
Ming Liu, Qinrui Song, Feifeng Zheng and Chengbin Chu

Arrival and departure joint scheduling in a multi-airport system with capacity restriction based on efficiency and penalty 430
Ming Liu, Zhihui Sun, Feifeng Zheng and Chengbin Chu

Production and Design Methods

An industrial perspective on workload control: Economic evaluation and suitable technologies 437
Massimo Bertolini, Giovanni Romagnoli, Dieter Uckelmann and Francesco Zammori

Methodological problems in defining costs used in industrial management decision models . . . 443
Vincent Giard

Using dominance relation to identify relevant generalized technical contradictions in innovative design 449
Ivana Rasovska, Roland De Guio and Sébastien Dubois

Industrial symbiotic relations as cooperative games 455
Vahid Yazdanpanah and Devrim Murat Yazan

Production Planning

Supply planning optimization for linear production system with stochastic lead-times and quality control 462
Oussama Ben Ammar, Belgacem Bettayeb and Alexandre Dolgui

Optimization of production and pricing strategy for a profit-maximization model 468
Yan-Chun Chen

Impact of lead-time sensitive cost on lead-time quotation and pricing problems 474
Abduh Sayid Albana, Ramzi Hammami and Yannick Frein

A MILP for strategic planning of production-inventory network manufacturing bulk materials 480
Nikolay Tchernev, Sylverin Kemmoé T. and Sabah Belil

Project Management

Development of professional competences for Industry 4.0 project management 487
Alberto Cerezo-Narvaez, Manuel Otero-Mateo and Andres Pastor-Fernandez

Coping with new business challenges: Analysis of the elements that affect the sustainability of continuous improvement programs 493
Orlando Chirinos, Georges Habchi, Pralus Magali and Zahir Messaoudene

Measuring the success of collaborative university-industry R&D funded contracts 499
Gabriela Fernandes, Eduardo B. Pinto, Madalena Araújo, Pedro Magalhães and Ricardo Machado

Improving project management practice: A car industry case study 505
Daniela Gomes, Gabriela Fernandes and Madalena Araújo

A performance evaluation methodology for decision support in industrial projects 511
Fan Li, Alain Etienne, François Vernadat and Ali Siadat

A review of software capabilities for project planning and control 517
Robert Pellerin and Nathalie Perrier

Supplier Selection and Evaluation

A supplier selection model development with Six Sigma quality indices for products with multiple quality characteristics 523
Kuen-Suan Chen, Ching-Hsin Wang and Kim-Hua Tan

Sustainable supplier monitoring problem: A fuzzy logic approach 529
Abdelhak Elidrissi and Sophie Hennequin

Evaluating green supplier development programs by a new decision interval-valued fuzzy model-based possibilistic statistical concept 535
Nazanin Foroozesh, Reza Tavakkoli-Moghaddam and S. Meysam Mousavi

An exact approach for solving the supplier selection problem in multi-service outsourcing 541
Hoai Pham Thi, Hoai An Le Thi and Nam Nguyen Canh

Supply Chain Coordination and Risk Management

A fuzzy approach for supply chain vulnerabilities assessement: Application to an agrifood supply chain . . . 548
Hatem Elleuch, El Mouloudi Dafaoui, Abderrahman El Mhamedi and Habib Chabchoub

Identifying and ranking of clearance chain management (CCM) strategies . . . 554
Slim Harbi, Rihab Mechmech and Atidel Hadj-Alouane

Investigating the interrelationships of risk factors affecting outsourcing logistics: An analysis with DEMATEL method in the pharmaceutical supply chain context . . . 560
Nadine Kafa, El Mouloudi Dafaoui, Asmae El Mokrini, Abderrahman El Mhamedi and Abdelaziz Berrado

Coordinating contracts for a capital-constrained supply chain with retailer effort . . . 566
Dinh Anh Phan, Thi Le Hoa Vo and Anh Ngoc Lai

Supply Chain Management

MPC approach for synchronized supply chains of perishable goods . . . 574
Tomás Hipólito, João Nabais, Rafael Carmona-Benítez, Miguel Ayala Botto and Rudy R. Negenborn

The selling mode in a supply chain with online retailing . . . 580
Wenchang Liu and Danqin Yang

A quality model for a multi-layered sourcing supply chain . . . 586
Jrjung Lyu and Chia-Wen Chen

Integrated maintenance and distribution plan for an Urban Distribution Centre (UDC) using switching strategy . . . 591
Nadia Ndhaief, Olivier Bistorin and Nidhal Rezg

Supply Chain Network Planning

A stochastic model for analyzing the combined effect of demand uncertainty and carbon tax within the context of green supply chain design . . . 598
Amina Chelly, Imen Nouira, Atidel Hadj-Alouane and Yannick Frein

Lagrangian-relaxation based bounding approaches and fast deterministic heuristics for a class of multi-stage fixed-charge transportation problems . . . 604
S.A. Ezhil and Chandrasekharan Rajendran

Modeling and optimization of biomass supply chain with two types of feedstock suppliers . . . 610
Duc Huy Nguyen, Haoxun Chen and Nengmin Wang

Joint location of facilities and mobile resources: An application to the sector of public works 616
Agri Utama Noor Permadi, Axel Grimault and Olivier Péton

Vehicle Routing

A particle swarm optimization for the multi-depots pick-up and delivery problems with time windows and multi-vehicles . . . 623
Essia Ben Alaïa, Imen Harbaoui, Hanen Buchriha and Pierre Borne

A multi objective optimization approach for the m-MDPDPTW . . . 629
Essia Ben Alaïa, Hajer Ben Mahmoud, Imen Harbaoui and Taieb Ben Romthane

Freight buses in three-tiered distribution systems for city logistics: Modeling and evaluation . 635
Zheng Chang, Haoxun Chen and Farouk Yalaoui

A linear program for the resource-constrained project scheduling problem with routing 641
Philippe Lacomme, Alain Quilliot, Nikolay Tchernev, Marina Vinot and Aziz Moukrim

Mathematical model for electric vehicles cleaning problem 647
Ming Liu, Juan Jiang, Feifeng Zheng and Chengbin Chu

Integration of timetabling, multi-type vehicle scheduling and user routing in public transit network considering fuel consumption 653
Ming Liu, Ying Li, Feifeng Zheng and Feng Chu

Rich vehicle routing problem with loading constraints and resources synchronization 659
Mayssoun Messaoudi, Louis-Martin Rousseau, Issmail Elhallaoui, Adil Tahir and Mohamed Anouar Jamali

A DCA approach for the stochastic shortest path problem in vehicle routing 664
Vinh Thanh Ho, Hoai An Le Thi and Ahmed Zidna

Assembly Line Planning

Bounds and bound sets for the bi-objective simple assembly line balancing problem

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Audrey Cerqueus, Xavier Delorme
Mines Saint-Étienne, LIMOS UMR CNRS 6158,
158, cours Fauriel, CS 62362
42023 Saint-Etienne, cedex 2, France
Email: audrey.cerqueus@emse.fr, delorme@emse.fr

***Abstract*—This paper focuses on the use of bounds and bound sets in the solving of a bi-objective version of the simple assembly line balancing problem. The two objective functions considered are the takt time and the number of stations. We propose a bi-objective solution method, returning all solutions that present an interesting tradeoff between the two objectives. We present bounds and bound sets and introduce them in a branch-and-bound method. The algorithm is demonstrated on a didactic instance.**

I. Introduction

Assembly lines were introduced to organize mass production. A paced flow line is composed of sequentially ordered workstations, also called stations. Each station executes a given set of tasks and then each part moves to the next station at each takt time. The line balancing problem deals with the assignment of the tasks to the stations. It is a crucial problem, since it determines the number of stations (which are an important part of the cost when designing an assembly line), the takt time (the time required by the stations to executes the operation, i.e. the rythm of production of one unit of product), but also impacts the working conditions, etc.

The Simple Assembly Line Balancing Problem (SALBP) considers a line with a single type of product. It is a fundamental and well studied problem [1], [2].

The SALBP considers that each task can be performed on any station and has to be executed entirely on a unique station. Moreover, the processing times are assumed to be deterministic and independent on the station the task is executed on. The tasks are subject to precedence constraints.

We can mainly distinguish three types of SALBP:

- SALBP-1 aims to minimize the number of stations m, the takt time c being given.
- SALBP-2 aims to minimize the takt time c, knowing the number of stations m.
- SALBP-E aims to minimize $c.m$.

The industrial grow interests in other objectives such as the well-being of their employees or their ecological impact for example. If optimizing only these aspects does not really make sense, it would be interesting to propose tradeoff solutions regarding several objectives, among which the decision maker could choose the one to apply. Recently, multi-objective approaches have been proposed for the design of assembly lines [3], [4].

In this paper, we present a methodology to solve a SALBP problem, minimizing two objectives: the takt time and the number of stations. We denote this problem BO-SALBP.

The tasks are indexed from 1 to n and their processing time is given by t_j. We denote P_j the set of tasks that are predecessors of the task j.

The decision variables are x_{jk}, binary, representing whether the task j is assigned to the station k or not. We denote X the set of feasible solutions, i.e. assignments of values to $x_{jk}, j, k \in \{1, \ldots, n\}$ respecting the constraints stated before.

In the next section, we introduce some notions of multi-objective optimization.

II. Multi-objective optimization

When considering a multi-objective problem, we optimize simultaneously p objective functions, $p \geq 2$. In this article, we present the definitions for the case $p = 2$. The vector of objective values of the solution x is $z(x) = (z_1(x), z_2(x)) \in \mathbb{R}^2$. $z(x)$ is called a point. For a point y, we denote by y_1 the first component of the vector and y_2 the second one, i.e. $y = (y_1, y_2)$. The objectives considered in a multi-objective optimization problem are generally conflicting. Thus there generally does not exist a single feasible solution optimizing simultaneously all the objective functions. The solutions presenting an interesting tradeoff between the objectives are called efficient. Their definition is based on the notion of Pareto dominance (Definition 1).

Definition 1. Let us consider two points $y^a, y^b \in \mathbb{R}^2$.

- y^a weakly dominates y^b, denoted by $y^a \leqq y^b$, if $y^a_k \leq y^b_k$ for $k = 1, 2$.
- y^a dominates y^b, denoted by $y^a \leq y^b$, if $y^a \leq y^b$ and $y^a \neq y^b$.

Two points are incomparable according to this componentwise order if there does not exist any dominance relation between them (i.e. if none of them weakly dominates the other). The dominance cone is defined as $\mathbb{R}^2_{\leqq} = \{y \in \mathbb{R}^2 : y \leqq (0,0)\}$. $y + \mathbb{R}^p_{\leqq}$ stands for the vector addition of $y \in \mathbb{R}^2$

and any vector of $\mathbb{R}^2_{\leqq}$, it thus defines the area whose points weakly dominate y.

When solving a multi-objective problem, we aim to obtain solutions such that it is not possible to improve one of the objective functions without degrading another. Such a solution is called an efficient solution (Definition 2).

Definition 2. Let us consider $\hat{x} \in X$

- $\hat{x}$ is efficient if there does not exist $x \in X$ such that $z(x) \leq z(\hat{x})$. $z(\hat{x})$ is said to be nondominated.

The set of efficient solutions in denoted by X_E and its image in objective space is the set of nondominated points Y_N.

Example 1 illustrates Definitions 1 and 2.

Example 1. Let us consider a bi-objective minimization problem and seven feasible solutions $x^1, \ldots, x^7$, whose objective values are represented on Figure 1.

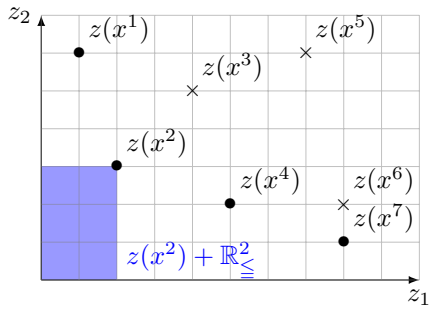


Fig. 1. Illustration of the notions of dominance and nondominated points. The nondominated points are depicted by •. The shaded area is $z(x^2) + \mathbb{R}^2_{\leqq}$

The point $z(x^3)$ dominates $z(x^5)$ and is dominated by $z(x^2)$. The solutions x^1, x^2, x^4 and x^7 are efficient.

Numerous branch-and-bound methods and dynamic programming methods have been developped for SALBP [5]. [6] presents a branch-and-bound method for a bi-objective assembly line balancing problem with choice of the equipment. These implicit enumerative solution methods bracket the objective value of the optimal solutions using bounds. An important number of bounds have been defined for SALBP, as attest [5], [7]. [8] notices that when dealing with a multi-objective optimization problem, a single point bound can be located far from the nondominated points. In order to obtain a finer estimation of the nondominated points, the notion of bound has been extended to bound sets, considering a set of points for bounding Y_N.

In this paper, we aim to elaborate bounds and bound sets to bracket the BO-SALBP. These bounds and bound sets are introduced in a branch-and-bound method, finding a minimal complete set of efficient solutions, i.e. subset $X' \subset X$ such that $z(X') = Y_N$.

III. The branch-and-bound method

In this section, we describe the developed branch-and-bound method and in particular the definition of the bounds and bound sets. Illustrations of the components are presented to help the understanding. A same instance is used for these examples. The description of this instance is given in Example 2.

Example 2. The instance used to illustrate the method involves 8 tasks. The precedence graph is given on Figure 2, an edge going from a node i to a node j means that the task i must be completed before j starts. The processing time of a task is written over the corresponding node.

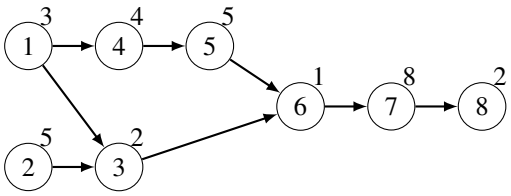


Fig. 2. Precedence graph of the example problem. The value written over a node j is t_j, the processing time of the task j.

In the branch-and-bound method, the feasible set is successively divided by fixing one or several variables. The partitioning of the feasible set can be represented by a tree whose root node is the global problem (i.e. the original problem to solve) and each node represents a subproblem, in which certain variables are fixed. We associate to each node a partial solution, composed only of the fixed variables. The partial solution associated to the root node is the empty solution (no task is assigned to any station).

The following notations are used:

- A solution or a partial solution will be denoted by the lists of the tasks indexes assigned to each stations (between [] symbols), the i-th list representing the i-th station.
- U is the set of tasks that are unassigned for a given partial solution.
- A is the set of tasks that can be immediatly assigned for a given partial solution (i.e. all their predecessors are assigned in the partial solution).
- $C_k(x)$ is the charge of the station k for the feasible or partial solution x, i.e. $C_k(x) = \sum_{j=1}^{n} x_{jk} t_j$.
- Let x be a feasible or partial solution, $m(x)$ is the number of stations in the solution and $c(x)$ is the takt time of the solution, i.e. $c(x) = \max_{k=1}^{m(x)} C_k(x)$.
- $t_{\max}(S)$ denotes the processing time of the longest task of the set S, i.e. $t_{\max}(S) = \max_{j \in S} t_j$.
- $t_{\max_i}(S)$ the i-th longest tasks of S.
- $t_{\text{sum}}(S)$ is the sum of the processing time of the tasks of the set S, i.e. $t_{\text{sum}}(S) = \sum_{j \in S} t_j$.

Example 3. The solution $x = [1, 2, 4, 5][3][6, 7, 8]$ assigns the tasks 1, 2, 4 and 5 to the first station, 3 to the second and 6, 7 and 8 to the third. Thus for this solution, the number of station is $m(x) = 3$ and the takt time is $c(x) = 17$.

A. The separation procedure

In this paragraph, we describe the construction of the search tree. It is based on a station-oriented strategy. At each node, the branching procedure chooses the assignment of the tasks to the next station. We consider only valid assignments, i.e. those satisfying the precedence constraints, similarly to the algorithm EUREKA [9].

The branches are identified according to the assignment chosen. The possible assignments are enumerated. Obviously, since the order on a same station is not important, we forbid to create two branches with the same set of tasks on the station, but in a different order.

Example 4 shows the first separation procedure.

Example 4. Figure 3 illustrates the separation procedure.

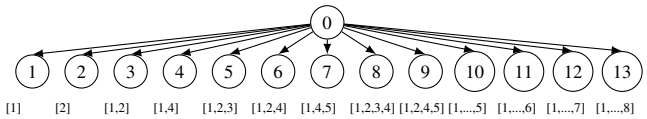


Fig. 3. Illustration of the separation procedure, on the first depth of the search tree. We denote $[1, \dots, i]$ the assignment of all tasks from 1 to i to the station ($[1, ..., 5] = [1, 2, 3, 4, 5]$).

For the node 4, the partial solution is $x = [1, 4]$, thus $U = \{2, 3, 5, 6, 7, 8\}$, $A = \{2, 5\}, c(x) = C_1(x) = 7$ and $m(x) = 1$.

Each subproblem is evaluated by lower bounds and/or bound sets (see Section III-C). If it is possible to prove that there is no new nondominated point in this branch of the search tree, then the node is closed, i.e. the separation stops in this branch. This proof is based on the comparison of the evaluation of the subproblem with the solutions found so far in the solving process, i.e. with the upper bound set that is explained in the next section.

B. Upper bound set

Classically the upper bound set is based on the potentially nondominated points found earlier in the execution, i.e. the images of the feasible solutions found filtered by dominance. We denote Y_{PN} the set of these solutions. Obviously, we are not interested in any solution which image in objective space is in the union of $y - \mathbb{R}^2_{\leqq}, y \in Y_{PN}$ since it would be dominated. An graphical illustration of this space is shown in dark red on Figure 4, considering that $Y_{PN} = \{y^1, \dots, y^5\}$.

This space can also be described using the local nadir points (Definition 3). In this definition, we consider $y^l, y^r \in Y_{PN}$. They are said to be adjacent if there does not exists $y \in Y_{PN}$ such that $y_1 \in [\min(y^l_1, y^r_1), \max(y^l_1, y^r_1)]$. In other terms, two points of Y_{PN} are adjacent if they are consecutive in the lexicographically ordered set. A local nadir point is defined for each pair of adjacent points in Y_{PN} (Definition 3).

Definition 3. Let us consider y^l and y^r adjacent in Y_{PN}, the local nadir point defined by y^l and y^r is $(\max(y^l_1, y^r_1), \max(y^l_2, y^r_2))$. We denote NP the set of nadir points.

On Figure 4, the local nadir points are depicted by a square.

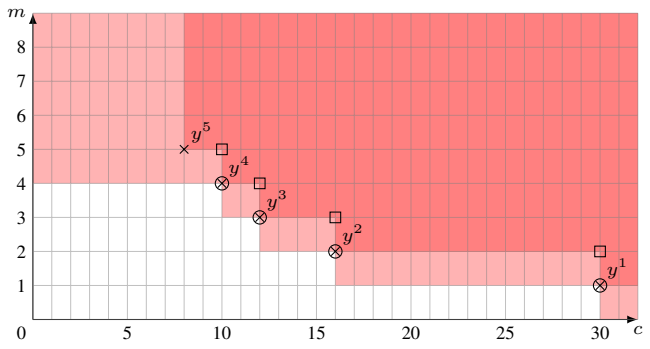


Fig. 4. Illustration of the potentially nondominated point $\times$, the associated nadir points $\square$ and shifted nadir points $\circ$.

The list of potentially nondominated points can be initialized with the image of any feasible solution. We initialize it with the image of the following solutions:

1) The trivial solution x^1 in which all tasks are assigned to the first station.
2) A solution x^M with a takt time equal to $t_{\max}(\{1, \dots, n\})$, which can be obtained by a constructive heuristics, such as the ranked positional weight method introduced in [10].
3) One solution for every m from 2 to $m(x^M)-1$. They can be generated by an iterative method, applying the ranked positional weight method [10] with a given value of takt time of $c = t_{\text{sum}}(\{1, \dots, n\})/m$. When no solution with m stations is found, the heuristic is executed again, with a larger takt time c, until a feasible solution with m stations is found.

The list of solutions generated is presented in Example 5.

Example 5. On the example instance, Y_{PN} is initialized by the image of the solutions:

- $x^1 = [1, 2, 3, 4, 5, 6, 7, 8]$ with $y^1 = z(x^1) = (c(x^1) = 30,\ m(x^1) = 1)$.
- $x^5 = [1, 4][2, 3][5, 6][7][8]$ with $y^5 = z(x^5) = (8, 5)$.
- $x^2 = [1, 4, 2, 3][5, 6, 7, 8]$ with $y^2 = z(x^2) = (16, 2)$.
- $x^3 = [1, 4, 2][5, 3, 6][7, 8]$ with $y^3 = z(x^3) = (12, 3)$.
- $x^4 = [1, 4][2, 5][3, 6][7, 8]$ with $y^4 = z(x^4) = (10, 4)$.

They are represented on Figure 4.

The second solution of the initialization of Y_{PN} is built such that the takt time does not exceed $t_{\max}(\{1, \dots, n\})$. Let us denote x^M this solution. We can affirm that a solution with a number of stations larger than $m(x^M)$ cannot be efficient. Indeed, the takt time cannot be less than the time required by the longest tasks, thus such a solution is always dominated by x^M. Moreover, a solution with a takt time larger than $t_{\text{sum}}(\{1, \dots, n\})$ will always be dominated by x^1.

Thanks to this observation, we can say that after the initialization of Y_{PN}, the only points in which we are interested are those dominating at least one of the local nadir points (Definition 3). Since we look for a minimal complete set of efficient solutions, we are not interested in equivalent efficient

solutions (solutions having the same objective values). Thus we can refine the zone of points we are interested in by defining the shifted local nadir points (Definition 4).

Definition 4. The set of shifted local nadir points SNP is defined by the the local nadir points to which the value of m has been reduced by one unit. $SNP = \{(c, m-1), (c, m) \in NP\}$. It exploits the fact that the number of stations is an integer value.

The shifted local nadir points are depicted by a circle on Figure 4. When evaluating the subproblems, we are only interested in the points dominating at least one of the shifted local nadir points. On Figure 4, these points are those located in the non-shaded area. The upper bound set in our method is the set of shifted local nadir points.

The upper bound set is the same for all subproblems, and it is updated each time a new nondominated point is found. The bounds and bound sets that are presented in the next section are computed for a given subproblem, i.e. for a partial solution.

C. Lower bound sets

We consider the following bounds and bound sets:

1) CM an upper bound set on the number of stations and the takt time,
2) LC the lower bound on the takt time,
3) UC the upper bound on the takt time,
4) LM the lower bound on the number of stations,
5) UM the upper bound on the number of stations.

CM is the continuous lower bound set defined by the equation $c.m \geq \sum_{j=1}^{n} t_j$, i.e. $m \geq \left\lceil \frac{\sum_{j=1}^{n} t_j}{c} \right\rceil$. It is an adaption of LM1 and LC1 described in [5] to the bi-objective case. For the evaluation on a partial solution x, the formula becomes $m \geq \left\lceil \frac{\sum_{j \in U} t_j}{c} \right\rceil + m(x)$. It is not possible to obtain a point that is below this bound set.

LC is the maximum of the value obtained for the three following formulas, for a partial solution x:

- $\max_{k=1}^{m(x)} C_k(x)$ the maximum charge among the stations,
- $t_{\max}(U)$ the longest task to process in U,
- $t_{\max_{UM-m(x)}}(U) + t_{\max_{UM+1-m(x)}}(U)$, an adaptation of LC2 described in [5].

For a partial solution x, LM is the maximum, rounded up, of:

- $m(x) + |\{j \in U, t_j > \frac{UC}{2}\}| + \frac{1}{2}|\{j \in U, t_j = \frac{UC}{2}\}|$, adapted of LM2 in [5],
- $m(x) + |\{j \in U, t_j > \frac{2UC}{3}\}| + \frac{2}{3}|\{j \in U, t_j = \frac{2UC}{3}\}| + \frac{1}{2}|\{j \in U, \frac{UC}{3} < t_j < \frac{2UC}{3}\}| + \frac{1}{3}|\{j \in U, t_j = \frac{UC}{3}\}|$, adapted of LM3 in [5].

The computation of LC and LM is essentially based on the values of UC and UM. Thus it is crucial to have a tight approximation of these two last values, which can be computed by the formula:

- $UC = \max\left\{\max_{k=1}^{m(x)} C(k), \sum_{j \in U} t_j\right\}$ i.e. the takt time of the partial solution and the charge of the next station if all tasks were assigned to it.
- $UM = \min\left\{m(x) + |U|, m(x^M) - 1\right\}$, the first part is the number of stations if the unassigned tasks were all on different stations and the second is the upper bound on the number of stations obtained during the initialization of Y_{PN} (Section III-B).

We have $UC \leq c(x^2)$, where x^2 is the solution generated during the initialization with $m = 2$. Indeed, a solution with a takt time strictly larger than $c(x^2)$ cannot be efficient: either $m(x) \geq 2$ and it is dominated by x^2, either $m(x) = 1$ and the only value possible for $c(x)$ is $t_{\text{sum}}(\{1, \ldots, n\})$ thus it is equivalent to the first solution of the initialization.

The values of UC and UM can be improved by comparing the shifted local nadir points and the lower bounds and bound sets.

Indeed, let us consider y^l and y^r adjacent in Y_{PN} such that $y^l_1 < y^r_1$ and y^l is located over or on CM, y^r is located under it and all $y \in Y_{PN}, y_1 > y^r_1$ and also under CM. Then it is not possible to obtain a new nondominated point with a takt time greated than y^l_1, thus $UC \leq y^l_1$. Example 6 gives an illustration of this assertion.

Example 6. Figure 5 represents the bounds and bounds sets for the node 6 on Figure 3. For the partial solution $x = [1, 2, 4]$, $C_1(x) = 12$ and $m(x) = 1$. CM is defined by the equation $m \geq \left\lceil \frac{\sum_{j \in U} t_j}{c} \right\rceil + m(x)$, i.e. $m \geq \left\lceil \frac{18}{c} \right\rceil + 1$. It is drawn in blue on Figure 5 and the strong blue shade represents the area in which it is not possible to have the image of a feasible solution for this subproblem.

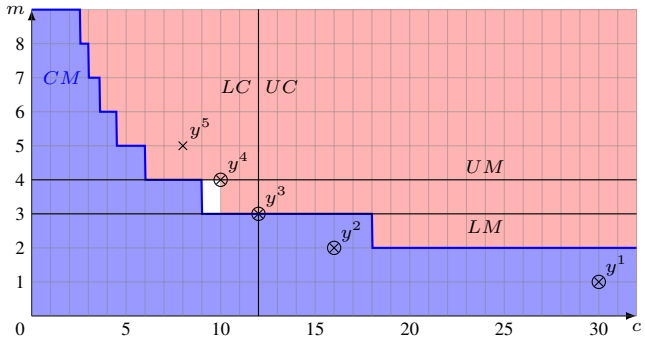


Fig. 5. Illustration the lower bounds and bound sets and the shifted local nadir points (○) associated to Y_{PN} (×) at the node 6 of Figure 3.

$UC = \max\{12, 18\} = 18$ and $UM = \min\{1 + 5, 5 - 1\} = 4$.

Since y^3 is on CM and y^2, y^1 are under it, we can tighten UC to 12.

Based on the values of UC and UM, we can compute LC and LM.

$LC = \max\{C_1(x), t_{\max}(U), t_{\max_3}(U) + t_{\max_4}(U)\}$
$= \max\{12, 8, 2+2\} = 12$

$LM = \max\{(1 + |\{j \in U, t_j > 6\}| + \frac{1}{2}|\{j \in U, t_j = 6\}|),$
$1 + |\{j \in U, t_j > 8\}| + \frac{2}{3}|\{j \in U, t_j = 8\}| + \frac{1}{2}|\{j \in U, 4 < t_j < 8\}| + \frac{1}{3}|\{j \in U, t_j = 4\}|\} = \max\{1+1, 1+\frac{2}{3}+\frac{1}{2}\}$, by rounding up $LM = 3$.

The values of LC, LM, UC, UM are represented in black on Figure 5.

D. Discarding rule

The comparison of the upper bound set and the lower bounds and bound sets allows to fathom nodes, i.e. to stop the separation in this branch.

To fathom a node by dominance, we check if it exists a point $y' \in \mathbb{R}^2$ such that $LC \leq y'_1 \leq UC$, $LM \leq y'_2 \leq UM$, y is above CM and y' dominates at least one of the shifted local nadir point, then, it is not possible to obtain a new nondominated point in this branch. Thus it could be fathom.

Example 7. In Example 6, the node 6 can be fathomed. Indeed there does not exist a point y' such that $LC \leq y'_1 \leq UC$, $LM \leq y'_2 \leq UM$, y' is above CM and y' dominates at least one shifted nadir point.

The node 5 of Figure 3 cannot be fathomed. Indeed, the point $y' = (10, 3)$, illustrated on Figure 6, is such that $LC \leq y'_1 \leq UC$, $LM \leq y'_2 \leq UM$, y' is on CM and it dominates y^3 and y^4.

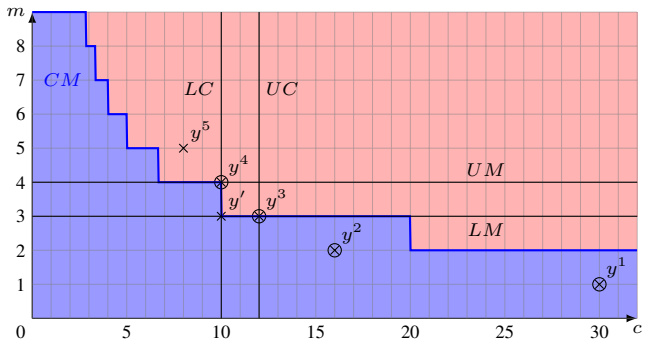


Fig. 6. Illustration the lower bounds and bound sets and the shifted local nadir points (∘) associated to Y_{PN} (×) at the node 5 of Figure 3.

Example 8. At the node 11 on Figure 3, the partial solution is $[1, 2, 3, 4, 5, 6]$, with $C_1 = 20$ and $m(x) = 1$. CM leads to $UC = 16$, see Figure 7. And since $C_1 = 20, LC = 20$. Thus we have $UC < LC$, this node can be fathom.

E. Dominance rules

If a node is not fathomed after the evaluation by the upper and lower bounds and bound sets, then the separation procedure is applied again. In order to reduce the search tree, at this step, we can apply dominance rules.

The maximum load rule is an adaptation of [11]. At a node, if it exists a task $j \in A$ such that $t_j \leq LC - C_{m(x)}$, thus this node cannot lead to an efficient solution. In the method, such a node is not created.

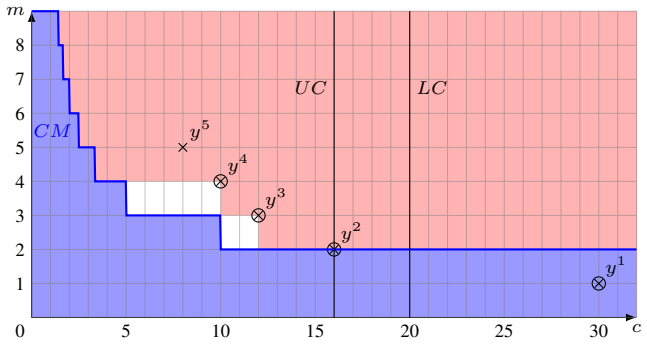


Fig. 7. Illustration the lower bounds and bound sets and the shifted local nadir points (∘) associated to Y_{PN} (×) at the node 11 of Figure 3.

If, before a separation, $\sum_{j \in U} t_j \leq LC$ then all the tasks can be assigned to the next station. The only node in this branch that can lead to a new nondominated point assigns all remaining tasks to the next station. The solution associated to this node is a feasible solution. Its objective values are computed and the upper bound set is updated. This is the remaining processing time rule.

Moreover, creating a node for which the takt time of the partial solution x is greater than the UC of its parent node would be fathomed by infeasibility. Indeed the value of UC in a chlid node cannot be greater than the one in the parent node, so we would have $LC \geq c(x) > UC$ in the child node. Thus such nodes are not considered. The same reasonning on the number of stations can be applied.

Example 9. On the instance of Example 2, for every node we have $LC \geq 8 = t_{\max}(\{1, \ldots, 8\}$. Thus the node 1 in Example 4 should not have been created, since $t_2 \leq LC - C_{m(x)}$ and $2 \in A$, the maximum load rule applies.

The nodes 9 to 13 in Example 4 should not have been created. Indeed, at the root node we have $UC \leq c(x^2) = 16$ and the partial solution associated to these nodes have a takt time greater than 16.

We can remark that the remaining processing time rule could have been applied on the node 11 if it was created.

IV. Illustration of the method

In this section, we run through the proposed algorithm on the example instance.

Figure 8, represents the search tree obtained at the end of the execution. The first level is smaller than in Figure 3. Indeed, the nodes 9 to 13 should not have been created (as explained in Example 9). Moreover, the nodes 1 and 2 should not have been created, because of the maximum load rule.

In this example, we analyze firstly the nodes for which the partial solution with the highest $\sum_{j \in U} t_j$. We can remark that this order impacts the efficiency of the method. Other orders for the nodes could be considered.

The nodes 6 to 8 of Figure 3 are fathomed by dominance. When evaluating the node 7, $UC = 12$ (by the comparison of Y_{PN} and CM), $UM = 4$, $LC = 12$ and $LM = 3$. The node is fathomed by dominance. Indeed, if there exists an efficient

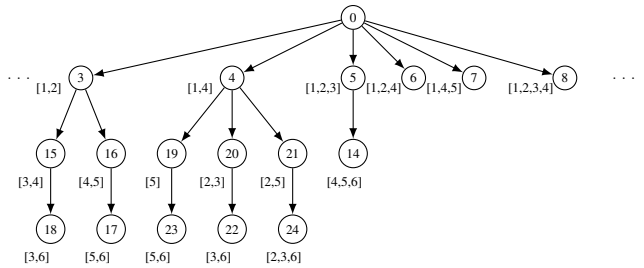


Fig. 8. Illustration of the search tree obtained by the method

solution in this node, its objective values are at best (12,3), we have $y^3 = (12, 3)$ and we are not interested in equivalent solutions.

The node 14, corresponding to the partial solution $[1, 2, 3][4, 5, 6]$ cannot be fathomed ($LC = 10, UC = 10, LM = 3, UM = 3$). The remaining processing time rule is applied and we obtain a new nondominated point $y^6 = (10, 3)$, associated to the solution $x^6 = [1, 2, 3][4, 5, 6][7, 8]$. The upper bound set is updated, $Y_{PN} = \{y^1, y^2, y^6, y^5\}$. The shifted local nadir points are $\{(30, 1), (15, 2), (10, 4)\}$. These two sets and the resulting upper bound set are represented on Figure 9.

Most of the leef nodes of the search tree (Figure 8) are fathomed by the remaining processing time rule, except for the nodes 6 to 8 which are fathomed by dominance and the node 14 for which the associated solution x^6 is feasible.

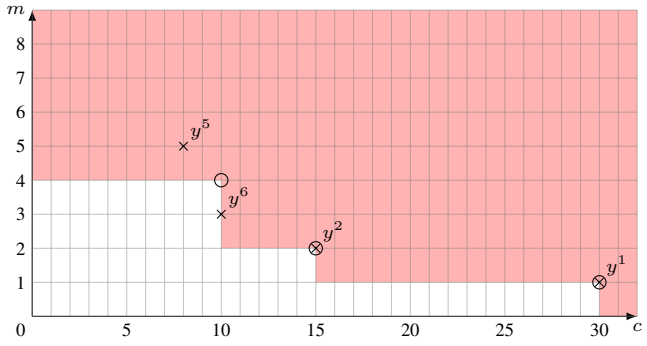


Fig. 9. The set of nondominated points (×), their shifted local nadir points (○) and the upper bound set (the shaded area).

The instance has four efficient solutions which are:

- $x^1 = [1, 2, 3, 4, 5, 6, 7, 8]$ with $z(x^1) = y^1 = (30, 1)$,
- $x^2 = [1, 4, 2, 3][5, 6, 7, 8]$ with $z(x^2) = y^2 = (16, 2)$,
- $x^6 = [1, 2, 3][4, 5, 6][7, 8]$ with $z(x^6) = y^6 = (10, 3)$,
- $x^5 = [1, 4][2, 3][5, 6][7][8]$ with $z(x^5) = y^5 = (8, 5)$.

The branch-and-bound method proved the efficiency of these solutions in 18 nodes in the search tree.

We can remark that x^1 is the optimal solution of the SALBP-1 version of this instance, x^5 is the one of SALBP-2 and x^1 and x^6 are both optimal for SALBP-E. x^2 is an efficient solution of BO-SALBP, which is not optimal for any of the problems SALBP-1, SALBP-2 and SALBP-E.

V. Conclusion

In this article, we proposed a solution method for a bi-objective version of the simple assembly line balancing problem. We presented bounds and bound sets, used in a branch-and-bound method. In order to reduce the number of subproblems considered in the search tree, we also adapted dominance rules. We illustrated the functioning of this algorithm on a didactic example.

Even if the algorithm is only illustrated on a didactic example here, preliminary experiments have shown that it can solve instances with up to 30 tasks in less than an hour. Further numerical experiments on a benchmark will be performed. To validate the efficiency of the algorithm, it will be compared to state of the art multi-objective solution method, such as the ε-constraint for example.

The method proposed in this paper aims to be a base solving block. Future studies will focus on the generalization of this method to extensions of the bi-objective ALBP to get closer to industrial problems. These extensions are generally more difficult to solve. They can consider a larger number of objective functions, which induces theoretical and computational difficulties when solving exactly the problem. More complex criteria can be used, for example, with non-integer values, which will forbid the use of the shifted local nadir points. The method would then need adjustments. Instead of a simple assembly line, the problem could concern more complex lines considering, for instance, several types of products or a flexible line.

References

[1] A. Scholl, *Balancing and Sequencing of Assembly Lines*, ser. Contributions to Management Science. Physica-Verlag HD, 1999.

[2] O. Battaïa and A. Dolgui, "A taxonomy of line balancing problems and their solution approaches," *International Journal of Production Economics*, vol. 142, no. 2, pp. 259–277, 2013.

[3] X. Delorme, O. Battaïa, and A. Dolgui, "Multi-objective approaches for design of assembly lines," in *Applications of multi-criteria and game theory approaches*. Springer, 2014, pp. 31–56.

[4] J. Oesterle and L. Amodeo, "Comparison of multiobjective algorithms for the assembly line balancing design problem," *IFAC-PapersOnLine*, vol. 49, no. 12, pp. 313–318, 2016.

[5] A. Scholl and C. Becker, "State-of-the-art exact and heuristic solution procedures for simple assembly line balancing," *European Journal of Operational Research*, vol. 168, no. 3, pp. 666–693, 2006.

[6] N. Pekin and M. Azizoglu, "Bi criteria flexible assembly line design problem with equipment decisions," *International Journal of Production Research*, vol. 46, no. 22, pp. 6323–6343, 2008.

[7] T. Pape, "Heuristics and lower bounds for the simple assembly line balancing problem type 1: Overview, computational tests and improvements," *European Journal of Operational Research*, vol. 240, no. 1, pp. 32–42, 2015.

[8] M. Ehrgott and X. Gandibleux, "Bounds and bound sets for biobjective combinatorial optimization problems," in *Multiple Criteria Decision Making in the New Millennium*, ser. Lecture Notes in Economics and Mathematical Systems, M. Köksalan and S. Zionts, Eds., vol. 507. Springer Berlin Heidelberg, 2001, pp. 241–253.

[9] T. R. Hoffmann, "Eureka: A hybrid system for assembly line balancing," *Management Science*, vol. 38, no. 1, pp. 39–47, 1992.

[10] W. Helgeson and D. Birnie, "Assembly line balancing using the ranked positional weight technique," *Journal of Industrial Engineering*, vol. 12, no. 6, pp. 394–398, 1961.

[11] J. R. Jackson, "A computing procedure for a line balancing problem," *Management Science*, vol. 2, no. 3, pp. 261–271, 1956.

Modeling kitting operations in a semi-robotic environment

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mohamed El Amine BOUDELLA, Evren SAHIN, Yves DALLERY
Laboratoire Genie Industriel
CentraleSupélec, Université Paris-Saclay
Chatenay-Malabry, France

Abstract—With increasing components variety in mixed-model assembly lines, kitting is an innovative line feeding mode to reduce congestion at the border of the line. This paper proposes a preliminary mathematical formulation of kitting activities for a hybrid robot-operator kitting system that delivers parts to a mixed-model assembly line. Indeed, elementary kitting operations are formulated, based on field assumptions, by distinguishing operations performed by the robot and the operator. Based on the formulation provided, the perspective of future research is to refine the model and to use the formulations in order to propose a quantitative model that would optimize the performance of the hybrid system.

Keywords—line feeding; kitting operations modeling; efficiency

I. Introduction

Kitting is one of the line feeding approaches used for supplying parts to mixed-model assembly lines. It consists in preparing a collection or "a kit" of parts needed to assemble a specific end product (EP) which is sequenced on the assembly line (see [1] and [2]). Hence, parts are physically placed together in one or more containers referred to as "kitting boxes", ready to be used on the assembly line.

Recently, due to technological improvements, hybrid robot – operator (human) kitting systems are introduced in industry. Indeed, both resources (robot and operator) working in series are involved in the preparation of kitting boxes. More specifically, the robot would start the kitting boxes preparation by delivering partially filled boxes. The operator then retrieves the preparation made by the robot and completes it with the remaining parts needed in the kit.

In such a system, one interesting question concerns tasks assigned to each resource and system performance assessment. In order to address this question, the elementary kitting operations have to be identified, by separating operations performed by the robot and by the operator. Then, assumptions stemming from discussions with company experts, technological solution provider and observations made in company current kitting processes enable to give a formulation of kitting operations modeled.

This paper is organized as follows: Section II gives a literature review related to the analyzed issue. Section III gives (i) an overview of the kitting system considered (ii) the formulations of operations carried out by the robot and the operator to deliver kits. Finally, we conclude the paper in Section IV.

II. Literature review

The literature on line feeding operations is expanding. However, most of existing models are developed for low-level picker-to-parts kitting systems where the picker is a human. The difference in our paper, is that we consider a hybrid kitting system working with both a robot and an operator.

In [1], authors develop the first descriptive model to quantify the trade-offs in material handling, space requirements in shop floors and work-in-process between two line feeding modes namely kitting and line stocking. In [3], authors develop a Mixed Integer Programming (MIP) that aims at minimizing the total labor cost by assigning components to line stocking or kitting (kit in its stationary form). Following a same approach for modeling operations related to line feeding, [4] introduces sequencing in addition to line stocking and kitting (traveling kit) and proposes an empirical assessment of operations' costs associated with the three line feeding modes. The aim of their study is to characterize the conditions (values of each parameter) that make a line feeding mode the least costly. In [5], authors rely on optimization to assign components to each of the three line feeding modes studied in their previous work. References [3], [4] and [5] all model operating costs pertaining to the studied line feeding modes (costs at the point-of-use in the assembly line, internal transport costs, costs for kitting and supermarket replenishment costs). In [6], authors consider annual costs for kitting and give an extension to previously cited works by integrating investment and error costs in addition to administrative costs related to pick lists preparation. Modeling kitting errors in manual kitting can be found in [7]. In [8], authors investigate the impact of parts features, namely size and cost, on the choice between kitting, line stocking and JIT policy.

Other research papers have addressed issues on ergonomics and time efficiency related to order picking activities. Authors in [9], show the impact of packaging type and size, angle of exposure of storage bins, height of storage racks and part size on picking efficiency. In [10], authors focus on packaging and show that picking efficiency can be impacted by picking from different sections in a pallet. Regarding kitting systems, authors in [11], analyze 15 case studies from the automotive industry and identify the most impactful factors on man-hour consumption. The parameters that have the biggest impact on kitting efficiency appear to be the type and size of bins, the type and design of storage racks, the batch size and the size of the picking area.

Limited research exists on the design and performance evaluation of robotic kitting systems. In [12], design factors and criteria to implement an effective robotic kitting system are also developed in accordance with the design process steps. Based on simulations, [13] shows quantitative results for the performance of six different robotic kitting configurations in terms of throughput, average time a kit spends in the system, and robot utilization.

The work presented in this paper is in the continuity of [14] that proposes a first formulation of operations performed by a robotic kitting cell in a given layout and a set of components already assigned to the robot. The current paper considers hybrid kitting system where a robot works in series with an operator to deliver kits. We propose the formulation of elementary kitting operations for both the robot and the operator, under the assumption of components are already assigned to either the robot or the operator. To our knowledge, this is among the first works that introduce the use of a robotic arm in existing manual kitting processes.

III. Modeling kitting operations

This section aims at giving the formulation of elementary activities taking place in the hybrid kitting system.

A. Hybrid kitting system description

As showed in Fig. 1, the hybrid kitting system involves a robot working in series with an operator. They prepare and deliver a kit (or a collection) of parts for each end product (EP) to be assembled. Parts needed by an EP are placed in kitting boxes before they are sent to the assembly line. We use the term "box" when referring to a kitting box. Physically, the size of a box is adapted to receive all parts needed at the same point-of-use (in the assembly line). For example, parts intended to be assembled on the left side of an EP are grouped into one box to ease their positioning close to the point-of-use. In Fig. 1, locations of parts associated with one kitting box are represented by the same color which means that in our case, we have 6 different kitting boxes (both resources deal with all box types). Physically, boxes are compartmented and the number and arrangement of compartments are specific to each box according to the number and shape of received parts.

In the robotic kitting area, components intended to be in the same box are grouped in zones (denoted as Zone 1 to 6 in Fig. 1) following a class-based storage policy as represented in Fig. 1. Hence, each box is of a certain "type" and is denoted as box 1 to 6 following the numbering of storage zones.

In the manual kitting area, parts are assigned following a volume-based strategy where highly-demanded components are located close to the buffer (i.e. close to the input/output point of the area). As we can see in Fig. 1, SKUs intended to be in a certain box are spread over the entire manual kitting.

Actually, components to assemble may have one (the parent component) or several variants based on the same component definition. Each EP uses one variant of a component. Those variants differ on one or more characteristics and have a unique SKU. When we speak about "parts" we refer to pieces of an SKU. As in [4], [7], [8] and [13], we consider elements at the variant (SKU) level since variants of the same component type may have different usage rates. The usage rate of a particular SKU is the percentage of EPs that use this SKU.

Each SKU is stored in a dedicated bin that may be of different sizes, materials and be foldable or not. Due to quality and transportation concerns, parts of an SKU may need specific inner packaging elements (plastic bags, cardboard and/or foam interlayers, dividers) to prevent damage to parts.

The robot starts by picking parts and placing them directly on a moving conveyor that drops them within large trays downstream the robotic kitting area. After that, the operator retrieves those parts, places them into boxes and completes the boxes with the remaining parts in their kitting area.

In our study, we only focus on operations that take place in the hybrid kitting system which is assumed to work at full speed without any failure (especially for the robotic kitting area). We also implicitly assume that everything works without failure outside the hybrid kitting system which means for instance that the replenishment of parts, the evacuation of empty bins (accumulated on evacuation ramps) and inner packaging elements are regularly carried out by external resources. Moreover, as a first approach, we assume that downstream the robotic kitting, the buffer (trays loop) does not constitute a bottleneck since parts are systematically collected by the operator to make trays available for a new preparation.

Fig. 1 presents a particular case of a hybrid kitting area where 6 boxes can be prepared for an EP. In fact, there are basic EP models that need common components while others with higher diversity, due to more design options, need more components (in addition to those of basic EPs). In terms of preparation, basic EPs only need 4 boxes to hold their parts while specific EPs need all the 6 boxes (boxes 5 and 6 being specific).

To prepare kits, different picking strategies can be adopted. In the single picking approach, a preparation cycle consists in picking parts associated with one EP. In the case of batch picking, the system prepares multiple EPs in one picking list.

1) Characteristics of the robotic kitting area

The robotic kiting area has a single narrow-aisle with 2-levels gravity flow racks on each side. Those racks are assumed to be of the same length with only one evacuation ramp for depleted bins per rack. The number of levels is limited to allow a better accessibility to the robot.

The moving conveyor belt, on which the robot places the picked parts one after the other, traverses the picking aisle so that parts are dropped within trays circulated in a closed loop. Each tray receives parts from one and only one class. Thus, to preserve parts integrity, the robot has to place the heaviest parts first. To make this easier, components are stored in each zone (class), following a dedicated storage strategy and a particular order from heaviest parts to more fragile ones.

2) Characteristics of the manual kitting area

The manual kitting area has 4-levels gravity flow racks. Unlike the robotic kitting area, the top level is entirely dedicated to the evacuation of bins.

The operator starts a kit preparation by loading, on their cart, empty boxes consumed at the assembly line and returned to the kitting area.

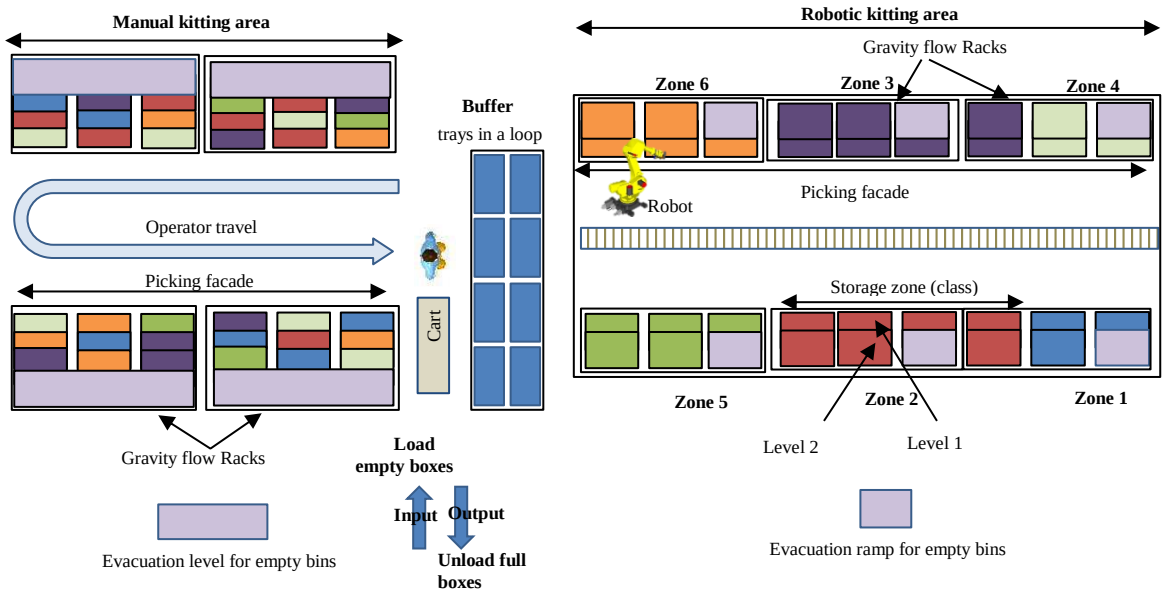


Fig. 1. Hybrid kitting system

After that, the operator retrieves, from each tray at a time, the preparation of the robot and places it in the destination box. For example, parts picked by the robot for box type 1 and held in one tray are retrieved by the operator and positioned in compartments of one box of type 1.

To complete each box, the operator travels to pick parts in their area and places them within the associated box. To help an operator in their tasks, there is a pick-to-light system that indicates the needed SKUs' location using light displays.

B. Formulation of kitting operations

Based on company common practices, discussions with experts and observations made on current kitting areas, the formulation of kitting cycle time components is developed in this section. Due to paper space limitations, details pertaining to the assumptions would not be developed.

Indeed, the robot and operator cycle times are the sum of elementary activities they perform each to prepare a kit of parts for a batch of EPs.

In more details, to give the formulation of kitting operations, we consider a representative preparation cycle over a reference time period. Hence, starting from picking lists in the reference period, we propose a calculation of an average cycle time for each resource (robot or operator) based on SKUs' usage frequency and average number of parts to pick per SKU. Indeed, with each EP to produce is associated a picking instruction (or list) that lists all SKUs needed in the kit and gives, for each SKU, the amount of parts to pick. Let i be the SKUs index. Thus, the usage rate τ_i of SKU i is the percentage of EPs in the reference time period that use this SKU. The average number of parts n_i to pick for SKU i is calculated by summing, for this particular SKU, BOM (Bill Of Material) coefficients of all EPs produced in the reference period and dividing the result by the number of EPs that use this SKU.

Each SKU is handled either by the robot or the operator. We denote the SKU index i, for SKUs assigned to the robot ($i = 1..C_R$) and the SKU index j, for SKUs assigned to the operator ($j = 1..C_O$).

Hence, the formulation of the robot cycle time is done for a set of C_R and C_O SKUs supposed to be already assigned respectively to the robot and the operator.

For the remaining of the paper, indexes and notations we adopt for the formulation of elementary operations' duration are given below:

- i, j: SKUs index;
- C_R : total number of SKUs assigned to the robot
- C_O : total number of SKUs assigned to the operator
- BS: batch size which is the number of EPs to prepare simultaneously i.e. a single picking strategy is adopted when $BS = 1$;
- K : number of common boxes per EP;
- K_{max} : maximum number of boxes per EP;
- We display "R"; which refers to the robot; and "O"; which refers to the operator; as superscripts in the formulation of activities done by both the robot and the operator.

C. Modeling the robot cycle time

The **robot cycle time** is an aggregation of all elementary activities that the robot performs to prepare a kit of parts for each batch of *BS* EPs. Equation (3) expresses the total cycle time of the robot.

$$CT^R = T^R_{pick/place} + T_{vision} + T_{tool} + T^R_{trav} + T^R_{B_rem} + T^R_{pack_rem} \quad (3)$$

$T^R_{pick/place}$ is the total **pick and place time**; T_{vision} is the total **image acquisition time** of storage bins content; T_{tool} is the total **grasping tool changing time**; T^R_{trav}, the total **travel time**; $T^R_{B_rem}$ is the total **time needed to return back empty bins** (pick bins, move them and place them on the evacuation ramp) and $T^R_{pack_rem}$ is the **total time to remove inner packaging elements**.

Equation (4) gives the total **pick and place time** for each preparation cycle.

$$T^R_{pick/place} = \sum_{i=1}^{i=C_R} [(1+\pi_i) * T^R_i * BS * \tau_i * n_i]. \quad (4)$$

T^R_i and π_i are respectively the time needed for the robot to pick and place a single part of SKU *i* and the proportion of failed picks associated with each SKU *i*. The failure rate provides information on the ability of the robot to properly pick parts at the first attempt. τ_i and n_i are respectively the usage rate of SKU *i* and the average number of parts needed for SKU *i*. Equation (5) expresses the average number of parts n_i.

$$n_i = \frac{\sum_{EP} BOM_Coefficient_i}{Total\ number\ of\ EPs\ that\ use\ the\ SKU\ i\ in\ the\ reference\ period} \quad (5)$$

Image acquisition consists in taking a picture of bins' content. Afterwards, the image is analyzed by the robotic kitting system to identify and propose potential parts to pick (generally parts on top). We assume that a part of this operation may be done as a background task i.e. in "hidden time" while the robot is traveling or rotating to drop a part on the conveyor. This is given in (6) as follows:

$$T_{vision} = \sum_{i=1}^{i=C_R} (1 - p) * BS * \tau_i * n_i * T_{cap} \quad (6)$$

T_{cap} is the technological time needed for a single acquisition and *p* is the part of the technological time executed in hidden time. Here, T_{vision} may be zero if *p* = 100% i.e. if the acquisition is completely executed in hidden time.

The average **time to travel** to pick all parts required for the preparation of *BS* EPs is given in

$$T^R_{trav} = \sum_{i=1}^{i=C_R} \left[\frac{2*BS - (F^R - Ps^R)}{Ps^R} * \frac{B_i}{v_R * N^R_{levels}} + BS * \tau_i * T^R_s \right]. \quad (7)$$

The robot travel time is given by the covered distance during a preparation cycle divided by the robot's velocity v_R as in (8) and the time needed to stop in front of bins as expressed by (9).

$$\frac{2*BS - (F^R - Ps^R)}{Ps^R} * \frac{\sum_{i=1}^{i=C_R} B_i}{v_R * N^R_{levels}} \quad (8)$$

The covered distance is given by the total number of roundtrips made by the robot $(1/Ps^R)*(2*BS - (F^R - Ps^R))$ multiplied by the total picking façade length $\sum_{i=1}^{i=C_R} B_i / N_{levels}$ as in (8).

F^R and Ps^R are respectively the number of façades in the robotic kitting area and the number of façades from which the robot picks parts at a time to prepare a box. B_i and N^R_{levels} are respectively the size of a storage bin that contains parts of SKU *i* and the number of storage levels (in a rack) in the robotic kitting sub-systems.

In each storage zone, the robot performs *BS* travels to pick parts in addition to $(BS - (F^R - Ps^R))$ travels to go back to the starting point of each storage zone which gives a total of $2*BS - (F^R - Ps^R)$ travels in a storage zone. This is equivalent to $2*BS - (F^R - Ps^R)$ travels along the robotic kitting aisle.

For example, in a batch picking strategy (*BS* >1), the robot picks all parts needed by an EP in a storage zone before moving to the next EP of the batch. If we consider a batch of four EPs, the robot starts moving inside the blue storage zone (in Fig. 1) to pick parts for EP 1, then he goes back to the beginning of this storage zone (to pick heavy parts first) to travel again and pick parts for EP 2 and so on until EP 4. At the end of this batch picking preparation cycle, the robot has prepared four kits.

$$\sum_{i=1}^{i=C_R} BS * \tau_i * T^R_s \quad (9)$$

The expression $BS * \tau_i * T^R_s$ in (9) refers to the average stopping time in front of the storage location of SKU *i* calculated for each batch of EPs.

Equation (10) gives the average **tool changing time**.

$$T_{tool} = T_{tool_chg} * \sum_{i=1}^{i=C_R} BS * \tau_i. \quad (10)$$

T_{tool_chg} is the technological time needed to perform a single tool change.

Equation (11) gives the average **time to remove bins** $T^R_{B_rem}$.

$$T^R_{B_rem} = (T_{tool_chg} + T^R_B + T_{Bm} + 2*T^R_s) * \sum_{i=1}^{i=C_R} \frac{BS*\tau_i*n_i}{P_i} \quad (11)$$

T_{tool_chg} is the time needed for a single tool change. The value of T_{tool_chg} considered in (11) is equal to the one in (10). T^R_B and T_{Bm} are respectively the time required to the robot for removing a bin from its storage location and placing it on the evacuation ramp and the time needed for moving the bin from its storage location to the evacuation ramp. T^R_s is doubled to include the first stop when the robot removes the bin and the second stop in front of the evacuation ramp. P_i is the number of parts of SKU *i* contained in a bin.

The average **time to remove packaging elements** (apart from plastic bags) is given by

$$T^R_{pack_rem} = T_{tool_chg} * \sum_{i=1}^{i=C_R} (CL_i * T^R_{CL} + D_i * T^R_D) * \frac{BS*\tau_i*n_i}{P_i}. \quad (12)$$

T_{tool_chg} is the time needed for a single tool change. It is also equal to T_{tool_chg} in (11) and (10). We assume that whatever the task (changing the tool to pick a part or a bin a packaging element), a tool change requires the same amount of time. T^R_{CL} and T^R_D are respectively the time to pick and place, on the conveyor, a cardboard interlayer sheet and a divider. CL_i and

D_i are respectively the number of cardboard interlayer sheets and the number of dividers contained in a bin of SKU *i*.

D. Modeling the operator cycle time

The **operator cycle time** is an aggregation of all elementary activities that the operator performs to prepare a kit of parts for each EP of the *BS* EPs in the batch. Equation (13) gives the total cycle time of the operator.

$$CT^O = T_{load} + T_{retrieve} + T^O_{pick/place} + T^O_{trav} + T^O_{B_rem} + T^O_{pack_rem} + T_{unload} \quad (13)$$

T_{load} is the total **time to load empty boxes** on the kitting cart when starting the preparation; $T_{retrieve}$ is the total **time to retrieve parts** prepared by the robot before completing with the remaining SKUs; $T^O_{pick/place}$ is the total **pick and place time**; T^O_{trav} is the total **travel time**; $T^O_{B_rem}$ is the total **time to return back empty bins** (pick bins, move them and place them on the evacuation ramp); $T^O_{pack_rem}$ is the total **time to remove inner packaging elements** and T_{unload} is the total **time to unload filled boxes** with parts from the kitting cart.

The average **total time to load empty** boxes T_{load} is given by

$$T_{load} = (BS^* T_{kb} + T^O_s)^* [K + \pi_{rich_EP} * (K_{max} - K)]. \quad (14)$$

T_{kb} is the time required to the operator for handling one box. The proportion of rich EPs over the reference period and the average number of boxes to prepare per EP are given respectively by π_{rich_EP} and $[K+\pi_{rich_EP} * (K_{max}-K)]$. A time to stop in front of each box type (before loading) is also considered.

Time to retrieve parts prepared by the robot is given in

$$T_{retrieve} = (BS^* T_{ret} + T^O_s)^* [K + \pi_{rich_EP} * (K_{max} - K)]. \quad (15)$$

T_{ret} is the estimated time to retrieve all parts contained in a tray and place them inside the appropriate box.

Equation (16) expresses the average **pick and place time**.

$$T^O_{pick/place} = \sum_{j=1}^{j=C_O} \frac{BS^* \tau_j * n_j}{\theta_j} * T^O_j. \quad (16)$$

T^O_j is the time required to execute a single movement of picking parts of SKU *j* and placing them in the boxes that are on the cart. θ_j is the number of parts the operator picks simultaneously. It would correspond to the minimum between the number of parts of SKU *j* needed, i.e. $BS * \tau_j * n_j$ and a_j the maximum number of parts the operator can pick simultaneously as in (17). The value of a_i depends on the size (small, medium, large) of SKU *j* and so does T^O_j.

$$\theta_j = min\,(BS * \tau_j * n_j,\ a_j) \quad (17)$$

The average **time to travel** along the aisle to pick all parts required for a set of *BS* EPs is given when *BS* = 1 by

$$T^O_{trav} = \sum_{j=1}^{j=C_O} \left(\frac{2 - (F^O - Ps^O)}{Ps^O} * \frac{B_j}{v_O * N^O_{levels}} + \tau_j * T^O_s \right). \quad (18)$$

For batch picking i.e. for *BS* > 1, the average **travel time** needed to prepare a batch of *BS* EPs is

$$T^O_{trav} = \sum_{j=1}^{j=C_O} \left([K + \pi_{rich_EP} * (K_{max} - K)] * \frac{2 - (F^O - Ps^O)}{Ps^O} * \frac{B_j}{v_O * N^O_{levels}} + \tau_j * T^O_s \right). \quad (19)$$

Indeed, for *BS* (>1) EPs in the batch, the operator prepares *BS* boxes of one box type (common or specific) at a time. Hence, the operator performs $[K+\pi_{rich_EP} * (K_{max}-K)]$ travels to prepare all $[K+\pi_{rich_EP} * (K_{max}-K)]*BS$ boxes needed for a batch.

The operator travel time can be divided into two terms. Firstly, the covered distance during a preparation cycle divided by the operator's velocity v_O as in (20) for *BS* = 1 and in (21) for *BS* >1. Secondly, the time needed to stop in front of bins as in (22).

$$\frac{2 - (F^O - Ps^O)}{Ps^O} * \frac{\sum_{j=1}^{j=C_O} B_j}{v_O * N^O_{levels}} \quad (20)$$

$$[K + \pi_{rich_EP} * (K_{max} - K)] * \frac{2 - (F^O - Ps^O)}{Ps^O} * \frac{\sum_{j=1}^{j=C_O} B_j}{v_O * N^O_{levels}} \quad (21)$$

The covered distance is given by the total number of roundtrips $(2 - (F^O - Ps^O)/Ps^O)$ when *BS* = 1, as in (20), or $(2 - (F^O - Ps^O)/Ps^O) * [K+\pi_{rich_EP} * (K_{max}-K)]$ roundtrips when *BS* > 1, as in (21), multiplied by the total picking façade length which is given by $\sum_{j=1}^{j=C_O} B_j / N^O_{levels}$ in both (20) and (21).

F^O and Ps^O are respectively the number of façades in the manual kitting area and the number of façades from which the operator picks parts at a time to prepare a box or multiple boxes (in case of batch picking). B_j and N^O_{levels} are respectively the size of a storage bin that contains parts of SKU *j* and the number of storage levels (in a rack) in the manual kitting sub-system.

As the operator only stops one time to pick all parts needed for a SKU, the average stopping time of the operator in front of the storage location of SKU *j* is given by $\tau_j * T^O_s$ as in (22).

$$\sum_{j=1}^{j=C_O} \tau_j * T^O_s \quad (22)$$

Equation (23) gives the average **time to remove bins**.

$$T^O_{B_rem} = \sum_{j=1}^{j=C_O} (T^O_B + T_{fold} * Fold_j) * \frac{BS^* \tau_j * n_j}{P_j}. \quad (23)$$

T^O_B is the time needed by the operator for removing an empty bin from its storage location and placing it on the evacuation ramp. In addition, if the bin is foldable, we add the second term $T_{fold} * Fold_j$ in (23) that corresponds to the time needed for folding a single bin. Not all bins are foldable that is why we introduce the parameter $Fold_j$ that equals to 1 when the bin, where SKU *j* is stored, is foldable and equals to zero otherwise.

Equation (24) expresses the average time to remove inner packaging elements.

$$T^O_{pack_rem} = \sum_{j=1}^{j=C_O} (CL_j * T^O_{CL} + FL_j * T_{FL} + D_j * T^O_D + Pb_j * T_{pb}) * \frac{BS * \tau_j * n_j}{P_j} \quad (24)$$

The total time needed to remove the packaging elements is given by the time to remove all elements in a bin multiplied by the number of bins consumed during a preparation cycle which is $BS*\tau_j*n_j/P_j$.

T^O_{CL}, T_{FL}, T^O_D and T_{pb} are respectively the time needed to evacuate (in a trash bin) a cardboard interlayer, the time needed to evacuate a foam interlayer, the time needed to evacuate a divider and the time needed to evacuate a plastic bag.

For each bin of SKU j, CL_j, FL_j, D_j and Pb_j are respectively the number of cardboard interlayers, the number of foam interlayers, the number of dividers and the number of plastic bags.

The average total **time to unload the filled boxes** (from the cart), before sending them to the assembly line, is given by

$$T_{unload} = (BS* T_{kb}+T^O_s)*[K+\pi_{rich_EP} * (K_{max}-K)] \quad (25)$$

T_{kb} is the time required to the operator for handling a box. We assume that this time is the same for loading and unloading a box. The operator unloads all boxes of a box type on a dedicated gravity flow lane. Thus, the operator needs to stop in front of each lane (i.e. for each box type).

IV. Conclusion and perspectives

The contribution of the paper is twofold. Firstly, by considering a hybrid kitting system, we clarify the activities performed by each resource (robot and operator) involved in the system. Second, based on company practices, discussions with experts and field observations, we propose a first mathematical formulation of kitting operations that take place in the system, for each resource.

The modeling approach developed in this paper is generic and can be used for analyzing other possible configurations (e.g. two robots in the robotic area, one for picking parts, the other for placing them into boxes, in order to reduce the recurrent operations of the operator). This is indeed possible by adapting the formulations provided in this paper to the new configuration of interest.

Besides, based on elementary kitting operations formulation, our future objective is to build a refined model that would optimize the performance of the hybrid system. We argue that the current paper provides an interesting relevant basis for the development of such an optimization model.

References

[1] Y.A. Bozer and L.F. McGinnis, 1992, “Kitting versus line stocking: A conceptual framework and a descriptive model”, International Journal of Production Economics, Vol. 28, pp. 1 – 19.

[2] H. Brynzér, and M.I. Johansson. "Design and performance of kitting and order picking systems." International Journal of production economics 41.1-3 (1995): 115-125.

[3] V. Limère, H. Landeghem, M. Goetschalckx, E.-H. Aghezzaf, and L. F. McGinnis, 2012, “Optimizing part feeding in the automotive assembly industry: deciding between kitting and line stocking”, International Journal of Production Research, Vol. 50, No 15, p. 4046 – 4060.

[4] M. Sali, E. Sahin and A. Patchong, 2015, “An empirical assessment of the performances of three line feeding modes used in the automotive sector: line stocking vs. kitting vs. sequencing, International Journal of Production Research, Vol. 53, No 5, p. 1439-1459.

[5] M. Sali and E. Sahin. "Line feeding optimization for just in time assembly lines: an application to the automotive industry." International Journal of Production Economics 174 (2016): 54-67.

[6] A.C. Caputo, P.M. Pelagagge and P. Salini, 2015, “A model for kitting operations planning”, Assembly Automation, Vol. 35 Iss 1 pp. 69 – 80.

[7] A.C. Caputo, P.M. Pelagagge and P. Salini, 2015, “Modeling errors in kitting processes for assembly lines feeding”, International Federation of Automatic Control, p 338 – 344.

[8] A.C. Caputo, P.M. Pelagagge, and P. Salini, "Selection of assembly lines feeding policies based on parts features." IFAC-PapersOnLine 49.12 (2016): 185-190.

[9] C. Finnsgård and C. Wänström, 2013, “Factors impacting manual picking on assembly lines: an experiment in the automotive industry.” International Journal of Production Research no. 51(6): 1789 -1798.

[10] R. Hanson, L. Medbo, P. Jukic and M. Assaf, 2016, “Manual Picking from Large Containers – Time Efficiency and Physical Workload.” Paper presented at the 8th IFAC Conference on Manufacturing, Modelling, Management and Control, Troyes, June 28-30 no 49(12): 1703 – 1708.

[11] R. Hanson and L. Medbo, 2016, “Aspects influencing man-hour efficiency of kit preparation for mixed-model assembly.” Paper presented at the 6th conference on Assembly Technologies and Systems no. 44: 353 – 358.

[12] K. Tamaki and S.Y. Nof, 1991, “Design method of robot kitting system for flexible assembly”, Robotics and Autonomous Systems, Vol. 8, p. 255 – 273.

[13] C.J. Sellers and S.Y. Nof, 1989, “Performance analysis of robotic kitting systems”, Robotics & Computer-Integrated Manufacturing, Vol .6, No. 1, pp. 15-24.

[14] M.E.A. Boudella, E. Sahin, and Y. Dallery, “A mathematical model to assess the performance of a robotic kitting system in an assembly plant”, 11th International Conference on Modeling, Optimization and Simulation, 2016, in press.

Impact of Smoothening and Synchronization of Components Flow on Staging Delay in Assembly Systems

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Shankar Saravana Kumar, Chandrasekharan Rajendran
Department of Management Studies
Indian Institute of Technology Madras
Chennai, India
shankarsaravana.ie@gmail.com, craj@iitm.ac.in

Rainer Leisten
Business Administration and Operations Management
University of Duisburg-Essen
Duisburg, Germany
rainer.leisten@uni-due.de

Abstract— Staging delay in assembly system is the waiting time of component of a final product or assembly till the arrival of all other components necessary for the final product in front of stage performing the assembly operation. In a fork-join model assembly system, staging delay of components in front of assembly stage is one of the contributors to the cycle time of the final product. Reducing Average Staging Delay (ASD) of components can potentially improve the throughput of assembly system, thereby impacting the cycle time. Staging delay is generally caused, among others, by variability and deficiency of synchronization of component arrivals at assembly stage. High variability prevents smooth flow of components. We demonstrate in this paper that variance and variability related performance measures in scheduling, when exploited in assembly system, can reduce the arrival variability and improve the component release synchronization. We observe that minimizing the variability of component completion time difference within component production, and completion time variance for components that are necessary for a final product can significantly reduce the Average Staging Delay of components in front of assembly stage. Numerical experiments are carried out to support our findings and are discussed in this paper.

Keywords— assembly system; scheduling; staging delay; flow smoothening and synchronization

I. Introduction

The main focus of this paper is to present the insights towards reducing the staging delay occurring in the assembly systems involved in producing components required for the final product and assembly operation that combines all the required components into the final product or assembly. The assembly system can be viewed as a two-stage system in which the first stage consisting of flowshops independently producing components simultaneosly namely the stage 1. The second stage performs the assembly operation which combines the components required into a final product. The staging delay occurs in front of the assembly stage. Staging delay is the waiting time of the component waiting for the arrival of its counterparts necessary for the final product/assembly. Staging delay is also a constituent of the cycle time of an assembly system and hence it influences its performance [1]. Basically, the staging delay is caused by the deficiency of synchronization of component arrivals to the assembly stage [1]. By the manufacturing law from [1] it is well known that three factors, namely, number of components being assembled, component arrival variability with respect to the arrival of components to the assembly stage and lack of coordination between component arrivals, affect the performance of an assembly system. An increase in these factors generally degrades the performance of an assembly system. Our paper focuses on providing some useful insights for reducing the component arrival variability ensuring smoothness improvement in component flow and lack of coordination between component arrivals which reduces the staging delay. The synchronization can be improved by the reduction of variability of component arrivals to the assembly stage or planning the releases of the components with more or less identical completion times from the first stage to the assembly stage [1]. In the perspective of scheduling, these two factors are well represented by the scheduling performance measures such as Completion Time Variance (CTV) and Completion Time Difference Squared Variability (CTDVB2), discussed briefly in [5]. For reducing the component arrival variability and lack of coordination between component arrivals, the following performance measures are used: (1) CTV of the components necessary for a particular final product across the flowshops (this improves synchronization of component release), and (2) CTDVB2 of the components within each individual flow shops (this improves the smoothness of components flow). Minimizing the two aforementioned flow-based scheduling performance measures in an assembly system can ensure variability reduction (smoothness) in the component arrivals and completion time synchronization of the components necessary for the final product. Minimizing the staging delay of components in assembly systems is vital and essential in assembly systems where the holding cost for process finished components/goods inventory is much more than the holding cost for work-in-process inventory.

II. Literature Review

Hopp and Spearman [1] mentioned that the staging delay in assembly system is one of the constituent of total cycle time. The authors also explained that the component arrival variability and lack of synchronization can degrade the performance of assembly systems. Russell and Taylor [2] proposed some sequencing rules and used a simulation analysis of a hypothetical job shop to evaluate the sequencing rules considering mean staging delay as the criterion of interest. Adam et al. [3] examined some priority assignment procedures that focus on staging delay in the lead time of job. Reeja and

Rajendran discussed the importance of staging delay and some dispatching rules for its reduction in assembly job shops in [4]. Leisten and Rajendran discussed briefly the role of variance and variabilty related scheduling performance measures in flowshop scheduling in [5]. Reference [5] was the first one to discuss the variability in flowshop scheduling with CTDVB2 and methods to improve (minimize) the same. Section III briefly describes the staging delay in assembly systems and its reduction in the system considered. Section III also explains the operations of the assembly system and the proposed models used, with a simple demonstration. Section IV provides details about the computational experiments. Section V concludes with some useful insights.

III. Staging Delay in Assembly Systems and its Reduction

Before starting the discussion of staging delay, we ascertain that all the flowshops producing components simultaneously in stage 1 and the assembly stage process components in a sequence common to the system, called the Common Sequence.

We use the following notations:

N Number of final products
K Number of flowshops in stage 1 (also equal to the number of components necessary for every final product)
M Number of machines in each flowshop (identical for all flowshops)
i Component index $(1 \leq i \leq N)$ (also index for final product)
j machine index $(1 \leq j \leq M)$ (in each flowshop)
k flowshop index $(1 \leq k \leq K)$ (in stage 1)
$t_{i,j}^{k}$ processing time required for component for the final product i on machine j in flowshop k
$s_k[i]$ start time of the component on the last machine in flowshop k for the final product sequenced in position i in Common Sequence
$f_k[i]$ finish time of the component on the last machine in flowshop k for the final product sequenced in position i in Common Sequence
$p_k[i]$ pickup time of the component produced in flowshop k for the final product sequenced in position i in Common Sequence by the vehicle serving between stage 1 and stage 2
$d_k[i]$ delivery time of the component produced in shop k for the final product sequenced in position i in Common Sequence by the vehicle to the stage 2
$a[i]$ time of availability of all the components required for assembling them in to the final product sequenced in position i in Common Sequence
$\theta[i]$ staging delay of the components required for the final product sequenced in position i in Common Sequence
α average staging delay

$a[i]$ can be calculated using (1) as follows:

$$a[i] = \max_{1 \leq k \leq K} \{d_k[i]\} \quad (1)$$

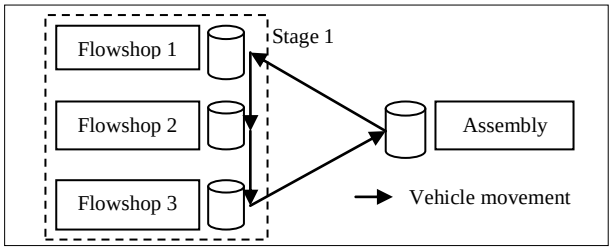
$\theta[i]$ can be calculated using (2) as follows:

$$\theta[i] = a[i] - \min_{1 \leq k \leq K} \{d_k[i]\} \quad (2)$$

A. Operation of Assembly System

Fig. 1 gives the schematic diagram of the assembly system considered in this paper with k=3. The movement of vehicle operating between the stage 1 and assembly stage is also shown in Fig. 1.

Fig. 1. Assembly system



The components are produced and assembled in the assembly system following the Common Sequence on all the flowshops in stage 1 and assembly stage. For finding the Common Sequence, we find the NEH sequence with the objective of total flowtime of components for each flowshop using the heuristic algorithm originally proposed for makespan objective by Nawaz et al. [6]. Also, we use increasing order of sum of processing times on all machines for the construction phase and first-tie rule for breaking the ties occuring in improvement phase in the algorithm. We simulate the assembly system with the NEH sequences for the individual flowshops with a static-deterministic-discrete event simulation model and choose the sequence that yields the minimum total flowtime of component production and their assembly i.e., total flowtime of components in the assembly system including the assembly operation as the Common Sequence. The number of components required for every single final product is identical. Hence the number of components required is equal to the number of flowshops in stage 1. A component can be produced in only one flowshop dedicated for its production and not in any other flowshop within stage 1. Components produced by the flowshops are stored in out-bound buffers of the respective flowshops. For the purpose of transporting the components from stage 1 to assembly stage, the assembly system uses a single vehicle which serves between the two stages as in milk-run concept. The vehicle starts from the in-bound buffer of assembly stage and visits each flowshop in stage 1 in a serial order, performs the pickup operation and returns back to the assembly stage and delivers the components in a single run. The serial order in which the vehicle visits the flowshops is maintained for all successive runs. In a single vehicle run, the vehicle picks up one component with the largest waiting time in the out-bound buffer of the flowshop and returns to the assembly stage. The vehicle does not wait at any flowshop for component arrival and it picks up the available component. If there is no component waiting in the out-bound buffer of a flowshop, the vehicle leaves without performing pickup operation. With respect to a flowshop in stage 1, the capacity of the vehicle is considered as equivalent to 1 component. The frequency of vehicle visiting the flowshops in stage 1 is

identical and constant for each run. The time between two successive visits of the vehicle for a given flowshop is fixed throughout the operation. Obviously, the time between two successive visits of the vehicle to assembly stage is also fixed.

The values of $s_k[i]$ and $f_k[i]$ are calculated by simulating the assembly system with the Common Sequence and they are used as parameters for the proposed models discussed in Section III. B. Actually, the values of $s_k[i]$ and $f_k[i]$ are used as the existing start time and completion time of component sequenced in position i in flowshop k.

The following algorithm is used to determine the frequency of vehicle visits to the flowshops and assembly stage, pickup time, delivery time and time of availability for the components:

Step 1: Calculate the difference between the completion times of the component produced in the last and first position of the Common Sequence on all the flowshops. Choose the difference with the largest value and let the same be represented as TD and calculated by (3).

$$TD = \max_{1\le k\le K}\{(f_k[N] - f_k[1])\} \quad (3)$$

For the purpose of simplicity, we assume that the time between two successive visits for a given flowshop and the time between two successive visits at assembly stage by the vehicle be identical and represented by IAT which is calculated by (4).

$$IAT = \left\lceil \left(\frac{TD}{(N-2)} \right) \right\rceil \quad (4)$$

Let, v_r^k be the visiting time of the vehicle at flow shop k for run r. Similarly, v_r^{asm} be the visiting time of the vehicle at the assembly stage for run r. For the first run, the visiting time of vehicle for flowshop 1 can be determined using (5).

$$v_1^1 = \min_{1\le k\le K}\{f_k[1]\} \quad (5)$$

Similary, the visiting time for other flowshops and the assembly stage for the first run can be determined using (6) and (7).

$$v_1^k = v_1^{k-1} + (IAT/(K+1)) \quad (6)$$

$$v_1^{asm} = v_1^K + (IAT/(K+1)) \quad (7)$$

Visiting times at each flowshop and assembly stage for the successive runs can be determined using (8) and (9).

$$v_r^k = v_{r-1}^k + IAT \quad (8)$$

$$v_r^{asm} = v_{r-1}^{asm} + IAT \quad (9)$$

Step 2: Determine $p_k[1]$, by comparing $f_k[1]$, with v_1^k. If $f_k[1]$ is less than or equal to v_1^k, then fix $p_k[1] = v_1^k$. Otherwise, compare $f_k[1]$ with v_2^k. Repeat this process of comparing until the condition $f_k[1] \le v_x^k$ is satisfied. Now fix $p_k[1] = v_x^k$.

Step 2a : In the same manner, determine pickup time for components sequenced in the successive positions in the Common Sequence such that $p_k[i] = v_x^k$ when $f_k[i] \le v_x^k$ and $v_x^k > p_k[i-1]$.

Step 3: Calculate $d_k[i]$ using (10) as follows:

$$d_k[i] = p_k[i] + \big((K-k+1) \times (IAT/(K+1))\big) \quad (10)$$

Step 4: Determine $a[i]$, $\theta[i]$ using (1) and (2).

Repeat steps 2a, 3 and 4 for all the N final products. Average Staging Delay (ASD) can be calculated using (12).

$$\alpha = \sum_{i=1}^{N} \theta[i] \Big/ N \quad (12)$$

B. Proposed Models

As mentioned earlier, we are interested in reducing the ASD in the actual status of the assembly system by using two proposed models. Model 1 and Model 2 described later in this paper are used for improving synchronization and reducing the arrival variability. In these two models, decision variables $f'_k[i]$ and $s'_k[i]$ are updated completion times and start times for component k required for product i respectively. In Model 1, decision variable λ'_i is the mean of updated completion times of k components required for product i. Model 2 is applied independently for each flowshop in stage 1.

Model 1: Minimization of the sum of CTVs of components required for the corresponding final products across all flowshops

Minimize

$$Z_1 = \sum_{i=1}^{N}\sum_{k=1}^{K} (f'_k[i] - \lambda'_i)^2 \quad (13)$$

subject to

$$f'_k[N] = f_k[N]\ (1 \le k \le K) \quad (14)$$

$$s'_k[N] = s_k[N]\ (1 \le k \le K) \quad (15)$$

$$f'_k[i] \ge f_k[i]\ \ (1 \le i < N)\ (1 \le k \le K) \quad (16)$$

$$s'_k[i] \ge s_k[i]\ \ (1 \le i < N)\ (1 \le k \le K) \quad (17)$$

$$\sum_{k=1}^{K} f'_k[i] = K \times \lambda'_i \qquad (1 \le i \le N) \quad (18)$$

$$f'_k[i] = s'_k[i] + t_{i,M}^k\ \ (1 \le i \le N), (1 \le k \le K) \quad (19)$$

$$s'_k[i+1] \ge f'_k[i]\ \ (1 \le i < N)\ \ (1 \le k \le K) \quad (20)$$

Model 2: CTDVB2 improvement within flowshop k (applied independently in each flowshop)

for k=1(1)K :

{

Minimize

$$Z_{2,k} = \sum_{i=1}^{(N-1)} (f'_k[i+1] - f'_k[i])^2 \quad (21)$$

subject to

$$f'_k[N] = f_k[N] \quad (22)$$

$$s'_k[N] = s_k[N] \quad (23)$$

$$f'_k[i] \ge f_k[i]\ (1 \le i < N) \quad (24)$$

$$s'_k[i] \ge s_k[i]\ (1 \le i < N) \quad (25)$$

$$s'_k[i] \ge f_k[i-1]\ (1 < i \le N) \quad (26)$$

$$s'_k[i] \geq f'_k[i-1]\ (1 < i \leq N) \quad (27)$$

$$f'_k[i] = s'_k[i] + t^k_{i,M}\ (1 \leq i \leq N) \quad (28)$$

}

Equation (13) represents the sum of CTVs of components produced across the flowshops required for all the final products. Equations (14) and (15) ensure that the updated completion times and start times of component sequenced in the last position in the Common Sequence in all the flowshops are unaltered. Equations (16) and (17) ensure that updated completion and start times of all components except the component sequenced in the last position in the Common Sequence are greater than or equal to the existing completion and start times of components in all the flowshops. Equation (18) ensure that the sum of k updated completion times for components is equal to the product for their mean and the number of flowshops k in stage 1 (or components required) for a particular final product i. Equation (19) ensures that the updated completion time of components is equal to their sum of corresponding updated start time and processing time in all flowshops. Equation (20) makes the updated start time of a component processing greater or equal to the updated completion of the component sequenced in previous position with respect to the Common sequence in all flowshops. Equation (21) represents the CTDVB2 of components produced in flowshop k. Reference [5] was the first to introduce variability related performance measures in flowshop scheduling and also mentions that allowing updating in the first and last positions makes the model more complex to solve. If we allow for updation in all positions except for last position in all the flowshops, the improvement of CTDVB2 might be better, the same is also shown in [5]. Allowing updation in first position encourages for further improvement of CTDVB2 and makes both Model 1 and Model 2 more comparable than fixing first and last position completion times in all flowshops. Equations (22) and (23) ensures that updation of completion time is not allowed in the last position in all the flowshops. Equations (24) and (25) ensure that the updated completion times and start times of components can be greater than or equal to the actual values respectively except for the last position. Equation (26) ensures that updated start time of a component is greater than or equal to the existing completion time of the component sequenced in the previous position in flowshop k. Equation (27) makes the updated start time of a component processing greater or equal to the updated completion of the component sequenced in the previous position in flowshop k. Equation (28) ensures that the updated completion time of components is equal to their sum of corresponding updated start time and processing time in all flowshops.

C. A Discussion on the Performance of Proposed Models with respect to Staging Delay being related to Smoothening and Synchronization of Component Flow

We now demonstrate the contribution towards the process of reduction of ASD of components necessary for final products in the assembly systems by the reduction of component arrival variability to the assembly stage and improving the synchronization level of the components' completion times necessary for every final product. This can be achieved by deploying the proposed models Model 1 and Model 2 presented in section III. B. Model 1 aims at reducing the sum of completion time variances of the components required for corresponding final products across the flowshops producing them. By this way, Model 1 intends to improve the synchronization of the components' releases towards the assembly stage buffer across the flowshops for each final product. Model 2 aims at the improvement of the completion time differences squared variability (with the fixed completion times of components sequence in the first and last position of the Common Sequence) of the components produced within each flowshop which inherently reduces the variability of components arrivals to the assembly stage. For illustration, let us apply the proposed heuristic algorithm and proposed models to problem with N=7, K=3 and M=5. Table I provides the processing time details for the flowshops in stage 1. The time required for assembly operation of a particular final product is the largest of the processing times required for components in respective the last machine of the flowshop. For instance, in Table I the time required for assembly operation for the final product 1 can be taken as 36 which is the maximum among the processing times 36 (component produced in flowshop 1), 25 (component produced in flowshop 2) and 6 (component produced in flowshop 3) required for its components in the last machine of the respective flowshops.

TABLE I. Processing times of components in flowshops 1, 2 and 3 in stage 1

			Component i						
			1	2	3	4	5	6	7
Machine j	1	$k=1$	55	8	39	97	53	76	57
		$k=2$	53	30	2	69	12	1	88
		$k=3$	40	70	39	28	41	39	8
	2	$k=1$	22	92	24	24	38	95	37
		$k=2$	34	3	93	21	94	99	84
		$k=3$	19	22	80	40	48	96	32
	3	$k=1$	76	67	44	85	72	63	61
		$k=2$	90	39	71	70	14	47	83
		$k=3$	7	79	49	53	1	56	34
	4	$k=1$	99	14	96	71	91	21	59
		$k=2$	3	67	2	99	1	65	43
		$k=3$	89	53	93	71	54	13	88
	5	$k=1$	36	90	50	72	20	63	25
		$k=2$	25	20	67	1	75	34	77
		$k=3$	6	71	10	32	73	65	95

Sequence 2-5-4-1-3-6-7 found by the NEH based approach discussed in Section III. A, is considered as the Common Sequence for processing components and their assembly operation for the problem with processing time data given in Table I. The objective function values of the proposed models are also calculated. The actual assembly system's completion times obtained by the simulation model are taken as the actual system status with existing start times $s_k[i]$ and completion

times $f_k[i]$. The actual system status is now improved for reducing the average staging delay with the help of proposed models. The proposed models provide updated start times $s'_k[i]$ and completion times $f'_k[i]$ for the schedule of flowshops in stage 1. The average staging delay is calculated from the updated completion times using the algorithm given in Section III. A. By intuition, we can also test the improvements obtained by the proposed models with performing right shifting of the schedule (right shifting performed for alignment of makespan of schedules in all the flowshops) in all the non-bottleneck flowshops with respect to the completion time of component processed in the last position of the Common Sequence in flowshops. The flowshop which possesses the largest completion time for component sequenced in the last position is considered as the bottleneck flowshop and the value of largest component completion time is taken as the limit for right shifting (on the completion time of component sequenced in the last position) of components processing in all the other flowshops (non-bottleneck flowshops). We believe that right shifting in non-bottleneck flowshops can sometimes help in achieving a better synchronization of components release from stage 1 to the assembly stage even after improving the assembly system performance by applying proposed models. But applying right shift for the updated component completion times obtained by Model 1 is not desirable as doing so can worsen the component release synchronization eventually resulting in the increase of staging delay already achieved by Model 1.

Table II reports the summary of ASD values for the problem considered with processing time data given in Table I. The important observation is that the CTDVB2 improvement within each flowshop using Model 2 sometimes does not achieve the reduction in ASD because the components sequenced in the first and last positions are excluded from the updating of completion and start times which achieves a relatively less amount of variability reduction than the variability reduction level required for reduction of average staging delay. But, CTDVB2 improvement using Model 2, sometimes followed by the right shift, offers a substantial reduction in the ASD as the right shift offers a better synchronization of component completion times across the flowshops. CTV improvement of completion times of components necessary for a final product across the flowshops using Model 1 achieves a better synchronization of component release towards the assembly stage. Unlike Model 2 with right shift, the Model 1 with right shift inflates ASD because right shift reduces the synchronization level already achieved by Model 1 and hence, Model 1 with right shift is undesirable and we will not discuss the same in this paper. The strategy of right shifting the actual completion times with respect to the bottleneck flowshop and then improving the CTV of components necessary for a final product using Model 2 might help us in gaining more component release synchronization level. But this strategy is not discussed in this paper.

IV. Computational Experiments

The proposed methodology for reducing average staging delay is tested with large N. Since we have considered flowshops in stage 1 in assembly system for component production, the new hard flowshop problem instances proposed by Vallada et al. [7] are used. We assume K=3. The (100 job-20 machine), (100 job-40 machine), (100 job-60 machine), (800 job-20machine) and (800 job-40 machine) problem instances are selected for computational experiments. Table III presents the average staging delay values obtained for actual, after improvements and after improvements along with right shift for (100 job-20 machine) and (800 job-40 machine) problem instances. Problem instance 100_20_1-2-3 in Table III represents the assembly system problem with N=100, M=40, problem instance 1 for flowshop 1, problem instance 2 for flowshop 2 and problem instance 3 for flowshop 3. The numerical comparison of ASD for existing NEH schedule and the proposed improvements through Model 1 and Model 2 with and without right shift are presented in https://www.dropbox.com/s/vo8ftp1bouswofl/Saravana_Dropbox_IESM_GERMANY_2017.pdf?dl=0.

TABLE II. Performance of proposed models with respect to ASD for the problem with processing time data in Table I.

Assembly system status		ASD
Existing NEH schedule		176.00
Model 1		57.14
Model 2	With out right shift	176.00
	With right shift	88.00

V. Conclusion

Through the computational experiments, it is observed that Model 1 is relatively more efficient in reducing the ASD than the Model 2 discussed in this paper because Model 1 aims at reducing the variance of completion times for the components necessary for every final product which ensures a nearly perfect component release synchronization of components across the flowshops in the assembly systems. Model 2 and Model 2 with right shift in a few cases inflates the ASD with respect to the existing NEH schedule. In many cases Model 2 with right shift reduces the ASD as Model 2 reduces the component arrival variability within each flowshop and a right shift of component completion times of non-bottleneck flowshops with respect to bottleneck flowshop further offers some synchronization level for component release. Therefore, Model 1 is relatively concluded as the best approach for reducing ASD in assembly systems.

The reason why Model 1 performs well is that, Model 1 ensures both component release synchronization and component arrival variability reduction in assembly systems simultanoeusly. Model 2 can ensure a better uniformity level in the component arrival to the assembly stage but with a lack of synchronization of components necessary for a final product across the flowshops. Right shifting of component completion times of non-bottleneck flowshops with respect to bottleneck flowshop in Model 2 can ensure component release synchronization but not the component harmonization level as ensured by Model 1. Model 1 can ensure more synchronization of components than Model 2 as Model 1 aims at reducing CTV across flowshops but, Model 2 is basically focussed within flowshops. We can also conclude that right shifting and then

applying Model 1 might increase the synchronization and harmonization level across the flowshops for reducing ASD in assembly systems but not discussed in this paper.

TABLE III. Comparison of ASD in existing status and after the proposed improvements using Model 1 and Model 2 and improvements along with right shift for (100 job-20 machine) and (800 job-40 machine) problem instances

Problem Instances	Existing NEH schedule	Model 1	Model 2	Model 2 with right shift
100_20_1-2-3	608.40	55.76	369.72	161.72
100_20_2-3-4	664.36	138.24	706.16	161.20
100_20_3-4-5	620.16	8.16	572.52	260.00
100_20_4-5-6	625.60	15.36	554.88	401.28
100_20_5-6-7	692.64	60.52	920.16	442.08
100_20_6-7-8	828.24	61.44	1211.52	515.04
100_20_7-8-9	653.48	39.68	906.24	473.28
100_20_8-9-10	668.44	55.04	1045.44	255.84
100_20_9-10-1	634.44	13.60	185.28	415.20
100_20_10-1-2	506.16	51.68	733.20	419.64
800_40_1-2-3	1914.38	99.54	2500.62	2001.54
800_40_2-3-4	1885.57	311.63	5330.94	2246.94
800_40_3-4-5	1809.75	247.88	5315.04	2342.34
800_40_4-5-6	1818.60	224.55	3948.10	727.35
800_40_5-6-7	1455.00	658.77	4191.46	1122.09
800_40_6-7-8	1748.78	860.79	3590.41	418.41
800_40_7-8-9	1720.88	54.53	4204.85	1140.10
800_40_8-9-10	1739.55	21.00	4204.72	1140.36
800_40_9-10-1	1893.82	88.65	5659.98	1724.04
800_40_10-1-2	2068.88	194.18	5625.78	2065.98

Acknowledgment

The first author wants to thank the DAAD exchange programme for supporting this work during June-December 2015.

References

[1] W.J. Hopp and M.L. Spearman, Factory Physics, 3rd ed., New York: McGraw-Hill, 2008.

[2] R.S. Russell and B.W. Taylor III, "An Evaluation of Sequencing rules for an Assembly Shop", Decision Sci, vol. 16, 1985, pp. 196-212.

[3] N.R. Adam, J.W.M. Bertrand and J. Surkis, "Priority Assignment Procedures in Multi-level Assembly Jobshops", IIE Trans, vol. 19, 1987, pp. 317-328.

[4] M.K. Reeja and C. Rajendran, "Dispatching Rules for Scheduling in Assembly Job shop - Part I", Int J Prod Res, vol. 38, 2000, pp. 2051-2066.

[5] R. Leisten and C. Rajendran, "Variability of Completion Time Differences in Permutation Flow shop Scheduling", Comput Oper Res, vol. 54, 2015, pp. 529-551.

[6] M. Nawaz, E.E. Enscore and I. Ham, "A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem", Omega-Int J Manage S, vol. 11, 1983, pp. 91-95.

[7] E. Vallada, R. Ruiz and J.M. Framinan, "New hard benchmark for flowshop scheduling problems minimizing makespan", Eur J Oper Res, vol. 240, 2015, pp. 666-677.

A MILP approach for robust transfer line design

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Aleksandr Pirogov*, Evgeny Gurevsky†, André Rossi‡ and Alexandre Dolgui*
* LS2N, IMT Atlantique, Nantes, France
† LS2N, Université de Nantes, France
‡ LERIA, Université d'Angers, France

***Abstract*—We address the design stage of transfer lines subject to precedence constraints, fixed cycle time, limited number of machines and uncertain tasks. The objective is to build the line configuration, so as to maximize its robustness, i.e., its ability to withstand task time variability. To do this, we formulate a mixed-integer linear program (MILP), where the concept of stability radius is used as robustness measure and calculated in two classic metrics, l_1 and l_∞. To handle these two MILP models, the commercial solver CPLEX 12.6.2 is used. The corresponding computational results are reported and discussed at the end of this paper.**

***Keywords*—transfer lines, robust optimization, stability radius**

I. Introduction

Transfer lines are automated flow-oriented manufacturing systems, dedicated to yield a large quantity of product and usually designed for a period of 7-10 years (7 days a week and 24 hours a day) of functioning. The machines of such a line are organized in a linear manner, connected by a conveyor belt without intermediate buffers. Each of them is equipped by one or several blocks, where production tasks are executed in parallel using multi-spindle heads. We consider the lines where the blocks of any machine are activated sequentially in a predefined order, one by one without significant interruptions. As concerns production parts, they start their execution, in a raw state, on the first machine and pass through all of them in the order of their location so as to be turned into a finished state. The machines work simultaneously, performing their own and same set of tasks on successive parts. The machine completes its execution on the current part when all its blocks have been activated and have accomplished their operation. Since the considered lines are paced, all parts situated on the conveyor belt are transferred simultaneously to the next corresponding machines, when the functioning of these latter are completed. The time period between two successive transfers is called the *cycle time*, which is equal to the working time of the most loaded machine and determines, among others, the moment for one outgoing and new one incoming parts.

As regards the production tasks, they can not be executed in an arbitrary manner, but subject to *precedence constraints*, usually described by a directed acyclic graph, where the set of vertexes (resp. arcs) represents the set of tasks (resp. a partial order of tasks execution). Thus, for example, the arc (i, j) means that the task j can not be performed before the task i, but i and j can be executed simultaneously inside a block.

The design of such a line can be usually viewed as an optimization problem, which, with respect to a given production goal and subject to known industrial constraints, consists in assigning all production tasks to blocks and allocate these latter to machines, determining inside them the order of blocks activation. The first works, appeared in international refereed journals and dealing with the representation of the design stage for transfer lines as optimization problem, were published by Dolgui et al. in [4]–[6], where the authors consider a cost-oriented objective function aiming to minimize a weighted sum of the numbers of used machines and blocks, respectively. These pioneer journal articles have finally motivated many other studies on this problem, among which we can cite the works of Battaïa et al. on reductions approaches in [2], [3], Guschinskaya et al. on meta-heuristic methods in [8], [9], and Dolgui and Ihnatsenka on a branch-and-bound algorithm in [7].

For all mentioned above papers, the task processing time was considered to be deterministic and the number of available machines was supposed to be unlimited, but in reality this is not necessarily truthful. Thus, for example, production space restrictions often impose the limits on the number of used machines. As for task processing times, they can be modified from their nominal values because of numerous delays and micro-stoppages during manufacturing or because of the changes of product specification required by customer. In this work, we show how to integrate these two important industrial aspects into the design stage of transfer lines through an original robust approach, which has a significant advantage for its application with respect to classic *stochastic* [1] or *fuzzy* [10] methods, consisting in the fact that there is no need to possess reliable historical data on the variability of task processing times. Namely, we model our purpose as optimization problem within the framework of mixed-integer linear programming (MILP), seeking a transfer line configuration with fixed number of machines, which is able to support, without violating its admissibility, the greatest amplitude of task time variability. The numerical value of this amplitude is customary called in the literature as the *stability radius* and usually calculated for two classic metrics, l_1 and l_∞. As in our previous paper [11], dealing with assembly lines, in this article, we continue to demonstrate that the stability radius can be also used as a useful and veritable measure of solution robustness, contrary to the popular works on scheduling problems with uncertain jobs, published in [12], where it was studied as an indicator of sensitivity for already known optimal solutions.

The rest of the paper is organized as follows. Section II describes the main notations of the studied optimization problem and formalizes the notion of stability radius. Section III introduces the necessary decision variables and presents a MILP model for each metric of stability radius computing. Computational results, provided by the commercial solver

CPLEX 12.6.2, are given in Section IV. Conclusion and perspectives are delivered in Section V.

II. Problem statement

Below, we introduce some useful notations, which are necessary to the further formalization of the studied optimization problem.

- $V = \{1, 2, \ldots, n\}$ is the set of necessary tasks;
- $\widetilde{V} \subseteq V$ is the set of all uncertain tasks, whose processing time can be modified during manufacturing;
- $W = \{1, 2, \ldots, m\}$ is the set of available machines;
- t_j is the nominal processing time of the task $j \in V$;
- ξ_j is a non-negative variation of the processing time of the uncertain task $j \in \widetilde{V}$;
- ρ_1 (resp. ρ_∞) is the value of stability radius, computed in the metric l_1 (resp. l_∞);
- T is the cycle time;
- $r_{\max}$ is the maximal number of tasks, which can be assigned to block;
- $b_{\max}$ is the maximal number of blocks, which can be allocated into one machine;
- $G = (V, A)$ is a directed acyclic graph representing the precedence constraints, where A is the set of arcs.

The value of $b_{\max}$ is not given *a priori*, but it can not be greater than $\max\left\{k \mid \sum_{i=1}^{k} t_{\pi_i} \leq T\right\}$, where π_i is the task index of the permutation of V with respect to the non-decreasing order of their processing times. The latter expression corresponds to the case of tasks allocation to machine, without violating the cycle time and requiring the greatest number of blocks. In what follows, we will use this upper bound as the value of $b_{\max}$.

The knowledge of $b_{\max}$ permits us to simplify the representation of line configuration as shown in Figure 1, where

- $U = \{1, 2, \ldots, mb_{\max}\}$ is the set of indices of all potential blocks;
- $U(p) = \{(p-1) \cdot b_{\max} + 1, \ldots, p \cdot b_{\max}\}$ is the set of indices of all blocks potentially belonging to the machine p, $p \in W$.

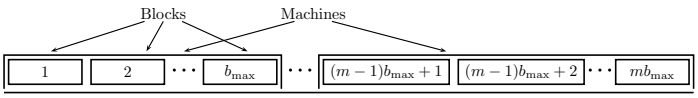


Fig. 1: Enumeration of the set of blocks

Based on the introduced notations, the studied problem can be formally viewed as the assignment of all tasks to blocks on the pre-defined line configuration, shown in Figure 1, subject to the cycle time and precedence constraints, which maximizes the value of ρ_1 (resp. ρ_∞) and preserves the assignment admissibility, whatever the total amplitude of uncertain tasks variations $\sum_{j \in \widetilde{V}} \xi_j$ (resp. $\max_{j \in \widetilde{V}} \xi_j$) not exceeding ρ_1 (resp. ρ_∞).

Let us consider an illustrative example, shown in Figure 2, where an allocation of the set of tasks $\{1, 2, \ldots, 6\}$ is given. For the presented example, the cycle time is equal to 8.5, indicated by a vertical dotted line, there are no precedence constraints and only one machine having two blocks is used. The first block contains the tasks 1, 2 and 5, while the second block is made of the tasks 2, 4 and 6. Each block is contoured by a dashed line. The tasks are represented by rectangles, whose length determines the corresponding processing time. The uncertain tasks $\{1, 4, 5\}$ are colored in gray. It is not difficult to see that for any total variation $\xi_1 + \xi_4 + \xi_5 \leq 2$, the allocation remains admissible, but a slightly greater value, where, for example, $\xi_1 = 2 + \delta$, $\xi_4 = 0$, $\xi_5 = 0$, violates its admissibility. This means that the stability radius in the l_1-metric equals 2. Regarding the l_∞-metric, it is easy to see that the considered allocation supports any variation ξ_1, ξ_4, and ξ_5 not exceeding 1.5, however it is not especially true for little bigger values, where, for example, $\xi_1 = 1.5 + \delta$, $\xi_4 = 1.5 + \delta$, $\xi_5 = 0$. The latter, following the definition of stability radius, indicates that $\rho_\infty = 1.5$.

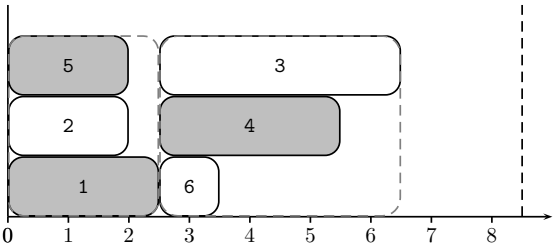


Fig. 2: Example of tasks allocation

III. MILP formulations

In this section, the MILP models for two metrics l_1 and l_∞, noted respectively by P_1 and P_∞, are described.

P_1 is formulated on the following decision variables:

- $\rho_1 \geq 0$ is the stability radius value to maximize;
- $x_{j,k}$ is a binary variable that is set to one if and only if the task j is allocated to the block k;
- y_k is equal to 1 if the block k is not empty and 0, otherwise;
- $\tau_k \geq 0$ determines the time of the block k;
- $\Delta_p \geq 0$ represents the minimal value of *save time* among all the blocks allocated into the machine p;
- a_p is a non-negative variable that is positive if the machine p has at least one assigned uncertain task.

The notions of block time and block save time are explained in Figure 3, where the block save time corresponds to the difference between block time and the longest processing time among the uncertain tasks assigned to this block. Thus, in Figure 3, the save time is equal to 1.5 for this first block and 0 for the second one.

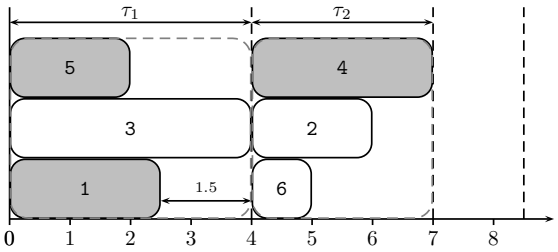


Fig. 3: Illustration of block time and block save time

An important result, simplifying the modeling of P_1, can be expressed by the following theorem.

Theorem 1. *The stability radius ρ_1 is calculated as follows*

$$\rho_1 = \min_{p \in \widetilde{W}} \left(T - \sum_{k \in U(p)} \tau_k + \Delta_p \right), \tag{1}$$

where $\widetilde{W}$ is a set of machines having at least one uncertain task.

P_∞ is formulated on the variables $x_{j,k}$, y_k and τ_k with the same sense as for P_1 and the following additional ones:

- $\rho_\infty \geq 0$ is the stability radius value to maximize;
- $\xi_{j,k}$ is the variation of the task j on the block k.

A. MILP formulation for P_1

Objective function:

$$\text{Maximize } \rho_1$$

Each task has to be allocated to exactly one block:

$$\sum_{k \in U} x_{j,k} = 1, \quad \forall j \in V \tag{2}$$

No more than $r_{\max}$ tasks can be included to block :

$$\sum_{j \in V} x_{j,k} \leq r_{\max}, \quad \forall k \in U \tag{3}$$

The block having at least one assigned task is considered as not empty, otherwise empty:

$$x_{j,k} \leq y_k, \quad \forall k \in U, \ \forall j \in V \tag{4}$$

$$y_k \leq \sum_{j \in V} x_{j,k}, \quad \forall k \in U \tag{5}$$

Any block has to be empty, if the previous block on the same machine is empty:

$$y_{k+1} \leq y_k, \quad \forall p \in W, \ \forall k \in U(p) \setminus \{pb_{\max}\} \tag{6}$$

The block time is the maximal processing time over the tasks allocated to this block:

$$t_j \cdot x_{j,k} \leq \tau_k, \quad \forall k \in U, \ \forall j \in V \tag{7}$$

The definition of the minimal save time for each machine:

$$\Delta_p \leq T \cdot (1 - y_k) + \tau_k - t_j \cdot x_{j,k}, \quad \forall p \in W, \ \forall k \in U(p), \ \forall j \in \widetilde{V} \tag{8}$$

The load of any machine can not be greater than the cycle time:

$$\sum_{k \in U(p)} \tau_k \leq T, \quad \forall p \in W \tag{9}$$

Precedence constraints:

$$\sum_{q=k}^{|U|} x_{i,q} \leq \sum_{q=k}^{|U|} x_{j,q}, \quad \forall (i,j) \in A, \ \forall k \in U \setminus \{1\} \tag{10}$$

Constraint (11) sets a_p according to its definition, and a_p is used in constraint (12) to compute ρ_1 by Theorem 1:

$$x_{j,k} \leq a_p, \quad \forall p \in W, \ \forall k \in U(p), \ \forall j \in \widetilde{V} \tag{11}$$

$$\rho_1 \leq T \cdot (2 - a_p) - \sum_{k \in U(p)} \tau_k + \Delta_p, \ \forall p \in W \tag{12}$$

To reduce the search space, it is possible to improve this model using the so-called assignment intervals:

$$x_{j,k} = 0, \quad \forall j \in V, \ \forall k \notin Q(j) \tag{13}$$

where $Q(j) = [l_j, u_j]$ is the interval of indexes of the blocks able to accommodate the task j, $j \in V$. Here,

$$l_j = \left\lceil \frac{|P(j) \cup \{j\}|}{r_{\max}} \right\rceil,$$

$$u_j = mb_{\max} + 1 - \left\lceil \frac{|S(j) \cup \{j\}|}{r_{\max}} \right\rceil,$$

$P(j)$ and $S(j)$ are the sets of predecessors and successors of the task j in the precedence graph, respectively.

B. MILP formulation for P_∞

The main idea of the MILP formulation for P_∞ is focused on the fact that the processing time of all uncertain tasks can be increased by ρ_∞ without loosing the feasibility for the optimal solution.

Constraints (2) – (6), (9) and (10) are the same for models P_∞ and P_1. The difference is expressed by constraints (14) – (17).

Objective function:

$$\text{Maximize } \rho_\infty$$

The block time is determined separately for blocks with uncertain tasks and without them:

$$t_j \cdot x_{j,k} + \xi_{j,k} \leq \tau_k, \quad \forall k \in U, \ \forall j \in \widetilde{V} \tag{14}$$

$$t_j \cdot x_{j,k} \leq \tau_k, \quad \forall k \in U, \ \forall j \in V \setminus \widetilde{V} \tag{15}$$

Variation of any uncertain task can not exceed the cycle time value, whatever its allocation:

$$\xi_{j,k} \leq T \cdot x_{j,k}, \quad \forall k \in U, \ \forall j \in \widetilde{V} \tag{16}$$

As mentioned above, the processing time of any uncertain task can be increased by the value of ρ_∞:

$$\rho_\infty = \sum_{k \in U} \xi_{j,k}, \quad \forall j \in \widetilde{V} \tag{17}$$

IV. Computational results

A set of 12 randomly generated benchmark instances has been used to examine the MILP model of P_1 and P_∞, where we distinguish 5 small and 7 medium size of them. Their main parameters are given in Table I, whose columns comprise the instance number $\#$, its number of tasks n, number of fixed machines m, the cycle time T and the order strength of the precedence constraints, $OS = \frac{2|A|}{n(n-1)}$, represented by the graph $G = (V, A)$ and expressed in percents.

TABLE I: Benchmark instances

$\#$	n	m	T	$OS, \%$
Small size				
1	8	4	25.5	75.00
2	11	6	10.5	58.18
3	9	5	9	83.33
4	11	4	67.5	60.00
5	7	4	9	52.38
Medium size				
6	29	11	37.5	50.74
7	28	8	162	22.49
8	21	7	19.5	70.95
9	32	9	2100	83.47
10	25	8	19.5	71.67
11	30	11	37.5	44.83
12	35	10	60	59.90

The computational results were carried out on a MacBook Pro having Intel Core i5 2.7 GHz and 8 GB RAM. The commercial solver CPLEX 12.6.2 was used through the Concert Technology C++ modeling layer for addressing the MILP models.

For each benchmark instance, six tests were generated by varying the number of uncertain tasks $|\widetilde{V}| \in \{\lceil \frac{n}{4} \rceil, \lceil \frac{n}{2} \rceil\}$ and the maximal number of tasks per block $r_{\max} \in \{1, 2, 3\}$. The results for each series are reported respectively in Tables II–VII, where for each metric, three columns indicate respectively the lower and upper bounds on the optimal value of stability radius, provided by CPLEX, as well as the corresponding CPU time (sec.), which is less than 600 sec., if an optimal solution was found before reaching this test time limit. In the case, where CPLEX was not able to find either lower or upper bound, the abbreviation "**n/a**" is put in the corresponding cell of the Table.

Analyzing the obtained results, we can remark that an optimal solution was found for all series of small size benchmark instances for P_1 and P_∞ within the limits of 2.51 sec. As concerns the medium size benchmark instances, the situation is different. Thus, for P_1, an optimal solution was found only for the $8th$ instance of each series, except the last one. For P_∞, we can note some improvements, but the necessary time for finding optimal solutions remains quite high. Another interesting observation is the greater the $r_{\max}$ value, the easier P_∞ to solve. The summary of found feasible and optimal solutions is given in Table VIII.

TABLE II: $|\widetilde{V}| = \lceil \frac{n}{4} \rceil$ and $r_{\max} = 1$

$\#$	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	4.5	4.5	0.02	4.5	4.5	0.04
2	1.5	1.5	0.29	1.5	1.5	0.31
3	1	1	0.03	1	1	0.03
4	41.5	41.5	1.07	21.5	21.5	1.04
5	0	0	0.03	0	0	0.01
6	1.5	74.6	600	6.75	12.5	600
7	40	409.9	600	33	55	600
8	6.5	6.5	175.7	4.25	4.25	321.57
9	148	4091.6	600	400	700	600
10	4.5	13	600	3.25	6.5	600
11	2.5	87.9	600	5.75	18.5	600
12	**n/a**	**n/a**	600	**n/a**	**n/a**	600

TABLE III: $|\widetilde{V}| = \lceil \frac{n}{4} \rceil$ and $r_{\max} = 2$

$\#$	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	13.5	13.5	0.09	13.5	13.5	0.02
2	4.5	4.5	1.24	4.5	4.5	0.13
3	3	3	0.05	3	3	0.04
4	55.5	55.5	1.61	55.5	55.5	0.3
5	4	4	0.06	4	4	0.01
6	12.5	74.6	600	12.5	12.5	55.99
7	55	444.8	600	55	92	600
8	10.5	10.5	243.97	10.5	10.5	10.01
9	454	4200	600	700	700	315.56
10	6.5	13	600	6.5	6.5	7.84
11	8.5	102.3	600	18.5	18.5	34.65
12	**n/a**	**n/a**	600	3	60	600

TABLE IV: $|\widetilde{V}| = \lceil \frac{n}{4} \rceil$ and $r_{\max} = 3$

$\#$	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	13.5	13.5	0.11	13.5	13.5	0.02
2	4.5	4.5	0.66	4.5	4.5	0.14
3	3	3	0.06	3	3	0.03
4	55.5	55.5	1.75	55.5	55.5	0.2
5	4	4	0.1	4	4	0.02
6	12.5	75	600	12.5	12.5	43.45
7	55	324	600	55	92	600
8	10.5	10.5	213.28	10.5	10.5	2.67
9	454	5719.1	600	700	700	98.87
10	6.5	37.3	600	6.5	6.5	10.42
11	14.5	75	600	18.5	18.5	14.62
12	**n/a**	**n/a**	600	1	174.4	600

TABLE V: $|\widetilde{V}| = \lceil \frac{n}{2} \rceil$ and $r_{\max} = 1$

$\#$	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	4.5	4.5	0.02	2.25	2.25	0.02
2	1.5	1.5	0.37	1.25	1.25	0.57
3	1	1	0.03	0.5	0.5	0.03
4	23.5	23.5	1.11	12.75	12.75	2.01
5	0	0	0.04	0	0	0.03
6	3.5	75	600	3.17	12.5	600
7	**n/a**	**n/a**	600	9.75	55	600
8	3.5	3.5	453.08	1.75	6.2	600
9	84	4599.7	600	80	700	600
10	1.5	12.5	600	3.17	12.5	600
11	3.5	74.3	600	3.75	12.5	600
12	**n/a**	**n/a**	600	**n/a**	**n/a**	600

V. Conclusion

This paper deals with finding a transfer line configuration having the greatest stability radius subject to various industrial constraints. The stability radius is evaluated in both l_1 and l_∞

TABLE VI: $|\widetilde{V}| = \lceil \frac{n}{2} \rceil$ and $r_{\max} = 2$

#	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	13.5	13.5	0.06	13.5	13.5	0.04
2	4.5	4.5	1.65	4.5	4.5	0.14
3	3	3	0.07	3	3	0.05
4	29.5	29.5	1.93	29.5	29.5	0.78
5	3	3	0.05	3	3	0.03
6	3.5	75	600	12.5	12.5	335.45
7	47	429.8	600	49.5	55	600
8	6.5	6.5	167.47	6.5	6.5	22.22
9	242	5369.6	600	700	700	250.02
10	6.5	13	600	6.5	6.5	9.46
11	6.5	75	600	12.5	18.5	140.05
12	**n/a**	**n/a**	600	**n/a**	**n/a**	600

TABLE VII: $|\widetilde{V}| = \lceil \frac{n}{2} \rceil$ and $r_{\max} = 3$

#	LB_1	UB_1	CPU	LB_∞	UB_∞	CPU
1	13.5	13.5	0.1	13.5	13.5	0.02
2	4.5	4.5	1.76	4.5	4.5	0.31
3	3	3	0.07	3	3	0.05
4	29.5	29.5	2.51	29.5	29.5	0.32
5	3	3	0.04	3	3	0.02
6	12.5	75	600	12.5	12.5	117.42
7	55	324	600	55	55	190.39
8	6.5	7.5	600	6.5	6.5	18.52
9	700	4200	600	700	700	89.29
10	6.5	36	600	6.5	6.5	21.27
11	5.5	102.6	600	12.5	12.5	39.15
12	1	174.2	600	4	20	600

TABLE VIII: Number of found solutions

$\lvert\widetilde{V}\rvert$	$\lceil \frac{n}{4} \rceil$			$\lceil \frac{n}{2} \rceil$		
$r_{\max}$	1	2	3	1	2	3
Feasible solutions for P_1	11	11	11	10	11	12
Optimal solutions for P_1	6	6	6	6	6	5
Feasible solutions for P_∞	11	12	12	11	11	12
Optimal solutions for P_∞	6	11	11	5	10	11

metrics. For each metric, the corresponding MILP model was elaborated. Numerical results show that the used commercial solver can find an optimal solution for small size instances in less than 3 sec., but unfortunately not able to provide satisfactory outcomes for medium size ones.

The proposed MILP models are a first attempt for applying robust optimization approaches for transfer line design. The second natural step of our future research is a detailed analysis of the studied problem in order to find its new structural and combinatorial properties permitting to develop more efficient resolution methods, based on appropriate reduction rules.

ACKNOWLEDGEMENT

This work was partially supported by the council of the french region “Pays de la Loire”.

REFERENCES

[1] K. Ağpak and H. Gökçen, *A chance-constrained approach to stochastic line balancing problem*, European Journal of Operational Research, 180(3): 1098–1115, 2007.

[2] O. Battaïa and A. Dolgui, *Reduction approaches for a generalized line balancing problem*, Computers & Operations Research, 39(10): 2337–2345, 2012.

[3] O. Battaïa, E. Gurevsky, F. Makssoud and A. Dolgui, *Equipment location in machining transfer lines with multi-spindle heads*, Journal of Mathematical Modelling and Algorithms in Operations Research, 12(2): 117–133, 2013.

[4] A. Dolgui, B. Finel, F. Vernadat, N. Guschinsky, G. Levin, *A heuristic approach for transfer lines balancing*, Journal of Intelligent Manufacturing, 16(2): 159–172, 2005.

[5] A. Dolgui, B. Finel, N. N. Guschinsky, G. M. Levin, F. B. Vernadat, *MIP approach to balancing transfer lines with blocks of parallel operations*, IIE Transactions, 38(10): 869–882, 2006.

[6] A. Dolgui, N. Guschinsky, G. Levin, J.-M. Proth. *Optimisation of multi-position machines and transfer lines*, European Journal of Operational Research, 185(3): 1375–1389, 2008.

[7] A. Dolgui and I. Ihnatsenka, *Branch and bound algorithm for a transfer line design problem: Stations with sequentially activated multi-spindle heads*, European Journal of Operational Research, 197(3): 1119–1132, 2009.

[8] O. Guschinskaya, A. Dolgui, N. Guschinsky and G. Levin, *A heuristic multi-start decomposition approach for optimal design of serial machining lines*, European Journal of Operational Research, 189(3): 902–913, 2008.

[9] O. Guschinskaya, E. Gurevsky, A. Dolgui and A. Eremeev, *Metaheuristic approaches for the design of machining lines*, International Journal of Advanced Manufacturing Technology, 55(1-4): 11–22, 2011.

[10] Y. Kara, T. Paksoy and C.-T. Chang, *Binary fuzzy goal programming approach to single model straight and U-shaped assembly line balancing*, European Journal of Operational Research, 195(2): 335–347, 2009.

[11] A. Rossi, E. Gurevsky, O. Battaïa and A. Dolgui, *Maximizing the robustness for simple assembly lines with fixed cycle time and limited number of workstations*, Discrete Applied Mathematics, 208: 123–136, 2016.

[12] Y. Sotskov, N. Sotskova, T.-C. Lai and F. Werner, *Scheduling under Uncertainty: Theory and Algorithms*, Belorusskaya nauka, Minsk, 2010.

Automatic Storage and Retrieval Systems

Design of an automated storage space for reticles in a semiconductor manufacturing facility: A simulation-based approach

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Ali Ben-Salem[1, 2]
Claude Yugma[1]
[1]Department of Manufacturing Sciences and Logistics
Ecole des Mines de Saint-Etienne, CMP Georges Charpak
880 route de Mimet F-13541 Gardanne, France
Ali.bensalem@emse.fr
Claude.yugma@emse.fr

Emmanuel Troncet[2]
Jacques Pinaton[2]
[2]STMicroelectronics Rousset
190 Avenue Coq Monnet, F-13106 Rousset, France
Ali.bensalem@st.com
Emmanuel.troncet@st.com
Jacques.pinaton@st.com

Abstract—In semiconductor domain, the integration of an Automated Material Handling System (AMHS) in the manufacturing facility requires the migration from traditional methods of material storage and management to more developed approaches. In this context, an automated storage process is strongly needed to better accentuate the improvement of the facility performance due to AMHS. This paper addresses the automated storage problem of an auxiliary resource (called "reticle") in the photolithography work area of a legacy 200mm wafer facility that was not initially designed and optimized to receive such automation. A decision support tool is developed to better design the automated storage space. The challenge is not only to determinate the storage size and their locations on the work area layout, but also to optimize the reticles management inside the storage space. A simulation-based approach is proposed to solve the problem. Using a real instance extracted from the company database, results demonstrate that the proposed design of the automated storage can improve the work area performance when an AMHS is integrated.

Keywords—semiconductor; reticle; storage; automation; simulation

I. Introduction

Semiconductor manufacturing is among the most complex industries due to the fabrication process specificity [1]. Starting from a thin disc of silicon (called "wafer"), several production steps (up to 700 steps) with different durations (from 5 minutes to 12 hours) are performed. At the end of the cycle time, we obtain a few hundreds of chips at the surface of the wafer. This stage is called the "Front-End". Then, the second main stage is the "Back-End" which consists on testing, sorting, assembling and packaging each die of the wafer. For more details on the wafer manufacturing, interested readers are referred to [2].

Depending on the wafer diameter (i.e., 200mm or 300mm), we mainly distinguish two types of semiconductor manufacturing facilities (called "fab"). The 200mm wafer fabs were originally designed and optimized to produce a High Volume with a Low Mix (HV-LM) of products [3]. These legacy facilities mostly operate in a manual or semi-automated way. This means that wafers handling (transportation and storage) and production tools setting are mainly performed by operators. Later, to reduce the production-cost, the wafer diameter was increased from 200mm to 300mm in order to get a rise on the number of chips per wafer. The 300mm fabs are rather modern and fully automated by integrating an AMHS for transporting and storing wafers initially when building the facility [4].

In this paper, we investigate the case of a 200mm fab of STMicroelectronics Rousset. Currently, this type of facilities requires modernization to remain cost-efficient and competitive in a very dynamic market [3]. For the considered fab, the modernization begun by the photolithography work area by implementing an AMHS to improve the fab performance by reducing the cycle time and increasing the utilization of production tools [5].

The design of an AMHS-based Overhead Hoist Transport (OHT) system for reticles in the photolithography work area of the considered fab was already discussed in [2]. Note that reticles are extremely expensive and the required quantity depends on the number of the layers and the production step. Furthermore, this auxiliary resource can easily become a constraint for production planning leading to the degradation of the cycle time. This study is an extension of [2] as we focus on the design of an automated storage space for reticles in the photolithography work area of the same 200mm fab. Several questions about the storage space capacity, location of stockers, as well as the management of reticles have to be addressed and responses have to be argued using relevant analysis. In fact, the design phase has a direct impact on the work area performance (throughput, cycle time, tool utilization…) [4] and on Return On Investment (ROI) of the whole project. Our paper contributes to the literature by developing a decision support tool using a simulation-based approach, whereby practical feedbacks about the weaknesses and strengths of a design are identified and evaluated before the implementation.

The remainder of the paper is organized as follows. The related works are reviewed in Section II. Section III describes the considered problem. The simulation-based approach used to solve the problem is discussed in Section IV. The design of the automated storage space is presented in Section V. Section

VI concludes the paper and gives some perspectives for the future studies.

II. Literature Review

The literature presents several approaches in various domains (e.g. scheduling [6], process operations [7] and material handling [8]) that have been developed to improve the photolithography area performance for the essential reason that it represents the main bottleneck work area in the fab.

A few papers have addressed the problem of reticles automated handling in the semiconductor manufacturing environment. The work of Yamagishi [9] is considered among the first papers studying the automation of reticle handling in the photolithography work area. A new AMHS-based OHT for the reticles was proposed and evaluated to replace the original Automatic Guided Vehicle (AGV) system. Based on simulation results, the author confirmed that the new handling solution considerably improved the area performance. Recently, Ndiaye, Dauzère-Pérès, Yugma, Rullière, and G. Lamiable (2016) [8] focused on the design of a reticle automated transport system using AGV in a 300mm fab. A discrete-event simulation method was proposed to characterize the system elements such as the layout of rail and the number of vehicles. The authors evaluated several layouts and dispatching policies (e.g., First Come First Serve, Nearest Vehicle etc.) based on a set of key Performance indicators (KPIs). In the same context, the implementation of an AMHS-based OHT for reticles in a 200mm fab was addressed in our previous paper [2]. We developed a relevant system design for a legacy fab that was not originally built to have automated transport. The paper examined the impact of the automation on the work area productivity. A combined multi agent-discrete event simulation model was developed to test different scenarios (i.e., vehicle number, rail layout, dispatching policy).

Given the complexity and the uncertainty of the production process in the fab, most of the encountered papers used a simulation-based approach to address problems linked to AMHS. In fact, the simulation is considered as a relevant tool, particularly, regarding AMHS design and optimization. It is among the most used estimation tool for complex manufacturing systems [10] compared to other tools such as deterministic [11] or queueing [12] methods. In fact, a set of assumptions (aggregations) are mainly used in deterministic or queuing concepts which may not be always realistic in a high complex manufacturing domain.

In the literature, several papers discussed the problem of automated handling in photolithography area. However, yet there is a lack of studies that focus on the storage issue, especially, for reticles. Indeed, generally, same papers tried to combine production, transport and storage process [13]. Yang, Huang, Yu, and Jan (2002) [14] developed an algorithm using queuing theory to determinate the storage space for the Work-In-Proceed (WIP) in a 300mm fab. The authors affirmed that 95% of the transport requirement could be satisfied by the calculated stocker dimension. In another study, the modelling and analysis of the stocker system were proposed by [15]. An analytical approach was used and results showed that the failure rate of the stocker robot and the size of the input/output buffer space are the most critical factors impacting the performance of the transport system. Rust, Patel, and Qu (2005) [4] proposed a simulation modeling method to evaluated different storage strategies based on under track storage and/or big stocker in a 300mm fab. The authors confirmed that the storage strategy has a significant impact on the delivery mission (retrieval and delivery), the congestion in the inter-bay and the vehicles traffic. Later, regarding reticles, Byrne (2007) [16] focused on the case of two different types of reticles. Three discrete-event simulation models were developed and a decision support tool was proposed to improve the reticles management. Murray and Miller (2003) [17] studied the problem of automated reticle handling in a photolithography work area. They evaluated two different strategies (distributed and centralized) for the storage of reticles using the simulation method. Results indicated that the distributed solution was better in term of reticle delivery.

From the literature review, an important observation is that storage process has a significant impact on the performance of the automated transport and consequently on the work area productivity. Moreover, the above works confirm also that the simulation-based approach is the most applied method for complex manufacturing systems as semiconductor wafer fabs. In this paper, different from all the obvious studies, we investigate the global problem of storage space design (location, size, management), which is the main contribution of the paper.

III. Problem Description

The photolithography work area of the considered 200mm fab contains more than 25 production tools which are different in term of technology and throughput. The work area layout can be divided in 4 main loops of tools. Actually, operators have to perform handling operations (loading/unloading, transport and storage) for both lots and reticles. Shelves are placed in many locations at the floor to store non-used reticles that are grouped in cassettes (called “pods”). For operators, handling tasks of more than 4000 reticles are tedious and challenging. Indeed, physical effort and continuous concentration are required to transport the right reticle to the right place without delay. Thus, reticles handling may impact the performance of the production tools due to lead time and transport errors.

An AMHS-based OHT for reticles was evaluated in [2] before the real implementation of the system. In the same direction, the decision makers of the company aim to automate the reticle storage process using centralized big stockers. Single pods (capacity of 1 reticle) will be used instead of multi-pods.

The challenge is to propose a relevant design for the storage space taking in to account many constraints such as the high number of reticles, the complexity of reticles management and limited space at the ground of the work area. In this study, we have to identify the stockers’ size while considering the financial impact of an oversizing. Moreover, given the tools layout, the possible locations of stockers are very limited. Another important issue to address is how to manage reticles inside the stocker to avoid bottleneck problems due to the handling time inside a stocker.

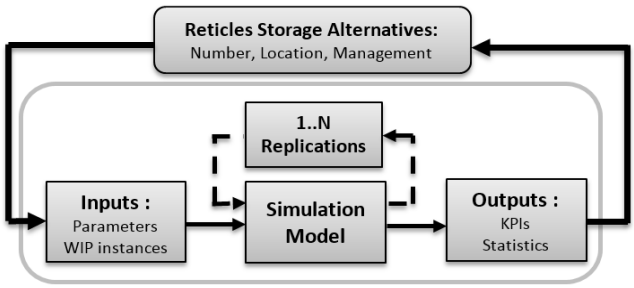
IV. Simulation-Based Approach

The rapid technology evolution in the AMHS for semiconductor manufacturing increases the pressure to accelerate modeling and analysis work by reducing the building time and improving the accuracy. Therefore, a high experience level is required to build a simulation model with some level of abstraction to reduce the complexity and to minimize the modeling cost. However, such a model has to contain the appropriate and the only necessary level of details of the system under study. That is the key factor that makes the difference between model builders [18].

In this paper, our first objective is to build and validate a relevant model with the main behavior of the considered system that allows us to represent the photolithography process. Fig.1 presents the proposed approach to solve the problem. The simulation model is alimented by real input instances of WIP. After that, several alternatives of storage space design (stocker's number, location and management) are evaluated based on KPIs.

The next section provides more details on the developed simulation model.

Fig. 1. Simulation-based approach



A. Simulation Model

A combined discrete event (DE) - agent based (AB) simulation model was developed with the main behavior of the considered photolithography area using ANYLOGIC software. This simulation software is becoming widely used in the research area, the industrial domain and service applications. This is due to its characteristics such as the possibility of combining several simulation approaches (discrete-event, agent-based, systems dynamic) in the same model [19]. The DE allows modelling the process as a sequence of operation. However, given the complexity of such manufacturing domain, using only DE is very complicate and time consuming especially for a large scale system. To overcome this problem, AB is used to give more flexibility in the modelling phase. Since the queuing concept, which the main issue of the manufacturing concept, is not defined in AB method, the combination of DE-AB methods is strongly recommended in the case of semiconductor manufacturing process.

The model was used in [2] to propose a relevant design the AMHS-based OHT. The main assumption was that only the automation of reticle handling operations was considered. In fact, the OHT system will provide only the reticle delivery. Thus, lot manual handling tasks are not considered in the model.

In this study, necessary model modifications were performed to focus on the storage process of reticles.

B. Simulation Model Validation

Model validation is an extremely important step in any simulation project to avoid errors and wrong conclusions. The validation method dependents on the nature and the objective of the developed model. In our case, we have to ensure that the model control logic (reticle location, destination, etc.) is identical to the real system logic (without automation). Although any AMHS cannot be separated from the production system (e.g. scheduling, maintenance, batching, etc.), the control interface of delivery missions can be dissociated from the production controller. In our case, WIP instances extracted from the company database are used to develop a from/to movement data which drives the model.

The validation is done through a comparison of a real matrix of from/to reticle movement to simulation outcomes matrix. In addition, real performance indicators of photolithography tools (i.e., throughput, utilization rate, etc.) are compared with model outputs. This allows us to confirm that our simulation model adequately represents the dynamic of the real system.

We note that the modelling phase including data collection, logic implementation and validation was complicate and time-consuming (around 4 months).

V. Automation of Reticle Storage Process

As previously mentioned, the OHT design (number of vehicles, layout of rails, etc.) was already performed in [2]. In this paper, we aim to find a relevant design for automated reticles storage. The required number and locations of stockers have to be discussed. Each stocker contains different zones. This makes the handling time (to enter or to take out a reticle) varies between 10, 60 and 100 seconds in average. That is why the assignment of reticles to each zone in a stocker has to be addressed to avoid bottleneck problem due to the handling time.

Although, the transport presents a non-value-added service, it is indispensable for the production process. Therefore, AMHS performance indicators (vehicle traffic, delivery time, vehicle utilization) have to be considered together with production KPIs in the design phase of the automated storage space. The simulation duration is chosen such that to guarantee that the steady state is reached (around 28 days). For each run, three replications are performed which was sufficient to reduce variability of outputs.

A. Number of Reticle Stockers

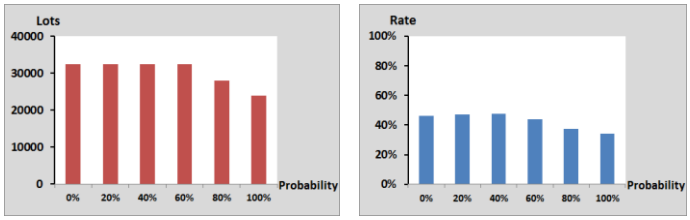
Given the current process conditions on the photolithography work area (activity level, number of reticles, etc.) and the high cost of a stocker, initially, the decision has been taken to implement only two stockers (with a capacity of about 1500 reticles per stocker). To take into account the handling time (10, 60 or 100 seconds), required modifications are performed in the simulation model. The concept of probability to assign a reticle to a zone when entering the stocker is adopted. The main idea consists on taking two types of decision linked to reticle at the entrance of a stocker:

- To keep the reticle in the single pod: it takes only 10 seconds, but it reduces the number of empty pods which can increase the handling time required to take out another reticle from the stocker. Moreover, exchanging and handling the pods can take 100 seconds.
- To remove the reticle from the single pod: it takes 60 seconds. However, it provides an empty pod which may reduce the handling time required to take out another reticle from the stocker.

For different probabilities to remove the reticle from the pod when entering to the stocker, Fig. 2 shows the work area throughput and the mean utilization of vehicles. Results from Fig. 2 confirm that for a probability higher than 60% (limit value), the stockers become a blocking point for the delivery service provided by OHT. Consequently, the work area throughput and the utilization of vehicles are degraded compared to the current values of indicators.

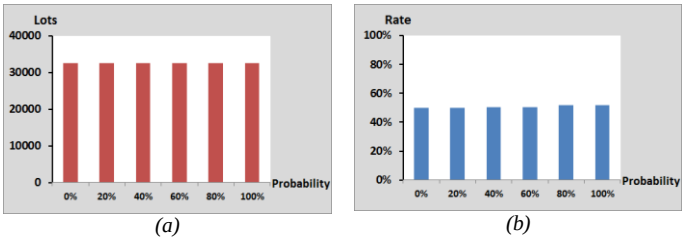
To handle this bottleneck issue, we propose to add a third stocker in order to reduce the workload. This scenario was implemented in the model. Simulation results are presented in Fig.3 for the case of 3 stockers.

Fig. 2. (a) Throughput and (b) mean utilization of vehicles for 2 stockers



(a) *(b)*

Fig. 3. (a) Throughput and (b) mean utilization of vehicles for 3 stockers



(a) *(b)*

According to Fig. 3, the limit value of probability is pushed from 60% (Fig. 2) to 100% when adding the third stocker. This confirms that the storage process can no longer became a blocking point for the AMHS.

B. Location Strategy of Stockers

In this subsection, the objective is to determine, using the simulation, the best location strategy of stockers in the work area. The proposed strategy should not degrade the performance indicators of both production system and transport service.

Different locations of the three stockers have been evaluated based on the traffic of vehicles and the delivery mission. Given the current layout of the work area, five possible configurations have been defined to receive the three stockers. Fig. 4 presents the proposed configurations.

Depending on the location of stockers, we distinguish three storage strategies: centralized (*Config 1*), partially distributed (*Config 2* and *Config 3*) and distributed (*Config 4* and *Config 5*). For all these configurations, a random assignment of reticules in each stocker was considered.

Fig. 4. Different configurations of storage space location

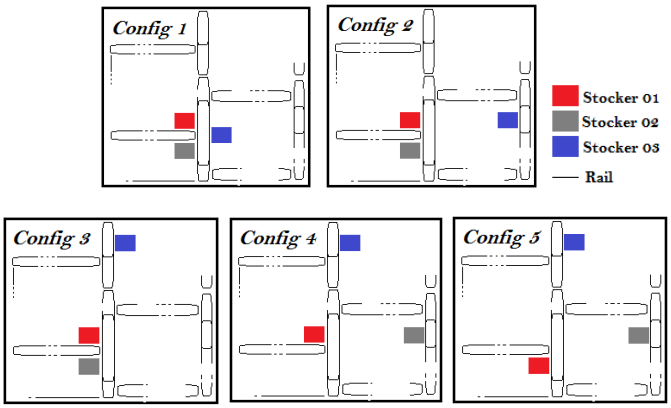


Table I indicates, by configuration, the proportion of the rail network used by the vehicles transporting reticles (only from stockers to tools or from tools to stockers, from tool to tool is negligible), and the maximum traffic rate.

TABLE I. Simulation Results for Different Configurations of Storage Location

Configuration of stockers	% Network used by vehicles	Max traffic rate(veh/min)
Config. 1	78.23%	1.842
Config. 2	80.91%	1.135
Config. 3	80%	1.197
Config. 4	**82.73%**	**0.892**
Config. 5	81.82%	0.934

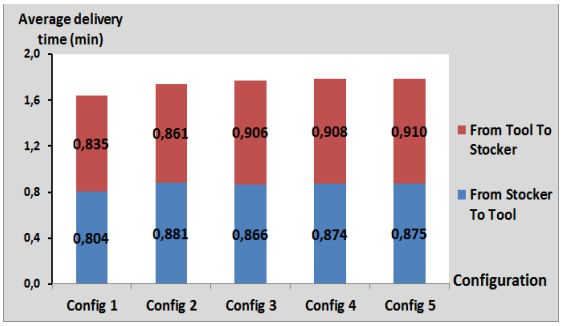
Fig. 5 presents the average delivery time for different configurations of stockers location. Results show that the location strategy has a direct impact on the performance of AMHS which may impact the performance of the photolithography process. From Fig. 5, for a centralized policy (*Config 1*), a little gain in the average delivery time is observed compared to the distributed and partially distributed strategies. However, such a strategy generates a traffic problem at the central rails as described in Table I.

By considering a partially distributed strategy (*Config 2* and *Config 3*), the traffic is reduced, particularly in the central rails. Furthermore, an increase in the delivery time is noted (see Fig. 5). Indeed, a vehicle has to travel a longer distance compared to the centralized strategy to retrieve and/or deposit a reticle to the stocker.

Regarding the distributed strategy (*Config 4* and *Config 5*), a better distribution of vehicles traffic on the rail network is observed (see Table I). Despite the little degradation of the delivery time, this strategy provides a more distributed risk in the case of rail unavailability (intervention, vehicle failure).

Thanks to this simulation-based study, we were able to orient the choice of the company towards a distributed strategy of reticule storage space (*Config 4*).

Fig. 5. Mean delivery time for the different configurations of location



C. Reticle Assignment Inside a Stocker

This section discusses the way reticles should be assigned to a zone in a stocker. The main idea consists of proposing and evaluating different assignment policies. The objective is to maximize the stocker throughput while taking into account the limit number of loading/unloading ports. The proposed policies are as follows:

- Policy 1: random policy; reticles are randomly assigned to each zone inside a stocker and no particular rules are implemented.
- Policy 2: random policy with mission parallelization; in addition to randomly assignment of reticles to each zone in a stocker (as Policy 1), we aim to exceed technical limitations of the stocker which can be an improvement trail of the throughput. For confidentially reason, we can only mention that the idea consists of taking advantage of the two load ports per stocker to parallelize the missions for stocker robot.
- Policy 3: smart policy; it is an intelligent way for reticle assignment to each zone inside a stocker based on the utilization of each reticle (high, mean, low) and the performance targets. To explain, depending on the production mix, all reticles are not used at the same frequency. Recall that to process a lot, reticle should be removed from stocker and transported by OHT system to the production tool. Therefore, the more reticles are used, the shortest handling time is required.

  For example, if we consider only 20 positions in zone A, and the handling time to remove a reticle from this zone is the shortest (e.g. 10 seconds), the 20th most used reticles are assigned to zone A. The next reticles in the utilization ranking are stored in zone B which has the second shortest handling time (e.g. 60 seconds). The proposed policy is more complicated given the industrial constraints in the work area.
- Policy 4: smart policy with mission parallelization; we combine the smart policy (Policy 3) with the possibility to parallelize the missions (Policy 2). The goal is to provide an intelligent assignment of reticles and at the same time to overcome the technical limitations of the stocker.

The same WIP instances are used for all policies to perform a comparative analysis. Results about only one stocker throughput are summarized in Table II.

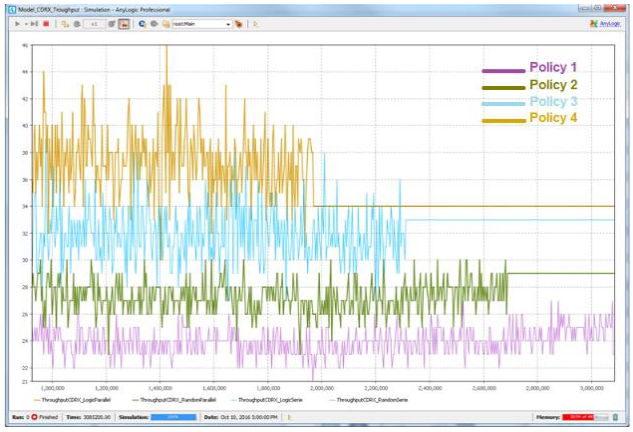
TABLE II. Simulation Results for Different Assignment Policies of Reticles

Policy	Stocker throughput (missions/hour)			Time to perform all missions (hour)
	Mean	Peak	Mean gain / *policy 1*	
Policy 1	23.53	27	-	856.4
Policy 2	26.98	31	12.78%	746.9
Policy 3	31.4	39	25.06%	641.7
Policy 4	36.83	46	36.11%	547.1

From Table II, the main conclusion is that the storage policy based on reticle utilization (Policy 3 and Policy 4) significantly improves the throughput of the stocker and consequently, reduces the required time to perform all the transport missions. Another important result is that the mission parallelization allows us to increase the throughput up to 12.78%.

Fig. 6 represents a screenshot from the simulation software that shows the evolution of the stocker throughput over time for different policies.

Fig. 6. Mean stocker throughput over time for different policies



From Fig. 6, we observe a variability concept of the throughput which is mainly due to the production process variability and the mix of products in the work area. In addition, we can confirm that the Policy 4 provides the best average throughput per stocker with the shortest end date.

VI. Conclusion

In this paper, the automation problem of the reticle storage space in the photolithography work area of 200mm fab was addressed. The decision of storage automation was taken by company decision makers considering an overall project of AMHS integration in the work area. The challenge is to propose a relevant design of an automated storage space to avoid the degradation of automated transport service which

may impact the work area performance. A simulation-based approach was used to solve the problem given many advantages of such approach compared to the analytic or mathematical methods.

Results affirmed that the automated storage of reticles can become a real bottleneck for the production process. Such finding is extremely critical for the company given the importance of this work area in the fab. To handle this problem, the required number of stockers was increased from 2 to 3 stockers despite the high financial impact. In addition, a depth study was performed to better locate the storage space considering the existing layout of tools. The simulation outputs confirmed that the distributed location strategy reduces the traffic of vehicles. The reticles management inside a stocker was also discussed. Indeed, four assignment policies were evaluated that allowed us to enhance the stocker throughput (up to 36.11%).

In the next step of this study, we will focus on the issue of reticules assignment between the different stockers. The assignment rules probably depend on the location of the stockers, the reticles utilization as well as the mix of WIP.

Acknowledgment

This study has been done within the framework of a joint collaboration of STMicroelectronics in Rousset, France, and the Center of Microelectronics in Provence (CMP) of the Ecole des Mines de Saint-Etienne (EMSE) in Gardanne, France. The authors would like to thanks both ANRT (Association Nationale de la Recherche et de la Technologie) which has supported this study and the engineers of STMicroelectronics Rousset for their help and availability.

References

[1] J.N. Gupta, R. Ruiz, J.W. Fowler, and S. J. Mason, “Operational planning and control of semiconductor wafer production”, Production Planning and Control, vol. 17, no 7, pp. 639-647, 2006.

[2] A. Ben-Salem, C. Yugma, E. Troncet, and J. Pinaton, “AMHS design for reticles in photolithography area of an existing wafer fab”, In Proceedings of SEMI Advanced Semiconductor Manufacturing Conference (ASMC2016), pp. 110-115, May 2016.

[3] H. Heinrich, G. Schneider, F. Heinlein, S. Keil, A. Deutschländer, and R. Lasch, “Pursuing the Increase of Factory Automation in 200mm Frontend Manufacturing to Manage the Changes Imposed by the Transition from High-Volume Low-Mix to High-Mix Low-Volume Production,” In Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC2008), pp. 148–155, May 2008.

[4] K. Rust, N. Patel, and P. Qu, “Modeling practical methods of material storage in a 300 mm fab”, In Semiconductor Manufacturing, ISSM 2005, IEEE International Symposium, pp. 25-28, 2005.

[5] F.K. Wang and J.T. Lin, “Performance evaluation of an automated material handling system for a wafer fab”, Robotics and Computer-Integrated Manufacturing”, vol. 20, no. 2, pp. 91-100, 2004.

[6] C.W. Chiou and M.C. Wu, “Scheduling of multiple in-line steppers for semiconductor wafer fabs”, International Journal of Systems Science, vol. 45, no. 3, pp. 384-398, 2014.

[7] M. Strobl, C. Hsu, Y.C. Lin, H. Chen, D. Chen, A. Cheng, and J. Kopp, “Immersion lithography process improvements by wafer edge inspection at 300mm DRAM manufacturing fab”, In Semiconductor Manufacturing (ISSM), 2010 International Symposium, pp. 1-4, October 2010.

[8] N.A. Ndiaye, S. Dauzère-Pérès, C. Yugma, L. Rullière, and G. Lamiable, “Automated transportation of auxiliary resources in a semiconductor”, In Proceedings of Winter Simulation Conference (WSC2016), IEEE, pp. 2587-2597, December 2016.

[9] T. Yamagishi, “The world's first automated reticle handling system using OHT”, In Proceedings of IEEE International Symposium on Semiconductor Manufacturing (ISSM’07), pp. 21-24, 2003.

[10] C. J. Huang, K. H. Chang, and J. T. Lin, “Optimal vehicle allocation for an automated materials handling system using simulation optimisation”, International Journal of Production Research, vol. 50, no. 20, pp. 5734-5746, 2012.

[11] P.J Egbelu, “The use of non-simulation approaches in estimating vehicle requirements in an automated guided vehicle-based transport system,” Material Flow, vol. 4, no. 1, pp. 17–32, 1987.

[12] J. M. A. Tanchoco, P. J. Egbelu and F. Taghaboni, “Determination of the total number of vehicles in an AGV-based material transport system,” Material Flow, vol. 4, no. 1-2, pp. 33-51, 1987.

[13] J.E. Kiba, G. Lamiable, S. Dauzère-Pérès, and C. Yugma, “Simulation of a full 300MM semiconductor manufacturing plant with material handling constraints”, In Winter Simulation Conference (WSC2009), pp. 1601-1609, December 2009.

[14] C.W. Yang, K.S. Huang, G. Yu, and D.Y. Jan, “Using queuing theory to estimate the storage space of stocker in automated material handling systems”, In Semiconductor Manufacturing Technology Work area, IEEE, pp. 102-104,December 2002.

[15] Y.J. Jang, G.H. Choi, and S.I. Kim, “Modeling and analysis of stocker system in semiconductor and LCD fab”, In Semiconductor Manufacturing (ISSM 2005), IEEE International Symposium, pp. 273-276, September 2005.

[16] P. J. Byrne, “An analysis of semiconductor reticle management using discrete event simulation,” In Proceedings of the 2007 Summer Computer Simulation Conference, pp. 593-600, 2007.

[17] A. M. Murray and D.J. Miller, “Automated reticle handling: a comparison of distributed and centralized reticle storage and transport,” In Proceedings of IEEE Winter Simulation Conference, vol. 2, pp. 1360-1365, December 2003.

[18] G. T. Mackulak, F. P. Lawrence, and T. Colvin, “Effective simulation model reuse: a case study for AMHS modeling”, In Proceedings of Winter Simulation (WSC1998), IEEE Computer Society Press, pp. 979-984, December 1998.

[19] P. A. Wai, and A. V. Ostroukh, “Development of Simulation Model Mixed System in the AnyLogic Software”, International Journal of Advanced Studies, vol. 4, no. 4, pp. 48-53, 2015.

Analytical models for a joint posture and fatigue analysis in order picking

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Martina Calzavara, Fabio Sgarbossa
Department of Management and Engineering
University of Padua
Vicenza, Italy
martina.calzavara@unipd.it, fabio.sgarbossa@unipd.it

Christoph H. Glock, Eric H. Grosse
Institute of Production and Supply Chain Management
Technische Universität Darmstadt
Darmstadt, Germany
glock@pscm.tu-darmstadt.de, grosse@pscm.tu-darmstadt.de

Abstract— This paper presents analytical models for assessing worker posture and fatigue in manual order picking activities and their performance impact. The OWAS (Ovako Working Posture Analysing System) index and the energy expenditure concept are combined to evaluate the order picking process with respect to health risks and to assess the performance impact of the order picking system from a human factors point of view.

Keywords—Order picking, Ergonomics, Performance, Warehouse, Human Factors, Pallet storage

I. Introduction

The advancement of e-commerce as well as increasingly challenging and diverse customer requirements, such as same-day-delivery or mass customization, put high requirements on internal logistics processes, in particular on warehousing. One of the most critical warehousing activities is order picking, which has been defined as the retrieval of items from storage locations to fulfil customer orders. It has frequently been characterized as a major determinant of delivery time and quality and ultimately customer satisfaction [1]. Although new technologies, such as automated guided vehicles (AGVs) or picking robots, facilitate automating order picking, most of the order picking systems that can be found in practice are still operated manually employing human workers who are responsible for retrieving items from their storage locations [2, 3]. Typically, these systems are operated according to the picker-to-parts concept where human operators travel along the warehouse aisles to retrieve demanded items using some kind of technical equipment. The major advantages of manual order picking are low investment cost and high flexibility with respect to changes in the product portfolio or customer demand, which explains why manual order picking systems are still dominant in practice. The high amount of human work required in manual order picking turns this activity into a very time- and cost-intensive process [4], however, which inspired the development of mathematical models that enable warehouse managers to plan efficient order picking operations [2].

One important characteristic of manual order picking are manual material handling activities, which are regularly associated with the lifting and carrying of heavy loads as well as stretching and bending in sometimes awkward postures. Such activities expose workers to a high risk of developing health problems, in particular musculoskeletal disorders (MSD). In 2015, for example, private sector warehouse workers in the US incurred the highest number of MSD cases with a days-away-from-work incidence rate of 289.4 cases per 10,000 full-time workers, which led to 56,550 days-away-from-work cases [5]. Many of the reported incidents were related to body-reactions and overexertion associated with repetitive material handling. To provide a work environment with minimal health risks to their employees and to comply with safety regulations, warehouse managers require decision support from an interdisciplinary perspective that considers both economic and ergonomic objectives in planning order picking operations [6]. Mathematical models for order picking that explicitly take worker health and performance aspects into account are still rare, however, which has been documented in two recent reviews of the literature [6, 7].

The paper at hand contributes to closing this research gap by developing mathematical models for a joint posture and fatigue evaluation of order picking activities using the OWAS (Ovako Working Posture Analysing System) index and the energy expenditure concept. The proposed models can be used by warehouse managers to quantify the ergonomic impact of different order picking activities and their possible influence on worker health. In addition, combining the developed OWAS assessment with the concept of energy expenditure helps to quantify the impact of different posture and fatigue conditions on order picking performance. To consider different practical scenarios, our analysis focuses on the picking of items from a) full pallets on the floor, b) half-pallets on the floor, c) half-pallets on the middle rank of a low-level shelf, and d) half-pallets on the upper rank of a low-level shelf. In a field study, data required for calculating the OWAS index was collected using a motion capture system. Despite contributing to the development of knowledge on the system design-performance link in order picking, the paper at hand supports warehouse managers in understanding the impact of the order picking system on human factors issues and how the configuration of the order picking system can impact workplace ergonomics and worker performance.

II. Literature review

Prior research on manual order picking had a strong focus on mathematical models that aim on minimizing order picking time or travel distance (see, for a review, [3]). The influence of the work environment (here: the warehouse and the order picking process) on the well-being of the order picker and the order

picker's performance have, in contrast, not received much attention in these modelling efforts so far [6, 7]. The number of published works in this area has been growing in recent years, however. Related works can roughly be divided into two research streams. The first stream of research investigates the impact of mental and psychosocial human factors on order picking performance. Some authors studied how humans learn in order picking and integrated so-called learning curves into models for scheduling storage reassignment processes [8] or for assigning a heterogonous workforce (i.e., workers with different learning rates) to different zones of an order picking warehouse [9], for example. [10], in turn, analyzed the impact of paperless information technologies on order picking quality and human errors. In the second stream, researchers focused on the "physical dimension" of human factors and studied health risks associated with the handling of loads in order picking. Works in this area developed integrated models that focus both on ergonomic (e.g., reducing health risks) and economic (e.g., reducing order picking time) objectives. One concept that has enjoyed some popularity in this respect is the energy expenditure concept developed by [11], which enables managers to estimate the metabolic cost associated with performing a manual task. Reference [12], for example, developed a bi-objective storage assignment model using Pareto-frontiers to balance the trade-off between order picking time and energy expenditure, while [13] proposed a decision support model to evaluate different rack layouts with respect to their influence on order picking time and energy expenditure. In another work on storage assignment, a bi-objective model to minimize cycle time and worker discomfort was proposed by [14] using Borg's CR-10 scale. Recently, [15] used the concept of human availability and rest allowance to develop a cost model that can guide managers in ergonomic interventions, e.g. by hiring stand-by operators to balance ergonomic conditions and hiring costs. Other authors also studied human factors and behavioral issues in order picking from an explorative perspective, such as [16].

Although human factors issues have not attracted much attention in works that propose mathematical models for order picking so far, the human factors engineering literature has focused on manual materials handling and warehousing activities for many years (albeit neglecting the economic perspective in many cases), see for example [17, 18, 19, 20, 21, 22]. The overview of the literature indicates that despite the high number of methods for an ergonomic evaluation of manual materials handling that have been proposed in the past (see also [23]), these methods have up to now only very infrequently been considered in the development of mathematical models for warehouse order picking. To contribute to closing this research gap, this paper proposes an assessment method consisting of the OWAS index and the energy expenditure concept to capture both the effects of awkward body postures and worker fatigue in performing order picking activities.

III. Problem description

The paper at hand investigates the manual picking of items from pallets in a warehouse that uses low-level racks. We assume that an order picker receives an order at the depot of the warehouse and then starts preparing the order picking process (setup step). After the setup of the order, the order picker travels along the shelves of the warehouse, possibly standing on a powered cart, to the storage locations where items need to be retrieved (travel step). Once the order picker arrives at the first location where items need to be retrieved, he/she searches for the item on the shelf (search step). Setup, travel and search are performed in standing posture or during upright walking. The warehouse worker then picks items from the identified shelf locations and places them on the cart (pick step). During picking, twisting, stretching and bending it may be necessary in addition to adopt a standing posture depending on the layout of the storage area and the type of racks used. After the completion of a pick, the order picker continues travelling along the aisles and picks items until the order list has been completed. It is obvious that each activity associated with the order picking process exposes the order picker to a different load level. We can expect that the actual pick activity, due to the possible need to adopt awkward body postures, has a particularly high influence on fatigue and injury risks, while the other activities (that are performed mainly during upright walking or standing) expose the order picker to lower load levels and injury risks.

The activities associated with the order picking process, their typical share in the total order picking processing time, possible (typical) body postures and the resulting energy consumption are summarized as assumptions for this paper in Table 1 [4, 6].

TABLE I.

Activity	Typical posture	Energy consumption and effect on fatigue
Setup and other	Standing, upright walking	Low energy consumption
Travel	Standing (on powered cart), or upright walking	Low energy consumption
Search	Standing, upright walking	Low energy consumption
Pick	Standing, twisting, stretching and bending – Several awkward postures depending on pallet location	High amount of energy consumption – possible effect on fatigue

a. Order picking activities with typical time components, body postures and energy expenditure

Items are stocked on pallets that are stored in the warehouse according to one of four possible rack layouts (see Fig. 1).

(1) Rack layout FP: Full pallets are stored on the ground floor of the rack;

Rack layout HP: Half-pallets are stored in the racks, which leads to three ranks per pick position, namely

(2) HP-0: Half-pallet positioned on the ground floor level of the rack;

(3) HP-M: Half-pallet positioned on the middle level of the rack at 0.85 meters from the floor;

(4) HP-1: Half-pallet positioned on the upper level of the rack at 1.4 meters from the floor.

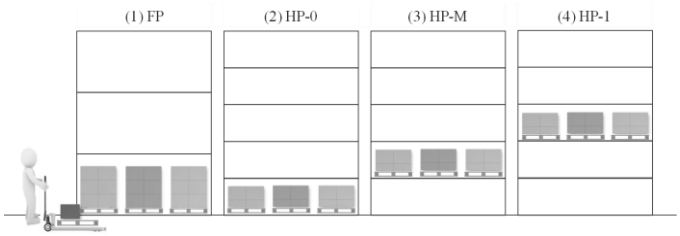


Fig. 1. Rack layout alternatives considered in this paper.

IV. Ergonomics evaluation of warehouse activities

A. OWAS index estimation using a motion capture system

For each activity defined in Section 3, the OWAS Index was calculated using a full-body motion capture system [24]. This system uses several inertial sensors that monitor the exact position of 17 specific reference points of the body for 60 times per second. These positions can be used as input values for several ergonomic analyses. In the work at hand, each posture observed during the field study was evaluated with respect to trunk, arms and legs position as well as handled load. For estimating the OWAS risk class of each analyzed posture, the data was processed using the matrix developed by [25] and classified according to the categories introduced by the OWAS method [26]. During the analysis of each activity, the global class risk value was estimated using the traditional OWAS Index (OI):

$$OI = (a \cdot 1 + b \cdot 2 + c \cdot 3 + d \cdot 4) \cdot 100 \quad (1)$$

Based on the table defined by [25], for each activity, *a* denotes the percentage of observations in the first risk class, and *b*, *c* and *d* the percentages for the second, third and fourth risk classes, respectively. The OWAS Index then adopts a value in the range [100, 400].

Table 2 reports the *OI* values for each major order picking activity estimated based on several tests carried out in a warehouse using a motion capture system and based on the literature [26]. For the actual picking activity, the *OI* was calculated for the different stocking locations outlined above as well as for two ranges of item weight. The *OI* value is constant for item weights ranging between 0 and 10 kg or between 10 to 20 kg. This is caused by the general concept of the OWAS method that considers three different categories for item weight, i.e. 0 to 10 kg (category 1), 10 to 20 kg (category 2), and items exceeding 20 kg (category 3). The paper at hand limits its analysis to items with a maximum weight of 20 kg (see, for recommended weight limits of lifting tasks, e.g. [27]). Table 2 shows that the process steps setup & others, travel and search have no impact on the posture analysis since no awkward postures have to be adopted during these activities. The pick activities, in turn, influence the posture analysis depending on item weight and pallet location. For the middle level, for instance, the *OI* is quite low because little or no bending or twisting of the back is required. For the ground floor level, the *OI* is higher due to necessary bending, while for the upper level, the *OI* also increases as it requires the order picker to move arms above the shoulders. According to the OWAS method, an activity becomes critical when *OI* > 200 [25].

TABLE II.

OWAS INDEX		
Setup & others	$OI_{s\&o} = 100$	
Travel	$OI_t = 100$	
Search	$OI_s = 100$	
Pick	Item weight w [kg]	
	$0 < w \leq 10$	$10 < w \leq 20$
Pick from FP	$OI^{10}_{p-FP} = 125$	$OI^{20}_{p-FP} = 150$
Pick from HP-0	$OI^{10}_{p-HP-0} = 140$	$OI^{20}_{p-HP-0} = 180$
Pick from HP-M	$OI^{10}_{p-HP-M} = 123$	$OI^{20}_{p-HP-M} = 145$
Pick from HP-1	$OI^{10}_{p-HP-1} = 139$	$OI^{20}_{p-HP-1} = 141$

b. OWAS results for each task in the order picking process

The above analysis shows that for the scenario considered here, none of the investigated activities would be assessed as critical by the OWAS method, as none exceeded the threshold value of 200. In fact, all major (time-consuming) activities of the order picking process, namely setup & others, travelling and searching, are performed mainly in standing position or during upright walking. The picking of items – especially for items stocked close to the ground floor – is executed in a bent position; however, the duration of the pick is short as compared to the other three main activities, which leads to a relatively low *OI* value (see also [13]). We note, however, that the above analysis may incorrectly reflect the influence of cumulative workloads on the order picker, such as worker fatigue, and therefore extend the above analysis by analyzing the metabolic cost of the order picking process at hand in the following.

B. Metabolic cost/energy expenditure

The metabolic cost, expressed as the calories consumed per minute in performing a particular task, is calculated to assess the fatigue level of the order picker in the following. One of the most popular methods for calculating metabolic cost was proposed by [11], who introduced several equations for estimating the metabolic cost or energy expenditure rate in kcal/min for elementary tasks such as lifting, lowering, standing, walking, or pushing. The proposed equations are functions of specific parameters of the analyzed task, such as the operator's gender and body weight and features of the elementary movements. Based on the method of [11], the metabolic cost of each activity, $\bar{\dot{E}}_{job}$, was calculated as the ratio of the total energy expenditure to the total time required for completing the analyzed activity, T. The total energy expenditure is the sum of I) a term related to the metabolic cost spent for maintaining body posture i, $\dot{E}_{pos-i}$, multiplied by the time spent in each body posture, t_i (where np is the total number of postures), and II) a term related to the elementary tasks performed during the analyzed activity, where ΔE_{task-j} is the energy consumption for performing an elementary task j, and n is the total number of elementary tasks:

$$\bar{\dot{E}}_{job} = \frac{\sum_{i=1}^{np} \dot{E}_{pos-i} \cdot t_i + \sum_{j=1}^{n} \Delta E_{task-j}}{T} \quad (2)$$

As for the OWAS Index in Section 4.1, we now calculate the metabolic cost (or energy expenditure rate) for each step of the order picking process (see Table 3). The required values were taken from the literature [13]. Figure 2 illustrates the energy expenditure rate for each pallet picking scenario considering different item weights. The energy expenditure rate for the picking activity is expressed as $\bar{\dot{E}}_{p-l}(w) = a_l \cdot w + b_l$, where $l \in \{L\} = \{FP; HP-0; HP-M; HP-1\}$. Parameters a_l and b_l depend on the location.

TABLE III.

ENERGY EXPENDITURE RATE [kcal/min]	
Setup & others	$\bar{\dot{E}}_{s\&o} = 2.2$
Travel	$\bar{\dot{E}}_t = 2.2$
Search	$\bar{\dot{E}}_s = 2.2$

Pick	item weight w [kg]					
	2	4	8	12	16	20
$\bar{\dot{E}}_{p-FP}(w)$	5.079	5.586	6.599	7.612	8.625	9.638
$\bar{\dot{E}}_{p-HP-0}(w)$	5.744	6.348	7.554	8.760	9.967	11.173
$\bar{\dot{E}}_{p-HP-M}(w)$	4.335	4.823	5.799	6.774	7.750	8.726
$\bar{\dot{E}}_{p-HP-1}(w)$	4.522	5.108	6.279	7.451	8.622	9.793

c. Energy expenditure rate for each step of the order picking process

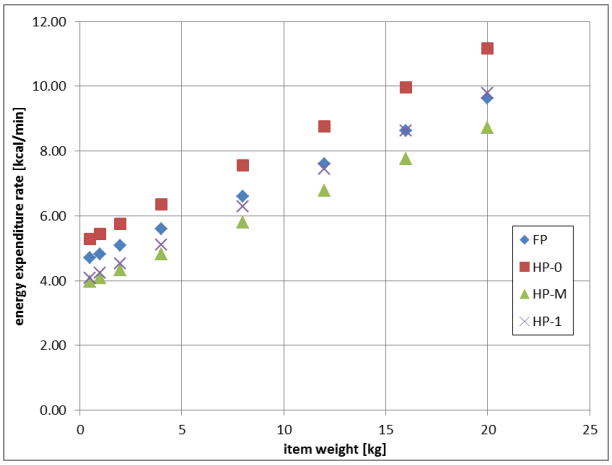


Fig. 2. Energy expenditure rate for each pallet picking scenario considering different item weights

The energy expenditure during the activities setup & others, travel, and search have been assumed equal to 2.2 kcal/min, as these activities are performed mainly in standing positions (1.8 kcal/min) and during upright walking (3 kcal/min). For the picking activities, the metabolic cost is a linear function of item weight as defined by [11]. Based on this concept, the next section proposes a simple equation for assessing the metabolic cost of picking items from different stocking locations. Generally, the energy expenditure rate during picking is quite high in contrast to the *OI* reported above, which may affect the fatigue level of the order pickers.

V. Models for an ergonomic assessment of the order picking process

This section develops an ergonomic measure for evaluating the order picking processes based on the *OI* and the energy expenditure concept. The resulting general equations provide an assessment of order picking based on a joint posture and fatigue analysis. We define the time required for performing a single basic task of a generic order as follows:

$t_{s\&o}$: the total time spent on setting up an order and for other administrative activities related to a single order;

t_t: the total time spent on travelling in the warehouse;

t_s: the total time spent on searching the stocking locations;

t_p: the total time spent on the actual picking activity.

We define t_p as $t_p = \sum_{l \in L} t_{p-l}^w$, where t_{p-l}^w is the total time spent on picking an item of weight w from stocking location l with $l \in \{L\} = \{FP; HP-0; HP-M; HP-1\}$. These terms are simplified to t_{p-l}^{10} and t_{p-l}^{20} for the estimation of the *OI* to match the load categories defined by this method. The following equations are introduced to estimate the total *OI* value, OI_{TOT}, and the total energy expenditure, E_{TOT}, for a generic order:

$$OI_{TOT} = OI_{s\&o} \cdot t_{s\&o} + OI_t \cdot t_t + OI_s \cdot t_s + \sum_{l \in L}\left(OI_{p-l}^{10} \cdot t_{p-l}^{10} + OI_{p-l}^{20} \cdot t_{p-l}^{20}\right) \quad (3)$$

$$E_{TOT} = \bar{\dot{E}}_{s\&o} \cdot t_{s\&o} + \bar{\dot{E}}_t \cdot t_t + \bar{\dot{E}}_s \cdot t_s + \sum_{l \in L} \bar{\dot{E}}_{p-l}(w) \cdot t_{p-l}^w \quad (4)$$

With reference to the values contained in Tables 2 and 3, where $OI_{s\&o} = OI_t = OI_s$ and $\bar{\dot{E}}_{s\&o} = \bar{\dot{E}}_t = \bar{\dot{E}}_s$, these formulas can be re-written as follows:

$$OI_{TOT} = \left(OI_f \cdot t_f\right) + \sum_{l \in L}\left(OI_{p-l}^{10} \cdot t_{p-l}^{10} + OI_{p-l}^{20} \cdot t_{p-l}^{20}\right) \quad (5)$$

$$E_{TOT} = \left(\bar{\dot{E}}_f \cdot t_f\right) + \sum_{l \in L} \bar{\dot{E}}_{p-l}(w) \cdot t_{p-l}^w \quad (6)$$

where $OI_f = OI_{s\&o} = OI_t = OI_s$, $\bar{\dot{E}}_f = \bar{\dot{E}}_{s\&o} = \bar{\dot{E}}_t = \bar{\dot{E}}_s$ and $t_f = t_{s\&o} + t_t + t_s$.

We define the total time required for completing a single generic order as t_o, and $\%t_f$, $\%t_{p-l}^{10}$, $\%t_{p-l}^{20}$, $\%t_{p-l}^w$ as the percentage of time spent on each basic activity. OI_{TOT} and E_{TOT} can now be calculated as follows:

$$OI_{TOT} = \left[\left(OI_f \cdot \%t_f\right) + \sum_{l \in L}\left(OI_{p-l}^{10} \cdot \%t_{p-l}^{10} + OI_{p-l}^{20} \cdot \%t_{p-l}^{20}\right)\right] \cdot t_o \quad (7)$$

$$E_{TOT} = \left[\left(\bar{\dot{E}}_f \cdot \%t_f\right) + \sum_{l \in L} \bar{\dot{E}}_{p-l}(w) \cdot \%t_{p-l}^w\right] \cdot t_o \quad (8)$$

VI. Application

The equations introduced in the previous section can be used, for example, to assess whether the order picking process has a relevant impact on fatigue, due to specific working postures and

weights of the items that need to be handled. They can be applied in workforce management, for example for assigning different kinds of orders to workers to prevent critical load and fatigue levels and to achieve an equal workload for the order pickers. Moreover, the equations can be useful in the design and operation of warehouses where the layout of the warehouse and the storage allocation policy need to be defined.

As already discussed, the posture analysis of the generic order picking process considered in this paper did not produce any critical result, since all activities led to an OWAS Index lower than the threshold value of 200. Thus, a fatigue analysis is performed in the following by comparing the total energy expenditure, E_{TOT}, with the recommended threshold value E_{T-rec}. Prior research (e.g., [28]) recommend that the average energy expenditure rate should not exceed 4.3 kcal/min to avoid high fatigue. Based on this value, E_{T-rec} can be calculated as $E_{T-rec} = 4.3 \cdot t_o$, where t_o is expressed in minutes. If E_{TOT} exceeds the threshold defined by E_{T-rec}, the fatigue level could induce higher injuries risks, lower operator performance, and lower process quality [28]. From a practical point of view, it would be useful to understand in which situation the operator's energy expenditure reaches this threshold value, expressed as a function of the item weight and the percentage of time spent on picking. Thus, the ratio of the total energy expenditure and the total recommended threshold value, R_E is defined in the following:

$$R_E = \frac{E_{TOT}}{E_{rec}} = \frac{\left[\left(\bar{E}_f \cdot \%t_f\right) + \sum_{l \in L} \bar{E}_{p-l}(w) \cdot \%t_{p-l}^{w}\right]}{4.3} \quad (9)$$

Figure 3 illustrates ratio R_E for different shares of the pick time in the total order completion time, $\%t_{p-l}^{w}$, different item weights, w, and for different pallet stocking locations, $l \in \{L\} = \{FP; HP-0; HP-M; HP-1\}$. Based on the resulting R_E-curves, we can define two areas of the graph. For points on the left side of the curves, the operator is exposed to an acceptable accumulated fatigue level, while for points on the right side of the curves, there is a risk of injuries, lower performance, and lower quality due to a fatigue level that exceeds the recommended threshold value.

Figure 3 also illustrates the trade-off between item weight and activity time: for example, if picking adopts a higher share of the total order picking time (e.g., $\%t_{p-l}^{w}$ is higher than 50%), the fatigue level could exceed the recommended value also for low item weights. On the other side, for standard picking from pallets stored on shelves, where $\%t_{p-l}^{w}$ could adopt a value at a maximum of around 30%, the total energy expenditure reaches the recommended value only for high item weights (here: greater than 14 kg for HP-0 when $\%t_{p-l}^{w}$=30%). Of course, such an evaluation has to consider the operating context. As reported by [4], in a traditional warehouse the actual picking activity is, on average, responsible for about 15% of the total order picking time. However, this percentage can reasonably increase, for example, when the warehouse is divided into different areas, with the highest required items stored together [29].

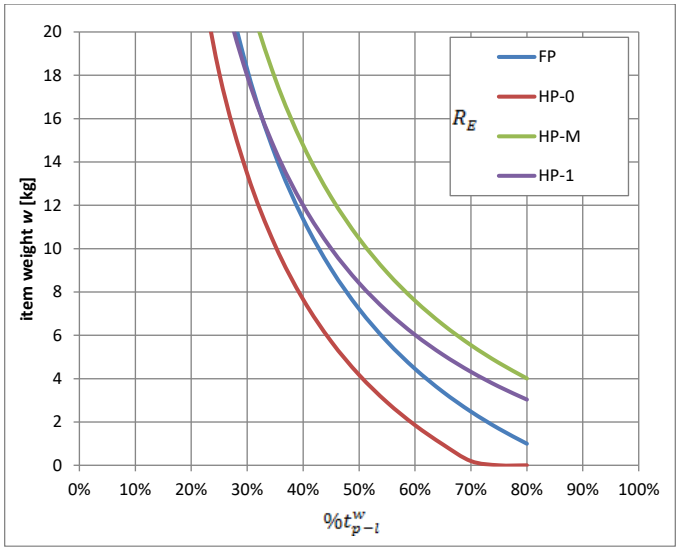


Fig. 4. Ratio of the total energy expenditure and the total recommended threshold.

VII. Conclusions

The high amount of manual material handling in order picking makes it a strenuous process for warehouse workers, who often need to handle heavy loads in awkward body postures while retrieving items, including stretching, bending and twisting. This can lead to increasing accumulated fatigue on the workers, which may increase the risk of injuries and discomfort. Although this has been realized in the human factors engineering literature for many decades, works that develop decision support models for managers have mainly focused on the economic aspects of order picking, neglecting ergonomics. While many analytical models exist in the order picking literature that guide managers in setting up efficient order picking processes, integrated analytical models for the assessment of ergonomic conditions in order picking are rare. This paper contributed to closing this research gap by developing analytical models for a joint posture and fatigue analysis in order picking.

In the proposed analytical models, the OWAS index and the energy expenditure concept are jointly used to evaluate the order picking process with respect to health risks and accumulated fatigue. These models can be of help as a decision support tool for managers in assessing ergonomics in their order picking warehouse using the OWAS method, which avoids costly field inquiries. Moreover, the developed models including the calculation of the energy expenditure rate can further be used in practice as a tool to assess if the working postures and item weights during order picking have a relevant impact on fatigue. This can facilitate the determination of workforce management, storage assignment methods or use of technical support systems.

One limiting factor of this paper is that the proposed models are, for now, not validated. Future research could thus focus on the validation of the developed models and on their improvement. A possible extension of this work could be the integration of travel time (in case workers walk through the warehouse using a hand trolley) in the calculation of accumulated fatigue, which may impact the results of this paper. Moreover, this paper could be used as a starting point for future

works that develop mathematical models for an integrated planning of order picking processes by considering both economic and ergonomic aspects.

References

[1] Battini, D., Calzavara, M., Persona, A., & Sgarbossa, F. (2015a). Order picking system design: the storage assignment and travel distance estimation (SA&TDE) joint method. International Journal of Production Research, 53(4), 1077-1093.

[2] de Koster, R., Le-Duc, T. and Roodbergen, K.J. (2007). Design and control of warehouse order picking: A literature review. European Journal of Operational Research, 182(2), 481-501.

[3] Napolitano, M. (2012). 2012 warehouse/DC operations survey: Mixed signals. Modern Materials Handling, 51 (11), 48-56.

[4] Tompkins, J. A., White, J. A., Bozer, Y. A., & Tanchoco, J. M. A. (2010). Facilities planning. John Wiley & Sons.

[5] Bureau of Labor Statistcs (BLS) (2016). Nonfatal occupational injuries and illnesses requiring days away from work, 2013. USDL-14-2246.

[6] Grosse, E. H., Glock, C. H., Jaber, M. Y., & Neumann, W. P. (2015). Incorporating human factors in order picking planning models: framework and research opportunities. International Journal of Production Research, 53(3), 695-717.

[7] Grosse, E. H.; Glock, C. H.; Neumann, W. P. (2017). Human factors in order picking: A content analysis of the literature. International Journal of Production Research, 55(5), 1260-1276.

[8] Grosse, E. H., Glock, C. H., & Jaber, M. Y. (2013). The effect of worker learning and forgetting on storage reassignment decisions in order picking systems. Computers & Industrial Engineering, 66(4), 653-662.

[9] Grosse, E. H., & Glock, C. H. (2015). The effect of worker learning on manual order picking processes. International Journal of Production Economics, 170, 882-890.

[10] Battini, D., Calzavara, M., Persona, A., & Sgarbossa, F. (2015b). A comparative analysis of different paperless picking systems. Industrial Management & Data Systems, 115(3), 483-503.

[11] Garg, A., Chaffin, D.B. and Gary D.H. (1978). Prediction of metabolic rates for manual materials handling jobs. The American Industrial Hygiene Association Journal, 39(8), 661-674.

[12] Battini, D., Glock, C. H., Grosse, E. H., Persona, A., & Sgarbossa, F. (2016a). Human energy expenditure in order picking storage assignment: A bi-objective method. Computers & Industrial Engineering, 94, 147-157.

[13] Calzavara, M., Glock, C. H., Grosse, E. H., Persona, A., & Sgarbossa, F. (2016a). Analysis of economic and ergonomic performance measures of different rack layouts in an order picking warehouse. Computers & Industrial Engineering. http://dx.doi.org/10.1016/j.cie.2016.07.001.

[14] Larco, J. A., de Koster, R., Roodbergen, K. J., & Dul, J. (2016). Managing warehouse efficiency and worker discomfort through enhanced storage assignment decisions. International Journal of Production Research, 1-16.

[15] Battini, D., Calzavara, M., Persona, A., & Sgarbossa, F. (2016b). Additional effort estimation due to ergonomicconditions in order picking systems. International Journal of Production Research, 1-11. http://dx.doi.org/10.1080/00207543.2016.1190879.

[16] Glock, C. H., Grosse, E. H., Elbert, R. M., & Franzke, T. (2016). Maverick picking: the impact of modifications in work schedules on manual order picking processes. International Journal of Production Research, 1-17. http://dx.doi.org/10.1080/00207543.2016.1252862.

[17] Garg, A. (1986). Biomechanical and ergonomic stresses in warehouse operations. IIE Transactions, 18(3), 246-250.

[18] Braam, I. T. J., van Dormolen, M., & Frings-Dresen, M. H. (1996). The work load of warehouse workers in three different working systems. International Journal of Industrial Ergonomics, 17(6), 469-480.

[19] Kadefors, R., & Forsman, M. (2000). Ergonomic evaluation of complex work: a participative approach employing video–computer interaction, exemplified in a study of order picking. International Journal of Industrial Ergonomics, 25(4), 435-445.

[20] St-Vincent, M., Denis, D., Imbeau, D., & Laberge, M. (2005). Work factors affecting manual materials handling in a warehouse superstore. International Journal of Industrial Ergonomics, 35(1), 33-46.

[21] Marras, W. S., Lavender, S. A., Ferguson, S. A., Splittstoesser, R. E., & Yang, G. (2010). Quantitative biomechanical workplace exposure measures: distribution centers. Journal of Electromyography and Kinesiology, 20(5), 813-822.

[22] Lavender, S. A., Sommerich, C. M., Johnson, M. R., & Radin, Z. (2010). Developing Ergonomic Interventions to Reduce Musculoskeletal Disorders in Grocery Distribution Centers, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54(15), 1229-1233).

[23] Nelson, N. A., & Hughes, R. E. (2009). Quantifying relationships between selected work-related risk factors and back pain: a systematic review of objective biomechanical measures and cost-related health outcomes. International Journal of Industrial Ergonomics, 39(1), 202-210.

[24] Battini, D., Persona, A., & Sgarbossa, F. (2014). Innovative real-time system to integrate ergonomic evaluations into warehouse design and management. Computers & Industrial Engineering, 77, 1-10.

[25] Louhevaara, V., Suurnäkki, T., Hinkkanen, S., & Helminen, P. (1992). OWAS: a method for the evaluation of postural load during work. Institute of Occupational Health. Centre for Occupational Safety.

[26] Calzavara, M., Glock, C. H., Grosse, E. H., Persona, A., & Sgarbossa, F. (2016b). Models for an ergonomic evaluation of order picking from different rack layouts. IFAC-PapersOnLine, 49(12), 1715-1720.

[27] Waters, T. R., Putz-Anderson, V., Garg, A., & Fine, L. J. (1993). Revised NIOSH equation for the design and evaluation of manual lifting tasks. Ergonomics, 36(7), 749-776.

[28] Price, A. D. (1990). Calculating relaxation allowances for construction operatives—Part 1: Metabolic cost. Applied Ergonomics, 21(4), 311-317.

[29] Bartholdi, J. J., & Hackman, S. T. (2016). Warehouse & Distribution Science: Release 0.97. Supply Chain and Logistics Institute.

A three-step heuristic for bi-directional flow-rack AS/RS with duration-of-stay storage policy

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Zhuxi Chen
College of Information Engineering, Yangzhou University
196 Huayang West Road, Yangzhou, Jiangsu, China
Email: zxchen@yzu.edu.cn

Jatinder N.D. Gupta
College of Business, University of Alabama in Huntsville
301 Sparkman Drive, Huntsville, AL, USA
Email: guptaj@uah.edu

***Abstract*—In a bi-directional flow-rack (BFR) automated storage and retrieval system (AS/RS), unit-loads can be stored or retrieved on both faces because the multi-deep bins in adjacent columns slope to opposite directions. This enables the handling machines work in dual-command cycle (DC) mode and restore the blocking unit-loads to same face. In a duration-of-stay storage (DOS) policy, the arrivals and the departures of the unit-loads within a planning horizon are known. Based on the turnover information, unit-loads can be carefully stored to appropriate bins to avoid blocking unit-loads. In addition, fast-turnover unit-loads should be stored to easy accessible bins, so that the overall travel time of warehouse operations can be remarkably reduced. This paper designs a three-step heuristic, which includes Grouping-Matching-Sequencing (GMS), for the storage assignment and operation generation in a BFR AS/RS with DOS storage policy. Experimental simulation results demonstrate that GMS can avoid blocking unit-loads, generate appropriate storage assignment plan and obtain high performance sequence of warehouse operations within acceptable computation time.**

I. Introduction

Since 2005, many multi-deep automated storage and retrieval systems (AS/RSs), including flow-rack [1], mobile-rack [2], 3D compact [3], double-deep [4], puzzle-based [5] and bi-directional flow-rack (BFR) [6], have been designed with the benefit of less number of aisles, comparing with single-deep AS/RSs, to mitigate the effects of the increasing land cost. In general, storage policies can be categorized as random (RAN), class-based (CB), full turnover-based (FTB) and duration-of-stay (DOS), which have been widely employed in normal (single-deep) AS/RSs [7]. Because the configurations of multi-deep AS/RSs are different from single-deep AS/RSs, novel methods for storage assignment, e.g. TB [8] and CB [9], and operation generation, e.g. dual-command cycle (DC) operation in 3D compact AS/RS [10], were proposed. In addition, cross aisle is deployed in mobile-rack AS/RS to reduce the weight of mobile-rack and the retrieval distance [11]. As well, novel heuristics for sequencing of retrievals have been designed for the puzzle-based system [12], [13].

In multi-deep flow-rack or bi-directional flow-rack AS/RSs, unit-loads follow first-in-first-out (FIFO) mode in each bin. As the consequence, outgoing unit-loads may be blocked by blocking unit-loads. Removing and restoring the blocking unit-loads incur enormous operational costs. Among the traditional storage policies, RAN is not suitable because it generates too many blocking unit-loads [1], [6], [14]. The storage assignment methods with CB [15] and DOS [16] storage policies might provide better performance. In DOS policy [17], the arrivals and departures of unit-loads within a planning horizon are known, the majority of unit-loads can share bins close to the pick-up/drop-off (P/D) station to reduce the travel time of operations [18]. Furthermore, a two-step heuristic was proposed [16], which assigns unit-loads into groups in the grouping step to minimize the number of blocking unit-loads and matches groups to bins to let as many as possible unit-loads be stored in easy accessible bins for the DOS storage policy in flow-rack AS/RSs. Experimental results demonstrated the effectiveness and efficiency of the two-step heuristic.

In the recently developed BFR AS/RS [6], the storage and retrieval operations are simultaneously conducted on both faces, which motivates the handling machines (HMs) perform DC operations, to reduce the operational costs. It can also remove and restore the blocking unit-loads on the same face. Since the operation generation method must choose the most appropriate available bins for restoring blocking unit-loads to minimize the travel time of removing blocking unit-loads, the restored bins must be determined before retrieving the requested unit-loads. Furthermore, because storage operations are generated on the storage face, retrieval operations are generated on the retrieval face and blocking unit-loads are removed to the restoring conveyor, the two-step heuristic cannot be employed in BFR AS/RSs directly. Therefore, a novel restoring policy and a DC generation mechanism have to be developed for BFR AS/RSs. In view of the above deficiencies of the two-step heuristic, this paper presents a three-step heuristic for the BFR AS/RS with DOS storage policy by carefully considering the configurations of BFR AS/RS, which not only generates DC operations but also introduces appropriate mechanism for restoring blocking unit-loads.

The rest of the paper is organized as follows. Section II introduces the problem to be addressed. A greedy-based heuristic and a three-step heuristic are designed in Section III. In Section IV, simulation experiments are conducted for the performance evaluation, followed by conclusions in Section V.

II. Problem description

Figure 1 illustrates an $L \times H \times M$ BFR AS/RS, which consists of L columns and H rows of M-deep slope bins. Normally, L is an even number to guarantee the two faces have the similar configurations. The set of bins is $\mathcal{K} = \{1, 2, ..., K\}$ where $K = LH$. Bin k locates at x_k^{th} column and y_k^{th} row where $x_k = k - \lfloor \frac{k-1}{L} \rfloor L$, $y_k = \lfloor \frac{k-1}{L} \rfloor + 1$ and $\forall k \in \mathcal{K}$. The notation $\lfloor x \rfloor$ stands for rounding down to the nearest integer.

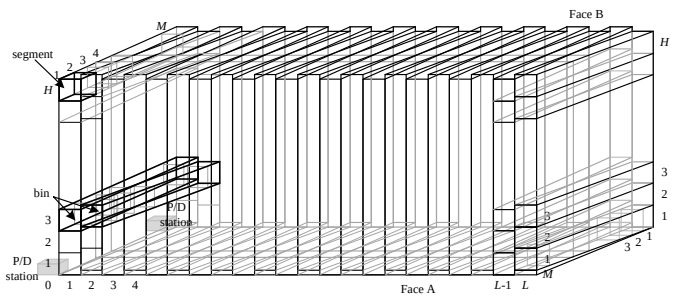


Figure 1. An $L \times H \times M$ bi-directional flow-rack

Two sets of bins are denoted as $\mathcal{K}_o = \{1, 3, ..., K-1\}$ and $\mathcal{K}_e = \{2, 4, ..., K\}$. Unit-loads slide from Face A (F_A) to Face B (F_B) in bins in $\mathcal{K}_e$ and slide from F_B to F_A in bins in $\mathcal{K}_o$. An available bin represents that a bin contains empty segments in $\mathcal{K}_e$ on F_A or in $\mathcal{K}_o$ on F_B. On each face, a pick-up/drop-off (P/D) station is arranged next to Bin $\#1$ with the coordinate of $(0, 1)$. Two HMs are deployed to F_A and F_B, respectively. The HMs move simultaneously in horizontal and vertical directions. Assume the horizontal and vertical travel times between two adjacent bins are t_h and t_v, respectively. For two coordinates $k_1 : (x_{k_1}, y_{k_1})$ and $k_2 : (x_{k_2}, y_{k_2})$, let $D_{k_1} = \max\{t_h x_{k_1}, t_v(y_{k_1} - 1)\}$, $d_{k_1 k_2} = \max\{t_h|x_{k_1} - x_{k_2}|, t_v|y_{k_1} - y_{k_2}|\}$ and $D_{k_1 k_2} = D_{k_1} + d_{k_1 k_2} + D_{k_2}$.

There are three types of warehouse operations: storage operation, retrieval operation and DC operation. In a storage operation, the HM picks up an incoming unit-load from the P/D station and stores it to an available bin. In a retrieval operation, the HM moves to a bin contains an outgoing unit-load, removes and restores blocking unit-loads to available bins on the same face, retrieves the outgoing unit-load and returns it to the P/D station. A storage operation and a retrieval operation can be merged to a DC operation, in which an incoming unit-load is stored and an outgoing unit-load is retrieved within a single operation. The dwell positions are the P/D stations.

In DOS storage policy, there are a number of T periods denoted as $1, 2, ..., T$ in a planning horizon. A unit-load can be described as a three-tuple with a form of (i, a_i, d_i) to represent that the unit-load i ($i \in \{1, 2, ..., N\}$) arrives at the beginning of period a_i and must be retrieved before the end of period d_i. At the beginning of a period t where $t \in \{1, ..., T\}$, a sequence of incoming unit-loads $S_t = ((i_1, t, d_{i_1}), ..., (i_m, t, d_{i_m}))$ where $d_{i_1} \leq ... \leq d_{i_m}$ and a set of outgoing unit-loads $R_t = \{(j_1, a_{j_1}, t), ..., (j_n, a_{j_n}, t)\}$ are established. All unit-loads in S_t must be stored and all unit-load in R_t must be retrieved within the period t. Generally, it assumes that all warehouse operations can be finished within each period. All unit-loads within a planning horizon can also be stored to the BFR, i.e., the number of simultaneous existing unit-loads is less than or equal to the number of segments. Based on the turnover information, unit-loads following FIFO mode stored in the same bin can avoid blocking unit-loads. What's more, it should store as many unit-loads as possible to bins close to P/D stations to reduce the operational cost of storage and retrieval. In summary, the purpose of storage assignment in BFR AS/RS with DOS storage policy is to assign unit-loads to appropriate bins to minimize the total travel time of all three types of warehouse operations within a planning horizon.

III. Proposed method

A. GREEDY heuristic

Apparently, the travel times of warehouse operations determine the operational cost and respond speed. The following ways can be considered to minimize the travel time. Firstly, store a unit-load to a bin close to the P/D station. This not only reduces the travel time of storage operation but also saves the travel time of retrieval operation. Secondly, minimize the distance between the storage location and the retrieval location. This can decrease the travel time of a DC operation. Lastly, minimize the number of blocking unit-loads, which can reduce the travel time of restoring in a retrieval or a DC operation. Based on above considerations, a greedy-based heuristic denoted as GREEDY is designed and illustrated as Algorithm 1.

Algorithm 1 GREEDY

Input: Unit-loads of the whole planning horizon
Output: Sequence of operations

$t \leftarrow 1$;
while ($t \leq T$) **do**
  Build S_t and R_t, $s_A \leftarrow head(S_t)$, $S_t \leftarrow S_t - \{s_A\}$, $s_B \leftarrow head(S_t)$, $S_t \leftarrow S_t - \{s_B\}$;
  while ($S_t \neq \emptyset || R_t \neq \emptyset$) **do**
    if (DC operation can be generated on F_A) **then**
      Choose an available bin k_s^A which contains fewest blocking unit-loads of s_A and a retrieval bin k_r^A where an outgoing r_A locates and is blocked by the fewest unit-loads with smallest $D_{k_s^A k_r^A}$;
    else if (Retrieval operation can be generated on F_A) **then**
      Choose a retrieval bin k_r^A where an outgoing r_A locates and is blocked by the fewest blocking unit-loads with smallest $D_{k_r^A}$;
    else if (Storage operation can be generated on F_A) **then**
      Choose an available bin k_s^A which contains fewest blocking unit-loads of s_A with smallest $D_{k_s^A}$;
    else
      Generate a dummy operation on F_A;
    if (DC operation can be generated on F_B) **then**
      Choose an available bin k_s^B which contains fewest blocking unit-loads of s_B and a retrieval bin k_r^B where an outgoing r_B locates and is blocked by the fewest unit-loads with smallest $D_{k_s^B k_r^B}$;
    else if (Retrieval operation can be generated on F_B) **then**
      Choose a retrieval bin k_r^B where an outgoing r_B locates and is blocked by the fewest unit-loads with smallest $D_{k_r^B}$;
    else if (Storage operation can be generated on F_B) **then**
      Choose an available bin k_s^B which contains fewest blocking unit-loads of s_B with smallest $D_{k_s^B}$;
    else
      Generate a dummy operation on F_B;
    if (s_A is stored) **then**
      $s_A \leftarrow head(S_t)$, $S_t \leftarrow S_t - \{s_A\}$;
    if (s_B is stored) **then**
      $s_B \leftarrow head(S_t)$, $S_t \leftarrow S_t - \{s_B\}$;
    if (r_A is retrieved) **then**
      $R_t \leftarrow R_t - \{r_A\}$;
    if (r_B is retrieved) **then**
      $R_t \leftarrow R_t - \{r_B\}$;
    Update stored unit-loads and empty segments;
  $t \leftarrow t + 1$;
return

In each cycle (Step 5-29), there is a warehouse operation to be performed on each face. The released segments cannot be employed for incoming and blocking unit-loads because the information of empty segments is not updated until Step 29. Dummy operation is introduced where the HM does nothing. On each face, a warehouse operation is generated based on the departures of incoming unit-load, outgoing unit-load and stored unit-loads. For storing and/or restoring, GREEDY checks the departures of stored unit-loads in available bins to minimize the number of blocking unit-loads. Normally, available bins with fewest blocking unit-loads should be chosen for storing/restoring. Removed blocking unit-loads are restored to nearest available bins on the same face. In order to balance the workload, GREEDY assigns incoming unit-loads to the two faces as equally as possible. For an incoming unit-load or a blocking unit-load, the time complexity of choosing an available bin is $O(KM)$. For generating a DC operation, the time complexity is $O(K^2M)$. Therefore, the overall time complexity of GREEDY is $O(K^2MN)$.

B. Grouping-matching-sequencing (GMS) heuristic

GREEDY minimizes the travel time of current generated warehouse operation, which only employs the departure of current incoming unit-load and departures of stored unit-loads. However, the arrivals and departures of all unit-loads within the whole planning horizon are known in DOS storage policy. Here is a more sophisticated method called GMS, which is designed to make good use of the turnover information for generating high performance warehouse operations.

In [16], two heuristics: forward-grouping and backward-grouping, were designed as the grouping step for the storage assignment in a flow-rack AS/RS with DOS storage policy. The proposed GMS heuristic utilizes the forward-grouping method as its grouping step with the time complexity of $O(N^2)$.

In the matching step (detailed in Algorithm 2), the unit-load groups are matched to bins. Unit-load groups are sequenced as $G = (g_1, g_2, ..., g_K)$ where $|g_1| \geq |g_2| \geq ... \geq |g_K|$ and $|g_k|$ is the number of unit-loads in g_k. Correspondingly, bins are separated into two sets $B_A = \{2, 4, ..., K\}$ and $B_B = \{1, 3, ..., K-1\}$ where unit-loads are stored to bins of B_A on F_A and to bins of B_B on F_B. The matching step is demonstrated in Algorithm 2 with a time complexity of $O(K^2)$.

Algorithm 2 Matching

Input: G, B_A, B_B
Output: Matching results

$n_A \leftarrow 0$, $n_B \leftarrow 0$, $k \leftarrow 1$;
while ($B_A \neq \emptyset \&\& B_B \neq \emptyset$) **do**
  if ($n_A \leq n_B$) **then**
    g_k is matched to bin $k_A \in B_A$ with minimal D_{k_A}, $B_A \leftarrow B_A - \{k_A\}$, $n_A \leftarrow n_A + |g_k|$, $k \leftarrow k+1$;
  else
    g_k is matched to bin $k_B \in B_B$ with minimal D_{k_B}, $B_B \leftarrow B_B - \{k_B\}$, $n_B \leftarrow n_B + |g_k|$, $k \leftarrow k+1$;
while ($B_A \neq \emptyset$) **do**
  g_k is matched to bin $k_A \in B_A$ with minimal D_{k_A}, $B_A \leftarrow B_A - \{k_A\}$, $k \leftarrow k+1$;
while ($B_B \neq \emptyset$) **do**
  g_k is matched to bin $k_B \in B_B$ with minimal D_{k_B}, $B_B \leftarrow B_B - \{k_B\}$, $k \leftarrow k+1$;
return

All unit-loads are matched to bins after the grouping and matching steps, where the storage assignment plan for the BFR AS/RS with DOS storage policy has been obtained. However, a sequence of warehouse operations must be generated to guide the HMs to execute the assignment plan with low operational cost. Therefore, the third step of the GMS heuristic should be carefully designed.

In period $t \in \{1, 2, ..., T\}$, incoming unit-loads can be separated into two sequences S_t^A and S_t^B to be stored to F_A and F_B, respectively. Outgoing unit-loads are organized as two sets R_t^A and R_t^B to be retrieved from corresponding F_A and F_B, respectively. Bins are separated into a total of $\frac{K}{2}$ pairs as $(k, k+1)$ where $k = 1, 3, ..., K-1$. For each pair, the blocking unit-loads removed from bin k are restored to bin $k+1$ on F_A (from bin $k+1$ to bin k on F_B) to ensure that blocking unit-loads can be restored to the nearest bin. In extreme case, "SWITCH" operation (in the last paragraph of Section 2.2 and illustrated in Figure 2 in [6]) is conducted to

Algorithm 3 Sequencing

Input: Matching results
Output: Operations

$t \leftarrow 1$;
while ($t \leq T$) **do**
  Generate S_t^A, S_t^B, R_t^A and R_t^B, $s_A \leftarrow head(S_t^A)$, $S_t^A \leftarrow S_t^A - \{s_A\}$, $s_B \leftarrow head(S_t^B)$, $S_t^B \leftarrow S_t^B - \{s_B\}$;
  while ($S_t^A \neq \emptyset || S_t^B \neq \emptyset || R_t^A \neq \emptyset || R_t^B \neq \emptyset$) **do**
    if ($s_A \neq NULL$ and bin k_{s_A} is available for s_A) **then**
      if ($R_t^A \neq \emptyset$) **then**
        Choose $r_A \in R_t^A$ to generate a DC operation with s_A;
      else
        Generate a storage operation which stores s_A to k_{s_A};
    else
      if ($R_t^A \neq \emptyset$) **then**
        Choose an appropriate outgoing unit-load r_A to generate a retrieval operation;
      else
        Generate a dummy operation;
    if ($s_B \neq NULL$ and bin k_{s_B} is available for s_B) **then**
      if ($R_t^B \neq \emptyset$) **then**
        Choose an outgoing unit-load r_B to generate a DC operation with s_B;
      else
        Generate a storage operation which stores s_B to k_{s_B};
    else
      if ($R_t^B \neq \emptyset$) **then**
        Choose an appropriate outgoing unit-load r_B to generate a retrieval operation;
      else
        Generate a dummy operation;
    if (s_A is stored) **then**
      $s_A \leftarrow head(S_t^A)$, $S_t^A \leftarrow S_t^A - \{s_A\}$;
    if (s_B is stored) **then**
      $s_B \leftarrow head(S_t^B)$, $S_t^B \leftarrow S_t^B - \{s_B\}$;
    if (a retrieval or a DC operation performed on F_A) **then**
      $R_t^A \leftarrow R_t^A - \{r_A\}$;
    if (a retrieval or a DC operation performed on F_B) **then**
      $R_t^B \leftarrow R_t^B - \{r_B\}$;
    Update stored unit-loads and empty segments;
  $t \leftarrow t+1$;
return .

ensure that all blocking unit-loads can be restored. Within a period, operations are separated into cycles to be performed. In each cycle, one operation is conducted on F_A and another operation is executed on F_B. Based on above mechanism, the sequencing step is illustrated in Algorithm 3.

Because unit-loads and bins are matched, the coordinate of storage location has been determined for the current incoming unit-load. For an incoming unit-load s_A on F_A, a DC operation or a storage operation should be generated if the bin k_{s_A} is available (Step 5). To generate a DC operation on F_A, an outgoing unit-loads must exist on F_A (Step 6). The choosing rule for the outgoing unit-load is as follow (Step 7): 1) if bin k_{s_B} is full with an outgoing unit-load in it, this outgoing unit-load is selected as r_A; else 2) if bin $k_{s_B^{next}}$ is full with an outgoing unit-load in it, this outgoing unit-load is selected as r_A (s_B^{next} is the following incoming unit-load of s_B); and else 3) choose an outgoing unit-load as r_A with minimal $D_{k_{s_A}k_{r_A}}$. If there is no outgoing unit-load on F_A, a storage operation for s_A is generated (Step 9). Retrieval or dummy operation is generated if the bin k_{s_A} is full (Step 12) or all incoming unit-loads stored to F_A in period t have been stored (Step 14). To generate a retrieval operation, the rule to choose the outgoing unit-load is similar to that in the generation of DC operation unless $D_{k_{r_A}}$ is employed as the selection measurement in 3). The operation generation on F_B is similar to that on F_A.

IV. Performance Evaluation

In this section, GREEDY and GMS are modelled with C++ on a laptop with Intel Core i7-4710MQ and 16GB DDR3L RAM. Let t_h be one second, t_v be two seconds and $N = 2LHM$ in this section. Four incoming modes, CENTERHIGH, DECREASING, INCREASING and RANDOM, are designed. The number of incoming unit-loads for each period is in the interval of $[1, \frac{2N}{T-1} - 1]$. The incoming modes are illustrated in Figure 2, in which blue bars represent the number of incoming unit-loads in each period. The duration of unit-load can be changed as well. Let d be the duration factor. Therefore, d_n is uniformly distributed in the interval $[a_n + 1, \min\{a_n + d, T\}]$.

For evaluating the impacts of flow-rack configuration, L, H and M take different values in this section. Let $\mathcal{T} = \max\{Lt_h, Ht_v\}$ be the normalization factor and $b = \min\{\frac{Lt_h}{\mathcal{T}}, \frac{Ht_v}{\mathcal{T}}\}$ be the shape factor of the working face of flow-rack. Obviously, the values of $\mathcal{T}$ and b are changed with L and H.

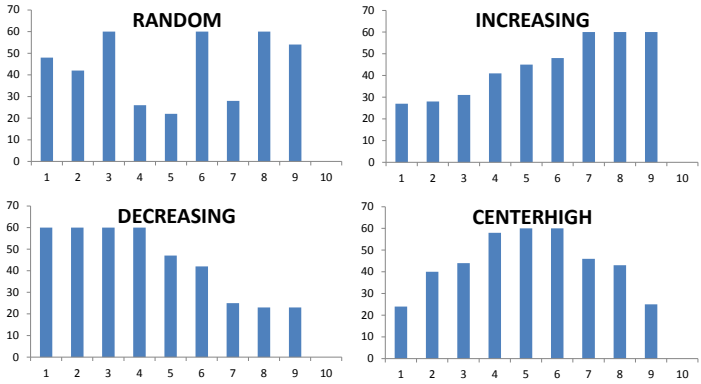


Figure 2. Incoming modes for simulation experiments

Each experiment repeats 100 times to obtain the average experimental results. In each repetition, let W_A^g and W_B^g be the total travel time of HM A and B obtained by GREEDY, respectively. W_A^G and W_B^G are the total travel time of HM A and B gained by GMS, respectively. The values of Average Relative Percentage Deviation (ARPD) are employed to evaluate the efficiencies of the two methods. For GREEDY, ARPDs are $\frac{W_A^g - \min\{W_A^g, W_A^G\}}{\min\{W_A^g, W_A^G\}}$ and $\frac{W_B^g - \min\{W_B^g, W_B^G\}}{\min\{W_B^g, W_B^G\}}$ for each repetition. For GMS, the ARPDs are similar. The CPU times and numbers of blocking unit-loads obtained by GREEDY and GMS of each repetition are recorded in the experiments.

A. Performance in different duration factors

From the definition of duration factor, larger value of duration factor implies more unit-loads have long duration, while smaller value of duration factor brings fewer long duration unit-loads. In addition, long duration unit-loads have high probability to block short duration ones. Therefore, the impacts of duration factors must be evaluated. Let $T = 10$, $L = 20$, $H = 10$ and $M = 5$ to avoid the influence of other factors. Experimental results are illustrated in Table I, in which $DC\%$ is the average percentage of DC operations, $\mathcal{B}$ is the average number of blocking unit-loads, $\mathcal{W}_A$ is the average value of W_A^g or W_A^G, $\mathcal{W}_B$ is the average value of W_A^g or W_A^G, $\mathcal{A}_A$ and $\mathcal{A}_B$ are the average values of corresponding ARPDs and "CPU" represents the average computation time of GREEDY or GMS in 100 repetitions.

With the increase of the duration factor, the number of long duration unit-loads increases. Correspondingly, the values of $\mathcal{B}$ demonstrate that GREEDY can avoid blocking unit-loads when $d = 2$ and $d = 4$ and obtains solutions with blocking unit-loads when $d = 6$ and $d = 10$. However, no blocking unit-load appears in INCREASING mode, since there are fewer long duration unit-loads in this mode than those in other modes. There is no blocking unit-load in the solution obtained by GMS, since it considers the arrivals and departures of all unit-loads and assigns these unit-loads into appropriate groups to avoid blocking unit-loads. Because unit-loads are stored to the two faces equally, GREEDY can generate more DC operations than those generated by GMS as illustrated by the values of $DC\%$. As well, the incoming modes affect the percentages of DC operations. The values of $\frac{\mathcal{W}_A}{N}$, $\frac{\mathcal{W}_B}{N}$, $\mathcal{A}_A$ and $\mathcal{A}_B$ illustrate that GMS obtains shorter travel time of operation than GREEDY does in most repetitions of experiments. One reason is that GMS avoids blocking unit-loads. Another reason is that GMS can assign as many as possible unit-loads to bins close to P/D stations. GMS requires much longer computation time than GREEDY does, because it considers arrivals and departures of all unit-loads. However, it is acceptable when GMS is able to obtain the solution for 2000 unit-loads in about 15 seconds for the reduction of operational costs.

B. Performance in different configurations of BFR

The configuration of BFR may affect the travel time of operation. Therefore, set L and H to different values with the constrain of $L \times H \approx 200$ to check the impacts of BFR configuration. To avoid influence, keep $T = 10$, $M = 5$ and $d = 10$.

Table I. PERFORMANCE OF GREEDY AND GMS WITH DIFFERENT VALUES OF d ($T = 10$, $L = 20$, $H = 10$ AND $M = 5$).

		GREEDY							GMS						
d		DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU	DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU
2	CENTERHIGH	56.61	0.00	12.57	12.49	15.53	12.09	0.33	51.99	0.00	10.89	11.15	0.00	0.00	12.78
	DECREASING	57.82	0.00	11.70	11.69	13.24	10.61	0.40	56.11	0.00	10.33	10.57	0.00	0.00	13.42
	INCREASING	56.49	0.00	13.24	13.07	17.72	13.58	0.32	48.46	0.00	11.26	11.52	0.00	0.00	14.92
	RANDOM	42.67	0.00	14.58	14.50	22.40	19.45	0.32	39.75	0.00	11.92	12.14	0.00	0.00	12.82
4	CENTERHIGH	40.82	0.00	15.67	15.64	6.01	4.31	0.28	36.48	0.00	14.79	15.00	0.00	0.00	13.36
	DECREASING	43.98	0.00	14.68	14.70	6.92	5.73	0.24	41.41	0.00	13.74	13.91	0.00	0.00	11.63
	INCREASING	39.33	0.00	16.41	16.28	5.47	3.18	0.34	33.01	0.00	15.57	15.79	0.00	0.00	15.68
	RANDOM	38.66	0.00	15.79	15.76	10.48	8.63	0.30	35.65	0.00	14.30	14.51	0.00	0.00	13.41
6	CENTERHIGH	35.46	0.45	17.15	17.15	3.29	2.05	0.23	31.18	0.00	16.61	16.80	0.02	0.01	13.48
	DECREASING	36.51	0.30	16.65	16.69	3.51	2.84	0.24	33.38	0.00	16.09	16.23	0.00	0.00	12.42
	INCREASING	35.08	0.00	17.53	17.47	2.78	1.46	0.25	29.14	0.00	17.06	17.22	0.00	0.03	15.34
	RANDOM	33.33	0.00	17.62	17.60	5.56	4.32	0.25	30.17	0.00	16.69	16.87	0.00	0.02	13.96
10	CENTERHIGH	34.29	1.38	17.34	17.33	5.02	3.87	0.24	30.36	0.00	16.53	16.70	0.00	0.00	12.08
	DECREASING	34.54	1.08	17.17	17.19	5.10	4.00	0.18	31.18	0.00	16.34	16.52	0.00	0.00	11.52
	INCREASING	37.18	0.00	17.12	17.07	5.91	4.15	0.28	32.00	0.00	16.17	16.39	0.00	0.00	13.95
	RANDOM	33.82	0.21	17.43	17.44	7.35	6.20	0.21	31.11	0.00	16.24	16.42	0.00	0.00	12.97

Table II. PERFORMANCE OF GREEDY AND GMS WITH DIFFERENT CONFIGURATIONS OF BFR ($T = 10$, $M = 5$, $d = 10$ AND $L \times H \approx 200$).

		GREEDY							GMS						
(L, H)		DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU	DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU
$(14, 14)$	CENTERHIGH	34.33	1.10	18.22	18.25	2.49	1.62	0.24	30.37	0.00	17.78	17.96	0.01	0.02	11.82
	DECREASING	34.47	0.91	18.12	18.13	2.95	1.77	0.24	31.06	0.00	17.60	17.82	0.00	0.00	11.08
	INCREASING	37.45	0.00	17.29	17.29	1.62	0.79	0.31	32.08	0.00	17.01	17.21	0.03	0.36	13.60
	RANDOM	34.14	0.10	17.69	17.76	3.61	3.07	0.29	31.47	0.00	17.08	17.23	0.00	0.02	12.26
$(16, 12)$	CENTERHIGH	34.34	1.45	17.13	17.15	3.31	2.57	0.19	30.41	0.00	16.58	16.72	0.00	0.01	11.14
	DECREASING	34.37	1.26	17.02	17.04	3.59	2.49	0.21	31.06	0.00	16.43	16.63	0.00	0.00	10.72
	INCREASING	37.41	0.00	16.44	16.44	2.52	1.42	0.24	32.02	0.00	16.04	16.24	0.04	0.17	12.70
	RANDOM	33.46	0.48	16.98	17.03	4.49	3.69	0.23	30.86	0.00	16.25	16.43	0.00	0.00	11.62
$(18, 11)$	CENTERHIGH	34.43	1.11	17.12	17.13	4.21	3.30	0.23	30.50	0.00	16.43	16.59	0.00	0.02	11.83
	DECREASING	34.45	1.45	17.00	17.02	4.18	3.17	0.29	31.05	0.00	16.32	16.50	0.00	0.00	11.35
	INCREASING	37.22	0.00	16.66	16.64	3.57	2.11	0.30	32.01	0.00	16.09	16.31	0.00	0.07	14.04
	RANDOM	34.07	0.25	17.04	17.06	5.30	4.44	0.29	31.37	0.00	16.18	16.34	0.00	0.00	12.29
$(20, 10)$	CENTERHIGH	34.29	1.38	17.34	17.33	5.02	3.87	0.24	30.36	0.00	16.53	16.70	0.00	0.00	12.08
	DECREASING	34.54	1.08	17.17	17.19	5.10	4.00	0.18	31.18	0.00	16.34	16.52	0.00	0.00	11.52
	INCREASING	37.18	0.00	17.12	17.07	5.91	4.15	0.28	32.00	0.00	16.17	16.39	0.00	0.00	13.95
	RANDOM	33.82	0.21	17.43	17.44	7.35	6.20	0.21	31.11	0.00	16.24	16.42	0.00	0.00	12.97

Table II illustrates the experimental results that the percentages of DC operations obtained by GREEDY are still higher than GMS does in different configurations. The values of $\mathcal{B}$ clarify that GMS can avoid blocking unit-loads. As well, there is no blocking unit-load in solutions of INCREASING mode obtained by GREEDY because of fewer long duration unit-loads in this mode than those in other modes. GMS obtains shorter travel time of operation than GREEDY does, which is demonstrated by the values of $\frac{\mathcal{W}_A}{N}$, $\frac{\mathcal{W}_B}{N}$, $\mathcal{A}_A$ and $\mathcal{A}_B$, even if GREEDY obtains more DC operations than GMS does. As expected, GMS requires longer CPU time than GREEDY does. The values of $\mathcal{T}$ and b are $(\mathcal{T}, b) = (28.0, 0.50)$, $(\mathcal{T}, b) = (24.0, 0.67)$, $(\mathcal{T}, b) = (22.0, 0.82)$ and $(\mathcal{T}, b) = (20.0, 1.00)$ when $(L, H) = (14, 14)$, $(L, H) = (16, 12)$, $(L, H) = (18, 11)$ and $(L, H) = (20, 10)$, respectively. The experimental results imply that let the value of b close to 1.00 can reduce the total travel time of warehouse operations when the total number of bins is determined.

C. Performance in different depths

Increase segments in each bin can reduce the number of bins, i.e., diminish the size of face, when the total number of segments is determined. Therefore, the impacts of depths are evaluated by setting $T = 10$, $d = 10$, $b = 1.00$, with the constrain of $L \times H \times M \approx 1000$ to avoid the influence of other factors. Experiments run with the cases of (L, H, M) being $(20, 10, 5)$, $(18, 9, 6)$, $(16, 8, 8)$ and $(14, 7, 10)$, respectively.

Table III demonstrates the experimental results that GMS can avoid blocking unit-loads in all incoming modes while GREEDY avoids blocking unit-loads only in INCREASING mode. The values of $DC\%$ clarify that GREEDY generates more DC operations than those generated by GMS. However, GMS obtains shorter travel time of operation than GREEDY does in most repetitions of all experiments. Again, GMS requires longer CPU time than GREEDY does but it is acceptable for overall operational cost reduction. The size of working face decreases, i.e., the number of bin decreases, with the increase of M. Correspondingly, the values of $\frac{\mathcal{W}_A}{N}$ and $\frac{\mathcal{W}_B}{N}$ gained by both heuristics decrease with the increase of M even if the number of blocking unit-loads obtained by GREEDY increases. Therefore, keep the overall segment unchanged by increasing number of segments in each bin and decreasing number of bins accordingly, can save the operational cost.

V. CONCLUSION

In this paper, a three-step heuristic contains grouping, matching and sequencing (GMS) steps, designed for the storage assignment and operation generation in a BFR AS/RS with DOS storage policy. The grouping step assigns unit-loads to groups, the matching step matches groups to bins

Table III. PERFORMANCE OF GREEDY AND GMS WITH DIFFERENT FLOW-RACK DEPTHS ($T = 10$, $d = 10$, $b = 1.00$ AND $L \times H \times M \approx 1000$).

M		GREEDY							GMS						
		DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU	DC%	$\mathcal{B}$	$\frac{\mathcal{W}_A}{N}$	$\frac{\mathcal{W}_B}{N}$	$\mathcal{A}_A$	$\mathcal{A}_B$	CPU
5	CENTERHIGH	34.29	1.38	17.34	17.33	5.02	3.87	0.24	30.36	0.00	16.53	16.70	0.00	0.00	12.08
	DECREASING	34.54	1.08	17.17	17.19	5.10	4.00	0.18	31.18	0.00	16.34	16.52	0.00	0.00	11.52
	INCREASING	37.18	0.00	17.12	17.07	5.91	4.15	0.28	32.00	0.00	16.17	16.39	0.00	0.00	13.95
	RANDOM	33.82	0.21	17.43	17.44	7.35	6.20	0.21	31.11	0.00	16.24	16.42	0.00	0.00	12.97
6	CENTERHIGH	34.30	1.80	15.63	15.62	5.03	3.86	0.28	30.43	0.00	14.88	15.04	0.00	0.00	11.14
	DECREASING	34.43	1.62	15.46	15.48	4.70	3.48	0.22	30.73	0.00	14.77	14.96	0.00	0.00	10.56
	INCREASING	37.19	0.00	15.36	15.31	5.47	4.04	0.19	31.98	0.00	14.57	14.72	0.01	0.02	12.73
	RANDOM	33.65	0.16	15.75	15.75	7.14	6.31	0.26	30.98	0.00	14.70	14.81	0.00	0.00	11.12
8	CENTERHIGH	34.26	3.54	13.84	13.84	4.26	3.36	0.15	30.24	0.00	13.28	13.39	0.00	0.02	11.70
	DECREASING	34.40	3.89	13.78	13.78	4.70	3.52	0.21	30.52	0.00	13.16	13.32	0.00	0.00	10.99
	INCREASING	37.14	0.00	13.74	13.70	5.84	4.16	0.24	31.92	0.00	12.99	13.15	0.00	0.00	13.58
	RANDOM	33.64	0.57	14.09	14.11	7.77	6.87	0.28	30.84	0.00	13.08	13.20	0.00	0.00	12.00
10	CENTERHIGH	34.47	5.33	12.07	12.03	4.06	3.12	0.25	30.46	0.00	11.60	11.67	0.00	0.03	10.40
	DECREASING	34.53	5.82	11.99	12.00	3.98	3.01	0.19	30.48	0.00	11.53	11.65	0.00	0.00	9.82
	INCREASING	36.98	0.00	12.10	12.05	5.69	4.54	0.20	31.58	0.00	11.45	11.53	0.03	0.04	12.05
	RANDOM	33.07	1.30	12.44	12.43	7.94	6.65	0.24	30.51	0.00	11.52	11.66	0.00	0.00	10.87

and the sequencing step generates the sequence of warehouse operations. Because the arrivals and departures of unit-loads within a planning horizon are fully employed, GMS can not only store as many as possible unit-loads to easy accessible bins on the BFR to minimize the travel times of warehouse operations, but also avoid blocking unit-loads to save travel times of restoring blocking unit-loads. Experimental results illustrate that GMS can gain low operational cost and achieve high performance warehouse operations within an acceptable computation time. The percentages of DC operations obtained by GMS can be further improved because the balance of workload between the two faces is only considered in the matching step.

The future work will focus on how to decrease the workload difference between the two HMs. Another interesting topic might be how to deploy multi-shuttle machine to the BFR AS/RS to improve the throughput performance.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation for Young Scientists of China [71602169] and by Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China [16KJB520045].

REFERENCES

[1] Z. Sari, C. Saygin, and N. Ghouali, "Travel-time models for flow-rack automated storage and retrieval systems," *The International Journal of Advanced Manufacturing Technology*, vol. 25, no. 9-10, pp. 979–987, 2005.

[2] T.-H. Chang, H.-P. Fu, and K.-Y. Hu, "Innovative application of an integrated multi-level conveying device to a mobile storage system," *The International Journal of Advanced Manufacturing Technology*, vol. 29, no. 9-10, pp. 962–968, 2006.

[3] R. B. De Koster, T. Le-Duc, and Y. Yu, "Optimal storage rack design for a 3-dimensional compact as/rs," *International Journal of Production Research*, vol. 46, no. 6, pp. 1495–1514, 2008.

[4] T. Lerher, M. Sraml, I. Potrc, and T. Tollazzi, "Travel time models for double-deep automated storage and retrieval systems," *International Journal of Production Research*, vol. 48, no. 11, pp. 3151–3172, 2010.

[5] K. R. Gue, K. Furmans, Z. Seibold, and O. Uludag, "Gridstore: A puzzle-based storage system with decentralized control," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 429–438, 2014.

[6] Z. Chen, X. Li, and J. N. Gupta, "A bi-directional flow-rack automated storage and retrieval system for unit-load warehouses," *International Journal of Production Research*, vol. 53, no. 14, pp. 4176–4188, 2015.

[7] K. J. Roodbergen and I. F. A. Vis, "A survey of literature on automated storage and retrieval systems," *European Journal of Operational Research*, vol. 194, no. 2, pp. 343–362, 2009.

[8] Y. Yu and R. De Koster, "Designing an optimal turnover-based storage rack for a 3d compact automated storage and retrieval system," *International Journal of Production Research*, vol. 47, no. 6, pp. 1551–1571, 2009.

[9] Y. Yu and M. De Koster, "Optimal zone boundaries for two-class-based compact three-dimensional automated storage and retrieval systems," *IIE Transactions*, vol. 41, no. 3, pp. 194–208, 2009.

[10] Y. Yu and R. De Koster, "Sequencing heuristics for storing and retrieving unit loads in 3d compact automated warehousing systems," *IIE Transactions*, vol. 44, no. 2, pp. 69–87, 2012.

[11] K.-Y. Hu, T.-H. Chang, H.-P. Fu, and H. Yeh, "Improvement order picking in mobile storage systems with a middle cross aisle," *International Journal of Production Research*, vol. 47, no. 4, pp. 1089–1104, 2009.

[12] N. Zaerpour, Y. Yu, and R. De Koster, "Small is beautiful: a framework for evaluating and optimizing live-cube compact storage systems," *Transportation Science*, vol. 51, no. 1, pp. 34–51, 2017.

[13] M. Mirzaei, R. De Koster, and N. Zaerpour, "Modelling load retrievals in puzzle-based storage systems," *International Journal of Production Research*, 2017. [Online]. Available: http://dx.doi.org/10.1080/00207543.2017.1304660

[14] Z. Chen and Y. Li, "Dual-command operation generation in bi-directional flowrack automated storage and retrieval systems with random storage policy," *Lecture Notes in Electrical Engineering*, vol. 360, pp. 125–131, 2016.

[15] O. Cardin, P. Castagna, Z. Sari, and N. Meghelli, "Performance evaluation of in-deep class storage for flow-rack as/rs," *International Journal of Production Research*, vol. 50, no. 23, pp. 6775–6791, 2012.

[16] Z. Chen, X. Li, and J. N. Gupta, "Sequencing the storages and retrievals for flow-rack automated storage and retrieval systems with duration-of-stay storage policy," *International Journal of Production Research*, vol. 54, no. 4, pp. 984–998, 2016.

[17] M. Goetschalckx and H. D. Ratliff, "Shared storage policies based on the duration stay of unit loads," *Management Science*, vol. 36, no. 9, pp. 1120–1132, 1990.

[18] L. Chen, A. Langevin, and D. Riopel, "The storage location assignment and interleaving problem in an automated storage/retrieval system with shared storage," *International Journal of Production Research*, vol. 48, no. 4, pp. 991–1011, 2010.

Throughput Analysis of Random Storage Systems: Closest Eligible Location Pays

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Anja Heßler, Christoph Schwindt
Clausthal University of Technology
Julius-Albert-Straße 2
38678 Clausthal-Zellerfeld, Germany
Email: {anja.hessler, christoph.schwindt}@tu-clausthal.de

***Abstract*—We study the performance of storage and retrieval systems executing single-command cycles under the closest eligible location rule. For each arriving storage or retrieval request, this strategy assigns a storage location with minimum cycle time. The performance of the storage and retrieval system is analyzed in terms of different key performance indicators and indices, such as expected cycle time, expected storage times, or storage-time imbalance. Assuming that the arrivals of the storage requests as well as the arrivals of the retrieval requests follow independent Poisson processes, we propose Markov models for the evolution of the spatial inventory distribution in the storage and develop closed-form expressions for the stationary probabilities of a given storage location being selected for a storage or retrieval request. Based on these probabilities, we derive formulas for the key performance indicators and indices and investigate characteristic curves of the expected cycle time and the storage-time imbalance in a numerical experiment. Some of the results obtained were unexpected and disprove popular premises stated in widely-used approaches for throughput analysis. Comparing our results to those of alternative approaches from the literature shows that the latter tend to significantly underestimate the expected maximum system throughput of storage and retrieval systems. The analysis also provides insights into the crucial role that the assignment of storage locations plays in the performance of storage and retrieval systems.**

I. Introduction

In warehouse design, appropriately dimensioning the storage and retrieval (S/R) system presupposes an accurate model of the system throughput under steady-state conditions. The expected maximum system throughput, calculated from the reciprocal expected cycle time, is largely influenced by the storage and retrieval strategy, which defines the way in which storage and retrieval requests are executed during warehouse operation. Given a set of requests to be processed, the strategy partitions the set into operation cycles of the S/R system and allocates appropriate storage locations to each request. The way in which the latter allocation is done will be referred to as the storage assignment strategy. Disregarding the time savings achieved by optimally allotting storage locations to storage and retrieval requests may heavily bias the throughput analysis. Based on continuous-time Markov chains, we derive analytical results for the expected cycle time and further key performance indicators (KPIs) and indices of random storage systems keeping multiple stock keeping units (SKUs). The storage assignment strategy considered in this paper is the closest eligible location (CEL) rule, which for each arriving request selects a storage location allowing for a minimum cycle time. Applied to storage requests exclusively, this strategy is also called the closest open location (COL) rule.

The remainder of this paper is organized as follows. In Sect. II we review the literature of performance evaluation models for random storage systems operated under some cycle-time aware storage assignment strategy. In Sect. III we introduce the Markov models and present the analytical results for different KPIs of S/R systems. Based on numerical experiments, in Sect. IV we investigate characteristic curves for the expected cycle time and the storage-time imbalance and compare our results to alternative approaches from the literature. A short summary and some remarks on future research avenues are given in Sect. V.

II. Literature Review

The estimation of expected cycle times for random storage systems has been studied extensively in the literature. According to the focus of our study, we only include papers in our review which consider storage and retrieval strategies aiming at small cycle times for random storage systems. In their seminal paper [1], Hausman et al. asserted that in case of stable capacity utilization over time, the COL rule can be approximated reasonably well by the pure random storage (PRS) strategy, which allocates storage locations to storage or retrieval requests in a non-optimized, purely random way. Since then, this assumption has been adopted in many modeling approaches for random storage systems, see, e. g., the handbook [2, pp. 659 f.] or the industry standards [3] and [4]. However, recent simulation studies comparing the COL rule with PRS show that the expected cycle times significantly differ when the storage level is below 100 %, see, e. g., [5, pp. 88 ff.].

The literature on analytical performance evaluation of cycle-time aware storage and retrieval strategies is quite sparse. Yamashita et al. [6] examine a unit-load single-SKU random storage system replenished according to an (s, q) reordering policy. After an exponentially distributed lead time, the q reordered units are stored according to the COL rule in the warehouse. The arrivals of retrieval requests follow a Poisson process, and the storage location to be allocated to an entering retrieval request is chosen at random among the currently occupied storage locations. Starting from these assumptions, the authors model the S/R system as a Markov chain and develop an aggregation approach to reduce the state space

and the computational effort for calculating the stationary distribution of the stochastic process. Park and Lee [7] consider a unit-load random storage system with Poisson arrivals of the storage requests, which are served under the COL rule. The authors premise i. i. d. storage times with a finite mean W_ℓ for each SKU $\ell = 1, \ldots, L$. Under these assumptions, the S/R system can be modeled as an Erlang loss system with ordered entry, i. e., an $M/G/N/N$ queueing system with each customer taking the first idle server among the ordered servers representing the N ordered storage locations. Note that the mean storage time of a loading unit only depends on SKU ℓ but is independent of the loading unit's storage location n. Moreover, constant mean storage times for SKU ℓ imply that the arrival rate of retrieval requests must be proportional to the current stock of SKU ℓ. The assumptions of the model may, e. g., be satisfied if for each SKU ℓ the arriving retrieval requests form a pure birth process with stock-dependent intensities and the retrieval requests are served according to the PRS strategy.

Besides these analytical models, there also exist heuristic approaches. Fukunari and Malmborg [8] consider Poisson arrivals of the storage requests and a known average storage time of the stored SKUs, which is again independent of the assigned storage location. To model this S/R system as an $M/M/N$ queueing system, they assume a hypothetical load consolidation scenario, in which all SKUs are relocated after each retrieval so that at any time, only the storage locations nearest to the I/O point are occupied. The occupancy probability of each storage location is then approximated based on the stationary probabilities for the number of customers in the queueing system. Due to the hypothetical load consolidation, this model typically only provides rough estimations for real-life storages. Malmborg [9] improves the accuracy of the model by estimating the expected number of free storage locations with a smaller distance to the I/O point than the furthest occupied location. This approach is based on linear regressions on the results of simulation studies.

To the best of our knowledge there is no analytical model available in the open literature that considers the widely-used storage and retrieval strategy minimizing the cycle time for both storage and retrieval requests. We refer to this storage assignment strategy as the CEL rule. The analytical approaches mentioned above only take the COL rule for storage requests into account but ignore the optimization potential of the analogous rule applied to retrieval requests. They assume random storage allocation for the retrieval requests or known mean storage times of the SKUs. Yamashita et al. [6] argue that the CEL rule is not appropriate for practical use because loading units at unfavorable locations may stay in the storage for a very long time. Although the latter observation is correct, this argument does not form an obstacle to the application of the CEL strategy in practice. To avoid obsolescence of inventory, it suffices to periodically retrieve loading units reaching critical storage times. Aside from these occasional FIFO retrievals, the S/R system can then be operated under the CEL rule, which allows for a significant increase in the maximum throughput of the S/R system. In the next section, we introduce the Markov models which form the basis of our throughput analysis.

III. Mathematical Model

We consider a rack storage under random storage strategy serviced by stacker cranes performing single-command cycles. We assume that storage and retrieval requests of different SKUs $\ell = 1, \ldots, L$ are released according to independent Poisson processes with intensities λ_ℓ and μ_ℓ, respectively, and are executed in the sequence of their arrivals. Storage locations are assigned to arriving storage and retrieval requests following the CEL rule. Note that since this rule selects the eligible storage location with minimum sum of the storage and the retrieval cycle times, it is the optimum online storage assignment strategy under our assumptions. If the storage is full or empty, storage and retrieval requests, respectively, are assumed to be lost. We further suppose that all requests refer to single loading units like pallets and that each storage location can hold exactly one loading unit of an arbitrary SKU. For what follows, we establish the convention that the N storage locations $n = 1, \ldots, N$ are numbered in order of nondecreasing cycle times.

In [10] we proposed an aggregation approach to calculate the storage and retrieval access probabilities of each storage location under these assumptions. To sketch the basic idea of this approach, we briefly recapitulate the model for the special case of homogeneous inventory of a single SKU. Let $\{Y(t) \mid t \geq 0\} = \{(Y_1(t), \ldots, Y_N(t)) \mid t \geq 0\}$ be the stochastic process with state space $E = \{0, 1\}^N$ modeling the evolution of the inventory distributed over the storage locations, where $Y_n(t)$ denotes the Bernoulli variable that equals 1 if storage location n is occupied at time t, and 0 otherwise. The feasible state transitions directly follow from the CEL rule. For each arriving request, the first feasible location in the sequence $n = 1, \ldots, N$ of the storage locations is selected, i. e., for a storage request the first free, and for a retrieval request the first occupied location is chosen. The respective transition rates correspond to the arrival rates λ and μ of storage and retrieval requests for the single SKU. Since the size of the state space grows exponentially with increasing number N of storage locations, the stationary distribution of the homogeneous and irreducible continuous-time Markov chain $\{Y(t) \mid t \geq 0\}$ can only be computed for very small instances. Similar to Yamashita et al. [6], we derive truncated and aggregated versions of process $\{Y(t) \mid t \geq 0\}$ to reduce the computational complexity. For fixed $n \in \{0, \ldots, N\}$, we consider the stochastic process $\{Z^{(n)}(t) \mid t \geq 0\}$ with random variables $Z^{(n)}(t)$ counting the number of occupied locations among the first n storage locations. This process corresponds to the throughput process of an $M/M/1/n$ queueing system with arrival rate λ and service rate μ and thus, the formulas for the stationary probabilities $\pi_k^{(n)}$ are known. According to the PASTA property (Poisson arrivals see time averages, see [11, p. 394]), arriving requests see the stationary distribution and we obtain the following formulas for the stationary probabilities $P_S(n)$ and $P_R(n)$ of storage location n being chosen for a storage and a retrieval request, respectively.

$$P_S(n) = \pi_{n-1}^{(n-1)} - \pi_n^{(n)} \tag{1a}$$

$$P_R(n) = \pi_0^{(n-1)} - \pi_0^{(n)} \tag{1b}$$

Storage location n is assigned to a storage request if and only if storage locations 1 to $n-1$ are occupied and the n-th storage location is free. This holds true exactly if the first $n-1$ but

not the first n locations are occupied. Analogously, storage location n is selected for a retrieval request if and only if storage locations 1 to $n-1$ are empty and the n-th storage location contains a loading unit.

In [10] we mentioned that these results extend to the case of multiple SKUs $\ell = 1, \ldots, L$. For given storage location $n \in \{0, \ldots, N\}$, we define the process $\{Z^{(n)}(t) \mid t \geq 0\} = \{(Z_1^{(n)}(t), \ldots, Z_L^{(n)}(t)) \mid t \geq 0\}$ with random variables $Z_\ell^{(n)}(t)$ counting the number of storage locations among the first n locations occupied by SKU ℓ at time t, which again is a homogeneous and irreducible continuous-time Markov chain. By $E^{(n)} = \{(j_1, \ldots, j_L) \mid \sum_{\ell=1}^{L} j_\ell \leq n\}$ we denote the state space of the process $\{Z^{(n)}(t) \mid t \geq 0\}$. The following product-form equations provide the stationary probabilities $\pi^{(n)}_{(j_1,\ldots,j_L)}$, where $\rho_\ell = \frac{\lambda_\ell}{\mu_\ell}$:

$$\pi^{(n)}_{(j_1,\ldots,j_L)} = \pi^{(n)}_{(0,\ldots,0)} \cdot \prod_{\ell=1}^{L} \rho_\ell^{j_\ell}, \ (j_1, \ldots, j_L) \in E^{(n)} \quad \text{(2a)}$$

$$\sum_{k=0}^{n} \sum_{\substack{j_1,\ldots,j_L: \\ j_1+\cdots+j_L=k}} \pi^{(n)}_{(0,\ldots,0)} \cdot \prod_{\ell=1}^{L} \rho_\ell^{j_\ell} = 1 \quad \text{(2b)}$$

According to (2b), the normalization constants

$$c(n) := \left(\pi^{(n)}_{(0,\ldots,0)}\right)^{-1} = \sum_{k=0}^{n} \sum_{\substack{j_1,\ldots,j_L: \\ j_1+\cdots+j_L=k}} \prod_{\ell=1}^{L} \rho_\ell^{j_\ell} \quad (3)$$

of $\{Z^{(n)}(t) \mid t \geq 0\}$ can be computed recursively using Buzen's convolution algorithm [12] for the normalization constant of a Gordon-Newell network. Applying the same arguments as for homogeneous inventory, the access probability of storage location n for storage requests is

$$P_S(n) = \sum_{\substack{j_1,\ldots,j_L: \\ j_1+\ldots+j_L=n-1}} \pi^{(n-1)}_{(j_1,\ldots,j_L)} - \sum_{\substack{j_1,\ldots,j_L: \\ j_1+\ldots+j_L=n}} \pi^{(n)}_{(j_1,\ldots,j_L)}. \quad (4)$$

Using the normalization constants, (4) can be rewritten as

$$P_S(n) = \frac{c(n-1)}{c(n)} - \frac{c(n-2)}{c(n-1)}. \quad (5)$$

Whereas the storage access probability of storage location n coincides for all types of SKU $\ell = 1, \ldots, L$, the retrieval access probability depends on ℓ. The probability of storage location n being chosen for a retrieval request of item type ℓ is given by

$$\begin{aligned} P_R(\ell, n) &= \sum_{\substack{j_1,\ldots,j_L: \\ j_\ell=0}} \pi^{(n-1)}_{(j_1,\ldots,j_L)} - \sum_{\substack{j_1,\ldots,j_L: \\ j_\ell=0}} \pi^{(n)}_{(j_1,\ldots,j_L)} \\ &= \rho_\ell \cdot P_S(n). \end{aligned} \quad (6)$$

For the retrieval access probability of storage location n, we obtain

$$\begin{aligned} P_R(n) &= \frac{1}{\sum_{\ell=1}^{L} \mu_\ell} \sum_{\ell=1}^{L} \mu_\ell \cdot P_R(\ell, n) \\ &= \rho \cdot P_S(n) \end{aligned} \quad (7)$$

with $\rho = \lambda/\mu = \sum_{\ell=1}^{L} \lambda_\ell / \sum_{\ell=1}^{L} \mu_\ell$. Recall that storage and retrieval requests are lost if they cannot be served immediately. Since the probabilities $P_S(n)$ and $P_R(n)$ refer to all arriving requests, the service levels of the S/R system for storage and retrieval requests are given by

$$\alpha_S = \sum_{n=1}^{N} P_S(n) = \frac{c(N-1)}{c(N)} \quad (8)$$

$$\alpha_R = \sum_{n=1}^{N} P_R(n) = \rho \cdot \alpha_S. \quad (9)$$

To obtain the expected cycle time, the storage and retrieval access probabilities referring to requests that enter the S/R system have to be considered. According to the service levels in (8) and (9), the expected cycle time for single-command cycles equals

$$\begin{aligned} \bar{t} &= \sum_{n=1}^{N} t(n) \left[\frac{\lambda}{\lambda+\mu} \cdot \frac{P_S(n)}{\alpha_S} + \frac{\mu}{\lambda+\mu} \cdot \frac{P_R(n)}{\alpha_R} \right] \\ &= \sum_{n=1}^{N} t(n) \cdot \frac{P_S(n)}{\alpha_S} \end{aligned} \quad (10)$$

with $t(n)$ denoting the mean of the single-command storage and the single-command retrieval cycle times for storage location n. Note that the normalized access probabilities coincide for storage and retrieval requests, i. e., $P_S(n)/\alpha_S = P_R(n)/\alpha_R$. This finding is plausible because for a specific storage location, storages and retrievals always alternate. The latter observation has been exploited in the analysis of Park and Lee [7].

The mathematical model presented above allows for the evaluation of further indices and KPIs for the steady-state S/R system. First, we consider the expected number of loading units of SKU $\ell = 1, \ldots, L$ stored in the warehouse. According to [11, p. 581], this number can be computed as

$$\bar{n}_\ell = \sum_{n=1}^{N} \rho_\ell^n \cdot \frac{c(N-n)}{c(N)}. \quad (11)$$

The mean occupancy rate is given by the ratio of the expected number of occupied storage locations and the capacity N of the storage:

$$\Gamma = \frac{1}{N} \cdot \sum_{\ell=1}^{L} \bar{n}_\ell \quad (12)$$

By applying Little's law, we obtain the expected storage time a loading unit of SKU $\ell = 1, \ldots, L$ spends in the S/R system:

$$W_\ell = \frac{\bar{n}_\ell}{\alpha_S \cdot \lambda_\ell} \quad (13)$$

While these indices refer to the S/R system as a whole, we now turn to location-specific parameters. Due to the assumption of unit-load locations and requests, the steady-state occupancy of storage location n is a binary random variable $Y^{(n)}$ and the occupancy rate of storage location n, similarly

to [13, p. 89], arises as

$$\begin{aligned}\gamma_n &= P\left(Y^{(n)} = 1\right) = E\left[Y^{(n)}\right] \\ &= E\left[\sum_{\ell=1}^{L} Z_\ell^{(n)}\right] - E\left[\sum_{\ell=1}^{L} Z_\ell^{(n-1)}\right] \\ &= 1 - \frac{1}{c(n)} \sum_{\nu=0}^{n-1} c(\nu) + \frac{1}{c(n-1)} \sum_{\nu=0}^{n-2} c(\nu). \end{aligned} \tag{14}$$

In (14), the occupancy rate γ_n equals the difference of the total expected number of occupied storage locations among the first n and $n-1$ storage locations, respectively. Based on the stationary analysis of the process $\{Z^{(n)}(t) \mid t \geq 0\}$ and some elementary calculus, these expected values can be written in terms of the normalization constants. Once the occupancy rate of each storage location n is known, the expected storage time of a loading unit in storage location n is obtained according to Little's law as follows:

$$w_n = \frac{\gamma_n}{\lambda \cdot P_S(n)} \tag{15}$$

Finally, we examine the special case of $\rho_\ell = 1$ for all SKUs $\ell = 1, \ldots, L$ in more detail. This case is particularly relevant because for most practical applications, in the long-run the reorder or production rate λ_ℓ should coincide with the demand rate μ_ℓ for each SKU ℓ. Based on our closed queueing network model, after some combinatorial calculus we obtain the following simplified expressions:

$$P_S(n) = P_R(n) = \frac{L}{(n+L-1)(n+L)} \tag{16a}$$

$$\alpha_S = \alpha_R = \frac{N}{N+L} \tag{16b}$$

$$\Gamma = \gamma_n = \frac{L}{L+1} \tag{16c}$$

$$w_n = \frac{(n+L-1)(n+L)}{\lambda(L+1)} \tag{16d}$$

The first formulas confirm the intuition of access probabilities decreasing in the index n of the storage location. This means that the larger the distance to the I/O point, the less frequently the location is selected for a storage or retrieval request. The storage and the retrieval service levels coincide and deteriorate with growing number L of SKUs. Furthermore, the occupancy rates are identical for all storage locations n and increase in L. Finally, the last equation shows that, as expected, the mean storage time grows with increasing index n of the storage location. However, the storage time remains quadratically bounded in n.

IV. Numerical experiments

In this section, we report on the results of numerical experiments that we performed using our Markov model. First, we present the characteristic curves for the expected cycle time and the storage-time imbalance as functions of the storage capacity N and the number L of SKUs. Then, we compare the outcome of our throughput analysis with the results of alternative approaches from the literature. The analysis is based on the storage rack of a high bay warehouse with 10 levels, 60 bays, square storage locations, and a single I/O point located in the bottom left corner considered in [10]. Assuming that the stacker crane can move simultaneously in horizontal and vertical directions with the same velocity in both directions, the travel times to the storage locations are calculated according to the Tchebychev metric. Furthermore, we suppose that $\lambda_\ell = \mu_\ell$ and hence $\rho_\ell = 1$ for all SKUs $\ell = 1, \ldots, L$.

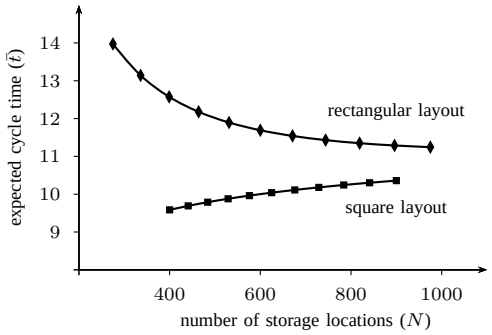


Fig. 1. Expected cycle time as a function of the total number of storage locations

A. Characteristic curves

In what follows, we interpret the KPIs and indices of a storage as functions of the dimensional parameters N and L. For certain KPIs and indices these functions depend on the selected storage and retrieval strategy. In particular, this holds true for the expected cycle time and the storage-time imbalance. We have chosen these two KPIs due to their relevance to storage design and because the CEL rule is intended to minimize the expected cycle time at the cost of an increased storage-time imbalance. In the second subsection we will compare the performance of the CEL rule with respect to these two KPIs against the scenario considered in the model of Park and Lee [7].

At first, we analyze the relationship between the expected cycle time and the storage capacity N. Since the cycle time depends on the arrangement of the storage locations, we consider a square as well as a rectangular storage rack. For the rectangular layout, we start with the storage rack containing 600 storage locations arranged in 10 levels and 60 bays. We then obtain further instances by iteratively adding or removing one level and one bay. In the same way, we construct the instances for the square layout starting with 25 levels and 25 bays. Fig. 1 shows the resulting characteristic curves for both layout types. Somewhat surprisingly, the lines show opposing trends for the two layout types. Whereas for the square storage rack, the expected cycle time grows with increasing capacity, it decreases for the rectangular layout. This effect can be explained as follows. In the square layout, each of the storage locations that is added to an existing storage rack is associated with a maximum cycle time. This observation is not true for the rectangular layout, in which typically the distances between the I/O point and some of the added storage locations are smaller than for certain storage locations contained in the previous configuration.

The characteristic curve in Fig. 2 displays the relationship between the expected cycle time and the number L of SKUs for the original rectangular layout with 600 storage locations. The results show that the expected cycle time is an increasing

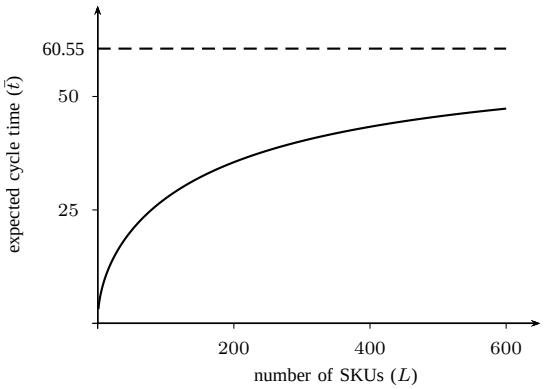


Fig. 2. Expected cycle time as a function of the number of SKUs

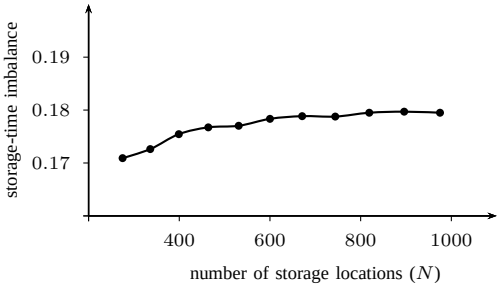


Fig. 3. Storage-time imbalance as a function of the total number of storage locations

and concave function in the number of SKUs stored in the warehouse. In [10] it is mentioned for $\rho_\ell = 1$ for all SKUs ℓ that $\bar{t}$ attains the expected cycle time of the PRS rule as L tends to infinity, the latter cycle time being independent of L.

Next, we turn to the expected storage time of each storage location n. As an aggregate KPI, we propose the storage-time imbalance, which we define to be the percentage of storage locations for which the expected storage time of SKUs exceeds the double of the total mean obtained by averaging over all storage locations. Storage capacity N is varied by considering the same rectangular layouts as before. Fig. 3 reveals that the storage-time imbalance is only little affected by the storage capacity. The curve shows a global trend towards increasing imbalance for larger storage racks. However, note that the function is not consistently increasing since the storage-time imbalance decreases with growing N until the storage time of a further storage location exceeds the threshold.

Finally, we investigate the relationship between the storage-time imbalance and the number L of SKUs for the scenario with 600 storage locations. The characteristic curve in Fig. 4 shows that this relation follows a decreasing, almost linear step function. The more types of SKUs are stored in the warehouse, the less storage locations contain loading units over a long time. Since for $L \to \infty$, the storage behaves like under the PRS rule, the storage-time imbalance tends to zero when the number of SKUs becomes large. In sum, the curves of Figs. 3 and 4 suggest that obsolescence of inventory does not pose a serious obstacle to the application of the CEL rule.

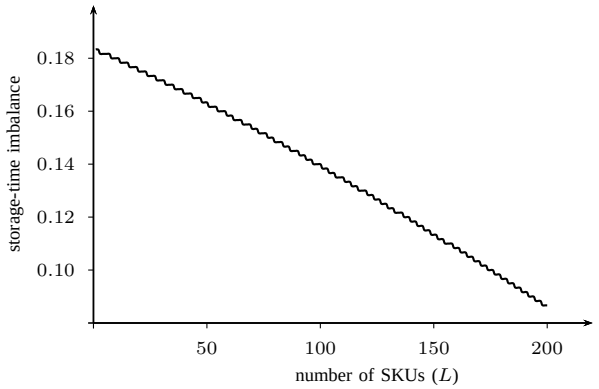


Fig. 4. Storage-time imbalance as a function of the number of SKUs

TABLE I. COMPARISON TO MODELS ASSUMING UNIFORM DISTRIBUTION

Model	MM(1)	MM(3)	MM(5)	MM(10)	MM(15)	UD
$\bar{t}$	3.12	5.34	6.86	9.61	11.69	60.55
δ_{tv}	0.92	0.87	0.83	0.77	0.73	–

B. Comparison with previous approaches

We proceed by putting the performance of the CEL rule into perspective with the results obtained for strategies where locations are, at least to some extent, chosen at random. At first, we revisit the experiments discussed in [10]. In this study, we compared Hausman et al.'s classical assumption of uniformly distributed access probabilities (UD model), i. e., an S/R system operated under the PRS strategy, with the access probabilities arising from our Markov model for the CEL rule (model MM(L), with L as the number of SKUs). Table I displays the expected cycle time $\bar{t}$ for the different models. Moreover, as a measure of inaccuracy for the UD model, we provide the total variation distances δ_{tv} between the uniform distribution and the normalized storage (and retrieval) access probabilities $P_S(n)/\alpha_S$. The results show that when the S/R system is operated under the CEL rule and the number of SKUs is relatively small, assuming a uniform distribution heavily underestimates the maximum system throughput, which seems to be contrary to the assertion of Hausman et al. mentioned in Sect. II. As a consequence, applying the industry standards [3] and [4] may lead to a largely oversized S/R system. We notice, however, that the expected cycle time significantly increases and hence the error of the UD model diminishes when the number L of SKUs increases, as visualized in Fig. 2. This observation is confirmed by the decrease of the total variation distance with growing number L of SKUs. Recall that the normalized storage (and retrieval) access probabilities converge to the uniform distribution as L tends to infinity. However, the convergence may be rather slow. In our example with 600 storage locations, an acceptable total variation distance of 5 % is only attained with more than 2700 types of items. The lessons learnt from the experiments can also be put differently. Given that the UD assumption is satisfied if the storage is operated under the PRS strategy, the large gaps between the expected cycle times provide evidence for the considerable optimization potential of the CEL rule.

Finally, we compare our model to the approach by Park and Lee [7], which represents an S/R system that is run applying

TABLE II. EXPECTED CYCLE TIMES OF PARK AND LEE'S MODEL

Model	PL(1)	PL(3)	PL(5)	PL(10)	PL(15)
$\bar{t}$	31.72	46.56	51.69	56.50	58.07

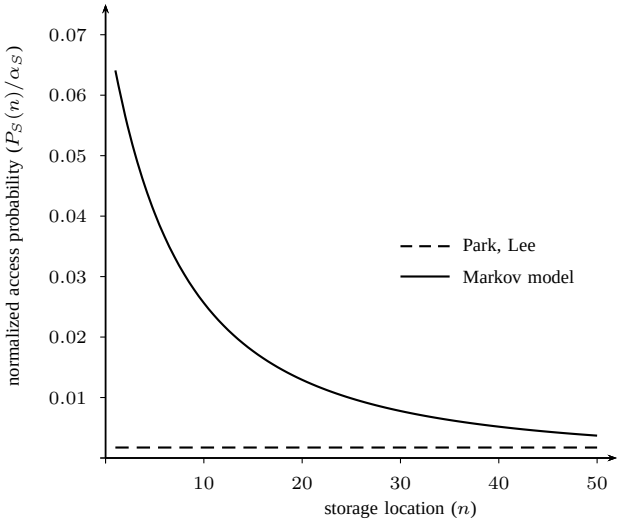


Fig. 5. Comparison of access probabilities

the COL rule to the storage requests. For the purpose of comparison, we calculate the mean storage time W_ℓ yielded by (13) of our model for each SKU ℓ. Subsequently, we insert the values of W_ℓ as parameters into the model of Park and Lee and compute the storage and retrieval access probabilities according to their formulas. The resulting expected cycle times $\bar{t}$ are displayed in Table II, where PL(L) stands for the model of Park and Lee with L SKUs.

The results show that the model by Park and Lee does not underestimate the maximum system throughput as much as the UD model, but the resulting expected cycle times are still significantly higher than those of our Markov model. For the instance with $L = 15$, a closer look to the normalized storage access probabilities $P_S(n)/\alpha_S$ of the first 50 storage locations in Fig. 5 reveals that in our model, the chances of favorable storage locations being assigned to a storage request are significantly higher than those of the Park and Lee model. It came as a surprise to us that the probabilities computed by the approach of Park and Lee only slightly deviate from the uniform distribution. This seems to indicate that their assumptions, especially with respect to i. i. d. storage times, do not comply with a storage and retrieval strategy minimizing the expected cycle time. As it is shown by (16d), the expected storage time w_n of a loading unit is a quadratic function in the location index n, which clearly contradicts the premise of identically distributed storage times.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed Markov models for the system throughput analysis of S/R systems executing single-command cycles under the CEL rule. We showed how the expected cycle time and further KPIs and indices of S/R systems can be calculated efficiently based on aggregate Markov models. As a major finding we could demonstrate that in contrast to a conjecture stated in the literature, inventory obsolescence is not a big issue when applying the CEL strategy. Comparing our results with alternative approaches from the literature reveals that the latter may significantly underestimate the maximum system throughput. Future research will be concerned with extending the models to dual-command cycles and considering batch arrivals of the storage and retrieval requests.

REFERENCES

[1] W. H. Hausman, L. B. Schwarz, and S. C. Graves, "Optimal storage assignment in automatic warehousing systems," *Management Science*, vol. 22, no. 6, pp. 629–638, 1976.

[2] D. Arnold, H. Isermann, A. Kuhn, H. Tempelmeier, and K. Furmans, *Handbuch Logistik*, 3rd ed. Berlin: Springer, 2008.

[3] Fédération Européenne de la Manutention, *FEM-Richtlinie 9.851: Leistungsnachweis für RBG-Spielzeiten*. Frankfurt a. M.: VDMA-Verlag, 2003.

[4] Verein Deutscher Ingenieure, *VDI-Richtlinie 3561: Testspiele zum Leistungsvergleich und zur Abnahme von Regalförderzeugen*. Düsseldorf: VDI-Verlag, 1973.

[5] M. Glass, "Schnellläuferstrategien in Lagern und dynamische Zonierung," Dissertation, Technische Universität Dresden, Dresden, 2008.

[6] H. Yamashita, H. Ohtani, and S. Suzuki, "An exact aggregation for inventory distribution in an automatic warehousing system," *Journal of the Operations Research Society of Japan*, vol. 41, no. 4, pp. 492–508, 1998.

[7] B.-C. Park and M.-K. Lee, "Closest open location rule under stochastic demand," *International Journal of Production Research*, vol. 45, no. 7, pp. 1695–1705, 2007.

[8] M. Fukunari and C. J. Malmborg, "A heuristic travel time model for random storage systems using closest open location load dispatching," *International Journal of Production Research*, vol. 46, no. 8, pp. 2215–2228, 2008.

[9] C. J. Malmborg, "Application of storage consolidation intervals to approximate random storage transaction times with closest open location load dispatching," *International Journal of Production Research*, vol. 50, no. 24, pp. 7242–7255, 2012.

[10] A. Heßler and C. Schwindt, "Markov models for system throughput analysis in warehouse design," in *Operations Research Proceedings 2016*, A. Fink, A. Fügenschuh, and M. J. Geiger, Eds. Springer, 2017 (in press).

[11] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton: Princeton Univ. Press, 2009.

[12] J. P. Buzen, "Computational algorithms for closed queueing networks with exponential servers," *Communications of the ACM*, vol. 16, no. 9, pp. 527–531, 1973.

[13] R. B. Cooper, *Introduction to Queueing Theory*, 2nd ed. New York: Elsevier North Holland, 1981.

Forecasting

Shelf Space Allocation during Promotion Periods

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Mahmoud El-Banna
Industrial Engineering Department,
German Jordanian University,
P.O. Box 35247, Amman 11180, Jordan,
Email: mahmoud.banna@gju.edu.jo

Ahmed Alabadleh
Logistic Sciences Department,
German Jordanian University,
P.O. Box 35247, Amman 11180, Jordan,
Email: ahmed.alabadleh@gju.edu.jo

***Abstract*—The need to have the right decisions related to the shelf space allocation during the promotion time is extremely important to maximize both the total gross profit and the customer service level. In this paper, multi-objective piecewise linearization model with an exponential function of demand in selling price, competitive price, and the number of faces is proposed to handle the shelf space allocation during the promotion time. The model was examined on small grocery hypermarket dataset and the results of the model were successfully verified.**

***Keywords*—Retail management, shelf space allocation, promotion, demand, multi-objective, piecewise linearization, competitor price.**

I. Introduction

A report produced by [1] stated that the total retail revenue for the top 250 retailers around the world reached nearly $4.3 trillion in 2012. They also mentioned that the retail industry continues to grow despite the difficulty in the economy back in 2011 in the United States and Europe. In addition, in March 2011, the retail industry in Japan emerged as a part of the economic growth during their recovery from the earthquake and tsunami. In spite of this growth, the average gross margins of grocery retail stores was about 28% of sales, but net profit after taxes was only about 3.0%. Because of the competition and low profit margins, retail chains are continuously searching for ways to reduce costs, increase sales, and improve customer service.

A study conducted on retail industry showed that between 2000 and 2009, the average number of items in store assortments increased by 30%, while the shelf space available at disposal was limited [2]. Therefore, efficient shelf space allocation is one of the key factors to gain an advantage in the highly competitive retail industry. Decisions involved in shelf space allocation are: (1) how much shelf space to allocate to a particular product category and (2) how much shelf space to allocate to each different product within a product category [3]. In this paper we will assume the category shelf space is determined, and we need to allocate the shelf space for each product within the category. Shelf space allocation decisions are extremely important as indicated by [4], since most customer decisions are made at the point of purchase, which indicates that the product choice of customers may be affected by in-store factors such as selling price, and shelf space allocated for that product, and out-store factors such as competitor price.

This paper focuses on how retailers can maximize their profit, meanwhile minimizing their stock-out rate, through the better arrangement of their products (i.e determining the number of faces), using the historical demand data, and its relation to the shelf space, the selling price, and the competitor price.

This work is an extension to Reyes model [3], where new demand function of selling price, competitor price, and shelf space is developed, multi-objective piece wise linearization approach instead of nonlinear multi-objective approach is used to solve the model.

The rest of this paper is organized as follows: Section II, we review the relevant literature. Section III, methodology is presented which includes the benchmarked model that has been extended, the demand sensitivity functions, the nonlinear multi-objective optimization model, and piecewise linearization approach. Section IV presents the computational experiments results, and Section V presents the summary of the proposed research.

II. Literature Review

In this section, presentation of the key literatures related to the shelf space allocation planning will be provided in the following areas: experimental studies, and optimization models. However, for the comprehensive work in this area , the recent works [5] and [6] are fruitful references.

At this stage, it is important to distinguish the difference between the assortment planning and shelf space planning. According to the category management definition reported in [5], the assortment planning is

> a listing of decisions based on consumer choice behavior and substitution effects, while shelf space planning is facing and replenishment decisions based on space elasticity effects, limited shelf space, and operational restocking constraints.

In the experimental studies domain, Gruen et al. [7] reported that when a customer cannot find a desired product on the shelf, 45% of these customers will substitute another product, 15% will delay purchase, 31% will switch to another store, and 9% will not buy the item at all. This examination was performed in response to stockouts across eight categories at retailers worldwide. In a large-scale experimental study on shelf management, [8] reported that the location of a product on the shelf is far more important to determining sales than the number of facings allocated to the product, as long as a minimum threshold to avoid out of stocks was maintained. This result was also emphasized by [9], who investigated the impact of location factors and derived optimal space allocation rules for different location profiles. This approach may be good for new established stores, while for the well-established

ones, changing the location of the products may disturb or annoy the customers who get used to finding their products in certain locations.
In terms of the optimization techniques used in assortment planning and shelf space allocation, Reyes et. al [3] used nonlinear integer goal programming formulation that considers both profitability and customer service factors. Brijs. et al [10] proposed a microeconomic model for product selection based on the use of frequent item sets obtained from association rule mining. More specifically, they integrated the notion of frequent sets with some important microeconomic parameters that are often used by retailers to support their product selection decision-making process into an integer programming model. Results indicated that their model was able to identify cross-selling effects implicitly by using frequent item sets, instead of having to estimate cross-selling parameters explicitly (as it is often done in product selection and shelf-space allocation models). The dynamic programming approach was proposed by [11] to select optimally among a given set of products and allocate integer shelf-space units to the selected products in supermarkets. The approach was designed to consider general objective-function specifications that account for space elasticity, costs of sales, and potential demand-related marketing variables. The optimization is subject to constraints due to product supply availability, 'block' product allocation, and operational requirements. Corstjens and Doyle [12] developed a geometric programming model considering space elasticity and cross elasticity of products and solved it using the branch and bound algorithm as proposed by Cole et al. [13]. They extended this model to a dynamic model [14] that incorporates the goodwill or carryover effects which retailers accrue from a given merchandise area over a period. Ryzin and Mahajan [15] proposed a newsboy model to determine the purchase quantities of different SKUs within a category. The paper makes use of a multinomial logit choice model to determine the individual purchase decisions. They showed that the optimal assortment has a simple structure for their stylistic model in which the unit prices are identical. Majahan and van Ryzin [15] also allowed dynamic substitution to occur when the SKU demanded was not in stock and used sample path analysis to obtain structural results.
Few integrated models jointly considered product assortment, shelf space allocation, and inventory replenishment decisions to maximize retailers profit under the given operating constraints [16]. More recently, several other approaches like game theory [17] and data mining [18] are used for product assortment and allocation problems in retailing.

III. Methodology

The proposed model is an extension to Reyes et. al model [3], which will be explained later in this section. It is a multi-objective model to solve the shelf space allocation problem, where the first objective measures the profitability, the second is related to the customer service level, and the third is related to the overstocking issue. As known, increasing the customer service level tends to increase the overstocking units and to decrease the profitability, therefore, trade off analysis is required to find the optimum solution. The proposed work differs from Reyes et. al [3] regarding the following points

- The Demand functional form of selling prices, competitor price, and the number of faces is exponential, while Reyes is a power functional form.
- Selling price, competitor price, and the number of faces model parameters are obtained from historical data by using nonlinear regression approach, while Reyes model is using previously assumed parameters.
- In the proposed model, a preemptive method is used to deal with the multi-objective problem, while Reyes model used weighted approach.
- Piecewise technique is used to convert the nonlinear model into a linear model, while Reyes model is a nonlinear model.
- Reyes model considered only the selling price, and the space, while in our model we add the competitor price in addition to the previously mentioned factors.

The proposed functional form of the demand versus the selling price, the competitor price and the number of faces will be presented next, followed by the optimization model used to solve the shelf location problem.

A. Demand Estimation Models

The main factors that are considered in demand estimation are the shelf space and substitution. The researchers [5], [19] reported four different types of demand estimation models: shelf-space, multinomial logit (MNL), exogenous, and locational choice models.
The shelf-space demand model is the demand of the product that is mainly affected by the provided shelf space for itself and the other products within the same category. MNL demand model [20] is a utility-based model assumes that customers associate a utility with each product, and they choose the highest utility alternative available. Exogenous demand model considers the demand of the product that rises because of activities outside the regular course (e.g. increase in ice cream demand due to a sudden temperature rise). In the locational choice model [21], products are treated as a set of attributes, and each product is specified by its location in attribute space. Individual preferences are defined by associating point in the attribute space with each customer. This preference is defined on the attributes rather than on the product themselves as in the MNL. In this paper exogenous and shelf-space models are coupled in one model.

B. Demand Sensitivity

The observed product demand (i.e. at the point of sales) is the net result demand of many influential factors such as selling price, competitor price, shelf space, perceived quality, brand reputation, personal loyalty to a brand, willing to try different products, promotions, advertisement, the location of the store and others [22]. Price, competitor price, and shelf space are considered among the most important factors that influence the demand. While no standard functional form that fits all cases between selling price, competitor price and demand, the demand and the space is represented by space elasticity relationship [14]. In this paper, a novel practical functional form is proposed for the selling price, the competitor price, and the shelf space with the demand. Before introducing the

proposed model, the model proposed by Reyes et. al [3] will be explained and the weakness points found in this model will be addressed.

1) Reyes Model: Reyes et. al model [3] presented the demand as a function of the price and shelf space as:

$$D_i = d_i \Pi \Sigma = d_i (\frac{p_{avg}}{p_i})^{\delta} (\frac{S_i}{S_i^o})^{\lambda} \quad (1)$$

where,
D_i: demand rate for product i
d_i: base level demand rate for product i , which is the demand rate obtained when all the prices in the category products are the same (i.e. $p_i = p_{av}$) and the base level shelf allocation is used (i.e. $S_i = S_i^o$)
Π: price sensitivity factor, which is equal to $(\frac{p_{avg}}{p_i})^{\delta}$, p_{av}: average category price, p_i: product selling price, and δ: price exponent value of 0.5.
Σ: space sensitivity factor, which is equal to $(\frac{S_i}{S_i^o})^{\lambda}$, S_i: actual self space allocation, S_i^o: base level allocation $= (d_i/d_{total})S_N$, and S_N: total category shelf space , and λ shelf space exponent value of 0.5.
Examining (1), to estimate d_i base level demand, price and space sensitivity factors should be equal to one, which means that selling price of the product should be equal to average category selling price (i.e. $p_{av} = p_i$), and the actual space allocation should be also equal to the base level allocation ($S_i = S_i^o$). Obviously, the application of this equation in real life is impractical, since the situation where $p_{av} = p_i$ is difficult to obtain in reality most of the times, and the determination of the base level space allocation depends on knowing the base level demand (d_i), which is not available at this stage yet. Therefore, the need for obtaining a practical and applicable function of the demand sensitivity and its influential factors is necessary, which is the proposed demand model of this paper.

2) Proposed model: In this model, the demand rate consists mainly of three terms; the selling price, the competitor price, and the shelf space. The functional behavior of the demand is the exponential type,(2), where the demand curve versus the selling price is decreasing exponentially, while versus the other two terms is increasing exponentially. The parameters values of the demand equation (i.e. θ, β,and γ) can be estimated using the historical data (i.e. nonlinear regression).

$$\begin{aligned} D = {} & \theta_1 - \theta_2 \exp^{(-\theta_3\, price)} \\ & + \beta_1 - \beta_2 \exp^{(-\beta_3\, competitor\ price)} \\ & + \gamma_1 - \gamma_2 \exp^{(-\gamma_3\, number\ of\ faces)} \end{aligned} \quad (2)$$

In contrast to Reyes demand model, there is no need for the base level demand, instead, the demand is expressed as a function of the saturated demand (i.e. $\theta_1, \beta_1, \gamma_1$) as shown in Figs.1, 2, and 3 while the other nonlinear regression parameters ($\theta_2, \theta_3, \beta_2, \beta_3, \gamma_2, \gamma_3$) are responsible for the slope and shift from y axis.

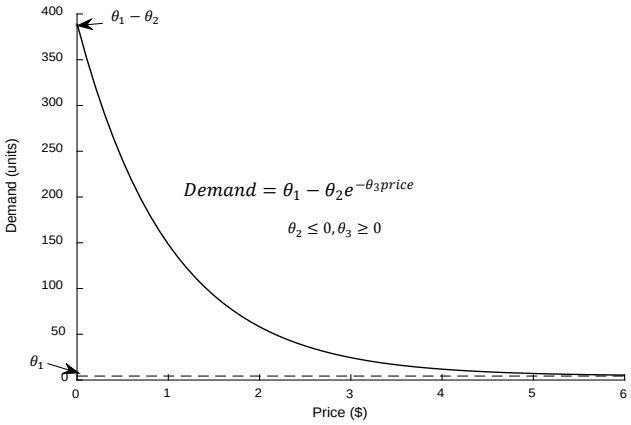


Fig. 1: Demand versus selling price

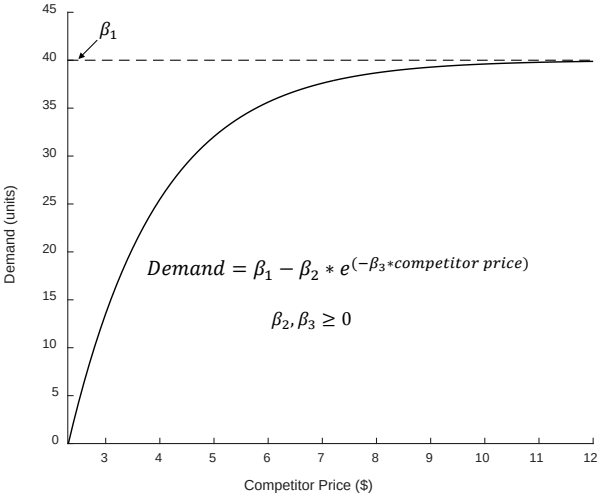


Fig. 2: Demand versus competitor price

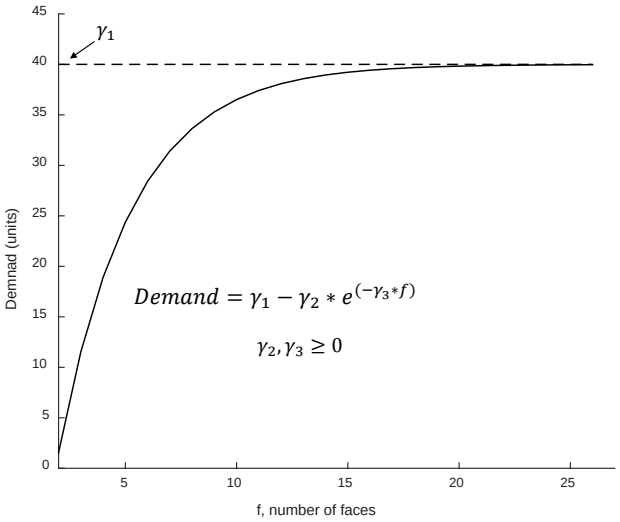


Fig. 3: Demand versus number of faces, f

C. Multi-Objective Piecewise Linear Model

A multi-objective model is developed to solve the space allocation problem, where lexicographic optimization approach is used to handle the multi-objective model. In lexicographic

optimization approach, the model is solved in successive steps, where the first objective function is used with the original constraints to get the first solution, then a new constraint is added to the original model to avoid deteriorating the first objective. The next step is to solve the new model (original constraints with the new constraint from the previous step) using the second objective, upon finding a solution, a new constraint is added to prevent the deterioration of the second objective. The process repeats itself in the same manner until the last objective function is solved.

Now, let's define the model assumptions, parameters, and variables that will be used to build the model.

Model assumptions:
The following assumptions are considered for the proposed model:

- Assortment planning is performed to select the products that will be placed on shelves from all possible SKUs.
- Cross elasticity among shelved products is ignored.
- With out loss of generality, the items will be placed on only one place or pre-specified shelf.
- The shortage (overstock) items on shelves are supplied (removed) during the night or at the early morning before the store door opens.
- The selling price, and the competitor price are assumed to be known in advance.
- Different products cannot share the same facing.

Parameters:
n: Total number of products in the category,
c_j: Purchased cost per unit j,
p_j: Selling price per unit j,
pc_j: Competitor selling price per unit j,
UL_j^o: Upper limit for the number of overstock units j,
UL_j^s: Upper limit for the number of shortage units j,
Cap_j: capacity of faces for the product j, i.e how many items in each face,
w_j: Face width for product j;
Sh_L: Shelf length specified for the category,
θ_1,θ_2,θ_3,β_1,β_2,β_3,γ_1,γ_2,and γ_3: Estimated nonlinear regression parameters from historical data, (where $\theta_2 \leq 0$ and other parameters are $+ve$),

Decision variables:
f_j: no. of faces for product j,
S_j^-: no. of shortage units for product j,
S_j^+: no. of overstock units for product j,

$$Maximize = \sum_{j=1}^{n} (p_j - c_j) * D_j \tag{3a}$$

$$Minimize = \sum_{j=1}^{n} S_j^- \tag{3b}$$

$$Minimize = \sum_{j=1}^{n} S_j^+ \tag{3c}$$

Subject to:

$$\begin{aligned} D_j = {} & \theta_{1j} - \theta_{2j} \exp^{(-\theta_{3j} p_j)} \\ & + \beta_{1j} - \beta_{2j} \exp^{(-\beta_{3j} pc_j)} \\ & + \gamma_{1j} - \gamma_{2j} \exp^{(-\gamma_{3j} f_j)} \qquad \forall j \end{aligned} \tag{4}$$

$$f_j * Cap_j + S_j^- - S_j^+ = D_j \qquad \forall j \tag{5}$$

$$S_j^+ \leq UL_j^o \qquad \forall j \tag{6}$$

$$S_j^- \leq UL_j^s \qquad \forall j \tag{7}$$

$$\sum_{j=1}^{n} w_j \frac{D_j}{Cap_j} \leq Sh_L \tag{8}$$

$$f_j, S_j^-, S_j^+ \geq 0 \qquad and \quad integer \qquad \forall j \tag{9}$$

Equations (3a),(3b), and (3c) represent the objective functions; where (3a) is a nonlinear objective function for the total margin profit occurred from selling the products, (3b) is the total number of shortage units, and (3c) is the total number of overstock units. Constraint (4) represents the exponential functional form of the desired demand versus the selling price, the competitor price, and the number of faces. Constraint (5) states the level of shelf inventory relative to the desired demand for each product (i.e. S_j^- for the shortage, S_j^+ for the overstock). Constraints (6) and (7) states the upper limit for the number of overstocks and shortages units for each product. Constraint (8) guarantee that the products located on the desired shelf will not exceed the pre-specified shelf length for them. Constraint (9) ensures that number of facings of products, shortage and overstock units are a nonnegative integer.

1) Piecewise Linearization: A nonlinear function can be approximated as a sequential of linear segments or pieces as shown in Fig.4. These piecewise linear models can be transformed into mixed integer linear programming, and sometimes into linear model [23]. In the proposed model, constraint (4) implied the nonlinear term of the decision variable f (number of faces) as $exp^{-\gamma_3 f}$. In order to convert the exponential term into piecewise linear function, let's denote the number of facing f_j of product j as μ_j^k where $k = 1..K, K = (Sh_L/w_j)+1$. Then the following adjustment for the original model are made as follows:

$$\begin{aligned} D_j = {} & \theta_{1j} - \theta_{2j} \exp^{(-\theta_{3j} p_j)} \\ & + \beta_{1j} - \beta_{2j} \exp^{(-\beta_{3j} pc_j)} \\ & + \gamma_{1j} - \gamma_{2j} \exp^{(-\gamma_{3j} \mu_j^k)} \qquad \forall j \end{aligned} \tag{10}$$

$$f_j = \sum_{k=1}^{K} \mu_j^k y_j^k \qquad \forall j \tag{11}$$

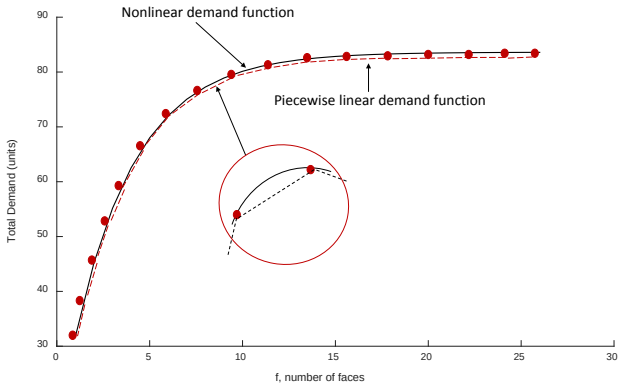


Fig. 4: Piecewise approximation for nonlinear demand function

$$\sum_{k=1}^{K} y_j^k = 1 \qquad \forall j \qquad (12)$$

$$y_j^k = \{0, 1\} \qquad \forall j, k \qquad (13)$$

The last three constraints above ensure the feasibility of f_j, while the revised constraint (10) is now converted into linear form.

IV. Results

In this section, data from local grocery hypermarket store were obtained, without loss of generality, demonstration for the proposed model performance will be based on small problem, consists from three products.

Four models will be presented; relaxed, integer, competitor price changing, and selling price changing models, to demonstrate the model performance.

Data analysis was performed using HP machine with a processor Intel (R) Core (TM) i7 CPU 2.2 GHz and 4.00 GB RAM. IBM ILOG CPLEX 12.6.1 was used for solving the optimization model.

A. Upper bound:Relaxed Model

In order to benchmark the proposed model, the integrality constraints of the model were relaxed to obtain the linear programming model. The solution of the relaxed model is the upper bound solution of the model [24], [25] (i.e. the integer model will provide a solution less than or equal to the relaxed model). Table I under the "Upper bound: Relaxed Model" column shows the parameters, number of faces, number of shortage (overstock) units, and the total profit obtained when solving the relaxed model.

B. Integer Model

In reality, the decision variables such as number of faces, shortage and overstock units must be an integer values. Therefore, the model is resolved with the integrality constraints included as shown in Table I, under the "integer Model" column. As expected, the total profit is less than the profit of the relaxed model.

C. Scenario 1: Changing the Competitor Price

In this scenario, the competitor price (pc) of product "B" is assumed to be changed from 2.6 to 2.2 \$/unit, with all other parameters are assumed to be unchanged as in Table I.
Table I under the "Scenario 1..." column, shows that the demand, number of shortage units, and the profit of product "B" are changed from 82 to 73 units, 4 to 0 units, and 98.8 to 95.6\$, respectively. The results are logically consisted, since reducing the competitor price will decrease the product demand, and hence the right action that should be taken from the store manager is to substitute the missing demand with the most profitable product.

D. Scenario 2 : Changing the Selling Price

Now, assume that the store manager decided to reduce the selling price (p) of product "C" from 2.5 to 2.4 $\$/unit$ (The reasons behind this reduction are many, for example this may be due to the large amount of inventory that the store have from this product and the need to sell some of them). If the model parameters remain unchanged as in Scenario 1, the new solution suggested by the model is to increase the number of faces of product "C" to fifteen in order to obtain \$88.2 as the total profit.

As seen from the above analysis, many different scenarios can be created and the proposed model will give insight about the best possible solutions to follow.

V. Conclusion

Shelf space allocation problem during the promotion period was investigated in this research. Due to the relatively low net profit in the grocery business (around 3%), retailers are required to carefully plan their promotions, since mis-estimation in the demand quantity can lead to either lost sales (due to the shortages) or increasing the holding cost (due to the overstock units). In this paper, demand estimation model as an exponential function of the selling price, competitor price, and a number of faces was proposed, while model parameters can be obtained using nonlinear regression approach from historical data to overcome the shortages found in the previous research domain. Multi-objective piecewise linearization approach was used to solve the shelf space allocation model, with real life data from the local grocery store was used to test the validity of the proposed model. Finally, the extension of this work can be in developing model for dealing with large size problems, including assortment, and inventory policies.

Acknowledgment

The authors would like to thank IBM company for providing them with an academic license for IBM ILOG CPLEX software.

TABLE I: Parameters Models and Results

	Upper bound: Relaxed Model			Integer Model			Scenario 1: Changing Comp.Price			Scenario 2: Changing Selling Price		
	Products			Products			Products			Products		
Parameters	A	B	C	A	B	C	A	B	C	A	B	C
p (\$/unit)	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	**2.4**
c (\$/unit)	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1	2.1
pc (\$/unit)	2.6	2.6	2.6	2.6	2.6	2.6	2.6	**2.2**	2.6	2.6	2.2	2.6
Cap (units)	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0
UL^o (units)	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0
UL^s (units)	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0	6.0
w (cm)	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
θ_1	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3
θ_2	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1	-385.1
θ_3	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
β_1	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0
β_2	160.0	160.0	160.0	160.0	160.0	160.0	160.0	160.0	160.0	160.0	160.0	160.0
β_3	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
γ_1	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0	40.0
γ_2	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0
γ_3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Sh_L (cm)	1237.0			1237.0			1237.0			1237.0		
f (no. of faces)	13.75	13.73	13.73	14	13	13	14	13	**14**	14	13	**15**
$Demand$ (units)	82.53	82.45	82.42	83	82	82	83	**73**	**83**	83	73	**86**
S^- (units)	0.00	0.00	0.00	0	4	4	0	**0**	**0**	0	0	0
S^+ (units)	0.00	0.00	0.00	1	0	0	1	**5**	**1**	1	5	**4**
$Total\ Profit$ (\$)	98.96			98.8			**95.6**			**88.2**		

REFERENCES

[1] Deloitte, "Global powers of retailing: The art and science of customers in 2017," Deloitte Touche Tohmatsu Limited., Report.

[2] F. M. Institute, "Slotting allowances in the supermarket industry in 2002," Food Marketing Institute, Report.

[3] P. M. Reyes and G. V. Frazier, "Goal programming model for grocery shelf space allocation," *European Journal of Operational Research*, vol. 181, no. 2, pp. 634–644, 2007.

[4] POPAI, "Consumers buying habits study in 1997," Point of Purchase Advertising Institute, Report.

[5] A. H. Hübner and H. Kuhn, "Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management," *Omega*, vol. 40, no. 2, pp. 199–209, 2012.

[6] M. Karampatsa, E. Grigoroudis, and N. F. Matsatsinis, "Retail category management: A review on assortment and shelf-space planning models," in *Operational Research in Business and Economics*. Springer, 2017, pp. 35–67.

[7] T. W. Gruen, D. S. Corsten, and S. Bharadwaj, "Retail out of stocks: A worldwide examination of extent, causes, and consumer responses," 2002.

[8] X. Dreze, S. J. Hoch, and M. E. Purk, "Shelf management and space elasticity," *Journal of retailing*, vol. 70, no. 4, pp. 301–326, 1994.

[9] K. Campo, E. Gijsbrechts, T. Goossens, and A. Verhetsel, "The impact of location factors on the attractiveness and optimal space shares of product categories," *International Journal of Research in Marketing*, vol. 17, no. 4, pp. 255–279, 2000.

[10] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "Using association rules for product assortment decisions: A case study," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 254–260.

[11] F. S. Zufryden, "A dynamic programming approach for product selection and supermarket shelf-space allocation," *Journal of the operational research society*, pp. 413–422, 1986.

[12] M. Corstjens and P. Doyle, "A model for optimizing retail space allocations," *Management Science*, vol. 27, no. 7, pp. 822–833, 1981.

[13] F. Cole, W. Gochet, F. Van Assche, J. Ecker, and Y. Smeers, "Reversed geometric programming: A branch-and-bound method involving linear subproblems," *European Journal of Operational Research*, vol. 5, no. 1, pp. 26–35, 1980.

[14] M. Corstjens and P. Doyle, "A dynamic model for strategically allocating retail space," *Journal of the Operational Research Society*, pp. 943–951, 1983.

[15] G. v. Ryzin and S. Mahajan, "On the relationship between inventory costs and variety benefits in retail assortments," *Management Science*, vol. 45, no. 11, pp. 1496–1509, 1999.

[16] M. A. Hariga, A. Al-Ahmari, and A.-R. A. Mohamed, "A joint optimisation model for inventory replenishment, product assortment, shelf space and display area allocation decisions," *European Journal of Operational Research*, vol. 181, no. 1, pp. 239–251, 2007.

[17] N. Amrouche and G. Zaccour, "Shelf-space allocation of national and private brands," *European Journal of Operational Research*, vol. 180, no. 2, pp. 648–663, 2007.

[18] M.-C. Chen and C.-P. Lin, "A data mining approach to product assortment and shelf space allocation," *Expert Systems with Applications*, vol. 32, no. 4, pp. 976–986, 2007.

[19] A. G. Kök, M. L. Fisher, and R. Vaidyanathan, "Assortment planning: Review of literature and industry practice," in *Retail supply chain management*. Springer, 2015, pp. 175–236.

[20] P. M. Guadagni and J. D. Little, "A logit model of brand choice calibrated on scanner data," *Marketing science*, vol. 2, no. 3, pp. 203–238, 1983.

[21] V. Gaur and D. Honhon, "Assortment planning and inventory decisions under a locational choice model," *Management Science*, vol. 52, no. 10, pp. 1528–1543, 2006.

[22] R. Roy Dholakia, "Going shopping: key determinants of shopping behaviors and motivations," *International Journal of Retail & Distribution Management*, vol. 27, no. 4, pp. 154–165, 1999.

[23] J. H. Owen and S. Mehrotra, "On the value of binary expansions for general mixed-integer linear programs," *Operations Research*, vol. 50, no. 5, pp. 810–819, 2002.

[24] J. Irion, J.-C. Lu, F. Al-Khayyal, and Y.-C. Tsao, "A piecewise linearization framework for retail shelf space management models," *European Journal of Operational Research*, vol. 222, no. 1, pp. 122–136, 2012.

[25] H. K. Gajjar and G. K. Adil, "A piecewise linearization for retail shelf space allocation problem and a local search heuristic," *Annals of Operations Research*, vol. 179, no. 1, pp. 149–167, 2010.

Obsolescence forecasting strategy of technological components - A comparative study of algorithms

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Yosra Grichi[1], Yvan Beauregard[2], Thien-My Dao[3], Imen Zaabar[4]
Department of mechanical engineering
École de technologie supérieure
Montreal, Canada
yosra.grichi.1@ens.etsmtl.ca, yvan.beauregard@etsmtl.ca, thien-my.dao@etsmtl.ca, zaabar.imen@gmail.com

Abstract—Obsolescence is a major issue that is driven by many factors, mainly through technological advancement, in which the life cycle of the components is often shorter than that of their systems. Basically, obsolescence problems are often sudden and not planned, causing delays and extra costs. Moreover, forecasting appears to be one of the most efficient solutions. By using a case study, this paper presents a comparison of the different algorithms which are: Back Propagation Neural Network (BP-NN), AdaBoost, Support Vector Machine (SVM), Random Forest (RF) and Decision tree that is capable of forecasting obsolescence risk of a large sample of electronic components with a high degree of accuracy. Using supervised learning, the data set is divided into 2 groups: 2/3 for training and 1/3 for testing and validation. In addition, a database is developed to collect all the data necessary for training. Finally algorithms are compared between them to estimate the best predictor based on confusion matrix.

Keywords— obsolescence, forecasting risk, BP-NN, SVM, RF, AdaBoost, component, memory.

I. Introduction

As defined by Moore's law, the rapid evolution of electronic components continues to grow, which stipulates that semiconductor density doubles approximately every 18 months [1-3]. In fact, when we talk about technological obsolescence, we immediately think of electronics, this evolution creates new electronic components every year with short lifetimes. In USA, the industry has grown at a rapid rate since the 1990s. New technologies are introduced in the market with increased speed.

Today, the short life cycle of technology and the lack of forecasting represent a challenge for several companies which need to take into account the risk of obsolescence; on the other hand, the modeling of obsolescence is considered as a complex problem that requires a good knowledge about the components affected by obsolescence. In the same way, it is necessary to know the lifetime of the components and their technological evolutions and therefore know all the factors that influence the obsolescence of components. Moreover, forecasting appears to be one of the most efficient solutions.

In the literature, there are two types of forecasting, which are forecasting of the obsolescence risk and forecasting of the obsolescence date. These prediction methods are based on a human knowledge that is used to estimate obsolescence or market survey; however these methods do not produce satisfactory results in most cases [4]. Forecasting obsolescence risk was a reactive nature based on the resolution of the problem once noticed. It's used to predict the probability that a component is still in production or not [5, 6]. Rojo (2012) has developed a methodology to evaluate the risk level of obsolescence based on availability of the stock, consumption rate and the number of sources available [6] .

In the literature, the majority of the works are based on the life cycle forecasting for a single type of which become obsolete every day. To resolve this problem, machine learning appears as an optimal method for forecasting. In fact, machine learning is a data analysis method that automates the recognition of data models without human manipulation [7] capable of forecasting of a large sample of electronic components with a high degree of accuracy. During the last decades, machine learning has attracted the attention of many researchers in several disciplines and has been applied in a number of areas. According to Wu and Kumar [8], ten top algorithms have been identified by IEEE « International Conference On Data Mining », include, AdaBoost, SVM, K-Means, decision tree, etc. These algorithms are proven as a good predictor.

The current challenges to solve this problem are described below:

- What are the main algorithms that are best adapted to obsolescence forecasting?
- How can these approaches forecast a large number of data?
- How is the obsolescence data applicable in the model?

The rest of this paper is organized as follows; Section II introduces the potential obsolescence forecasting strategies. Section III presents the proposed framework of obsolescence forecasting. The case study is explained in Section IV. Experimentation and analysis of results are defined in the same section. Finally, section V presents the conclusion and future research.

II. Potential obsolescence forecasting strategies

A. Back Propagation Neural Network (BP-NN)

These networks are capable of assimilating complex relationships between several variables [9] . BP-NN is often used for forecasting because of its non-linear modeling capability to forecast a large data set. According to the literature, neural networks show good stability, better accuracy and the ability to predict and manage a large data [10, 11]. The algorithm is considered as follows:
We start by initializing the weights to a random value. Then we iterate for each of the examples to make a complete iteration of updating of the neural network. For each example, a forward propagation is done to calculate the activation function for each network and a back propagation to calculate the Deltas. Finally, we update the weights by adjusting them in the BP-NN.

Input: examples, a set of examples, each with input vector x and output vector y, a multilayer network with L layers, weights w_{ij} , activation function g
Local variables: Δ, a vector of errors, indexed by network node
For each weight w_{ij} in network **do**
$w_{ij} \leftarrow$ a small random number
Repeat
For each example (x,y) in examples **do**
/* Propagate the inputs forward to compute the output*/
For each node i in the input layer **do**
$a_i \leftarrow x_i$
For l = 2 to L do
For each node j in layer l do
$in_j \leftarrow \sum_i w_{ij}\ a_i$
$a_j \leftarrow g(in_j)$
/* Propagate deltas backward from output layer to input layer */
For each node j in the output layer do
$\Delta[i] \leftarrow g(in_i)\ (1-g(in_i)) \sum_j w_{ij}\ \Delta[j]$
For l= L-1 **to 1 do**
For each node i in layer l **do**
$\Delta[i] \leftarrow g(in_i)\ (1-g(in_i)) \sum_j w_{ij}\ \Delta[j]$
/* Update every weigh in network using deltas */
$w_{ij} \leftarrow + \alpha * a_i * \Delta[j]$
Until some stopping criterion is satisfied
Return network

B. Support Vector Machine (SVM)

The support vector machine has received a particular intention since 2000, which is applied in several areas such as forecasting and optimization. SVM is considered one of the best precise and robust approaches within all the recognised algorithms. It needs only some examples for training, and it is not depending on the number of dimensions [8]. The objective of this algorithm is to search for an optimal hyperplane that can separate the data into two classes. To determine the separator hyperplane, the SVM approach must maximize the margin called M between the patterns of the two classes. Geometrically, the margin corresponds to the shortest distance between the closest data points and the hyperplane. $M= d_+ + d_-$, d_+ is the shortest distance between the hyperplane and the positive patron, and d- represents the shortest distance between this hyperplane and the nearest negative patron.

C. AdaBoost

AdaBoost is a method that uses up to one classifier by improving its accuracy. [12]. The AdaBoost algorithm introduced by Yoav Freund and Robert Schapire is one of the most powerful approaches, since it has a good theoretical basis, and a very precise prediction with a great simplicity.

D. Decision Tree

Decision tree learning is used as a predictive model. It is used in particular in data mining and in automatic learning.
In these trees structures, the leaves represent the values of the target variable and the branch lines correspond to combinations of input variables that lead to these values. In decision analysis, a decision tree can be used to represent explicitly the decisions made and the processes that lead to them. In learning and data mining, decision tree describes the data but not the decisions themselves. [13].

E. Random Forest

Random forests are an integration of tree predictors so that every tree rely on the values of a random vector sampled separately and through similar distribution for the whole trees in the forest.
The tree classifiers of a forest has a generalization error which rely on the strength of each single trees in the forest and the correlation between them [14].
Using random selection of features to split each node yields error rates that compare favourably to Adaboost, but are more robust with respect to noise.
Significant improvements in classification accuracy have resulted from increasing a group of trees and allowing them elect the most popular class. An early example is bagging [14] at which place to raise each tree a random selection (without replacement) is came from the training set examples. Additional example is random split selection in which at every node the split is chosen randomly from between the K best splits.

III. Proposed framework

Throughout this research, the literature has showed how to evaluate the most promising approaches to obsolescence prediction based on selection criteria such as the capacity and stability of the algorithms. Based on these criteria, all the algorithms have been the subject of this research.
In fact, regardless the field of application, the learning stages are always the same:
Step 1 is the extraction of the data either by collecting data from the surveys or by extracting data from the available

databases. Step 2 is to extrapolate the data by choosing the attributes and specifications of the data that the model will require. Step 3 is the distribution of the sample used to estimate the error. The sample will be split into 2 groups: 2/3 for training and 1/3 for the test. Note that it is possible to subdivide the data into 3 parts: learning, validation and testing. Step 4 often used as an adjustment sample, and it is used to adjust the parameters of the data mining methods. Note that the absence of a variable to explain, learning is unsupervised, otherwise learning is supervised.

Once the data is collected, the next step (step 5) the learning set will be introduced into the algorithms to create the predictive model. Then, the parameters of the model will be estimated, as the learning duration, the number of variables, etc. After that, supervised learning algorithm is developed to train the data and create predictive models. This algorithm must identify the status (active or obsolete) for each component in the test set. The precision of the algorithms is calculated by comparing the current state with the state predicted by the model. Finally, step 6 is the optimization of these parameters and the comparison of the optimal models will be concluded from the estimation of the forecast error for each model.

IV. Case study

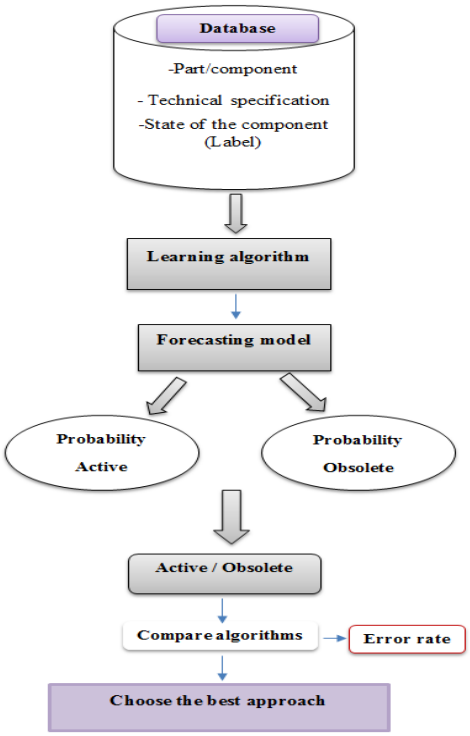
A. Collection data

The data is collected for the case study from www.micron.com. Micron technology is a global leader in the semiconductor industry that provides information on the technical specifications (Fig.1), technology, Brand and the status of (production/ end-of-life) of different electronic components. The data contains more than 700 components with a known label, including many technical specification show in figure 2 such as voltage (v), cycle time (ns), clock rate (HZ), depth (Mb), etc. Once the database is built, then the data is divided into two random groups divided into training (2/3) and test (1/3) to calculate the accuracy of the model. Fig.2 shows an overview of the adopted methodology to build the forecasting models.

Fig.1.Sample of case study data

Technology/ Memories	Availability	Voltage	Package	Depth
DDR4	End of life	1.2	FBGA	256Mb
DDR4	production	1.2	TFBGA	256Mb
SDRAM	Production	1.2	TFBGA	256Mb
DDR4	production	1.35	TFBGA	512Mb
SDRAM	production	1.35	FBGA	512Mb

Fig. 2.Methodology adopted



Once the data is constructed (Fig. 3), after that the next step is the training of the data set using the five algorithms chosen and applied in the case study: BP-NN, SVM, AdaBoost, RF and Decision Trees. These algorithms have been proven as the most relevant in prediction.

K-means will not be studied in this case because we are in the case of supervised learning (output label are known), while K-means is based on unsupervised learning.

Fig. 3.Construction of training data sets

```
set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset) # 734 observations
crs$sample <- crs$train <- sample(nrow(crs$dataset),
0.7*crs$nobs) # 513 observations
crs$validate <- NULL
crs$test <- setdiff(setdiff(seq_len(nrow(crs$dataset)),
crs$train), crs$validate) # 221 observations
crs$input <- c("Density", "Depth", "Width",
    "Voltage", "Package", "Pin.Count", "Clock.Rate",
    "Cycle.Time", "Op..Temp.", "CL", "Data.Rate")
crs$numeric <- NULL
crs$target  <- "Output"
crs$ignore  <- c("Part", "Brand")
crs$weights <- NULL
```

B. Experimental results

First, we distinguish the predictive variables (INPUT = 13 variables) and the target variable (OUTPUT = end-of-life / production). 734 observations, of which 70% (513) are for the learning sample data and 30% (221) are for the test.

The first algorithm used was BP-NN. The classification of the neural networks was used with R3.3.3 using the "nnet" package. The neural networks used in this study were constructed with 3 hidden layers. The probability that each part is available or obsolete is considered as output.

The following algorithm applied was SVM. This algorithm used the SVM classification function from the "e107" package with R3.3.3.

The third algorithm was AdaBoost. The classification of this algorithm was done with R3.3.3 from the "ada" package that implements the boost algorithm.

Fig.4. Build the Boost model

```
set.seed(crv$seed)
crs$ada <- ada::ada(Output ~ .,
                    data=crs$dataset[crs$train,c(crs$input,
crs$target)],
                    control=rpart::rpart.control(maxdepth=30,
                                                 cp=0.010,
                                                 minsplit=20,
                                                 xval=10),
                    iter=50)
```

After, RF was built with R3.3.3 from the 'random Forest' package. Type of random forest is classification, number of trees = 500, Number of variables tested at each split = 3

Finally, Decision trees algorithm was developed from the 'rpart' package.

The accuracy of the model was initially present by a confusion matrix (Table I), for training size of 70% and 30% for testing. The error for the model was retained to 21% (BP-NN), 22% (SVM), 25% (AdaBoost), 19% (RF) and 24% (Decision tree).

The confusion matrix shows how many components were correctly classified. The numbers in the cells (production, production) and (end-of-life, end-of-life) are correctly classified and all other cells present a classification error.

TABLE I. Confusion matrix of algorithms

		Predict		
	Actual	**End-of-life**	**Production**	**Overall Error**
BP-NN	**End-of-life**	12	9	21%
	Production	16	60	
SVM	**End-of-life**	9	12	22%
	Production	9	67	
AdaBoost	**End-of-life**	53	26	25%
	Production	32	110	
RF	**End-of-life**	10	11	19%
	Production	7	69	
Decision tree	**End-of-life**	53	26	24%
	Production	27	115	

C. Results analysis

Table II represents an average accuracy to evaluate model performance contain the prediction accuracy of the classification algorithms for different training data set sizes, from 50% to 90%, to see how much of the data portion influence the accuracy of the model.

The results show that RF shows better accuracy compared to the other algorithms at 70% of training size. Therefore, the prediction model created using SVM has such high precision for a 90% of training size. Note that all the algorithms show an increase of accuracy when the training size increases.

Fig.5 shows a graphical display of confusion matrix of algorithms which shows that RF represents the best classifier with a slight difference with the other algorithms.

TABLE II. Average accuracy of prediction for training size from 50% to 90%

	BB-NN		**SVM**		**BOOST**		**RF**		**Decision tree**	
Training size	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)
50%	94%	69%	87%	77%	90%	72%	94%	77%	85%	73%
60%	93%	70%	84%	74%	89%	73%	93%	77%	83%	77%
70%	91%	74%	85%	78%	89%	74%	91%	81%	82%	76%
80%	91%	76%	84%	78%	88%	78%	91%	84%	89%	80%
90%	90%	74%	92%	91%	89%	84%	90%	88%	90%	89%

Fig.5. Accuracy graph of algorithms

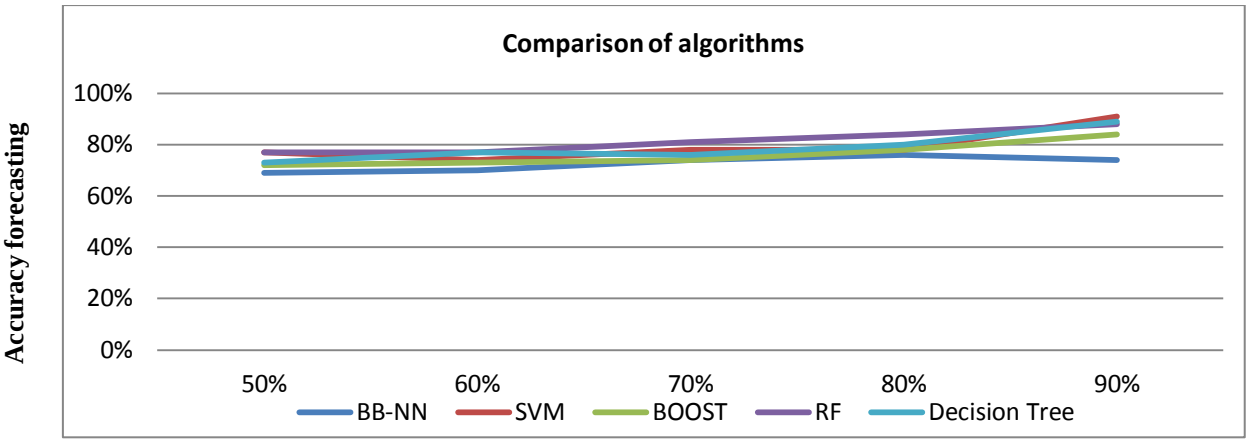


Fig.6 shows the measurement of the linearity relationship between the actual values of the obsolescence and the predicted values. The Predicted Versus Observed plot is relevant for regression models (predicting a continuous value rather than a discrete value). It will display the predicted values against the observed values. Two lines are also plotted, one being a linear fit to the actual points, and the other being the perfect fit, if the predicted values were the same as the actual observations. The Pseudo R-Squared is a measure that tries to mimic the R-Squared.

Fig.7 shows the results of precision models versus Recall. Recall represents the number of true classifications of the total number of classifications (true positive vs. false negative).

RF and SVM shows the best precision compared to all the other algorithms.

According to the results, Random Forest is chosen as the best predictor for Obsolescence risk of electronic components.

Fig.6. Predicted values vs. Observed

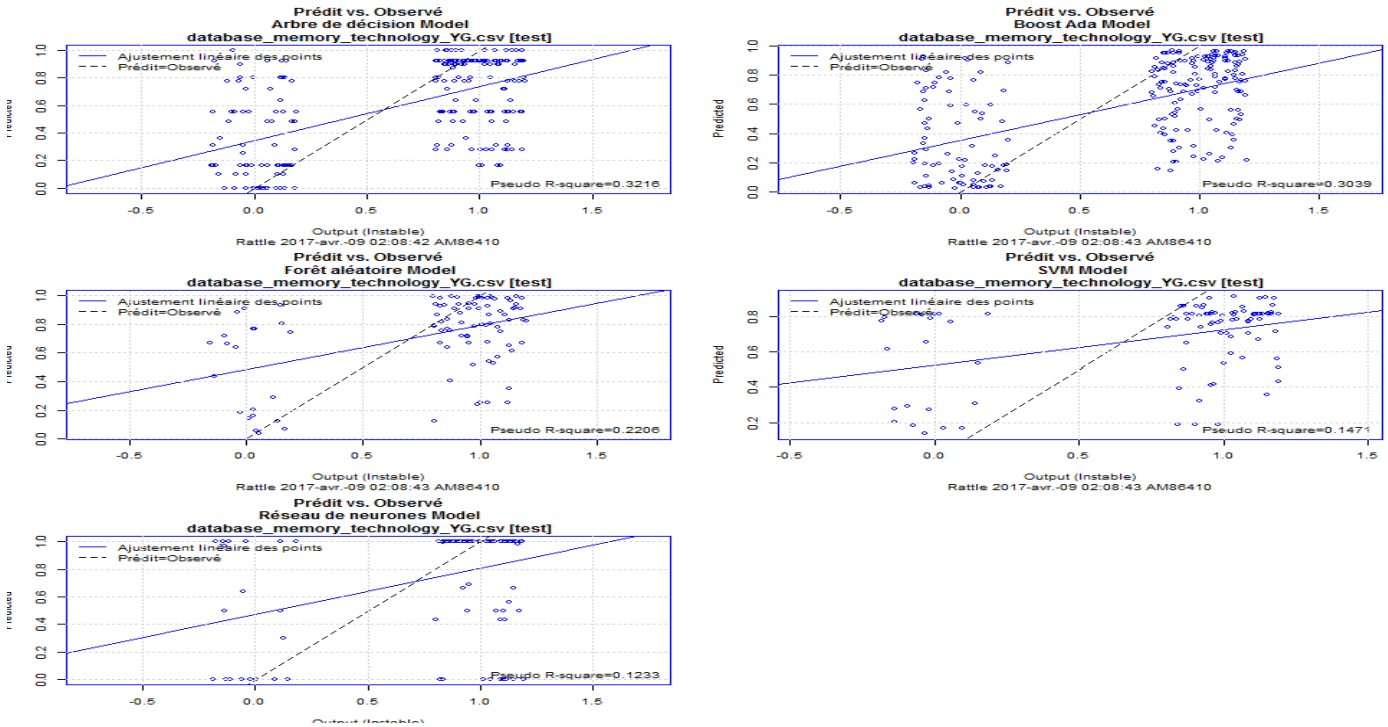
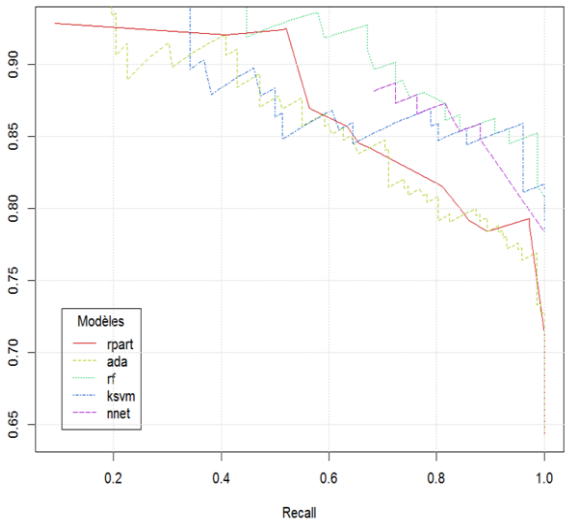


Fig.7. Accuracy curve of algorithms



V. Conclusion and future research

Research on obsolescence is growing, because of the serious challenge to complex systems. Machine learning methodology presented in this research can predict the risk of obsolescence with a high degree of accuracy without the need for human manipulation or market survey. Random Forest appears as the best predictor to forecasting obsolescence risk. However, the exploitation of the model can be carried out on new data.

- Limitations and future research

First, the sample is small and several information are missing. The larger the data the more the model gives more accurate results. Information can be gathered from a larger sample to minimize the error rate.

The standard BP-NN has a major disadvantage: "Overfitting problem". This means that the network can lose its capability. It happens when the neural network learns the specific details of the inputs and not their general characteristic found in the present and future data.

To resolve this problem, the Bayesian-Multilayer Perceptron (B-MLP) will be developed. The Bayesian regularization method is used to reduce the error rate. This integration of the error function will require to Neural Network to obtain smaller weights and biases to reach a greater capacity of generalization. B-MLP will be compared to RF model to estimate the best predictor.

VI. Acknowledgements

The authors would like to thank their partner company for supporting this research, as well as acknowledge funding from NSERC (CRD 45068513) and CRIAQ (Lean501) for conducting this research.

VII. References

[1] Tomczykowski, W.J. A study on component obsolescence mitigation strategies and their impact on RM. in The International Symposium on Product Quality and Integrity; Transforming Technologies for Reliability and Maintainbility Engineering, January 27, 2003 - January 30, 2003. 2003. Tampa, FL, United states: Institute of Electrical and Electronics Engineers Inc.

[2] Sandborn, P., Strategic management of DMSMS in systems. DSP Journal, 2008: p. 24-30.

[3] Homchalee, R. and W. Sessomboon. A comparison of forecasting models using multiple regression and artificial neural networks for the supply and demand of Thai ethanol. in 2013 IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2013, December 10, 2013 - December 13, 2013. 2014. Bangkok, Thailand: IEEE Computer Society.

[4] Jennings, C.P., Forecasting obsolescence risk and product lifecycle with machine learning. 2015, Iowa State University.

[5] Van Jaarsveld, W. and R. Dekker. Estimating obsolescence risk from demand data to enhance inventory control - A case study. 2011. P.O. Box 211, Amsterdam, 1000 AE, Netherlands: Elsevier.

[6] Rojo, F.R., R. Roy, and S. Kelly. Obsolescence risk assessment process best practice. in Journal of Physics: Conference Series. 2012. IOP Publishing.

[7] Zurada, J.M., Introduction to artificial neural systems. Vol. 8. 1992: West St. Paul.

[8] Wu, X., et al., Top 10 algorithms in data mining. Knowledge and Information Systems, 2008. **14**(1): p. 1-37.

[9] Maschio, C. and D.J. Schiozer, Bayesian history matching using artificial neural network and Markov Chain Monte Carlo. Journal of Petroleum Science and Engineering, 2014. **123**: p. 62-71.

[10] Tu, K.-J., et al. Predicting the operation and maintenance costs of apartment buildings at preliminary design stage: Comparing statistical regression and artificial neural network methods. in 25th Inaugural Construction Management and Economics: 'Past, Present and Future' Conference, CME 2007, July 16, 2007 - July 18, 2007. 2007. Reading, United kingdom: Taylor and Francis Ltd.

[11] Yun, C., L. Ping, and Y. Li. Aftermarket demands forecasting with a Regression-Bayesian-BPNN model. in 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering (ISKE 2010), 15-16 Nov. 2010. 2010. Piscataway, NJ, USA: IEEE.

[12] Alfaro, E., et al., Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. Decision Support Systems, 2008. **45**(1): p. 110-122.

[13] Rokach, L. and O. Maimon, Introduction of Decision trees. Data Mining with Decision Trees: theory and application, 2007: p. 5.

[14] Breiman, L., Random forests. Machine learning, 2001. **45**(1): p. 5-32.

Hybrid forecasting model for sales promotion: application in the modern retail supply chain

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Rihab Mechmech, Slim Harbi
Laboratory OASIS – ENIT
University of Tunis El Manar
Tunis, Tunisia
slim.harbi@enicarthage.rnu.tn
ryhabmechmech@gmail.com

Mohamed Bahroun, Hanen Bouchriha
Laboratory ACS – ENIT
University of Tunis El Manar
Tunis, Tunisia
Bahrounm@hotmail.com ,
hanen.bouchriha@enit.rnu.tn

Abstract—Forecasts are often produced using a Forecasting Support System that integrates statistical forecasting with managerial judgment. Forecasting sales under promotional activity is one of the main reasons to use expert judgment. This exploratory and theoretical research focuses on the promotion sales forecasting in the Tunisian retail supply chain. The paper aims to combine previous contributions in operation and logistics management (sales forecasting) with marketing (consumers' behavior in promotion period). We propose a novel model for sales promotion forecasting based on a new formulation of the promotion index. The performance of the model is compared against other forecasts models based on real data from a case study.

Keywords— Sales forecasting, promotion index, judgmental adjustment, consumer behavior

I. Introduction

Over the last two decades, the retail sector in Tunisia has grown considerably, with the development of several large hypermarkets. This situation brings with it a significant change in the management practices of distributors and manufacturers as well as in the behavior of consumers [1]. Indeed, new strategies has emerged regarding marketing activities and customer loyalty (loyalty cards) besides the adoption of various techniques to forecast sales promotion.

The majority of papers on sales promotion agree on the following definition: "Sales promotion is all marketing activities undertaken to create new behaviors in customers and to boost sales of the product or service" [2]. Promotion is an important driver for retailers to build consumer loyalty and attract new customers. In recent years, promotional intensity has increased and promotion techniques are more sophisticated (price discount, free percentage offered, three for two, loyalty bonus, etc.). The promotion management process has become an important factor in the company's performance [3]. Based on an empirical study, [4] shows how sales promotion has become strategic. However, for a retailer, with thousands of products, promotion is not as simple as it seems. The forecasting of sales in presence of promotion, the relationship with suppliers and manufacturers and the ordering process during promotional period represent a complex and important decision process. According to [5] this complexity is due to several factors such as the increased competition and aggressive marketing. Moreover, the accuracy of forecasts for the promotional periods has a direct impact on the effectiveness of supply chain relationships [6]. Thus, a "good" forecast of sales promotion is necessary. The first models used by retailers are very simple and naive and often inappropriate to the real world situations [7], such as forecasting the same quantity sold in a similar promotional campaign. During the last years, several techniques and models have been developed with a higher accuracy, in order to reduce costs, especially for shortage and over-stocks [8]. The promotional period is characterized by a significant increase in the volume of sales over a limited period. It can be noted that promotion leads to more customers attracting changes in brand or product sales even after the promotional period [9]. This major change in consumer behavior often make the promotional forecast more difficult.

The consumers' behavior during a promotional period has been the subject of several researches. Many studies have used surveys and experiments to understand the change in the customers behavior and to identify the influencing factors, such as promotional techniques, price sensitivity, etc. [9] [10]. Therefore, in Tunisia, retailers need a better understanding of consumer behavior during the promotional period in order to anticipate its changes, meet its needs and forecast sales promotion. The emergence of hypermarket retailing in Tunisia since 2001 has changed the whole retailing industry. Retailers are facing increasing competition. Promotion is one of the essential marketing mix to increase their market share. Consequently, a good sales forecast for this period will have a significant impact on the company's performance since a large part of the revenue comes from promotional sales. However, these forecasts are strongly influenced by the consumer behavior, which makes their accuracy sometimes low. Taking into account the Tunisian context, our main research questions are: (a) how has promotion sales changed the way consumers behave? (b) How to integrate sales historical data and consumers changing behavior over time in the forecasting sale models under promotional activity?

In this research, we first carried out a literature review on promotional sales models and empirical studies on consumer behavior during promotion. Next, we conducted a survey based on a case study aimed to understand purchasing behavior during promotional period. The result of this survey was used for the development of a promotion sales forecasting model that takes into account the Tunisian consumer profile in order to increase forecast accuracy. Finally, in order to examine the performance of the proposed model, an illustration and computational comparison was conducted based on a real case data.

This paper is organized as follows: In section 2, we analyze the relevant literature to the topic. Section 3 details the results of our survey on consumer behavior in presence of promotion. We present in section 4 three promotional sales forecast models. A comparison of these models based on a real case data is illustrated in section 5. Finally, section 6 provides some concluding remarks and insights for further research.

II. Literature review

The review of the relevant literature covers two main areas of research: (a) marketing: through the analysis of consumer behavior during the promotion period and (b) industrial management: through the analysis of sales promotion forecast models.

A. Consumer behavior during the promotion period

“Sales promotion is a temporary change in supply, aiming to have a direct impact on the behavior of the customers and on the sales force" [11]. The consumer's sensitivity to sales promotion is defined as "the degree to which a consumer is influenced by promotion in terms of its frequency of purchase, choice of brand, store or quantity" [12]. Previous research has focused on the following three criteria to study the customer behavior during a promotion: price sensitivity, brand loyalty, and type of promotion and promotion sensitivity.

1) Price sensitivity

A retailer when determining its optimal pricing and marketing strategy for the product often uses the price sensitivity of a product within its target market. Price sensitivity is the degree to which the price of a product affects consumers purchasing behaviors [13]. Promotional depth specifies the size of the promotion to which consumers are exposed. Price promotion, which is widely used, is a common marketing tool [35]. For a price discount during a promotion, each consumer has its price reference and acceptability margins that influence its decision-making: buy or not, increase the quantity of purchase or not. According to [35] under the promotion strategy, the retailer sets a regular price, a promotion price, and a promotion frequency while incorporating the effects of a reference price. Previous research suggests that depth and frequency have a distinct effect on promotional effectiveness.

2) Brand loyalty

"The concept of brand loyalty can be described as the commitment of consumers to a brand" [10]. Store and product brands have witnessed considerable growth in the last few decades, especially in grocery products [14]. In general, there are two kinds of consumers: those who do not change brand and buy only their usual product and those who change their product and migrate to another product for some reason. This reason can be promotion. Therefore, for a product that is on promotion sale, its loyal consumers are the first beneficiaries of this offer and most of them increase their usual quantities [15]. Nevertheless, there are other opportunistic consumers, who simply want to profit from promotion and make financial benefits [15] [16] [17]. Thus, whether or not the customer is loyal to a brand or a product, promotion creates a change in his behavior.

3) Promotion technique sensitivity

Several promotional techniques have been adopted in order to attract the maximum number of buyers, such as price discount, gratuities, gifts and games. In order to understand consumer changes in promotion, influencing techniques and different responses, researchers have used surveys and empirical studies [12] [10]. Many studies have shown that consumers adopt different behavior towards each promotional technique. Consumers are sensitive to promotion in general but the degree of sensitivity varies according to the promotional technique [17]. The problem of sensitivity is no longer towards promotion in general but has become towards promotional techniques [18].

B. Sales promotion forecasting models

Sales forecasting models cover both qualitative (judgment-based) and quantitative (statistics based) techniques. A review of the various techniques is detailed in [19]. Several studies have focused on the integration of judgmental and statistical approaches. The objective is to improve the accuracy of forecasts [20]. Demand forecasting models are still an ongoing research area [12] [17] [7] [21] [22]. However, research on sales promotion forecast models is much scarcer.

Researches focusing on promotion-specific forecasting proposed several models for predicting promotional sales. These models include simple prediction models (naïve model), time series with adjustments that reflect the increase in purchase due to promotion, models that deal with promotional forecasting in the case of a lack of data such as promotional indices, and hybrid models in which they use a combination of SARIMA models and neural networks [9]. One of the major problems that forecasters faced in the case of promotion is to integrate consumer behavior to forecast the volume of sales in promotion. Indeed, a promotion event influences consumers and causes a significant disruption in the time series of sales. Thus, conventional approaches cannot be used directly. A few specific models have been proposed for the retailing industry sector [23] [8] [21].

In practice, the approach used to forecast the quantity to sell during promotion was to make a simple estimation called a naive method [19]. This estimation will be roughly a quantity close to the quantity sold during a similar promotion period [8]. The most forecasting models mentioned in the literature are based on the classic time series and add adjustment. In practice, large retailers use a base-lift factor to predict sales promotions [23] [22] [24]. Recent studies have focused on how to choose and calculate adjustment in order to improve forecast accuracy. These models estimate adjustment based on a large number of variables related to price, store, history data on previous promotions, and information on manufacturers and product categories [22]. However, such information is often unavailable to small and medium-sized retailers. Among the main factors, influencing sales, emerging from the literature, the frequency of promotion, advertising modes (newspapers, TV, radio, etc.), category or characteristics of the product group, competitor promotions, radio promo promotions, previous promotions of the same product, the number of variants of the product on sale, the weather, etc. [7]. A review of the causal models is presented in [21]. Well known models cited in the literature include the SCANpro and its extensions. This model decomposes sales for a brand in inter and intra effects of price, marketing, shelving, period and store [25]. Reference [23] proposes a "PromoCastTM" promotional sales model based on the use of historical data to construct a regression model with 67 variables.

Reference [26] developed a model called "CHAN4CAST". This model uses a regression model that captures the effects of these variables such as past sales, trend, price, promotion variables, and seasonality. Other models are directly interested in information related to promotion. It is a multiple regression model, "data mining", "machine learning" (support vector machine), neural networks, genetic algorithms and hybrid models [23] [27] [28] [29] [30] [21] [7]. Among the different promotion sales forecasting models cited in the literature, we will focus on three categories of models.

1) Promotional index models

The idea of these models is to calculate a promotion index $I_{p,t}$, which determines the increase in sales during promotion period compared to the usual period sales [31] [7]. The promotion index is calculated according to the following formula:

$$I_P = \frac{1}{k} * \sum_{i=1}^{k} \frac{y_i}{\omega(t)_i} \quad (1)$$

With "y" represent the sales during the promotion period and ω (t) is the sales forecast for the same period without promotion, and k is the number of promotions periods considered. They also propose a forecasting approach of the promotion index using a simple exponential smoothing:

$$I^*_{p,t+1} = \alpha . I_{p,t} + (1-\alpha) . I^*_{p,t} \quad (2)$$

With $I^*_{p,t+1}$ is the promotion index for period t +1, $I_{p,t}$ is the promotion index for the last promotion period t and $I^*_{p,t}$ is the forecast of the promotion index of period t. α is the smoothing coefficient chosen between 0 and 1.

2) Benchmark models

The benchmark model proposed by [21] is based on exponential smoothing with a lift adjustment. During no promotional period, the forecast is the smoothed value of the sales of the previous non-promotional periods; otherwise, the forecast is the sum of the last increase in sales observed and the smoothed value obtained. Reference [22] proposes the following similar model:

$$F_t = \begin{cases} M_t & \text{if no promotion} \\ M_t + adjustment & \text{if promotion} \end{cases} \quad (3)$$

$$M_t = (1-\alpha) M_{t-1} + \alpha S_{t-1} \text{ if no promotion}$$

With F_t is the sales forecast for week t, M_t is the baseline forecast for week t generated by exponential smoothing. S_{t-1} is the value of non-promotional sales in week t-1, α is the smoothing parameter. The adjustment is calculated based on the additional sales during the previous promotion period. In the same way, [32] propose an identical model with an adjustment based on the last promotion.

3) Judgmental adjustment models

Expert judgment techniques play an important role in forecasting sales promotions. These methods sometimes represent the only alternative when a rapid change in consumer behavior occur or in case of the introduction of a new product or brand. Judgments are sometimes based on analogous product analysis or on consumer intentions surveys [33]. Reference [34] compares the accuracy of forecasting models based on consumer intentions. They conclude that the intention-based model is more accurate and that a combination of these models improves the accuracy of the forecast. Reference [5] shows the effectiveness of judgmental adjustment methods that improve the baseline forecast for the promotion period.

III. Customer behavior during promotion period

We conducted a survey of consumer behavior in relation to promotion. This survey targeted 3700 consumers and we have collected 609 answers. The method of collecting responses combined online questionnaires and direct in-store interviews.

The results of to the survey show that, 22% of respondents do their shopping several times a week (2 or 3 times a week), but 78% of respondents do them once a week. 72.6% of respondents spend less than 50 TND per shopping. Those with values greater than 50 TND per shopping represent only 27.4% of which 9.2% have high expenditure values, greater than 100 TND. The survey also shows that the degree of penetration of loyalty cards is quite high. Indeed, 63% of respondents have loyalty cards. However, those who are loyal to only one store brand represent only 16.51%; others have more than a loyalty card.

A. Customer behavior in presence of promotion

The results of the survey show that 46.8% of respondents are attracted by promoted product even though it is not their usual product. Moreover, during the promotion (case of price discount), 71.6% of respondents switch to expensive products (or brands) that they were not accustomed to buying.

The results obtained are therefore similar to those of a study carried out on Unilever brand products, which showed that promotions tend to convince consumers to change brands and buy more [9].

It can be noted that during promotion, the behavior of the consumer varies according to the technique used. Indeed, it appears that price discount for a single product are the more attractive technique to the Tunisian consumer with 42% of the respondents. However, discount price technique for the second or third product purchased is the least preferable with 17% of the respondents.

Regarding price sensitivity, the results show that the quantity purchased depends on the discount percentage. Between 10% and 20% and 30% to 50%, 72% of customers double their purchased quantities, 26.6% buy the same quantity and only 6.2% triple.

B. Concluding remarks

We can conclude that:

- Promotion has several effects on product sales. Indeed, during the promotion period, there is an increase in the quantities sold for the products in promotion,
- The increase in promotional sales (promotion lift) depends on the promotion technique. In the case of a price discount, this increase is price sensitive,

- The increase in promotion sales results from the acceleration of purchases (increase in quantities) by consumers loyal to the product, as well as the switch of other buyers to the product under promotion (brand change),
- The switch of other buyers to the promoted product causes a decreasing in sales of non-promotional substitutes (cannibalization effect).

IV. Proposed models

In this section, we present three models of promotion sales forecasting. The first model relies exclusively on the analysis of the sales history (promotional sales (PS) and non-promotional sales (NPS) through the calculation of a statistical promotion index (SPI). The second model is based on the development of an empirical promotion index (EPI) calculated by referring to the results of the survey on consumer behavior during promotion. Model 3 combines the two previous approaches using expert judgment by the proposition of an adjusted promotion index (API). A comparison of the performance of the three models on a real case is presented in Section 5.

A. Model 1 : Statistical promotion index

The promotion index represents a multiplication factor, which reflects the increase in sales during the promotion period "T" compared to sales during non-promotion periods. First, we define a non-promotional sales reference called baseline (BL) or "reference value". This BL value represents the arithmetic mean of the n periods just preceding the promotion period, denoted BL(n).

$$BL(n) = \frac{1}{n} * \sum_{i=t-n}^{t-1} NPS_i \qquad (4)$$

Where NPS_i are sales during period i. In our case, the promotion is over two periods (2 weeks). Thus, the promotion indexes represent weekly variations. For each promotion, we define two indices, one for the first week and the other for the second.

$$I_{PS}^{1} = \frac{1}{K} * \sum_{i=1}^{K} \frac{PS_i^1}{BL(n,i)} \qquad (5)$$

$$I_{PS}^{2} = \frac{1}{K} * \sum_{i=1}^{K} \frac{PS_i^2}{BL(n,i)} \qquad (6)$$

Where:

- PSi1 are the promotional sales (PS) of the first week.
- PSi2 are the promotional sales of the second week.
- BL(n, i) is the base value calculated over n periods preceding promotion number i.
- K represents the number of previous promotions (eg over a year).

Sales forecasts are calculated using the following formula:

$$F(t) = F^*(t) + z * (I_{PS} - 1) * F^*(t) \qquad (7)$$

Where

- F (t): The forecast for period t.
- F * (t): The forecast of non-promotional sales in period t with a forecasting method (eg exponential smoothing).
- I_{PS}: promotion index considered
- z: variable that is equal to 1 during promotion and 0 otherwise.

Thus, during the promotion, the forecast is calculated according to the following formula:

$$F(t) = I_{PS} * F^*(t) \qquad (8)$$

B. Model 1 : Emperical promotion index

This model is similar to the previous one and differ on how to calculate the promotion index. We propose an empirical promotion index (EPI) calculated based on the consumer purchasing intension as mentioned in the result of our survey on consumer behavior during promotion. This method can also be used if there is no available sales history or the introduction of a new promotion technique.

As mentioned in our survey, the increase in promotion sales results from the increase in quantities by buyers loyal to the product, and the switch of other buyers to the product on promotion, and also the purchase by occasional consumers who follow all kinds of promotion. We have noticed that the volume of sales by the third kind of consumer is quite low. Thus, the promotion index contains two components.

$$I_{PE} = I_{PE1} + I_{PE2} \qquad (9)$$

I_{PE1} represents the increase in the quantity PS by the usual consumers loyal to their product (or brand).

I_{PE2} concerns those who switch to promoted products; they are not loyal to any type of product.

Application:

I_{PE1} is calculated directly from the analysis of the survey results. For a price discount of 10% to 20%.

- The probability that a consumer buys the same quantity (a single package) is 0.42,
- The probability that a consumer doubles or triple the quantity (buys 2 to 3 packages) is 0.53,
- The probability of a consumer buying five or more packages is 0.05.

Thus, to calculate the value of I_{PE1}, we multiply the different probabilities by the corresponding quantities, as shown below:

$$I_{PE1} \cong 0{,}42 + (0{,}53*2{,}5) + (0{,}05*5) = 1{,}995$$

For I_{PE2}, as mentioned above, this index represents the sales of consumers who switch to the discounted product. We define:

- P_i: the promotional product
- P_k: the substitute products for P_i.
- π: the probability that a consumer switch to the promotional product P_i.

- NPS_i: Non-promotional sales of Pi
- NPS_k: Non-promotional sales of P_k
- R: the ratio between the usual sales of substitute products P_k and the usual sales of P_i.

Thus, we get:

$$I_{PE2} = \pi.R \quad (9)$$

$$R = \frac{NPS_k}{NPS_i} \quad (10)$$

Application

For the calculation of the probability π that a consumer switch to the product in promotion, we examine the result of the survey as follows:

- 35% of consumers change cleaning products when promoting their regular product.
- 72 % of consumers switch to expensive products during promotion.

Therefore, in our example: π = 0.72 * 0.35 = 0.252

Thus, I_{PE} is equal to:

$$I_{PE} = I_{PE1} + I_{PE2} = 1.995 + (0.252 * R)$$

C. Model 3 : Adjusted promotion index

In practice, forecasters tend to adjust forecasting models based on their experience. Thus, we propose an adjusted promotion index (hybrid) that combines the two previous approaches. In this case, it is calculated as follows:

$$I_{PA} = \alpha * I_{PS} + (1-\alpha) * I_{PE} \quad (12)$$

α is a coefficient chosen by the decision-makers to combine the statistical approach and the empirical approach.

V. Computation and analysis

In this research, we applied the three forecast models proposed to an example from a real case of our partner company in this project, namely a high-end item belonging to the family of laundry products. The confidentiality of the information forces us to hide the actual data and work with values that reflect a real sales pace. Three measures of the quality of forecasts were used: MAD (Mean Absolute Deviation), Mean Forecast Error (MFE) and MAPE (Mean Absolute Percent Error).

We chose n = 2, so for each promotion we calculated BL(2). Therefore, before each promotion, one must have the weekly sales of the two weeks preceding it. Since the number of promotions considered is five promotions / year, we need 15 sales: 5 promotional sales and 10 NPS (2 NPS that precede each promotion). The sales forecasting algorithm was implemented using Matlab. Figure 1 and Table I show PS forecasts and the various forecast errors (MFE, MAD and MAPE).

Fig. 1. Promotion sales forecasting

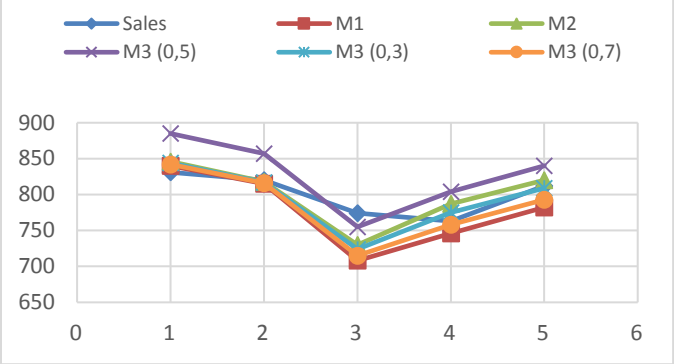


Table 1 presents the model 3 forecast errors for different values of α.

TABLE I. Forecast errors

	MAD	MFE	MAPE
α=0	12.52	-5.81	1.6
α=1	19.47	16.2	2.53
α=0,3	10,38	0,79	1,33
α=0,5	33,69	-33,69	4,19
α=0,7	13,79	9,6	1,76

It can be noted that all forecast errors with the I_{PE} method are lower than those of the I_{PS}. The result shows that this method provides better forecast with better accuracy. Thus, the experimental promotional index better aligns the consumer's profile; therefore, it explains a significant representation of its promotional behavior. Model 3 is based on the same data as the previous two models. We present the results for three values of α (0.3, 0.5 and 0.7). Table I shows that α = 0.3 presents the best forecast. The choice of this parameter will depend on the expert judgment. This depends on the type of product as well as the promotional event and promotion period.

VI. Conclusion

Through this study, we proposed three models of promotion sales forecasting. In our case study, the hybrid model (adjusted) yielded the best results. This model cannot guarantee the best forecast in other circumstances. Indeed, this depends on the choice of the coefficient α that reflects the judgment of the experts. Thus, we can conclude that the proposed method of improvement based on consumer intension is specific, depending on the type of promotion and the product in question, and that each item has its own formula for calculating its promotional index. This work can be investigated further by analyzing the effect of promotion by product families. We propose to combine the questionnaire analysis with a more detailed description of the consumer profile by analyzing data from loyalty cards using "datamining" and "vector machine" techniques.

References

[1] M. Bahroun and S. Harbi, "Risk management in the modern retail supply chain: Lessons from a case study and

literature review", 6th IESM Conference, October 2015, Seville, Spain.

[2] P. Desmet, « Promotion des Ventes », Paris, Nathan, 1992.

[3] S. Ogden-Barnes, and S. Minahan, "Sales Promotion Decision Making: Concepts, Principles, and Practice", Business Expert Press, LLC, 2015.

[4] Ç., Martichoux, « La promotion des ventes en pratiques ». Éditions d'Organisation, 2004.

[5] J. R. Trapero, D. J. Pedregal, R. Fildes, N. Kourentzes, "Analysis of judgmental adjustments in the presence of promotions", International Journal of Forecasting Vol. 29, Issue 2, April–June 2013, p. 234–243.

[6] H. Stadtler and Ch. Kilger, "Supply Chain Management and Advanced Planning", Springer, 2010.

[7] M. Patak, and V. Vlckova, "Promotional Indices Application in Business Forecasting of Promotion Events", Proceedings of the 1st International Conference on Management, Marketing, Tourism, Retail, Finance and Computer Applications (MATREFC'13), Dubrovnik, Croatia June 25-27, 2013.

[8] M. Trusov, , A.V Bodapati, and L.G Cooper, "Retailer Promotion Planning: Improving Forecast Accuracy and Interpretability", Journal of Interactive Marketing, 2006, Vol.20, No.3-4. pp. 71-81.

[9] M. O. Ansah and K. Poku, "Investigation into Consumer Response to Sales Promotional Activities: The Case of Unilever Ghana Limited". International Review of Management and Marketing, 2013, Vol. 3, No. 4, pp.134-145.

[10] L. Achour, « La relation entre la satisfaction et la fidélité à la marque. Une étude empirique auprès des consommateurs Tunisiens de yaourts », La Revue des Sciences de Gestion, 6 / 2006 (n°222), p. 61-68.

[11] P. Chandon. « Dix ans de recherches sur la psychologie et le comportement des consommateurs face aux promotions », Recherche Et Applications En Marketing, 1994, 9 (3), pp 83-108.

[12] R. C. Blattberg and S. A. Neslin, "Sales promotion – Concepts, methods, and strategies", Englewood Cliffs, NJ, Prentice Hall, 1990.

[13] M. Zollinger. « Le concept de prix de référence dans le comportement du consommateur: D'une revue de la littérature à l'élaboration d'un modèle prix de référence-acceptabilité ». Recherche Et Applications En Marketing, 1993, 8 (2), pp 61-77.

[14] R. Sethuraman and K. Gielens, "Determinants of Store Brand Share", Journal of Retailing Volume 90, Issue 2, June 2014, pp 141–153.

[15] S A. Neslin, "A Market Response Model for Coupon Promotions", Marketing Science, 1990, 9 (2), p. 125-145.

[16] L. Froloff, « La sensibilité du consommateur à la promotion des ventes : De la naissance à la maturité ». Recherche Et Applications En Marketing, 1992, 7(3), pp 69-88.

[17] H L. Fangué Ndjiozé, V. Tsapi, and N. Fidèle, "Promotional offers and the sensitivity of the consumer to sales promotion", International Journal of Innovation and Applied Studies, Vol. 9 No. 3 Nov. 2014, pp. 1392-1400.

[18] B E. Kahn and R. Jagmohan, "Effects of Price Promotions on Variety Seeking and Reinforcement Behavior", Marketing Science, 1991, 10 (4), p. 316-337.

[19] S. Makridakis, S. C. Wheelwright, , R. J. Hyndman, "Forecasting: Methods and Applications", J. Wiley & Sons, 1998.

[20] J. S. Armstrong, and F. Collopy, "Integration of statistical methods and judgment for time series forecasting: Principles from empirical research", in G. Wright and P. Goodwin (Eds.), Forecasting with Judgment. Chicliester:John Wiley. 1998.

[21] Ö. Gür Ali, S.Sayin, T. van Woensel, and J. Fransoo, "SKU demand forecasting in the presence of promotions", Expert Systems with Applications, 2009, 36(10), pp 12340–12348.

[22] T. Huang, R. Fildes, and D. Soopramanien, "The value of competitive information in forecasting FMCG retail product sales and the variable selection problem", European Journal of Operational Research, 2014, 237(2): 738-748.

[23] L.G. Cooper, P. Baron, W. Levy, M. Swisher, and P. Gogos, "PromoCast: A New Forecasting Method for Promotion Planning", Marketing Science, 1999, vol.18, No.3, pp. 301-316.

[24] R. Fildes, P. Goodwin, M. Lawrence, and K. Nikolopoulos, "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning", International Journal of Forecasting, 2009, 25, 3-23.

[25] R. L. Andrews,. I. S. Currim, P. Leeflang, and J. Lim, "Estimating the SCAN*PRO model of store sales: HB, FM or just OLS?", International Journal of Research in Marketing, 2008, 25, 22-33.

[26] S. Divakar, B. T. Ratchford, and V. Shankar, "CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods", Marketing Science, 2005, 24, 334–350.

[27] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods". Journal of Retailing and Consumer Services, 2001, 8, 147–156.

[28] G. P. Zhang and M. Qi,. "Neural network forecasting for seasonal and trend time series". European Journal of Operational Research, 2005, 160, pp 501–514.

[29] , J. V. Hansen, J. B. McDonald, and R. D. Nelson, "Some evidence on forecasting time-series with support vector machines". Journal of the Operational Research Society, (2006, 57, pp 1053–1063.

[30] L. Aburto, and R. Weber, "Improved supply chain management based on hybrid demand forecasts", Applied Soft Computing, 2007 7, pp 136-144.

[31] M. Patak, and V. Vlckova, "Promotional Forecasting Based on Promotional Indexes in Yoghurt Production, DVANCES in ECONOMICS, RISK MANAGEMENT", POLITICAL and LAW SCIENCE, Czech Republic September 20-22, 2012.

[32] J. R. Trapero, N. Kourentzes, R. Fildes, "On the identification of sales forecasting models in the presence of promotions", Journal of the Operational Research Society (2015), 66, pp 299–307.

[33] W. Infosino, "Forecasting New Product Sales from Likelihood of Purchase Ratings", Marketing Science, 1986, 5(4), 372-384.

[34] J.S. Armstrong, V.G. Morwitz, and V. Kumar, "Sales Forecasts for Existing Consumer Products and Services: Do Purchase Intentions Contribute to Accuracy?", International Journal of Forecasting. 2000, vol. 16(3). p.383.

[35] Q. Geng, C. Wu, and K. Li "Pricing and Promotion Frequency in the Presence of Reference Price Effects in Supply Chains", California Journal of Operations Management, Volume 8, Number 1, pp 74-82.

Components' demand analysis in Assemble-to-Order systems: An analysis method

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mohammed Hichame Benbitour, Evren Sahin, Yves Dallery
Laboratoire Génie Industriel, CentraleSupélec
Grande Voie des Vignes
92290 Châtenay-Malabry, France
Email: mohammed-hichame.benbitour@centralesupelec.fr
evren.sahin@centralesupelec.fr
yves.dallery@centralesupelec.fr

Abstract— In this paper, we propose a method to analyze components' demand in an Assemble-to-Order system. This method uses past demand data of finished goods and the Bill of Materials in order to find the probability distribution of components demand and to evaluate forecast accuracy. Demand data is organized in a simple format we call the "Demand Summary Table" to facilitate visualization, manipulation and analysis. The proposed method has been applied to the case of an auto parts supplier and revealed to be simple and efficient. In addition, it is easily implementable in calculation software. It includes an analysis of the Bill of Materials and the different characteristics of components. We recommend to integrate the Demand Summary Tables in the ERP system, hence demand data can be collected and cleaned continuously.

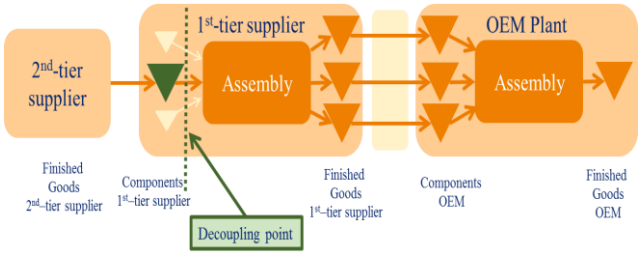
Keywords—demand analysis; assemble-to-order; methodology;

I. Introduction

The analysis of the demand of goods -a good can be a finished good (or FG), a sub-assembly or a component- in production/inventory systems and inventory systems is an important task. The demand pattern obtained from this analysis can be used as input information in other processes such as: Inventory management, Production planning, Capacity planning, Marketing strategy, etc. For example, the probability distribution of components demand is used in the estimation of safety stocks required to face stochastic demand. We define demand analysis as the method enabling to: 1) Find a statistical model that describes the randomness, seasonality, trend and cyclical fluctuations of demand and 2) Evaluate the forecasts (if available) accuracy by comparing them to realized demand.

In this paper, we propose a method to analyze the demand of components in an Assemble-to-Order (ATO) system. As defined by [7], an ATO system includes components and finished goods; demands occur for finished goods, but the system keeps only inventory of components. Finding a probability distribution of components demand is required to develop an efficient inventory control model in ATO systems. The ATO system we consider is inspired from the case of a first-tier supplier in the automotive industry. As described in the figure below, this supplier works on a Just-in-Time (or JIT) setting with the original equipment manufacturer (or OEM); the decoupling point between customer demand and supply is at the level of components in the first-tier supplier.

Fig. 1. An example of a 1st-tier supplier ATO system



The proposed method aims at finding a probability distribution that fits the demand of components, and to evaluate forecasts accuracy. This method takes as inputs the past demand data of FGs and the Bill of material (or BOM). The historical demand data contains demand (realized) and forecasts pertaining to FGs. First, we start by cleaning the input data in order to define the final sample size that will be used in the analysis. After cleaning the input data of FGs, the demand data of components is generated using the BOM. The BOM is analyzed and characteristics relative to components are highlighted. We propose a simple format to represent demand data; this format allows fast computational analysis and gives clear visualization of information. The demand of components is analyzed directly without analyzing the demand of FGs. The proposed analysis method was applied to the case of an auto parts supplier; the case study is presented in section 4 of this paper.

The reminder of this paper is organized as follows. We present a short literature review on ATO systems and demand analysis methods in section 2. The proposed analysis method and the case study are given in sections 3 and 4 respectively. We provide conclusions and directions for future research in section 5.

II. Literature review

Our work is related to two streams of literature: ATO systems and demand analysis techniques. There are two papers

Ref. [7] and Ref. [1] that summarize literature in the ATO systems research field. Ref. [7] classified papers into two categories depending on the addressed problem: 1) Performance evaluation and defining operating policies and 2) Design of ATO systems. They consider that there are two important issues to resolve in ATO systems: coordination (timing and quantity of supply orders) and allocation of components. The assumption relative to demand probability distribution of components in studied papers has a considerable impact on the analysis of ATO systems. In the other hand, [7] have pointed that data estimation (especially in large data sets) is not a trivial task. In our paper, we propose a simple method to collect, organize and clean large data sets in order to find the probability distribution of components demand.

Ref. [1] classified papers into periodic-review and continuous-review models; they also separated them depending on the number of FGs (one or multiple). In most studied systems, components demand is assumed to follow Poisson and compound Poisson probability distributions. The nature of the BOM has also an important impact on the performance of ATO systems. Literature generally considers simple BOM configurations. In our paper, we propose a detailed analysis of the BOM where different characteristics of components are identified

Many references have been developed in order to understand the nature of demand and its decomposition into different patterns. Ref. [4] was used to review the different characteristics of time series (demand of components is a time series) and the relation between forecasts and realized demands. The classic Ref. [2] was used for understanding the usual probability distributions, their characteristics and the different processes behind them. References [3], [5], [8] give some interesting insights about demands and forecasts in the automotive industry. The relation between first-tier suppliers and OEM is also studied in those references.

III. Analysis Method

We consider an ATO system with multiple (FGs) and multiple components. The ATO system has one customer (this customer is not essentially the final customer of the supply chain and can be any link in the chain). The demand of each FG is random and independent from the other FGs. Furthermore, customer demand must be satisfied totally and it cannot be backlogged.

Without loss of generality, we assume that at the beginning of each day, the customer sends a requisition (i.e., firm order) to the ATO system and this order is satisfied totally during the same day. The customer sends also at the beginning of the first day of each week the demand forecasts for the future days including the day of sending those forecasts. Sending the forecasts at the beginning of each week is a common practice in the automotive industry.

We assume that the probability distribution of FGs and components demand is unknown. At the same time, knowing the probability distribution of components demand in ATO systems is an essential task in advanced planning, operations management, inventory control…etc. In this study, we propose an analysis method to find the probability distribution of each component demand, and to estimate the parameters of this distribution.

In order to implement this analysis method, the following inputs are required: 1) demand data of FGs and 2) the BOM. Demand data consists of FGs requisitions and forecasts. FGs requisitions represent the firm orders of customer; they stand also for the ATO system sales (because all firm orders are totally satisfied and cannot be backlogged). The proposed analysis method was developed, based on a case study about an auto parts (car seats) maker.

In order to find the probability distribution of components demand, we start by collecting and cleaning the past requisitions and forecasts (demand data) of FGs, then the requisitions and forecasts of components are calculated using the BOM. Requisitions of each component is a time series. Hence, it can be decomposed into deterministic (trend, seasonal and cyclic) and non-deterministic (random) parts (see Ref. [4]). We use a method to find the probability distribution that fits the random part of requisitions without neglecting the deterministic part. The parameters of this distribution can be calculated from the past requisitions and/or forecasts depending on the nature of requisitions and the accuracy of forecasts.

The proposed analysis method passes through two types of analysis. It starts by the Preliminary analysis where: past requisitions and forecasts of FGs are extracted, then a suitable time unit to analyze demand data is defined and the Demand Summary Table (a table which contain requisitions and forecasts) of each FG is created. After finishing the Preliminary analysis, we go to the Detailed analysis where: the Demand Summary Tables of FGs are cleaned and as a result, the Demand Summary Tables of components are created using the BOM. Finally, we finish the Detailed analysis by finding the best fitting probability distribution of each component demand, and by analyzing the forecasts accuracy. In other words, the FGs requisitions and forecasts, and the BOM are used intelligently by the proposed analysis method to find the best fitting probability distribution of each component demand, and to estimate the parameters of this distribution.

Part 1: Preliminary analysis

Step 1: Extract the FGs requisitions and forecasts for a representative time horizon

First, extract the FGs demand data (requisitions and forecasts) from the information system of the plant (e.g., ERP, customers' invoices…) for a time horizon (interval) that is representative of the behavior of demand. There is no general rule to define the best extracting time horizon. From the experience and the know-how of the ATO system managers, a suitable time horizon can be defined. For example, if demand is seasonal, then the time horizon of extraction must contain at least one demand season.

For instance, an extraction is made for a time horizon composed of T weeks (from week 1 to week T), where week 1 is the most recent week, called also the "last extraction week" and week T is the "first extraction week". Then, we have for each week (From 1 to T) and for each FG the requisitions that occurred during this week and all the forecasts made at the first

day of this week for future delivery dates. Consequently, the extracted demand data in a week w ($1 \leq w \leq T$) and for all FGs can be represented by Table 1. For each requisition and forecast, we know the quantity and the corresponding delivery date. For requisitions, all delivery dates are within week w while they can be ulterior for forecasts.

Table 1. Extracted data in week w for all FGs (number of FGs is n)

		Quantity	Delivery date
FG 1	Forecasts of FG 1 made in week w		
	Requisitions of FG 1 made in week w		
FG 2	Forecasts of FG 2 made in week w		
	Requisitions of FG 2 made in week w		
…	…	…	…
FG n	Forecasts of FG n made in week w		
	Requisitions of FG n made in week w		

Step 2: Find a suitable time unit (time slot) to analyze demand

Find a suitable time unit to represent FGs requisitions. Requisitions can be expressed per day, week, month or another customized time unit. Generally, the best time unit to choose is the smallest possible time unit because it gives the biggest sample size. We say possible time unit because if requisitions are announced per day without specifying the hour of delivery, then the time unit cannot be the hour. However, it is not always better to use the smallest possible time unit. In some cases, aggregation of demand (from small to large time unit) can give interesting insights. The most suitable time unit would be the one that allows to depict clearly and completely the demand behavior (trend, seasonal, randomness…). In the next steps, we assume that the time unit is the day without loss of generality.

Step 3: Generate the Demand Summary Table for each FG

From extracted FGs demand data, find the "first delivery date" and the "last delivery date". The first delivery date corresponds to the earliest announced delivery date in forecasts and requisitions of all FGs, and the last delivery date is the most recent one (similarly, in forecasts and requisitions). The last delivery date can be more recent than the last extraction week in the case of a forecast of a future delivery date.

We use the Demand Summary Table (henceforth DST) explained in the next example to summarize the demand information related to each FG. The DST is an organized representation of the extracted requisitions and forecasts obtained in step 1. It has a matrix format in order to facilitate visualization, cleaning and manipulation.

Example: As shown in Table 2, the DST has as columns the delivery dates and as rows the requisition and forecasts for each delivery date. Forecasts can be made in the same week of delivery date (denoted by "Forecast W-0"), and can be made in prior weeks also. For instance, we denote by "Forecast W-1" the forecasts made one week before the delivery date. In order to replenish the DST of each FG, we start by generating the list of all the days which exist between the first and last delivery dates. For each delivery date, we fill the corresponding requisition and forecasts. If the requisition is unknown (it doesn't exist in the extracted requisitions), we put "NR" in the DST. "NF" is put in the case of an unknown forecast. The number of all delivery dates represents the sample size of the DST. For simplification, we say that forecasts can be provided at least 20 weeks prior to the delivery date. In short, the DST is a matrix which contains the requisition and forecasts (quantities) of each delivery date.

Table 2. Example of demand summary table (DST) of a FG

Year	First delivery date	…	2017	2017	2017	2017	2017	2017	2017	…	Last delivery date
Month			3	3	3	3	3	3	3		
Week			10	10	11	11	11	11	11		
Day			9	10	13	14	15	16	17		
Requisition	28		38	35	NR	29	31	27	NR		NR
Forecast W-0	29		38	33	NF	30	32	32	NF		NF
Forecast W-1	NF		37	NF	NF	NF	30	29	33		NF
Forecast W-2	NF		NF	NF	27	32	NF	30	31		NF
Forecast W-3	NF		30	37	32	29	32	NF	31		35
Forecast W-4	NF		NF	38	30	25	30	27	28		32
Forecast …	…		…	…	…	…	…	…	…		…
Forecast W-20	NF		NF	42	29	35	40	34	NF		30

Part 2: Detailed analysis

Step 4: Analyze the total FGs requisitions and clean the Demand Summary Tables

After generating the DSTs of FGs from the extracted requisitions and forecasts in step 1, it is often noticed that the DSTs may contain incoherent information (either about forecasts or about requisitions). The objective of this step is to clean the DSTs of FGs in order to have rigorous results in the next steps (garbage in, garbage out). Incoherent information in a DST can have different forms. For example, some delivery dates have requisitions which are very different than the other delivery dates. Those delivery dates constitute exceptions that

should be justified. Most importantly, those delivery dates do not represent the normal demand behavior and should be deleted from all the DSTs.

In addition, the cases "NR" and "NF" have to be explained. In fact, the information "NR" or "NF" can have two significations: 1) Requisitions or forecasts have not been made because it was an exceptional day (no-working day, strike, production problem,…), or 2) Requisitions or forecasts are null.

In order to identify the delivery dates which cannot be exploited in the demand analysis, we propose to analyze the total FGs requisitions (the sum of requisitions of all FGs). Delivery dates with "NR" are considered as zeros in the sum. In fact, the system managers know the total requisitions (the mix) better than the detailed requisitions (per FG), and they can identify the incoherent delivery dates of the mix more quickly than the delivery dates of each FG.

After deciding the delivery dates to keep in the analysis, the DSTs of FGs must be cleaned. The information "NR" in the new tables is replaced by zero (null requisition).

We propose a conventional method to correct forecasts (replace the information "NF" by a real number). The information "NF" made in a week *w* is replaced by the most recent existing forecast made in weeks prior to week *w*. In the example of Table 2, Forecast W-2 to be delivered the 09/03/2017 takes the value 30. Forecasts W-1 and W-2 to be delivered the 10/03/2017 take the value 37. When the information "NF" cannot be corrected, it takes the value zero.

Step 5: Find the representative components and generate their DSTs

After cleaning FGs demand data, components DSTs can be calculated using the BOM. The BOM is a list of components and sub-assemblies required to construct the FGs. The BOM appears generally in a hierarchical format, with the highest level representing the FG and the bottom level representing the individual components required by this FG. In our analysis method, we study the individual component only (because the stock of individual components is held in the ATO system, and there is no stock of sub-assemblies).

In ATO systems, the number of individual components is usually large and it is time consuming to analyze the demand of each one. Some components are similar and can be represented by one component (we call it a representative component). Two components *i* and *i'* are said to be similar if the assembly coefficients of component *i* are the same, multiple or divisor of the assembly coefficients of component *i'*. Instead of calculating the DSTs of all components, only the DSTs of representative components are calculated.

Step 6: Find representative components with abnormal demand behavior and correct them

Start by depicting the graphs of requisitions of representative components and find components with abnormal demand behavior. The abnormal demand behavior can have many forms. For example, a component can be replaced by another one. The replaced component and its substitute must be considered as the same component.

Step 7: Find the best fitting probability distribution of each representative component demand and analyze forecast accuracy

Requisitions of each component is a time series. We assume that there is only a linear trend pattern in the deterministic part $R_t = \alpha t + \beta$. Use a decomposition procedure in order to remove linear trends. Matlab or R (software for statistical computing) can be used to remove trends (function detrend).

After removing the trend, calculate for each component the basic statistical indicators of requisitions: Average, standard deviation (*Std*), Coefficient of variation, *Max*, *Min*, *Max-Min*, $\frac{Max-Min}{Std}$ to understand the behavior of each component. In order to find the best fitting probability distribution: distributions like Binomial, Poisson, Normal or Geometric can be tested using goodness-of-fit tests like Pearson's chi-squared test and Kolmogorov–Smirnov test.

After defining the probability distribution of each component, the accuracy of forecasts has to be evaluated. Forecasts are used in the estimation of the distribution law parameters. For example, when there is a trend in a component demand, then the average demand can be estimated using forecasts. Simple accuracy indicators like: The quotient $\frac{Average[F]}{Average[R]}$, The error $Average\ [F-R]$, The absolute error $Average\ [|F-R|]$, The relative error $\frac{Average[|F-R|]}{Average[F]}$, The quadratic error $Average\ [(F-R)^2]$ give interesting insights about the quality of forecasts. Those indicators are calculated for each forecast week (W-1, W-2,…).

IV. Case Study

We applied the proposed analysis method to the case of an auto parts maker plant situated in the west of France. This plant manufactures customized seats for two five-door car models. A finished good represents the seats (front and rear) of one car. There are 264 FGs. The plant works with a relatively constant production rate of 332 cars (FGs) per day.

Preliminary analysis:

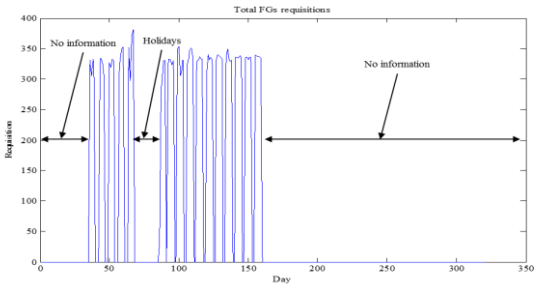
We started by making an extract of FGs demand data (requisitions and forecasts). The first extraction week is week 42 of year 2014 and the last extraction week is week 13 of year 2015. The chosen time unit is the day. The first delivery date is the 13/10/2014 and the last delivery date is the 28/08/2015. The initial sample size is 320 days. We filled the DSTs of all FGs using an Excel VBA application (from extracted data to DSTs).

Detailed analysis:

We studied the total FGs requisitions depicted in Fig. 2. We did not use an upper and a lower limits because the total FGs requisitions were relatively steady (except the case where requisitions were null). It can be seen that requisitions in the right half of the graph are null. This is due to the fact that all requisitions after the last extraction week are unknown (the delivery dates after the last extraction week are announced by

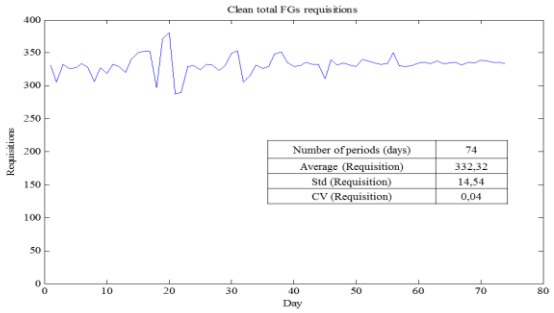
forecasts). During the end of the year (holidays), the plant was closed and for that reason there is no demand. In the left part of the graph, the requisitions are unknown because they were not recorded. The other days where requisitions are null represent no-working days.

Fig. 2. Total FGs requisitions-Case study



After removing all delivery dates with null total requisitions, the sample size became 74 days and the average demand became equal to 332 units/day (see Fig. 3). The DSTs of all FGs are updated (only the 74 days are kept). We correct forecasts using the conventional method given in step 4 of our analysis method.

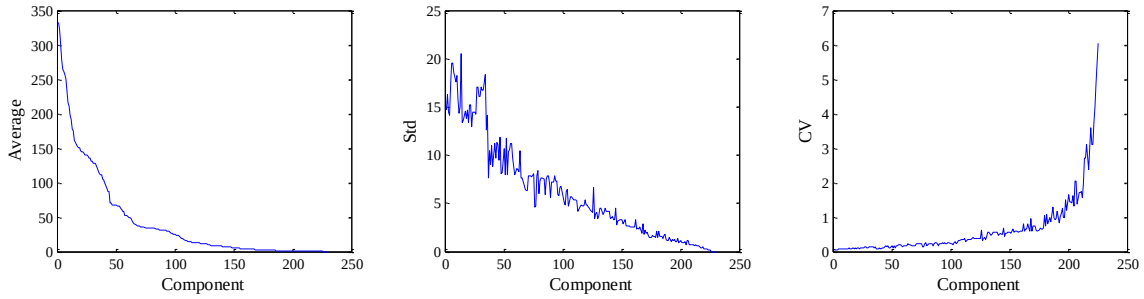
Fig. 3. Clean total FGs requisitions-Case study



Once the DSTs of FGs are cleaned, we analyzed the BOM. There are 689 individual components, 236 of them can be representative components. The DSTs of all representative components are generated (requisitions and forecasts are divided by the average assembly coefficient of each component). We found that there were five components replaced by others. The number of studied components became 231.

The linear trend is removed from components using the function "detrend" in Matlab. Once components requisitions are ready to the last step, we start by calculating the basic statistical indicators. The average requisition, standard deviation *Std* and coefficient of variation *CV* are depicted in Fig. 4. There are 123 components which have less than 5% of the average total demand (332 units/day), 48 components with a demand between 5% and 15%, and 60 components with more than 15%.

Fig. 4. Variations of Average requisition, Std and CV



Because the sample size is 74 (relatively small), we found that more than one probability distribution can be accepted by the goodness-of-fit tests. On the other hand, the total demand is relatively constant and each FG has a percentage of use (Average requisition of this FG divided by the total demand). Hence, the demand of each FG can be viewed as a binomial process. This analogy is explained in the table below:

Table 3. Analogy between a binomial demand and a FG demand

	Binomial demand	FG demand
Number of experiments	Defined number of Bernoulli trials	Total demand (332 units/day)
Probability of success	Constant probability of success	Use percentage
Result (random variable)	Number of successes	Demand per day

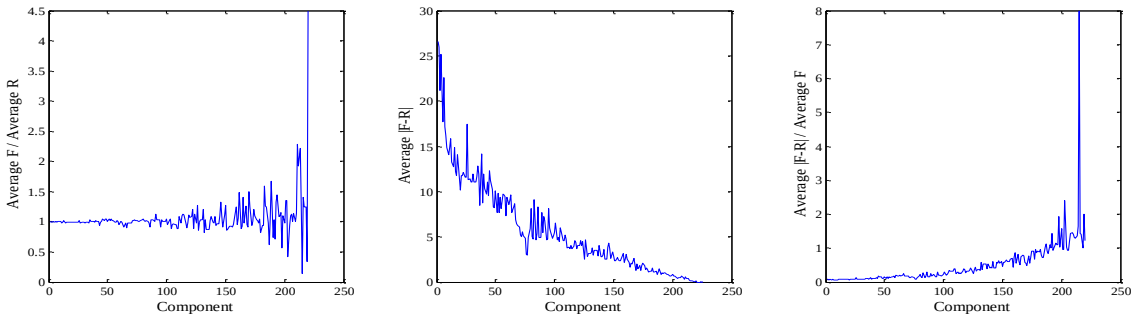
Because the use percentage of FGs is small, the binomial process can be approximated by a Poisson process. Hence, we can say that FGs demand follow a Poisson distribution law. In addition, the requisitions of each component are expressed by the sum of requisitions of FGs that use this component, and the sum of Poisson processes is a Poisson process. So, we can say that the requisitions of components follow a Poisson process. In order to test this hypothesis, we compare $\sqrt{average\ requisition}$ to *Std*. In a Poisson process, $\frac{\sqrt{average\ requisition}}{std} = 1$.

The min and max of quotient $\frac{\sqrt{average\ requisition}}{std}$ are 0.69 and 1.9 respectively. There are 31 components with a quotient greater than 1.30. The average quotient for all components is 1.11 with a coefficient of variation equal to 0.17. We can

conclude that the Poisson distribution can be a good approximation for components' requisitions. More precisely, components have a compound Poisson process when their average assembly coefficient is greater than one.

The forecasts accuracy is also evaluated. The figures below show the indicators $\frac{Average[F]}{Average[R]}$, $Average\ [|F-R|]$ and $\frac{Average[|F-R|]}{Average[F]}$ for all components (components are ranged in a decreasing order of average requisition). Fig. 5 represents forecasts made three weeks before requisition. It can be seen that forecast accuracy seems to be good especially for high and medium runners. To conclude the case study, we can say that the compound Poisson distribution is a good approximation to components demand. The parameter (mean requisition) of this distribution can be calculated using the forecasts.

Fig. 5. Forecasts made three weeks before requisition



V. Conclusions and Future Research

In this paper, we proposed a method to analyze components' demand in an ATO system. The proposed method has as inputs the past demand data of FGs and the BOM, and as outputs the probability distribution of components demand and the accuracy of forecasts. Instead of analyzing the demand of FGs and then deducing the demand of components, we analyze directly the demand of components. First, the past demand data of FGs is cleaned and represented as a matrix we call the "Demand Summary Table" (or DST). Then, the DST of each component is calculated and analyzed.

The proposed method is applied to the case of an auto parts supplier with 264 finished goods and 689 components. We find that the components demand probability distribution can be approximated by a compound Poisson law. The forecasts accuracy is shown to be very good. The proposed method has shown its efficiency for many reasons: it allows to clean quickly the input data, it is easily implementable in computer calculation software (for instance, Matlab and Excel VBA were used in the case study), and it facilitates the visualization and the manipulation of data.

Even if this method was used in the case of an ATO system, it can be easily implemented in the case of Make-to-Stock systems. In the latter system, there is an inventory of FGs and components. The DSTs of components and FGs must be analyzed separately in order to find probability distribution of each one. In this paper, we took the assumptions that the time unit is the day and that all forecasts are received at the beginning of week. The calculation of DSTs holds the same whatever the time unit or the arrival process of forecasts are.

An important aspect in using this method and analyzing demand is related to the extraction of useful input data. We propose to integrate the DSTs of FGs in the ERP system in order to be filled in a continuous way. The continuously filled DSTs can be improved by adding comments to each filled day (column), hence data analyzers can interpret more quickly incoherent information. For future research, we propose to include the analysis of customer behavior in the analysis of demand. In fact, the behavior of customers and the way they send orders determine the probability distribution of demand.

References

[1] Atan, Z., Ahmadi, T., Stegehuis, C., de Kok, T., & Adan, I. (2017). Assemble-to-order systems: A review. *European Journal of Operational Research*.

[2] Feller, W. (1968). *An introduction to probability theory and its applications: volume I* (Vol. 3). New York: John Wiley & Sons.

[3] Giard, V., & Mendy, G. (2006). Amélioration de la synchronisation de la production sur une chaîne logistique. *Revue Française de Gestion Industrielle*, *25*(1), 63.

[4] Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts.

[5] Meyr, H. (2009). Supply chain planning in the German automotive industry. In *Supply Chain Planning* (pp. 1-23). Springer Berlin Heidelberg.

[6] Orlicky, J. A., Plossl, G. W., & Wight, O. W. (2003). Structuring the Bill of Material for MRP. *Operations management: critical perspectives on business and management*, *58*.

[7] Song, J. S., & Zipkin, P. (2003). Supply chain operations: Assemble-to-order systems. *Handbooks in operations research and management science*, *11*, 561-596.

[8] Zhang, X., & Chen, R. (2006). Forecast-driven or customer-order-driven? An empirical analysis of the Chinese automotive industry. *International Journal of Operations & Production Management*, *26*(6), 668-688.

The impact of macroeconomic leading indicators for tactical sales forecasting on SKU inventory management

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Yves R. Sagaert*, Stijn De Vuyst*, Nikolaos Kourentzes† and El-Houssaine Aghezzaf*
*Department of Industrial Management
Ghent University, Technologiepark 903, 9052 BE, Zwijnaarde
Email: yves.sagaert@ugent.be
†Department of Management Science
Lancaster University Management School, LA1 4YW UK, Lancaster

***Abstract*—An accurate sales forecasting has indispensable effects on the supply chain management as this input is essential in the decision making process. Macroeconomic leading indicators can provide early indications of global changing economic dynamics. By including this external information, the global tactical sales forecasting can be improved. This paper wants to quantify the impact on inventory level, where decisions are typically taken on an individual product base. For this, the high-level forecast needs to be disaggregated to the product level. Techniques that make use of the hierarchical structure present can benefit from pooling individual forecasts on different hierarchical levels. We propose an empirical technique to reconcile the forecast distributions of different aggregation levels in a hierarchical structure. We focus on the first and second moment of the forecasting distribution, the mean and variance. We evaluate our proposed method on inventory and service level via inventory simulations.**

I. Introduction

Recent research has shown that macroeconomic leading indicators have the potential to detect early warning signals from economic activity. Furthermore, [13] shows that this information can improve tactical sales forecasting. The tactical level is typically formulated on a global scale, where the economic dynamics are expected as well. The sales forecasts steer the business decision process on global budget, global capacity pooling, as well as on production planning and inventory management. The latter typically happens on an operational level. In the studied case in this research, inventory management is done on production plant level, which are geographically distributed. Furthermore, inventory decisions are taken on Stock-Keeping-Units (SKU) level, as each product or material has an individual stock.

The methodology proposed in [12] avoids to forecast the macroeconomic indicators individually by shifting these indicators in time prior to insert them in the sales forecasting model. As the forecast model uses then the already realised values of the indicators, this would bring less uncertainty to the calculated point forecast. As a consequence, the prediction interval around the point forecast is more narrow on a tactical level. In this research, we investigate whether the formulation of more narrow prediction intervals around the point forecast on global level can be beneficial for the lower levels of forecasting. Our paper aims to link the tactical forecast to its impact on inventory level. This is in line with the literature in sales forecasting that argues for evaluation of forecast models where they are used [2], [9], [17]. As the forecasting models are used for inventory management, the evaluation should also occur on inventory performance. However, inventory performance is largely impacted by the shape of the demand distribution [19]. Here, inventory simulation models are proposed to quantify the impact of different forecasting models [15].

In order to quantify the impact on inventory, it is necessary to address the first and second moment of the forecast distribution. This means that both the mean and the variance of the high-level forecast need to be disaggregated to the lower levels. This is in contrast with traditional techniques for disaggregating, as these focus mainly on the calculated mean of the forecast [5]. Here, we use hierarchical forecasting via reconciling to link all levels of the company structure. This approach was first proposed by [7], who found that reconciliation can improve the forecast accuracy on several levels of the hierarchical structure. Here, all the time series are forecasted independently at all levels of the hierarchy. Next, these forecasts are combined and reconciled using a regression model. This approach allows to combine extrapolations of patterns in historical sales on lower levels with forecasts on highest level that are augmented with exogenous indicators. In this way, information of all levels is addressed.

We propose to reconcile the uncertainties on different hierarchical levels through an empirical method. By reconciling the prediction intervals across different hierarchical levels, we investigate whether more narrow prediction intervals on high level can result in lower stock in inventory management on SKU level. For this, the reconciled prediction intervals on SKU level are used as input for an inventory simulation with real case data.

II. Literature review

There are different strategies to handle time series with several levels of aggregation [10]. The most simple approach ignores the structure completely and generates a direct forecast. [4] gives an overview of direct versus aggregation derived forecasts. He concludes that forecasting via aggregation results

in superior forecast performances in a majority of the research, even if an information loss may exist by summing product time series. Forecasting approaches that account for this hierarchical structure can be divided in four categories. First, top-down forecasting calculates a direct forecast at the top level time series and generates lower level forecasts by dividing the top level forecast proportional to the lower level sales volume. Alternatively, the volume of the sales forecast on the lower level can be used to disaggregate the top level forecast. Second, a sales forecast is formulated for each of the lowest level time series and the top level forecast is obtained by simply summing these forecasts. This is known as bottom-up forecasting. There is clear trade-off between these two approaches, noted by [18]. Top series may be easier to forecast, but more univariate patterns can be captured on low level series. Third, a combination of both top-down and bottom-up can be used, sometimes called middle-out. This approach is often used in practice [5]. Here, a direct forecast is calculated on an intermediate level, and higher and lower levels are calculated using aggregation and disaggregation. Fourth, combinatorial forecasting creates a direct forecast at all levels of the hierarchy and combines them. [1] proposes a new approach for this combination process, by reconciling the individually formulated forecasts using a regression model. This combination process allows for to alter the individual forecasts on all levels of the hierarchical structure. Alternatively, the forecasts on all hierarchical contribute to the final forecast. Furthermore, forecast reconciliation is flexible, as it can be used across multiple hierarchies, such as product, geographic or customer hierarchies [14].

Most of the literature focusses on the point forecasts in these hierarchical time series. However, [7] notes the theoretical formulation to formulate prediction intervals of the reconciled forecasts and defers the optimisation of this formula to future work. In this paper, we propose an approach to reconcile the empirical prediction intervals across the hierarchical structure in a similar matter as the reconciliation takes place for the point forecast.

III. Methodology

In our methodology, the exogenous macroeconomic indicators are identified on a high aggregation level of the sales, as we do not expect an individual product to have the same dynamics as the global or national economy. However, the impact of the forecasting models happens on the product level, in terms of the inventory control for this product. Therefore it is essential that the tactical forecasts on high level are linked to the lower levels of the hierarchical structure. Furthermore, the top levels do not exhibit any low level historical pattern, such as product-specific seasonality. This type of information will be captured by low level models. As both levels can add information to the final forecast, hierarchical reconciliation is used as an approach to combine the different hierarchical levels. Next, the performance on the product level is evaluated via production simulation and inventory performance.

A. Macroeconomic indicators selection and forecasting model

On a global level, the sales forecast is modelled using the external information of macroeconomic indicators. These are selected using the methodology described by [13] that is based on the Least Absolute Shrinkage and Selection Operator (LASSO) [16]. For the aggregated time series on top level, univariate information such as seasonality and auto-regressive process are combined with external indicators. The relevant indicators are selected in a fully automatic way, simultaneously with their leading effect. This means that the optimal shift in time is determined for each indicator separately. Furthermore, the used framework is set up to formulate unconditional forecasts. By design, the indicators are only used when their leading effect is higher than or equal to the forecast horizon. While this complicates the computation as for each forecast horizon an individual LASSO model needs to be formulated, it makes the application in practice extremely relevant.

The amount of potential indicators p is very large, while the number of historical data point n is typically limited, creating a $p \gg n$ problem. This problem expands even further as for each indicator, the optimal leading effect needs to be addressed as well. [13] shows that LASSO in this context was capable to retain useful and relevant indicators automatically.

The prediction interval around the point forecasts is formulated empirically for the h-step ahead forecasts by $\hat{Y}_h \pm k\hat{\sigma}_h$, where k is the appropriate percentile point of the standard normal distribution. Here, we assume that the error has a Gaussian distribution with zero mean and standard deviation σ. The estimation $\hat{\sigma}_h$ is done via the Root Mean Squared Error (RMSE) on the historical sales. For the univariate models, we re-estimate the RMSE for each forecast horizon, while for the LASSO model a different model is formulated for each forecast horizon already.

B. Hierarchical reconciliation

Once the individual forecasts are formulated, these can be combined across the hierarchical structure. Naturally, bottom-up aggregation cannot benefit from the leading indicator forecasts as this approach generates higher time series forecasts by summing up the lower levels. [14] notes that traditional techniques for aggregation often have weak performance. For example, top-down disaggregation yields often good forecasts at higher aggregation levels, but worse forecast performance at lower levels. In this research, we combine the forecasts on different aggregation levels, via hierarchical reconciliation. Using a linear regression on the different individual time series, we obtain the coherent reconciled forecasts for the product hierarchy. The weights for this linear combination can be derived from the structure of the hierarchy, as discussed by [7]. This work argues that the optimal combination is independent of the data, and that the weight of an aggregated forecast is set equal to the amount of lower level series.

C. Inventory simulation

The designed simulation allows to simulate the production and inventory build-up, based on the sales forecasts and their prediction intervals. The product forecasts with their respective prediction intervals are an central input in this. These are the reconciled forecasts at the lowest level of the hierarchical structure. Here, we assume all products to be produced via a Make-to-Stock policy. In the inventory simulation is as follows: at the end of each month, the actual realised sales is taken into account, and the remaining stock is reviewed. The inventory is controlled by how much of each product

should be produced in the upcoming month that is planned. To approximate the real world inventory management, the simulation enforces a production stand-off of 5 months. This condition makes that the planned amount to produce in the month after the production stand-off is equal to the total sales in the next 6 months, reduced with the production of the next 5 months and the current inventory of that product.

To be more specific, we considering one product time series at time t. Let I_t be the inventory at the end of period t, P_t the amount of items produced in a month t, and $\hat{Y}_{t+h}$ the forecasted demand for horizon $t+h$. The inventory position at time $t+h$ can then be estimated as

$$\hat{I}_{t+h} = I_t + \sum_{i=1}^{h} (P_{t+i} - \hat{Y}_{t+i})\,. \tag{1}$$

The demand at $t+h$ will be satisfied entirely whenever

$$I_{t+h-1} + P_{t+h} > Y_{t+h}\,. \tag{2}$$

If we choose to produce P_{t+h} items at time $t+h$, the probability that the demand in period $t+h$ is not entirely satisfied can be estimated as

$$\hat{\alpha}(P_{t+h}) = \text{Prob}\left[\hat{I}_{t+h-1} + P_{t+h} < \hat{Y}_{t+h}\right] \tag{3}$$

Given the (continuous) distribution function

$$F^*_{t+h}(x) = \text{Prob}\left[\sum_{i=1}^{h} \hat{Y}_{t+i} \leq x\right], \tag{4}$$

of the summed forecasted demand and a certain acceptable shortage probability α_{t+h} for period $t+h$, the production quantity P_{t+h} in that period can be allocated as

$$P_{t+h} = F^{*-1}_{t+h}(1-\alpha_{t+h}) - I_t - \sum_{i=1}^{h-1} P_{t+i}\,, \tag{5}$$

The production is determined in the simulation via (5). This equation takes into account the inventory on-hand and the allocated production up to $h-1$ (due to the production stand-off). Furthermore, it accounts via (4) for the estimated demand up to period h and the related shortage probability, shown in (3), over the period 1 to h. As a consequence, the simulation allows for a certain acceptable shortage, which is evaluated by the fill rate at product level.

IV. Performance measures

In order to evaluate the effect of the empirical formulated prediction intervals in the hierarchical structure of the sales, we will analyse the forecast results at each stage of our methodology.

We compare the forecasting performance from an accuracy standpoint using the Mean Absolute Percentage Error (MAPE) in (6), as is widely used in practice and relevant to the case company.

$$MAPE_h = \frac{1}{n}\sum_{t=1}^{n} \frac{\left|Y_{t+h} - \hat{Y}_{t+h}\right|}{Y_{t+h}}, \tag{6}$$

where Y_{t+h} is the actual product demand and $\hat{Y}_{t+h}$ is the forecast for time period $t+h$.

The uncertainty around the point forecast is evaluated via $\hat{\sigma}_h$, as this is the key element of the prediction interval. We define $AvgRel\hat{\sigma}_h$ as:

$$AvgRel\hat{\sigma}_h = \sqrt[m]{\prod_{p=1}^{m}\left(\frac{\hat{\sigma}_h^A}{\hat{\sigma}_h^B}\right)}, \tag{7}$$

where $\hat{\sigma}_h^A$ refers to the evaluated forecast via hierarchical reconciliation and $\hat{\sigma}_h^B$ to the benchmark model on low level with purely historical data. In literature, the $\hat{\sigma}_h$ is traditionally obtained by multiplying the estimated demand variance by the length of the forecast horizon. [11] notes that this approach is flawed, as forecast errors for different periods of the forecast horizon are positively correlated. In contrast to the traditional methods, we determine the $\hat{\sigma}_h$ on all levels of the hierarchical structure empirically.

Finally, we display the effect of the proposed method and the univariate benchmark on inventory management via an inventory simulation. To judge how well a forecasting model performs on product j, we use average on-hand inventory $\bar{I}_\alpha^{+(j)}$ and the fill rate $\overline{FR}_\alpha^{(j)}$ (FR) achieved over the test periods as performance metrics. Here α is a tuning parameter in the decision of the production quantity. The fill rate refers to the fraction of demanded items which can be obtained immediately from stock without backordering. To assess inventory performance, a weighted average is taken across the different products:

$$\bar{I}_\alpha^{+} = \sum_{j=1}^{14} w_j \bar{I}^{+(j)}\,, \qquad \text{and} \qquad \overline{FR}_\alpha = \sum_{j=1}^{14} w_j FR^{(j)}\,, \tag{8}$$

with weights according to the actual demanded volumes per product j

$$w_j = \frac{\sum_{t=n+1}^{n+T} Y_t^{(j)}}{\sum_{j=1}^{m}\sum_{t=n+1}^{n+T} Y_t^{(j)}}\,. \tag{9}$$

V. Results

The dataset for the company case contains one business unit from a tire manufacturer for the period 2005–2015. This experiment considers three hierarchical levels from the business unit structure. The highest level is the overall sales for the business unit. The intermediate level is the plant level, which consists of 5 manufacturing plants. The lowest level differentiate all the customers that are served from this plant. The customers are Business-to-Business users and typically order only one material from an individual plant. The lowest level consists of 14 different time series of customer–product combination. In our experiment, we have omitted the customers who do not order frequently, as these customers result in intermittent demand patterns.

Figure 1 shows a generic hierarchical structure of three levels. The top level is the aggregated sales of the business unit. The intermediate level are the individual manufacturing plants and the lowest levels are the individual customers that are served from each plant. Customers who order from several plants are shown multiple times on the lowest level, as the

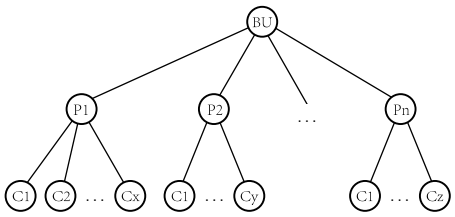


Fig. 1. Hierarchical structure that exists of three levels: Business Unit (BU), Manufacturing Plant (P) and Customer level (C)

time series on this level represent an unique combination of product and customer.

The models are trained in the period 2005:01-2012:12 and their performance is tested in the period 2013:01-2014:12 over a rolling origin experiment. The design of our experiment is as follows: the sales forecasts are generated for the next 6 months to allow to plan production. However, the planning decisions are made with a production stand-off of 5 months. The benchmark on the lowest level is exponential smoothing (ETS) as proposed by [6].

TABLE I. MAPE FOR FORECASTING SETUP

	1	2	3	4	5	6
ETS	31.2	33.2	33.6	33.9	33.9	37.4
Hierarchical	30.2	31.4	31.8	31.7	31.7	35.3

Table I compares the individual forecasts on each aggregation level with the reconciled forecasts. The MAPE for each horizon h is aggregated across forecasts from 13 origins and 14 products to form the MAPE displayed. The hierarchical reconciled forecasting method that makes use of external data outperforms the benchmark technique on the product level. Here, the forecast horizon $h = 6$ is of special interest for the inventory simulation. For this horizon, we see that the reconciled forecast outperforms the benchmark model substantially.

TABLE II. EVALUATING FORECAST UNCERTAINTY VIA $AvgRel\hat{\sigma}_h$

	1	2	3	4	5	6
Hierarchical	1.9	1.7	1.6	1.5	1.4	1.3
ETS	1	1	1	1	1	1

Table II exhibits the $AvgRel\hat{\sigma}_h$ for each forecast horizon, aggregated via the geometric mean across time series and rolling origins. The demand forecast on the top level that is formulated with macroeconomic indicators, has a more narrow prediction interval than the benchmark model ETS. This is intuitive as these formulated forecasts use realised values of the indicators. As a consequence, the empirical reconciled $AvgRel\hat{\sigma}_h$ is expected to decrease via the reconciliation. However, the results in table II suggest the opposite. A crucial point in the hierarchical reconciliation is the weighting scheme in the linear combination of the different aggregation levels. The modelling technique of LASSO differs substantially of the nature benchmark, which has an impact on traditional weighting scheme for hierarchical reconciliation as proposed by [7].

The inventory simulation performs the final evaluation on product-customer level for the different forecasting models. Figure 2 exhibits the simulation results for one specific

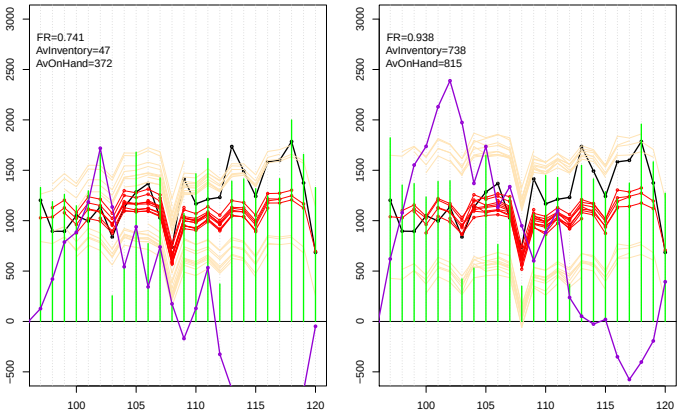


Fig. 2. Production simulation for ETS and LASSO over a rolling origin with the actual demand (black), the point forecasts (red), 20% and 80% quantiles of the forecast (orange), the production quantities (green) and the inventory position (purple). Production stand-off $h = 4$ and parameter $\alpha = 0.05$.

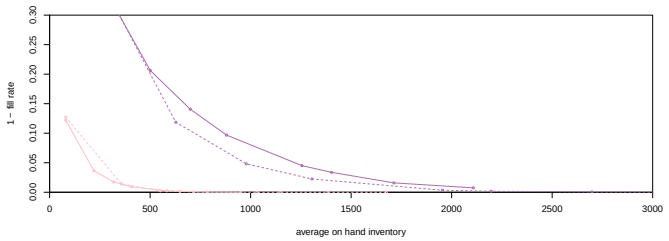


Fig. 3. Results of production simulation for LASSO (solid line) and ETS (dotted) for both $h = 0$ (light) and $h = 4$ (dark). On each line, the parameter α for the production allocation decreases from 0.5 (top left) to 0.001 (bottom right). The vertical axis is $1 - \overline{\text{FR}}_\alpha$, and on the horizontal axis is the average on-hand inventory $\bar{I}^+_\alpha$.

product. In this figure, the black line represents the actual demand. The green bars indicate the allocated production, and the purple line is the inventory position for this product. In red, the consecutive forecasts are shown, with their respective prediction intervals in orange.

Figure 3 shows the resulting performance curves from the inventory simulation. Given a desired fill rate, the average on-hand inventory is lower for the reconciled forecasts on the lowest levels, compared to univariate models. We can see that the hierarchical reconciliation method outperforms the benchmark technique for a given service level of the fill-rate.

VI. CONCLUSION

Incorporating exogenous information in top level sales forecasts can improve these forecasts. In our methodology, we select macroeconomic indicators and their leading effect fully automatic from a large pool of indicators. We found that the forecasting accuracy is improved and that the prediction intervals are more narrow on the top level. However, supply chain decision are often taken on a lower level of aggregation. It is interesting to quantify the difference of this exogenous information on an operational level, such as inventory management. For this, it is essential to link the top level to the lowest individual product level, as inventory stock is kept at this level. Dealing with hierarchical structures in sales forecasting

is described in literature [5], [7], but is mainly focussed on the point forecast. For inventory decision, the uncertainty around this point forecast (prediction intervals) is equally important. In this paper, we propose a new empirical approach to reconcile prediction intervals in a hierarchical structure.

We evaluate our proposed approach on each stage of our methodology. The results of the reconciled forecasts and prediction intervals show the following: (i) the forecast accuracy on the product level improves by reconciling the different hierarchical levels. (ii) while we expect the prediction intervals to become more narrow by reconciling across different hierarchical levels this is not the case, and (iii) The inventory simulation exhibits lower stock levels for the same fill rate. Whereas the LASSO prediction intervals on the top level produce more narrow prediction intervals, this improvement does not translate to lower levels of aggregation. A key aspect here is that the weighting of reconciliation occurs via the MSE on the train sample. As the LASSO technique has a different cost function, it is questionable if this weighting scheme is appropriate. Investigating other weighting schemes for this purpose is put as further research.

REFERENCES

[1] G. Athanasopoulos, R. A. Ahmed and R. J. Hyndman, *Hierarchical forecasts for Australian domestic tourism*, International Journal of Forecasting, Volume 25, Issue 1, 2009, Pages 146-166

[2] M. Babai, M. Ali, J. Boylan, and A. Syntetos, *Forecasting and inventory performance in a two-stage supply chain with ARIMA(0,1,1) demand: Theory and empirical analysis*, International Journal of Production Economics, Volume 143, Issue 2, 2013, Pages 463–471

[3] A. Chen, J. Blue, *Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands*, International Journal of Production Economics, Volume 128, Issue 2, 2010, Pages 586-602

[4] G. Fliedner, *An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation*, Computers & Operations Research, Volume 26, Issue 10, 1999, Pages 1133–1149

[5] G. Fliedner, *Hierarchical forecasting: issues and use guidelines*, Industrial Management & Data Systems, Volume 101, Issue 1, 2001, Pages 5–12

[6] R. Hyndman, A. B. Koehler, J. K. Ord and R. D. Snyder, *Forecasting with exponential smoothing: the state space approach*, Springer Science & Business Media, 2008

[7] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, *Optimal combination forecasts for hierarchical time series*, Computational Statistics & Data Analysis, Volume 55, Issue 9, 2011, Pages 2579–2589

[8] A. Kerkknen, J. Korpela and J. Huiskonen, *Demand forecasting errors in industrial context: Measurement and impacts*, International Journal of Production Economics, Volume 118, Issue 1, March 2009, Pages 43–48, ISSN 0925-5273

[9] N. Kourentzes, *Intermittent demand forecasts with neural networks*, International Journal of Production Economics, Volume 143, Issue 1, 2013, Pages 198–206

[10] S. Moon, C. Hicks and A. Simpson, *The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean Navy - A case study*, International Journal of Production Economics, Volume 140, Issue 2, 2012, Pages 794–802

[11] D. Prak, R. Teunter, A. Syntetos, *On the calculation of safety stocks when demand is forecasted*, European Journal of Operational Research, Volume 256, Issue 2, 2017, Pages 454–461

[12] Y. R. Sagaert, E.-H. Aghezzaf, N. Kourentzes and B. Desmet, *Tactical sales forecasting using a very large set of macroeconomic indicators*, European Journal of Operational Research, 2017

[13] Y. R. Sagaert, E.-H. Aghezzaf, N. Kourentzes and B. Desmet, *Temporal Big Data for Tactical Sales Forecasting in the Tire Industry*, Interfaces, 2017

[14] A. A. Syntetos, Z. Babai, J. E. Boylan, S. Kolassa, K. Nikolopoulos, *Supply chain forecasting: Theory, practice, their gap and the future*, European Journal of Operational Research, Volume 252, Issue 1, 2016, Pages 1-26

[15] L. Tiacci and S. Saetta, *An approach to evaluate the impact of interaction between demand forecasting method and stock control policy on the inventory system performances*, International Journal of Production Economics, Volume 118, Issue 1, March 2009, Pages 63–71

[16] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society Series B (Methodological), 1996, Pages 267–288

[17] X. Wang and F. Petropoulos, *To select or to combine? The inventory performance of model and expert forecasts*, International Journal of Production Research, Volume 54, Issue 17, 2016, Pages 5271–5282

[18] G. Zotteri, M. Kalchschmidt and F. Caniato, *The impact of aggregation level on forecasting performance*, International Journal of Production Economics, Volume 93, 2005, Pages 479-491

[19] G. Zotteri *The impact of distributions of uncertain lumpy demand on inventories*, Production Planning & Control, Volume 11, Issue 1, 2000, Pages 32–43

Information and Communication Technologies for Manufacturing and Services

Smart city projects and improvement of public service supply chains

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Gautier AUBOURG[1,2,3]
[1]Direction Déléguée au Développement et à la Coordination des Programmes
Toulouse Métropole
Toulouse, France
gautier.aubourg@toulouse-metropole.fr

François GALASSO[2], Jacques LAMOTHE[2]
[2]Industrial Engineering Center, Université de Toulouse
Ecole des mines d'Albi-Carmaux
Albi, France
jacques.lamothe@mines-albi.fr, francois.galasso@iut-rodez.fr

Bernard GRABOT[3]
[3]Laboratoire Génie de Production, Université de Toulouse
INP-ENIT
Tarbes, France
bernard.grabot@enit.fr

Abstract—The involvement in smart city initiatives is a key factor of changes in modern cities. It drives the integration of ICT, the development of collaborative behaviours, and the involvement of people in the life of the city. This paper is interested in the impact of such initiatives on the performance of public services and on the supply chains that provide these services. Based on the example of the Toulouse metropolis in France, it shows that smart city initiatives are becoming a vector for instilling continuous improvement in public services.

Keywords—Interoperability and sustainable enterprise; Internet of Services and Service Science; Smart City; ICT.

I. Introduction

The mutation of the public service through innovative methods, or methods coming from other domains remains a major issue for modern societies.

Five factors [1] are at the origin of the mutations required for public service: 1) the increasing urbanization creates new challenges for managing a city community. 2) The satisfaction of users of public services depends on the role of the concerned people (users, citizens, consumers etc.) but requires more and more customization. 3) The governance pressure for decreasing the public budgets while maintaining service quality. 4) The interest of the citizens for a better involvement in the life of the city is a potential lever for public management and economic development. 5) The direct transposition of industrial improvement approaches (BPM and Lean) to public service has encountered a massive resistance.

With the development of supply chain management and business process management, the industrial sector has demonstrated its capability to improve processes using realistic and holistic approaches. But, in France, programs such as the Planning Programming Budgeting System (PPBS) introduced in the early 1960, "Rationalization of Budgetary Choices" (RCB) in 1970's, New Public Management (NPM) in the 1990's, "Organic Law Relating to the Laws of Finance" (LOLF) in 2000's regularly reminded the requirement of public services efficiency and the improvement of underlying processes but rarely resulted in important changes [2].

Recently, the involvement of citizens, users and companies throughout ICT (Information and Communication Technologies) has benefited to quality of life in cities through various experiences [3]. The concept of "Smart City" is implemented at an international level [4] and its label allows publicizing the realised progress [5] and comparing cities performances.

Now, the "smart city" initiatives are centred on the requirement to develop added value for citizens with realistic investments. They come back to the principles of service for citizens and create an opportunity to instil changes in public processes if coupled with appropriate business processes management approaches.

This paper is centred on that idea. It analyses, in the Toulouse French metropolis, the impacts of a portfolio of "smart city" initiatives on the improvement of public processes.

II. State of the Art

A. Industrial methods in Public Services

Some attempts for using industrial methods in public services have shown the interest of such approaches [6], [7] that nevertheless require to interpret and re-structure the expectations of the public service, and to create dedicated tools and methods allowing to better answer to the users' needs.

In 2007, the "Direction Générale de la Modernisation de l'État Français" (DGME) decided to spread continuous improvement principles in various public services (prefectures, hospitals, social security treasury, courts, etc.) through the General Revision of Public Politics ("Révision Générale des Politiques Publiques"; RGPP) [8]. Tools like

JIDOKA and POKA YOKE were tested in prefectures, Value Stream Mapping in courts, or 5S in operating rooms of hospitals [9]. The principle of a participative system has attracted the state agencies involved in these projects, but the integration of continuous improvement has often not survived the test phase, the performance assessment principles always associated with these methods having been badly perceived by the agents [10].

The concept of SCM appeared in the 90's. Supply chains are networks of organisations linked by upstream and downstream flows, producing value through products and services delivered to customers [11].

To our best knowledge, the hospital sector has been the first public service that studied and took in consideration approaches coming from Supply Chain Management (SCM) in its daily management [12], taking advantage of their specialisation on one type of service: health. Studies have shown the important benefit resulting from the introduction of the SCM dimension in the hospital sector (decrease up to 46% of the expenses, collaborative supplies, decrease of wastes and wasted times, improved availability of medicines, increased availability and productivity of critical resources) [13].

In 2015, South Africa published a report [14] on the integration of SCM in its public organisations. This report shows the benefits expected from the method: better planning, transparency, decreased costs and wastes, decrease of corruption. Nevertheless, it does not explain precisely the way SCM principles are integrated in public services but shows a success in their integration.

B. Specificities of the city Public Sector

A Public service can be defined as an action performed by a public authority in order to address needs of general interest [15], [16]. The concept of general interest, based on values shared by the members of a society, is a first difference with the industrial sector. In public services, the shared common values rely in the importance of a service provided for the general interest. While the industrial sector focuses on the financial aspect and market segments, public service is concerned with quality of result and equity [17]. The public service has the moral but also legal imperative to satisfy its users with justice and equity, giving a right of recourse to the users against the institutions [18]. In this context, the identification of users expectations and the definition of the services to provide are major challenges.

When considering a city or a metropolis, a broad spectrum of services has to be considered: transportation, energy, education, healthcare, buildings, physical infrastructures, food, water, public safety [19]. ISO37120 that addresses the sustainable assessment of a city identifies 16 domains of public activities: support for economy, education, environment, finance, fire and emergency response, governance, health, recreation, safety, shelter, water and solid wastes management, telecommunication, transportation, urban planning, sanitation. This diversity of public intervention creates a management complexity. Moreover, [20] underlines that these domains must not be considered as independent but “should be treated as an organic whole in a holistic way”.

Present metropolises are more and more in competition and are eager to capture up-to-date tendencies with a "marketing" objective until recently absent from public service. This may be done through the concepts of "Smart City" and sustainability making the city more attentive and attractive to individual needs [21], [22].

C. Smart City and sustainability frameworks

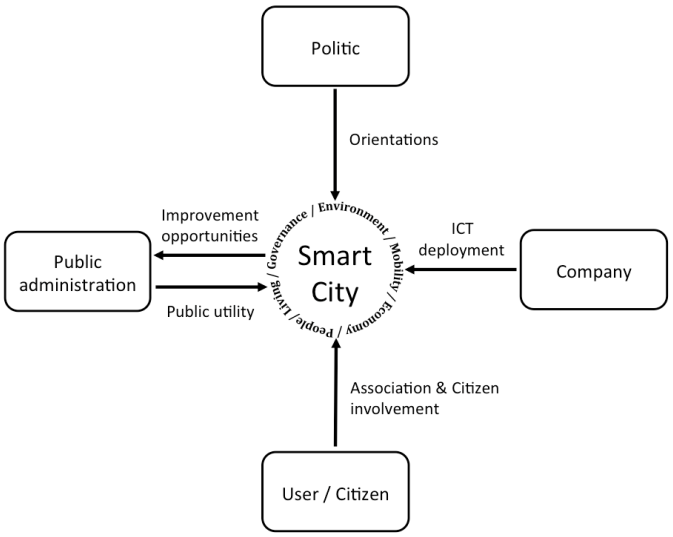
Cities face a rapid growth of their population [23] and have to fight against wastes for a better sustainability [24]. In that sense, the "lean", the sustainability and the supply chain management related concepts are consistent with the principles of "Smart Cities" and sustainability. Defining the concept of Smart City is nevertheless difficult. In an extensive review of the literature, [25] highlights the discrepancies between the definitions of Smart and Sustainable concepts, as the former is primarily focused on ICT but has evolved to integrate the latter. According to [25] a Smart City looks for a better quality of life through an enrolment of all the stakeholders thanks to ICTs.

As suggested by [26], the "Smart City" supposes also a participative model for involving the citizens in the urban life, like the continuous improvement model. In order to measure the improvements, smart indicators are set and usually grouped in six categories: (1) smart economy (e.g. Public expenditure on R&D, Unemployment rate…), (2) smart people (e.g. Foreign languages or computers skills…), (3) smart governance (e.g. Number of universities, e-Government use by individuals…), (4) smart mobility (e.g. Sustainable, innovative and safe public transportation…), (5) smart environment (e.g. efficient use of water and electricity…) and (6) smart living (e.g. Number of public libraries, Theatre and Cinema attendance…) [20]. They are different from those included in the usual norms used for quality management in industry (namely the ISO 9000 standards), and are closer to those related to corporate social responsibility (ISO 26000), environment management (ISO 14001) or energy management (ISO 50001). However, none of these smart standards covers all the needs of nowadays metropolises and a brand new set of indicators (not listed here) has been published in the ISO 37120 standard dealing with sustainable development of public communities [27]. Both smart and sustainable indicators provide a performance level but are not sufficient to define and position improvement actions within a metropolis organisation and regarding its stakeholders.

Moreover, like SCM, the "Smart City" principles promote transversality and interoperability between actors and could also play a role in industrial supply chains with the implementation of new technology, urban logistics and traffic regulation [28].

Figure 1 summarizes the state of the art of this paper. The stakeholders communicate through the Smart City located in the center. All the actors have the common ambitions to improve the six dimensions that characterizes the Smart City. The multiplicity of Smart City projects seems to have an impact on public service processes. This paper aims to confirm these impacts on the jurisdiction of the local government.

Fig. 1. Representation of Smart City impacts on public service processes



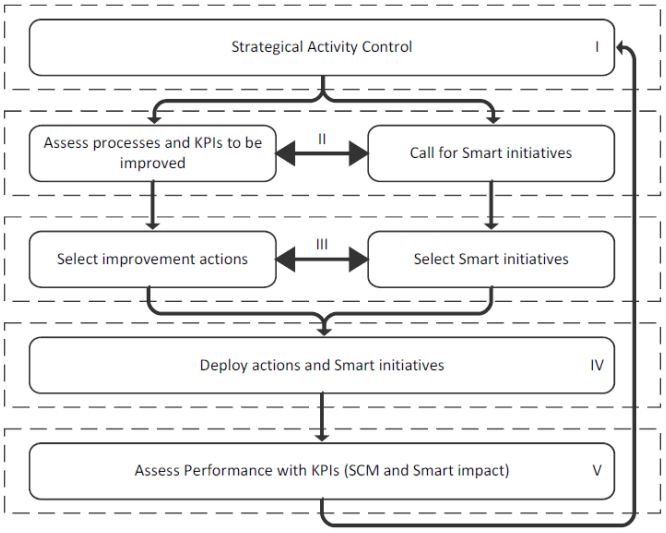
III. New management structure of public action

This section aims to develop a new methodology to control the public activities. This new method of control, depicted in Fig. 2, is based on continuous improvement tools, SCM, performance assessment and smart city initiatives. SCM improves collaboration between actors while continuous improvement tools and performance assessment drive the improvement process. Smart city initiatives consider the city as a global eco-system of which public activities are only a little part. Such initiatives first aim at improving the quality of life but they can also profit to facilitating public services supply chain.

The method is divided into five steps:

- First step, the political environment is in charge of providing the guidelines for the public services within a political program.
- Second step consists in the identification of the impacted processes thanks to a process analysis.
- In the third step, the processes are redesigned taking into account the new potential of smart initiatives acting as performance levers.
- Fourth step consists in deploying the solution with the project management tools.
- Fifth step is to evaluate the performance and compare it with the initial objectives concerning the strategic impulse given by the political vision.

Fig. 2. Main steps of the implementation method



In a previous communication, this methodology description and the examples focused on internal management of public activities (left side of Fig 2) [29]. This article will focus on the right part of Fig. 2, i.e. on the management of innovation through Smart City opportunities.

In the first paragraph, we will introduce the Smart City approach applied by the city of Toulouse. In addition, we will highlight the impacts of Smart City projects on the internal processes of the Public Service. In the following, we will present our summary table to characterize the public action in line with the Smart City approach and sustainability.

A. The impact of Smart City projects on public processes; Toulouse example

The metropolis of Toulouse has the particularity of having completely melted the precepts of Smart City and sustainability. The durability designs are also linked with green supply chain or lean to decrease non-added values.

The goal of Smart City in Toulouse is to build with citizens the smart city of tomorrow: more fluid, friendly, innovative, dynamic, attractive, responsible, sustainable. A city that takes advantage of the most efficient technologies to make life easier for citizens and make public action more effective. This approach is called "Toulouse l'Open Métropole" [30].

The “Open Métropole” is built around five ambitions: an adaptable, efficient and breathable city; Easier and smooth mobility; An international city and rooted in its roots; A city of good living, more warm and intergenerational; A more beautiful, clean and safe city.

To work on these ambitions fifteen groups of work have been implemented. From these fifteen projects, eight projects (demonstrators) were implemented in the framework of partnerships between public and private actors on a wide range of themes (energy, water, parking, autonomy of the elderly,

urban congestion, autonomous and connected vehicle, etc.). There also resulted in smaller projects called "experiments". The objective of the experiments is the emergence of new products and behaviours, with the help of local start-ups.

In this paper, we will focus on experiments. Since the launch of Smart City, eighteen experiments have been carried out on the territory. They all aimed at promoting or extending services to users in the Smart City's frame of mind. We noted that some have impacts on public processes that produce public services (see Tables I and II). For example, the following three experiments will support our purpose.

"MyFeelBack": It is a mobile application enabling users to declare road incidents, incivility or cleanliness problems. This experimentation changes the role of the citizen in the process of continuous improvement of the public action: generation of alerts, Assesment of the satisfaction rate (success or not of the answer to the alert) of the citizens. But also, the service is generated on demand and driven by the need: instead of organizing turns for surveying infrastructure.

"Qui Dit Miam" is a mobile application to increase communication around School meals of the City of Toulouse (menu composition, allergens, origin, push alert for menu changes). This new service has resulted (i) in an improvement in Customer Relationship Management by the utilities; (ii) a change of supplier following the putting forward of a non quality by the consumers; (iii) as well as the re-definition of the processes of delivery of goods (evaluation of organic bread by consumers has changed the supply circuit).

"Nacelles": It provides connected urban furniture that supports social networks and physical meeting between people at the level of a street, a district, a city. Such urban furniture have quickly demonstrated impacts on citizen consultation: it enables the design of open meeting, consult citizen for surveys in the street, make easier the online sharing of physical demonstrations, give a quick access of users to online services. The implementation allowed to increase the customer and citizen relationship management process. Moreover, the emergence of this tool brings continuous improvement in the civil service process by placing the citizen as an actor that can be easily consulted.

Of the eighteen experiments that were launched on the Toulouse metropolitan territory, twelve experiments demonstrated to have an impact on the processes of the public service [31]. We have concatenated all these experiments in two Tables (Table I & II). The first table informs about the different experiments with their impacts on the dimensions of Smart City for the city.

TABLE I. The impact of Smart city projects on the city

Projects	Smart city dimensions	Smart city impact
Spotminder	Smart People	Propose a digital social wall at 4 cultural events in order to involve those who are not present.
MyFeelBack	Smart Environment	Citizens declare problems on wastes and infrastructure and assess the quality of the answer. Improve the quality of the services rendered by the public service by analysing the level of user satisfaction.
Qui Dit Miam	Smart Environment & Smart Living	A mobile application to increase communication around School meals of the City of Toulouse (menu composition, allergens, origin, push alert for menu changes).
Helpiness	Smart Living & Smart Environment	Propose a mobile application to facilitate the exchanges and environment of the person assisted (friends, professional helpers, invoice, etc.)
Coovia	Smart Mobility	Propose an innovative digital carpooling solution for everyday trips and multimodal solutions in metropolitan France, with secure carpool points.
Parkisseo	Smart Mobility	Deploy parking sensors and develop a tool to optimize parking in Toulouse.
CityMéo	Smart Economy	Develop interactive digital communication to disseminate useful information to citizens via information screens: information about the city, traffic, weather, etc.
Ze Watt	Smart Mobility	Provide a global charging service for electric vehicles to companies, visitors, administrations and other users who wish to.
Nacelles	Smart Governance	Propose a connected urban furniture to favor the interaction with the citizens and the increase the collective consciousness of the territory
Oabike	Smart Living	Propose a mobile application specializing transport and safety in bicycle.
ThirtyOne	Smart Mobility	Test an experimental system of hybrid motorization for electrically assisted bicycles.

Table II depicts the experiments impacts on the processes of the public administration of Toulouse.

TABLE II. The impact of Smart city projects on public process

Projects	Public process impacted	Nature of impact
Spotminder	Cultural affairs department	Customer relationship management (statistics on the citizen touched by an event, publicity of the event)
MyFeelBack	Direction tranquillity and public safety	Integrate citizen in the continuous improvement, Statistical evaluation
Qui Dit Miam	Direction for Solidarity, Education and Sport	Customer relationship management Evaluation of the quality of service Fighting waste
Helpiness	Direction for Solidarity, Education and Sport	Improving the cooperation in the supply chain of the public partners Integrating the consumers into the supply chain
Coovia	Direction mobility and network	Improving the interoperability of mobility services, design of secure car pool points.
Parkisseo	Direction mobility and network	Instrumentation of production (sensor), make control turns on sensors alerts, change the pay process
CityMéo	Direction local Democratic	Enrichment of Customer relationship, reduce setup for communication change
Ze Watt	Direction	Change city employees movements to

	mobility and network	electric vehicles
Adveez	Direction for Solidarity, Education and Sport	Instrumentation of production (sensor) Customer relationship management
Nacelles	Direction local Democratic	Citizen Relationship Management Customer relationship management, Integrate citizen in the continuous improvement
Oabike	Direction mobility and network	Incident reporting on bicycle paths, Evaluation of the cycling transport plan, Employee relationship management
ThirtyOne	Direction mobility and network	A lever to increase the internal development of electric transport

The experiment being still ongoing, a precise assessment of the various impacts of projects is hard. Nevertheless, it can be noted in Table II that major impacts on the public activities can be underlined:

- Enriching the bilateral relationship with customers and citizens (CRM & CiRM effect): Spotminder, Qui Dit Miam, CityMéo, Adveez and Nacelles.
- User goes back directly to the recording of the problems and the evaluation of the response (citizen in the loop of continuous improvement): MyFeelBack, Qui Dit Miam, Nacelles and Oabike.
- Service driven by need instead of pushing, customers at the heart of supply chain services requirements: MyFeelBack, Qui Dit Miam, Nacelles and Oabike.
- Adaptation of physical infrastructures: Coovia and Ze Watt.
- Better sustainability of public processes: Qui Dit Miam, Helpiness, Coovia, Parkisseo, CityMéo, Ze Watt, Nacelles and ThirtyOne.
- Sensors equiping the system managed by the public process: Parkisseo and Adveez.

Now that we have demonstrated the impact of Smart City projects on the internal processes of the public service, we must succeed in managing all Smart City projects in parallel with the daily actions of the public service. The impacts of Smart City projects on internal processes can be divided into two main categories: Supply Chain Management and Continuous Improvement. Indeed, the concept of Smart City is based on ICT and multi-actor collaboration, it makes sense to find strong impacts on the management of the information system (CRM, sensors). The integration of citizens and the digitization of the logistic chain naturally bring continuous improvement to the public administration. But it also offers the capability to move to pulled management of the service.

B. Assessment of the public action and the impact on the smart city and sustainable approach

For the assesment of public action, the paper is based on the internal practices of the Toulouse administration. The survey of various practices in different directions from the Toulouse administration results in a distribution of evaluation of public action into four categories and sixteen dimensions. The dimensions associated with the categories are shown in Table III.

TABLE III. The dimensions of public service management in Toulouse

Category	Fields
Outcome	Innovation
	Effectiveness / Satisfaction
	Personnalization
	Environment
	Societal
	Adaptability
Project management	Relevance
	Transversality / Collaboration
	Action plan
	Quality
Efficiency	Economy, cost of service
	Resource performance
Constraint	Ethics
	Traceability
	Law / Justice / Equity
	Formalization

Table III summarizes the internal practices and emphasizes their consistency with the standards of the Smart City and sustainability. In the state of the art, we have identified dimensions that define Smart City and sustainability. In these dimensions we find the Smart Governance, Economy, Environment, etc. which are also present in the synthesis of our internal practices. Smart City approaches and sustainability incorporates dimensions such as mobility, health, education, etc. Which are competencies of the public service. The improvement of public performance thanks to our new public management method will have a direct impact on the performance of the Smart City approach and of sustainability within Toulouse.

The evaluation of projects on the impact of public service processes through our synthesis of internal practices will enrich Smart City's models and sustainability. Until now Smart City projects have been conducted as opportunities without identifying input to the public service. In the future the evaluation of our actions with our synthesis of practices will promote Smart City projects to improve the performance of the public.

IV. Conclusion

This insight and analysis demonstrates the impact of Smart City projects on the processes of public administrations. The analysis of the twelve experiments revealed several types of impacts on public service processes. These impacts can be categorized into three management processes: Customer Relationship Management, Supply Chain Management and Continuous Improvement. One next step will be to associate KPIs to these impacts in order to make them visible, measurable and more systematically used by public service managers.

This paper validates the model (Figure 2) of our suggested method for managing public service. The work on the identification of internal evaluation practices must be

consolidated and then shared by all the public actors and fully integrated into our approach.

The future perspectives of our research will be to apply the method to new projects of different sizes, to check the adaptablity of the method with respect to the projects. Using this method, we hope to successfully implement and adapt industrial methods (SCM and continuous improvement) within the Public Service.

References

[1] J.L. Silicani, "Livre blanc sur l'avenir de la fonction publique", Ministère du budget des comptes publics et de la fonction publique, 2007, pp. 57-77.

[2] A.M. Brocas, and F. von Lennep, "Contrepoint : de l'utilité (et des limites) du recours aux indicateurs pour le pilotage des politiques publiques", *Revue française des affaires sociales*, N° 1-2, 2010, pp. 149-160.

[3] Debnath, Ashim Kumar, Hoong Chor Chin, Md. Mazharul Haque, and Belinda Yuen. "A Methodological Framework for Benchmarking Smart Transport Cities." Cities 37 (April): 47–56, 2014, doi:10.1016/j.cities.2013.11.004.

[4] D. Washburn, and U. Sindhu,"Helping CIOs understand "Smart City" initiatives, Report by Forrester Research", Inc, 2010.

[5] V. Ihamäki, E. van Ooij, and S. van der Panne, "Green Public Procurement in the European Union and the Use of Eco-Labels", 2014, https://www.maastrichtuniversity.nl/sites/default/files/2014/Green_Public_Procurement_in_the_European_Union_and_the_Use_of_Eco-Labels.pdf

[6] F. S.B. Manuel, and J. Ramis-Pujol, "Implementation of Lean-Kaizen in the human resource service process", *Journal of Manufacturing Technology Management*, Vol. 21(3), 2010, pp. 388-410.

[7] Z. Radnor, and P. Walley, "Learning to Walk Before We Try to Run: Adapting Lean for the Public Sector", *Public Money & Management*, Vol. 28(1), 2008, pp. 13-20.

[8] F. Dreyfus, "La révision générale des politiques publiques, une conception néolibérale du rôle de l'Etat ? ", *Revue française d'administration publique*, Vol. 136, 2010, pp. 857-864.

[9] F. Parisot, "Gouvernance: Le lean, c'est bon pour le service public", *L'usine Nouvelle*, Vol. 3302, 2012.

[10] E. Buschor, "Performance management in the public sector: past, current and future trends", Tékhne, Vol. 11(1), 2013, pp. 4.

[11] M. L. Christopher, "Logistics and Supply Chain Management: Strategies for Reducing Costs and Improving Services", Financial Times, 1992.

[12] M, Chabrol, P. Féniès, M. Gourgand, and N. Tchernev, "Un environnement de modélisation pour le système d'information de la Supply Chain : application au Nouvel Hôpital d'Estaing", *Ingénierie des Systèmes d'Information (ISI)*, Vol. 11(1), 2006, pp. 137-162.

[13] S. Landry, and M. Beaulieu, "La logistique hospitalière : un remède aux maux du secteur de la santé? ", *Gestion*, Vol. 26, 2001, pp. 34-41.

[14] F. Lungisa, "Public Sector Supply Chain Management Review", *Report National Treasury Republic of South Africa*, online at: http://www.treasury.gov.za/publications/other/SCMR%20REPORT%202015.pdf, 2015.

[15] "Fiche Métier de La Fonction Publique : Mission, Statut, Salaires." 2017. Carrières Publiques. Accessed February 7. http://www.carrieres-publiques.com/metier-fonction-publique.

[16] A, Intaher Marcus, and J. A. Badenhorst-Weiss, "Managing and Controlling Public Sector Supply Chains". INTECH Open Access Publisher. http://cdn.intechweb.org/pdfs/15532.pdf, 2011.

[17] G.J. Van Helden, and C. Reichard, "A meta-review of public sector performance management research", Téknhe, Vol; 11, 2013, pp. 10-20.

[18] P. Warin, "La performance publique : attentes des usagers et réponses des ministères". *Politiques et management public*, Vol. 17(2), 1999, pp. 147-163.

[19] S. Dirks, and M. Keeling, "IBM Global Business Services Government Executive Report IBM Institute for Business Value A vision of smarter cities How cities can lead the way into a prosperous and sustainable future", Copyright IBM Corporation 2009 http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03227usen/GBE03227USEN.PDF

[20] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," Journal of Urban Technology, vol. 22, no. 1, 2015 pp. 3–21.

[21] R.E. Hall, B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Von Wimmersperg, "The vision of a Smart City", In: *Procs of the 2nd Intl life extension technology workshop*, Paris, France, Sept. 28, 2000.

[22] F. Bertossi, "Villes intelligentes, "smart", agiles: enjeux et strategies de collectivités française", *Ministère de l'Environnement, de l'Energie et de la Mer*, 2016, pp. 1 - 58.

[23] A. Young, "Inequality, the urban-rural gap and migration", *The Quartely Journal of Economics*, Vol. 128(4), 2013, pp. 1727-1785.

[24] I. Vojnovic, "Urban sustainability: research, politics, policy and practice", *Cities*, Vol. 41, 2014, pp. 30-44.

[25] S. Bibri, and J. Krogstie, "Smart sustainable cities of the future: An extensive interdisciplinary literature review", *Sustainable Cities and Society*, Volume 31, May 2017, pp 183–212.

[26] M. Bartenberger, and V. Grubmuller-Régent, "The enabling effects of open government data on collaborative governance in Smart City contexts", *JeDEM*, Vol. 6(1), 2014, pp. 36-48.

[27] ISO37120. "ISO 37120: Sustainable Development of Communities – Indicators for City Services and Quality of Life", *International Organization for Standardization*, First Edition, 2014-05-15, ISO37120:2014(E).

[28] B. Mattoni, F. Gugliermetti, and F. Bisegna, "A Multilevel Method to Assess and Design the Renovation and Integration of Smart Cities." Sustainable Cities and Society 15 (July): 105–19. doi:10.1016/j.scs.2014.12.002, 2015.

[29] G. Aubourg, F. Galasso, B. Grabot, and J. Lamothe, "Support to the Public Services Mutation through Continuous Improvement in a French Metropolis", APMS international conference advances in production management systems. September 3rd – 7th 2017, Hamburg, Germany, unpublished.

[30] Toulouse Métropole, "Toulouse, l'Open Métropole : stratégie smart city 2015-2020", 2015, http://documents.toulouse.fr/SmartCity/master-6/sources/indexPop.htm.

[31] Toulouse Métropole, "Toulouse, l'Open Métropole : bilan smart city 2016", 2016, http://www.toulouse-metropole.fr/documents/10180/13170594/OpenMetropole_Bilan2016.pdf/4c1fa599-fa14-47e3-a7ec-0e702f0d7a50.

Concept and Possible Application of an Automated Framework to Influence Production Dispatch Based on Supply Chain Events

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Dennis Bauer, Florian Maier, Thomas Bauernhansl
Fraunhofer-Institute for manufacturing engineering and automation
Stuttgart, Germany
dennis.bauer@ipa.fraunhofer.de

Bernd Waschneck
Graduate School advanced Manufacturing Engineering (GSaME) – University of Stuttgart
Stuttgart, Germany
bernd.waschneck@gsame.uni-stuttgart.de

Thomas Ponsignon, Daniel Gürster
Infineon Technologies AG
85579 Neubiberg, Germany
thomas.ponsignon@infineon.com

Bernhard Oberegger, Andreas Felsberger, Gerald Reiner
University of Klagenfurt
9020 Klagenfurt, Austria
bernhard.oberegger@aau.at

Abstract— We present in this paper an automated framework for linking supply chain events to production management. While events in the supply chain are detected using statistical process control, the control loop to production is closed using ideas inspired from run-to-run control. The proposed framework is expected to offer a higher level of flexibility with respect to demand changes. This results in increased on-time delivery for customers. While the semiconductor industry faces the challenge of volatile markets, long cycle times and short product life cycles, systems for demand planning and production management are nowadays often decoupled. Therefore, the proposed framework will be applied to semiconductor manufacturing to link demand planning with production management.

Keywords—Statistical Process Control; Advanced Process Control; Automated Decision Making; Advanced Dispatch Control

I. Introduction

In a volatile market with short product lifecycles but long development and production times, an increase in flexibility to meet the customer demands is a significant competitive advantage. To avoid large inventories and possible write-offs, forecasts, production planning and production control need to be closely linked and well aligned. In Industry 4.0, this concept is called horizontal integration [1]. It ensures a high transparency, fast reaction times and a cross-system optimization.

The semiconductor industry faces the challenges of long cycle times up to several months, short product life cycles and a volatile market. As early as 2001, the consortium of the International Technology Roadmap for Semiconductors described their vision of Factory Integration [2]. Today, this vision is extended to a global virtual factory, which links the production network with the supply chain planning processes [3].

In this paper we describe an automated framework for linking supply chain events, e.g., changes in demand planning, to production management. Today these systems are often decoupled: after the demand is released to production, it enters into a freeze fence in the planning systems and it cannot be changed. The production commits to a delivery date with respect to the cycle time. During the freeze fence period, demand changes have no impact on production.

Our new integrated dispatching approach addresses this challenge to increase the flexibility of companies acting in volatile markets. Supply chain events are automatically detected via Statistical Process Control (SPC) adapted to supply chain processes. These events are then incorporated into the Manufacturing Execution System, specifically in the dispatching process in complex job shop environments.

In the next section we will review literature on SPC in Supply chain applications as well as the coupling between production control and the planning process. In Section III, our proposed approach is introduced. In Section IV, the usage of these events in factory scheduling is described. In Section V, the concept is applied to the planning and production system of a large semiconductor manufacturing company (Infineon), which faces the described challenges.

II. Process monitoring and control

A. State of research in SPC

We provide an overview about the current state of research in the field of process monitoring and SPC, i.e., we identify patterns, themes and issues in available academic literature [4]. This approach allows us to identify relevant research work based on pre-defined criteria [5]. The development of search strings and the determination of included criteria is an essential

The work has been performed in the project Power Semiconductor and Electronics Manufacturing 4.0 (SemI40), under grant agreement No 692466. The project is co-funded by grants from Austria, Germany, Italy, France, Portugal and - Electronic Component Systems for European Leadership Joint Undertaking (ECSEL JU).

part of a literature review. The metadatabase Web of Science (Thomson Reuters) is scanned for scientific papers that fulfil the criteria for inclusion (cf. Fig. 1).

The results of the search string applied in the metadatabase Web of Science resulted in 107 academic papers which fulfill the pre-defined search criteria. Finally, the total number of 107 papers could be reduced by certain refinement criteria such as several Science Citation Index indicators, thematically fitting research areas and a timespan from 1990-2016, to a final sample of 28 papers.

In the manufacturing field, SPC is a well-known and established concept for monitoring the performance of the system and the quality of the products. Originally developed by Walter Shewart in the 1930s [6], the concept is defined as a continuous monitoring of a process to determine whether or not it is in control. It is assumed that a process has natural variations, which have a common source, and variations caused by special sources, which should not occur in the process. Common variations can be deemed to be in-control since they are process-specific and produce a stable distribution, whereas special variations cause out-of-control signals which occur sporadically and unpredictably. Special causes can either be single outliers or permanent deviations of the process as a whole.

In general, the literature review shows that SPC has emerged from the manufacturing field and has been extended to different kinds of applications. From the health industry [7, 8] through commercial industries [9] and environmental applications [10] to supply chain processes [11–13]. Moreover, another essential part of these concepts is the utilization of control charts in inventory control, due to the awareness that inventory replenishment policies are considered one of the major causes of the bullwhip effect in supply chains [14–20]. Most of these approaches have the fundamental assumption that the observed variables are normally distributed. To obtain these assumptions, various proceedings can be found in the literature. The basic approach is the simple assumption of normality of the used data as found in [8, 9, 13]. Only very few industrial processes are actually perfectly normally distributed [21]. A more precise method to verify the assumption of normality was used by [7, 11] by considering the Central Limit Theorem. Reference [10] used an advanced approach and transformed the data with a logarithmic transformation if the observed variables lacked normality. An improved representation of the underlying distribution in reality was done by [12]. On the other hand, due to the reason of increasing quality standards and the fact that in reality only a few processes are normally distributed, non-parametric concepts have gained more attention in the recent years [22–25].

Database	Search string
Web of Science (Thomson Reuters)	*TS= (statistical process control OR spc) AND TS= (control chart AND chart*) AND TS= (perform* OR monitor* OR inventory*) AND TS= (forecast* OR demand* OR supply chain)*

Fig. 1: Metadatabase search string.

B. SPC vs APC

Traditionally, there have been two distinct approaches to process control. On the one hand, there is the already described SPC, originating from the parts industry where the process output is monitored to detect out-of-control processes. On the other hand, there is advanced process control (APC), sometimes also referred to as engineering process control (EPC), where important process variables are measured to incorporate a feedback loop to control the process. The feedback loop uses a mathematical model to adjust process inputs based on these measurements. Therefore, the mathematical model needs to link variability in the output variable to an input control variable. APC originated from the process industry [26–28]. While both SPC and APC aim at reducing the variability and therefore improving the process quality, they seek to achieve these goals in different ways. Partially, this is a result of from their origination of different industries [29, 30]. Table 1 compares SPC and APC.

Some industrial processes, such as those in semiconductor manufacturing, are characterized by certain aspects of the parts industry and others of the process industry. Within these processes, the sharply drawn lines dividing the parts industry and the process industry have begun to disappear [27]. Addressing these special needs, the concept of run-to-run (R2R) control emerged. The R2R approach combines techniques from both SPC and APC to minimize process drift, shift and variability. Achieving this, the product recipe with respect to a particular machine and process is modified at an ex-situ process-to-process-level [28, 31]. Therefore, SPC and APC complement each other very effectively [30]. However, there are also other approaches in R2R than relying solely on SPC. Research aimed in this direction includes works concerning neural networks, expert systems and fuzzy logic controllers [31].

TABLE 1: Comparison of SPC and APC [31], enhanced by [27, 29].

	SPC	**APC**
Goal	*Minimize variability*	
Concept	*Minimize variability by detecting and removing process upset*	*Minimize variability by adjusting the process to counteract process upset*
Function	*Monitor the process*	*Control the process*
Application	*Expect stationary process*	*Expect continuous process drift*
Automated Process Adjustment	*None*	*Semi-automated to automated*
Reaction to	*Statistically significant changes*	*Continuous changes*
Implementation	*Downward*	*Upward*
Results	*Process improvement*	*Process optimization*

III. Detection of supply chain events

The digitalization and globalization of supply chains lead to comparable levels of complexity and innovation speed in the incorporated processes and the manufacturing process alone [32]. In contrast to manufacturing processes, supply chain processes do not necessarily follow a certain, simply assessable, distribution. A major difference is the influence of human interactions in the processes and specific supply chain challenges like the bullwhip effect. Further differences of supply chain and simple manufacturing processes are compared in Table 2. Consequently, a simple transfer of SPC from the manufacturing level to supply chain processes is not possible since these specific characteristics have to be considered and integrated into the monitoring concept. Especially the semiconductor manufacturing, with its complex processes and the accompanying challenge of the long cycle times compared to the short life-cycle of the products containing semiconductors, necessitates an accurate monitoring system for all kinds of different supply chain processes.

A control chart design has been formulated that is able to incorporate the specifics of supply chain processes. The proposed control chart fulfils the following requirements: (i) The basis variable for the control limits and the basis variable for the centerline are differentiated; (ii) The control limits are not placed based on a predefined distribution of the observed variable, but based on a percentage of the cumulative distribution function assessed from a second variable that reflects the individual human behavior in the observed process; (iii) The upper and lower control limits do not have to be placed at the same distance to the centerline to take asymmetric planning behaviors into account; and (iv) the control chart incorporates a flexible centerline whose position may vary over time – since supply chain processes often do not have an overall average value, which could be represented by the centerline, the centerline is adaptable to the underlying planning situation of the observed process. An example of the proposed control chart design is depicted in Fig. 2.

Supply chain processes can be set up according to the Supply Chain Operations Reference (SCOR) model [33]. This reference model is an inter-industrial approach, which enables organizations to describe, analyze and improve their processes within a standardized framework and provides a common base for benchmarking between companies [34]. The six main processes are Plan, Make, Source, Deliver, Return and Enable. The standardized definitions of these processes enable the description of complex as well as simple supply chains on a common basis and can and should be applied to all links of a given supply chain [33]. Particularly with regard to the complexity of the semiconductor industry, the application of this reference model is useful to obtain standardized processes and a reasonable set up of the global supply chain. Breaking

TABLE 2: Comparison of the data structures.

Manufacturing	Supply Chain Processes
Constant Behavior	Trends, Seasonal Effects
Taking samples possible at every time	Relies on database setup and data granularity
Same process step is repeated very often	Comparability: e.g. cycle time of given products
Univariate data	Multivariate data

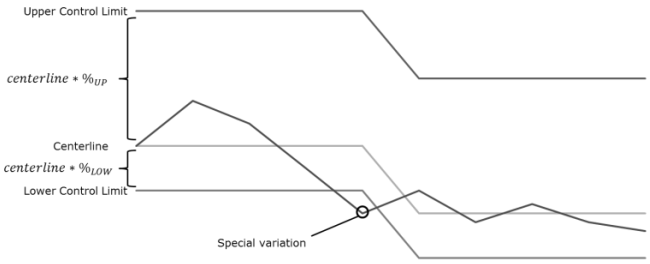


Fig. 2. Example of a control chart.

down the Plan process according to the SCOR model, five sub-processes, which we will exemplify using Infineon Technologies AG in the following section, can be determined as depicted in Fig. 3.

The proposed concept monitors the processes within these areas and triggers corrective actions via events in another planning field. This accelerates the reaction to occurring events and supports early warnings for the responsible planners. In a further step, the information is used to trigger automated intervention in the manufacturing execution system. This supports the planners and reduces the workload with better performance at the same time.

IV. Automated production dispatch influence

Due to criteria such as re-entrant production flows or sequence-dependent setup times semiconductor manufacturing can be modeled as a complex job shop. Within complex job shops, scheduling and dispatching highly impacts the performance of the manufacturing process. Scheduling describes the planning process of allocating resources to tasks over given time periods with the goal of optimizing one or more objectives. In the context of job shop scheduling, jobs will be assigned to machines for a specific time period in the future, where the definition of future depends on the planning horizon. This procedure aims at an effective and efficient use of the available resources [35, 36]. Dispatching, on the other hand, is a just-in-time decision. When a machine becomes available, the dispatching algorithm assigns the job with the highest priority from a queue of waiting jobs to the machine. The order of jobs may be determined by schedules or by dispatching rules [36]. Nowadays, rule-driven dispatching is still the dominant shop floor control method for semiconductor manufacturing [35, 36]. Advantages of dispatching rules are the real-time capability due to a small calculating effort and the fact, that results are easily comprehensible for operators [36]. In theory as well as practical application there are a lot of dispatching rules. Among them are popular ones such as First In First Out (FIFO), Shortest Processing Time (SPT) or

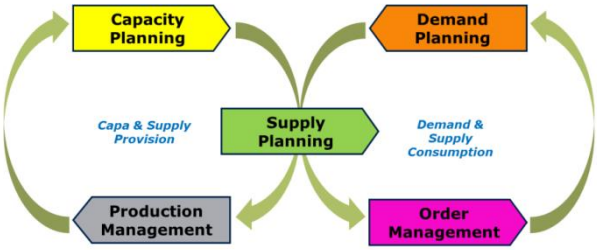


Fig. 3. Supply chain planning sub-processes according to the SCOR model.

Operation Due Date (ODD) which are described very well in the literature [36, 37].

Supply chain and logistics processes have reached a comparable level of complexity and innovation speed to the manufacturing processes themselves (cf. Section II). Therefore, they need to be closely monitored as well, which has been achieved by the concept proposed in Section III. In addition, ideas described by Industry 4.0 aim at decentralized and automated decision making [38]. This motivates the approach of advanced dispatch control (ADC) in transferring R2R control from manufacturing to dispatching as the interface between supply chain and manufacturing. Thereby, planning on the supply chain level and corrective actions on the manufacturing level are linked. The basis of this concept is an event-driven approach and a respective architecture (cf. Fig. 4). Events from supply chain SPC serve as triggers for corrective actions. These events will be further analyzed and enriched with additional data. Based on this data, corrective actions will be made automatically to influence job order by dispatching.

Key components of the ADC architecture, which are integrated by a manufacturing service bus (MSB) [39], are:

- Analyzer: The analyzer receives events from SPC and enriches them with additional data from manufacturing IT systems such as ERP or MES. While events already need to contain some information such as data on the violation of a limit or an estimated date of action, manufacturing IT systems could provide additional data. Among them are data on priority corridors or current work-in-process of the respective product.
- Rules Engine: Using the data supplied by the analyzer the rules engine decides on corrective actions. Corrective actions mean prioritization or de-prioritization of product groups via dispatching. Therefore, the rules engine correlates with the controller in the R2R concept. While R2R in manufacturing aims at minimizing variability in processes, ADC aims at grading output which is to some extent similar.

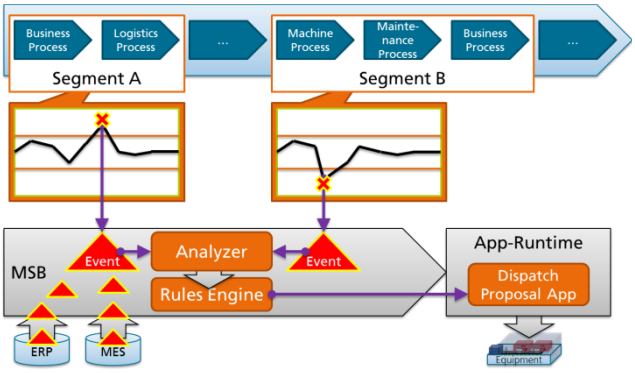


Fig. 4. Concept of ADC.

V. Demand Monitoring & Benefits

The planning landscape at Infineon Technologies AG is built according the SCOR model and specified for each sub-process as depicted in Fig. 5. Further explanations can be found in [40].

The developed concept is applied inside this planning landscape. The monitoring approach is used in Demand Planning where critical events are identified by means of the introduced control charts. For the monitored variable, the forecasts regarding the expected order quantities are used, which are extracted right before entering the planning of Operational Demand. This external information, sent by the customers, is one of the main inputs for the planning processes and therefore a major influencing factor in the system. Critical events in this context could for example be changes in the demand quantity or changes in the desired delivery date by the customer. Another critical event would be the appearance of a totally new order, which would make an adaption of the planned scenario necessary. The detection of a relevant deviation triggers a process in the system for Production Management which automatically adjusts the production to the new situation by changing parameters via the ADC. This will mainly be performed by applying short term decisions such as changing the priority of specific lots or changing the current flow factor of the production of a specific product.

Fig. 6 depicts the data exchange between the monitoring concept, which detects events like described above, and the ADC system. The triggered exchange should contain all relevant information necessary for the application of automated rules in the short-term production planning. Especially the knowledge about the product and the customer is relevant for the importance of the adaptability of the production.

The applied concept enables a reduction of the impact of unpredictable events on the supply chain‘s performance. Deviations to the as-is planning situation are recognized as soon as they occur, while the false alarm rate is reduced due to the consideration of the human behavior in the control charts. This leads to a more flexible and automated planning and production process and an adequate reaction to demand changes and customer requests. As a further consequence, the inventory levels can be reduced, since the safety stocks can be lowered due to the more accurate planning in advance.

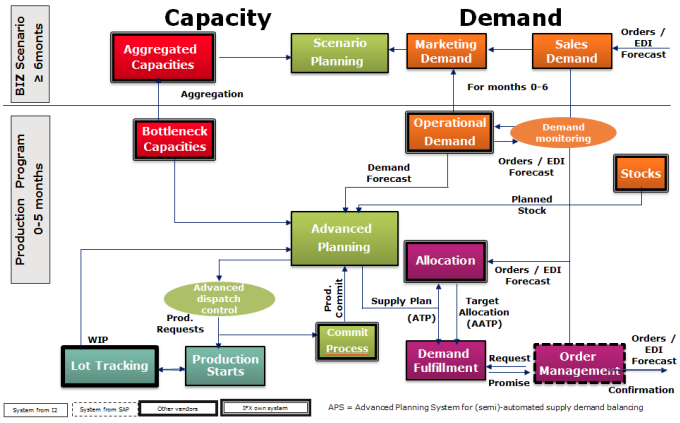


Fig. 5. Planning Landscape at Infineon Technologies AG.

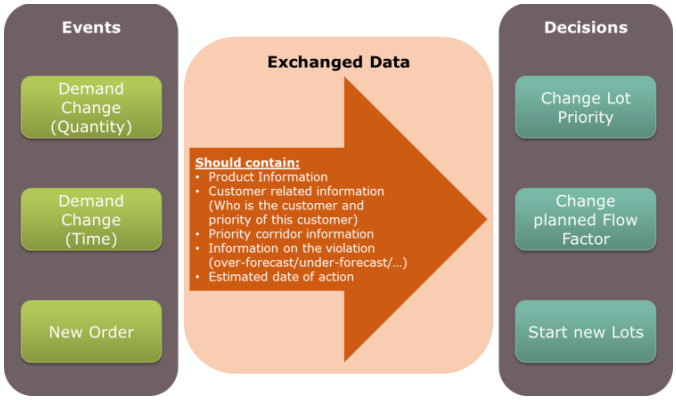


Fig. 6. Data exchange between monitoring and ADC systems.

VI. Conclusion and Outlook

In this paper we present a framework for automatically detecting supply chain events and acting on them by influencing the production schedules. The first step builds on SPC techniques which were transferred and applied to supply chain processes. The second step – the adjustment of production dispatching – implements a complete APC system. This results in a closed control loop as current work-in-process levels of the factories are considered in supply chain planning. With this approach, the IT systems used for supply chain planning and MES are closely integrated. We expect the system to offer a higher level of flexibility with respect to demand changes. This results in an increased on-time delivery rate for customers. The practicality and benefits have been shown in a feasibility analysis at Infineon Technologies AG. SPC for supply chain will be tested as notifications for the supply chain planners. With their feedback, the parameters of the SPC can be fine-tuned. In the next step, we want to implement the complete framework in a simulation model containing supply chain planning and production elements and quantify the benefits.

In conclusion, the framework supports the vision of a fully integrated factory as developed in Industry 4.0 by connecting supply chain planning and production with a closed feedback loop.

Acknowledgment

We thank all participants in Tasks 2.2.2 and 2.2.4 in the EU-ECSEL-Project SemI40 for fruitful discussions: Norman Schmitz, Felix Rotthowe and Dr. Roland Krumm (ELMOS Semiconductor AG), Dr. Andreas Kyek (Infineon Technologies AG) as well as Walter Laure and Herbert Forstner (Infineon Technologies Austria AG).

References

[1] J. Bischoff, Exploraton of Industry 4.0 application potential in SMEs (Erschließen der Potenziale der Anwendung von Industrie 4.0 im Mittelstand). [Online] Available: https://www.agiplan.de/fileadmin/pdf_dokumente/Studie_Industrie_40_BMWi_gesamt.pdf. Accessed on: Mar. 07 2017.

[2] ITRS, International Technology Roadmap for Semiconductors. [Online] Available: https://www.dropbox.com/sh/vxigcu48nfe4t81/AACuMvZEh1peQ6G8miYFCSEJa?dl=0&preview=1_Executive+Summary.pdf. Accessed on: Mar. 07 2017.

[3] K. Gruber, C. Pophal, and H. Ehm, “Infineon: Integrated Supply Chain Architecture to Support Sustainability,” in CSR, Sustainability, Ethics & Governance, Sustainable value chain management: Delivering sustainability through the core business, M. D'heur, Ed., Cham: Springer International Publishing, 2015, pp. 381–391.

[4] S. Seuring and M. Müller, “From a literature review to a conceptual framework for sustainable supply chain management,” Journal of Cleaner Production, vol. 16, no. 15, pp. 1699–1710, 2008.

[5] N. Kunz and G. Reiner, er,meta G. Reine of humanitarian logistics research,esJrnl Hum Log and Sup Chn Mnage, vol. 2, no. 2, pp. 116–147, 2012.

[6] W. A. Shewhart, Economic control of quality of manufactured product. New York: van Nostrand, 1931.

[7] C. D. Callahan and D. L. Griffen, “Advanced statistics: applying statistical process control techniques to emergency medicine: a primer for providers,” Academic emergency medicine, vol. 10, no. 8, pp. 883–890, 2003.

[8] D. T. Boe, W. Riley, and H. Parsons, “Improving service delivery in a county health department WIC clinic: An application of statistical process control techniques,” (eng), American journal of public health, vol. 99, no. 9, pp. 1619–1625, 2009.

[9] A. Scordaki and S. Psarakis, “Statistical Process Control in Service Industry An Application with Real Data in a Commercial Company,” Proceedings of the 7th Hell. Eur. Conf. Comput. Math. Its Appl., 2005.

[10] C. J. Corbett and J.-N. Pan, “Evaluating environmental performance using statistical process control techniques,” European Journal of Operational Research, vol. 139, no. 1, pp. 68–83, 2002.

[11] S. A. Darestani, M. Y. Ismail, N. Ismail, and R. Mohd Yusuff, “An investigation on supplier delivery performance by using SPC techniques for automotive industry,” Journal of American Science, vol. 6, no. 4, pp. 5–11, 2010.

[12] M. Bourne, C. Morgan, and A. Dewhurst, “Using SPC to measure a national supermarket chain's suppliers' performance,” Int Jrnl of Op & Prod Mnagemnt, vol. 27, no. 8, pp. 874–900, 2007.

[13] F. Costantino, G. Di Gravio, A. Shaban, and M. Tronci, “Inventory Control System Based on Control Charts to Improve Supply Chain Performances,” Int. j. simul. model., vol. 13, no. 3, pp. 263–275, 2014.

[14] C. A. Watts, C. K. Hahn, and B. Sohn, “Monitoring the Performance of a Reorder Point System: A Control Chart Approach,” Int Jrnl of Op & Prod Mnagemnt, vol. 14, no. 2, pp. 51–61, 1994.

[15] H.-C. Pfohl, O. Cullmann, and W. Stolzle, “Inventory management with statistical process control: Simulation and evaluation,” Journal of Business Logistics, vol. 20, no. 1, p. 101, 1999.

[16] H. T. Lee and J. C. Wu, “A study on inventory replenishment policies in a two-echelon supply chain system,” Computers & Industrial Engineering, vol. 51, no. 2, pp. 257–263, 2006.

[17] J.-C. Cheng and C.-Y. Chou, “A real-time inventory decision system using Western Electric run rules and ARMA control chart,” Expert Systems with Applications, vol. 35, no. 3, pp. 755–761, 2008.
[18] T. Kurano, K. N. McKay, and G. W. Black, “Proactive inventory policy intervention to mitigate risk within cooperative supply chains,” 10.5267/j.ijiec, vol. 5, no. 2, pp. 249–264, 2014.
[19] F. Costantino, G. Di Gravio, A. Shaban, and M. Tronci, “A real-time SPC inventory replenishment system to improve supply chain performances,” Expert Systems with Applications, vol. 42, no. 3, pp. 1665–1683, 2015.
[20] F. Costantino, G. Di Gravio, A. Shaban, and M. Tronci, “SPC forecasting system to mitigate the bullwhip effect and inventory variance in supply chains,” Expert Systems with Applications, vol. 42, no. 3, pp. 1773–1787, 2015.
[21] P. Castagliola and F. Tsung, “Autocorrelated SPC for Non-Normal Situations,” Qual. Reliab. Engng. Int., vol. 21, no. 2, pp. 131–161, 2005.
[22] C. Li, A. Mukherjee, Q. Su, and M. Xie, “Optimal design of a distribution-free quality control scheme for cost-efficient monitoring of unknown location,” International Journal of Production Research, vol. 54, no. 24, pp. 7259–7273, 2016.
[23] A. Mukherjee and S. Chakraborti, “A Distribution-free Control Chart for the Joint Monitoring of Location and Scale,” Qual. Reliab. Engng. Int., vol. 28, no. 3, pp. 335–352, 2012.
[24] Q. Chen, R. J. Wynne, P. Goulding, and D. Sandoz, “The application of principal component analysis and kernel density estimation to enhance process monitoring,” Control Engineering Practice, vol. 8, no. 5, pp. 531–543, 2000.
[25] N. Chen, X. Zi, and C. Zou, “A Distribution-Free Multivariate Control Chart,” Technometrics, vol. 58, no. 4, pp. 448–459, 2016.
[26] T. F. Edgar et al., “Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities,” Automatica, vol. 36, no. 11, pp. 1567–1603, 2000.
[27] G. Box and T. Kramer, “Statistical Process Monitoring and Feedback Adjustment: A Discussion,” Technometrics, vol. 34, no. 3, p. 251, 1992.
[28] J. Campbell, “Run-to-Run Control of Photolithography Processes,” Yield Management Solutions, vol. 2000, no. 3, pp. 65–68, 2000.
[29] D. C. Montgomery, B. Keats, G. C. Runger, and W. S. Messina, “Integrating statistical process control and engineering process control,” Journal of quality Technology, vol. 26, no. 2, pp. 79–87, 1994.
[30] K. Nishina, M. Higashide, H. Kawamura, and N. Ishii, “On the Integration of SPC and APC: APC Can Be a Convenient Support for SPC,” in Frontiers in Statistical Quality Control 10, H.-J. Lenz, W. Schmid, and P.-T. Wilrich, Eds., Heidelberg: Physica-Verlag, 2012, 109–119.
[31] J. Moyne, H. Etemad, and M. Elta, “A run-to-run control framework for VLSI manufacturing,” Microelectronic Processing '93, pp. 379–390, 1993.
[32] H. Ehm, T. Ponsignon, and T. Kaufmann, “The global supply chain is our new fab: Integration and automation challenges,” in 2011 22nd Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC), 2011, pp. 1–6.
[33] APICS Supply Chain Council, Supply Chain Operations Reference Model: Quick Reference Guide. [Online] Available: http://www.apics.org/docs/default-source/scc-non-research/apicsscc_scor_quick_reference_guide.pdf. Accessed on: Mar. 07 2017.
[34] H. Werner, Supply Chain Management: Basics, Strategies, Tools and Controlling (Grundlagen, Strategien, Instrumente und Controlling), 5th ed. Wiesbaden: Springer, 2013.
[35] L. Mönch, J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose, “A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations,” J Sched, vol. 14, no. 6, pp. 583–599, 2011.
[36] B. Waschneck, T. Bauernhansl, T. Altenmüller, and A. Kyek, “Production Scheduling in Complex Job Shops from an Industry 4.0 Perspective: A Review and Challenges in the Semiconductor Industry,” in i-KNOW, 2016.
[37] A. K. Gupta and A. I. Sivakumar, “Job shop scheduling techniques in semiconductor manufacturing,” Int J Adv Manuf Technol, vol. 27, no. 11-12, pp. 1163–1169, 2006.
[38] T. Bauernhansl, “Fourth industrial revolution: Road to a value-creating production paradigm (Die Vierte Industrielle Revolution: Der Weg in ein wertschaffendes Produktionsparadigma),” in Springer Reference Technik, vol. 4, Handbook Industry 4.0: Basics (Handbuch Industrie 4.0: Allgemeine Grundlagen), B. Vogel-Heuser, T. Bauernhansl, and M. t. Hompel, Eds., 2nd ed., Berlin, Heidelberg: Springer Vieweg, 2017, pp. 1–31.
[39] D. Bauer, D. Stock, and T. Bauernhansl, “Movement towards service-orientation and app-orientation in manufacturing IT,” Procedia CIRP, no. 62, pp. 199–204, 2017.
[40] German Electrical and Electronic Manufacturers’ Association (ZVEI), *Guideline Supply Chain Management in Electronics Manufacturing.* [Online] Available: https://www.zvei.org/en/press-media/publications/guideline-supply-chain-management-in-electronics-manufacturing. Accessed on: Mar. 23 2017.

Some elements regarding the relevance of using a fuzzy performance measurement framework for industrial visual management.

(presented at the 7th IESM Conference, October 11 - 13, 2017, Saarbrücken, Germany)

Lamia Berrah, Laurent Foulloy, Vincent Clivillé
LISTIC Polytech Annecy-Chambéry,
BP 80439 Annecy le Vieux Cedex, France
lamia.berrah@univ-smb.fr, laurent.foulloy@univ-smb.fr,
vincent.cliville@univ-smb.fr

Abstract- This study focuses on some fuzzy proposals for expressing industrial performance under symbolic signs that can be, on the one hand, easily understood by human operators as well by robots and, that can, on the other hand, identify the result of the handling of the industrial data that may characterise the way of manufacturing 4.0. After a state of the art that briefly describes the works that have been proposed in the literature in the industrial performance measurement field, we present the major ideas of the formal fuzzy framework that we have progressively defined over the last twenty years. In particular, by handling both numerical and linguistic terms is shown *via* the Zadeh "*precisiation*" concept, temporal performances are described, on the basis of the qualification of temporal trajectories performances. Lastly, symbolic fuzzification of conventional performance measurements is presented, leading to the translation of the obtained results to emoticons and other symbols that is typical of the currently practiced visual management. Some concluding remarks are proposed, highlighting the potential use of the presented development to the context 4.0.

Keywords: Performance measurement - Fuzzy handling - Visual management - Emoticons - industry 4.0.

I. Introduction

Currently, one can observe an extensive mutation of the industrial organisations, from standardised automated manufacturing systems, which is described as the industry 3.0, to "digital" manufacturing systems, which is codified as the industry 4.0 [1]. The movement towards the industry 4.0 thus leads practitioners and researchers to think about revisiting their way of designing and manufacturing the products, managing the teams and organising the business, and also handling the company information system… and in particular, the corresponding performance measurement sub-system.

This study deals with the evolution of the *performance measurement framework 3.0* into a *performance measurement framework 4.0*. The first framework deals with the automatic era and is essentially based on the cost- quality- delay performance criteria. Given that what we call the framework 4.0 remains imprecisely known and not completely defined, we choose here to analyse the framework 3.0 in order to highlight the possible extensions of some functionalities of it to the framework 4.0. This analysis will be made on the basis of the description of what has been proposed with regards to the development of the so-called PI's and PMS's, namely Performance Indicators and Performance Measurement Systems.

Indeed, handling the performance 3.0 concept has led to numerous analyses. Models and frameworks have been developed, dealing with one or more than one specification(s) of the performance paradigm. Some researchers have focused on the performance *definition*, other researchers have focused on the performance *measurement* mechanism, while some of them have been interested in the performance *improvement* dimension. Models dealing with performance definition were subscribing to conceptual and methodological aspects. Performance measurement frameworks focused on rendering coherent the set of performance indicators that are associated with the strategy of companies. Finally, works dealing with the performance improvement aspect were more related to the processes control and reactive reactions requirements.

Moreover, beyond the performance measurement analysis, we choose to focus on the visual management opportunities for expressing performance measurements under symbolic formats. Our idea is to look to define a formal way of providing visual scorecards from computed performance expressions. Such symbols offer the advantages of being easily understood by anyone, in a single and immediate manner, in the current context, as well as in the future context. Even if visual management is well-practiced nowadays, its integration in PMS frameworks remains implicitly done.

This paper is organised as follows. The second section will be devoted to a brief state of the art on the performance works, by particularly focusing on performance measurement points of view. We summarise in the third section our proposed fuzzy framework, by distinguishing the elementary performance expression mechanism and the aggregated performance expression mechanism. Throughout this analysis, some thoughts about the use of the mentioned points in the context 4.0 are proposed. The fourth section will be devoted to some recent extensions of the framework that handle temporal considerations and visual management principles. Before concluding, a focus on the elaboration of visual scorecards will be presented.

II. Historical considerations

From an epistemological point of view, the performance relies on the "perform" verb, which signifies "to carry out" or "to do" (an action) [2]. At the same time, the Collins dictionary defines the "performance" as any accomplishment as well as a reward that is associated with such an accomplishment. This explains that performance can be related to the obtained results

as well as to success or failure associated with them. Performance intrinsically becomes related to "something" that makes possible the carrying out of any action and the assessment of the obtained result possible. Performance becomes related to what is called "criterion". Objectives and Measurements will thus be the "parameters" of the action and its accomplishment. Objective is what we want and measurement is what we get. According to these considerations, a company is "performant" if it reaches its objectives, knowing that such objectives are coherent with the current industrial context.

One can distinguish between the Taylorian performance and the Post-Taylorian performance. Industrial performance has been defined in the Taylorian period (industry 2.0) as being related to the financial reports [3]. The major criteria were related to the productivity and the efficiency. The objectives were synonymous with maximising such quantities. Measurements were obtained by means of the so-called Taylorian *ratios* [4]. In the post-Taylorian period (industry 3.0), performance overcame the financial aspect, leading to discussions and works looking to integrate the "technical" dimension [5] [6], and more recently, a sustainability dimension [7]. In particular, in the current modern period, even if the industrial criteria remain the same, performance becomes quite different in the way the measurements are increasingly obtained in a digital way, knowing the emergence of the concepts that are inherent to "digital revolution" (the big-data, cloud, real-time, IoT (Internet of Things)…) [8].

A. *The Performance Indicator - PI*

Performance Indicators (PI's) which before were simple and constant have been studied in order to tackle financial criteria as well as technical criteria [9]. The PI's ware deployed throughout not only on the strategic decisional level but also on the tactical level and on the operational level. Handling such specifications has led the ABC/ABM (Activity Based Costing/ Activity Based Modeling) philosophies to go beyond the accounting approach and to propose methodologies for implementing such PI's [10] [11]. Standards have been developed such as the ENAPS (European Network for Advanced Performance Studies) [12], which has proposed PI lists. PI's for SCM (Supply Chain Management) [13] [14] and sustainability [15] have also been listed.

Dealing with the Bitton et *al.* "controllability" concept [16], the PI has been defined as being constituted of three parameters: the objective, the measurement and the criterion. Performance expression becomes more generally the result of a mathematical comparison of the objective and the measurement.

B. *The Performance Measurement System - PMS*

Dealing with the multicriteria and the multi-level aspects of the performance, Performance Measurement System (PMS) has been introduced as a tool which integrates the PI's of the considered system. Such integration relies on two main aspects that can be distinguished: the methodological aspect, *i.e.* how a PMS is implemented, and the measurement aspect, *i.e.* how the performances are expressed.

PMS's methodological frameworks were introduced from the beginning of the nineties [17] [18] [19]. The aim of such contributions was the integration of the whole life cycle of a PMS, from its design to its updating. For instance, let us mention the Balanced ScoreCard (BSC) model that is developed from five steps: "*1. Reach an agreement about the terms and their meaning, 2. Understand what results are expected to be achieved, 3. Organize proper discussion around the strategy, 4. Add KPIs and action plans, 5.°Make sure a cultural shift was made*" [17]. During the last two decades, numerous works have contributed to specify the PMS organisation and its functionalities [13][17][20][21][22][23].

The last works in this field concern both performance measurement and management aspects due to the IoT. Thus the PPAM (Production Performance Analysis Method) deals with real time manufacturing. From Bluetooth and Wi-Fi, companies track and trace the real-time manufacturing activities. From the distributed manufacturing (big) data obtained, the PPAM gives a dynamic sense for a real time operations management control [24]. Otherwise the IoT allows the companies to have a better understanding of the creating value process and consequently to develop the "servitisation"[1] processes with the data obtained by the IoT based solutions [25]. The specificities of the performance criteria of companies using IoT are identified, and a model of performance measurement using cloud and allowing daily measurement is proposed. As a transition with the measurement aspect of the PMS, the QMPMS (Quantitative Model Performance Measurement System) [26] is based on the MCDA (MultiCriteria Decision Analysis) school [27]

The measurement aspect of PMS's deals with the performance expression. The PMS is defined as: "*the set of metrics used to quantify both the efficiency and effectiveness of actions*" [28]. The question thus becomes related to the definition of methods to express both the elementary performances (metrics) that are associated with the PIs of the PMS and the overall performances. Historically, the first approaches including the BSC and the ISO 9000 standard [ISO 9000 (2000)] focused on the definition of the objectives and the gathering of the measurements, leading the decision-makers to implicitly interpret the achieved performances. ECOGRAI has considered the aggregation concept in order to reduce the size of the involved metrics [29]. ENAPS has adopted computation standards in the form of ratios. The QMPMS proposed a quantified performance model that is based on utility functions. This approach has inspired our works as well as the quantified model that is defined in the VR-PMS (Value Risk-PMS). Both elementary performances and overall performances are expressed, according to the highlighted hierarchy of criteria. Aggregation is achieved by using a more or less corrected Weighted Arithmetic Mean WAM. Furthermore, the performance community has underlined the necessity of handling interaction between PI's and metrics. Such a handling has led to the following definition: "*a performance measurement system is a multicriteria instrument, made of a set of performance expressions (also referred to as "metrics", i.e. physical measures as well as performance evaluations), to be*

[1] Servitisation consists of supplying both services and products for a better answer to the customer requirements.

consistently organized with respect to the objectives of the company" [30].

However, even if awareness about real-time and big data performance measurement is gotten, investigations just began in this sense.

III. Fuzzy performance measurement framework

A. *The elementary performance expression*

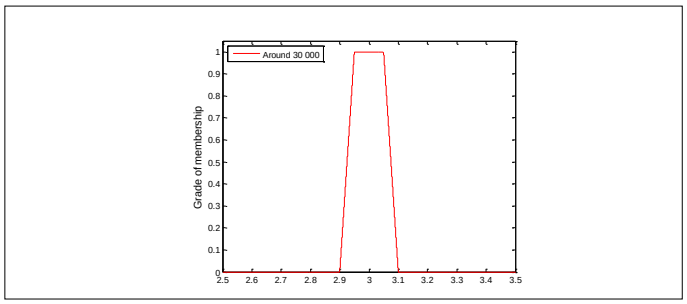
In the late nineties, we were interested in the extension of the Taylorian *ratios*. As explained before, the performance expression that is provided by a PI is considered to be the result of a mathematical comparison of the objective and the measurement. Such a definition led to questions such as: How to declare imprecise objectives? How to deal with uncertainty in measurements? What does comparing objectives and measurement in such cases mean? What are the semantics of the expressed performance when the objective is imprecise or the measurement is uncertain?

The industrial objectives are declared by decision-makers who often use natural language in this sense. For example, one may declare an objective related to a production line capacity in the following form: *a capacity of around 30000 components per week at the end of the year*. In such a declaration, should we consider that the desired capacity is exactly *30000 components;* or could we accept that the objective is reached if the capacity is about *30000 components per week;* and the closer to *30000 components per week* the higher satisfaction? Nuances should be taken into account when declaring objectives. Several remarks can also be addressed regarding measurements. Depending on measurement tools, nuances can also be introduced. On the one hand, it is interesting to propagate sensor uncertainties in the performance expression. On the other hand, human "sensors" show a natural ability to use additional information taken from the natural language, like *almost, close to, quite*… These considerations have previously led us to use the fuzzy subset theory [31] for the handling of both the objective and the measurement. This theory makes it possible to represent, from a mathematical point of view, the potential for an element to belong partially to several sets. Hence, the concept of gradual membership to a class is introduced. Later, Zadeh introduced the possibility theory and its link with fuzzy sets [32], offering an alternative to the probability theory to deal with uncertainties [33].

Transforming the declaration of flexible or imprecise objectives expressed in natural language into a mathematics based representation has been emphasised by Zadeh as *Precisiated Natural Language* [34]. Hence, the previous objective can be precisiated as "*capacity(Production line) is around 30000 components per week*" (Fig.1). The membership function is interpreted as a performance function providing a satisfaction in the objective achievement. In the case of a precise measurement, the performance is obtained from the membership function. For example, if the number of produced components is *29 450*, the performance is 0.9.

In the first works dealing with imprecision and uncertainty in the elementary performance, we were enthusiastic about the representation of respectively, categorical, flexible, imprecise objectives, either in the numerical form or the symbolic form[35]. The representation of measurement uncertainties by means of possibility distributions was made. Later, we gave a synthesis of what was done in a unified fuzzy framework [36].

Fig. 1."Precisiation" of "*around 30000 components".*



Such a fuzzy framework makes it promising to cope with the wide variety of data types coming from the industry 4.0. Indeed, numerical as well as linguistic information can be processed to provide a homogeneous way of representing performance expressions. Thus the "servitisation" process, useful for the customer creating value, can be aided by supplying adapted performance.

B. *The aggregated performance expression*

The second aspect of the developed framework has concerned the overall performances expression. The purpose of our investigations was to define a model for computing the overall performance, by considering the involved elementary performances. The relevance of such a model was in the handling of the "complexity" of the overall objective associated with the considered PMS. And formally speaking, we were looking for a homogeneous handling of this computation and what has been done for the elementary performance expression. According to the propositions of Globerson or Bititci and their teams, we have considered the aggregation way. The idea was simple: to synthetise several comparable numerical pieces of information, the elementary performance expressions, as a single piece of information, the so-called overall performance expression [37] [26]. The used mechanism has consisted of, on the one hand, choosing an aggregation operator, and on the other hand, numerically expressing the elementary performances, according to the same scales, generally on [0,1]. Usually, the aggregation operator was the WAM, which highlights the relative importance of the criteria that are associated with the elementary performances, namely their "weights". In this case, performance aggregation is based on the elicitation of the aggregation operator parameter and the elementary performance expressions. However, by considering only the weights of the elementary performances, an aggregation operator such as the WAM can partially take into account the PMS complexity, with the interactions between the elementary performances being neglected. The importance of the interaction notion in the PMS has been mentioned [17] [21] [26]. Rather than correcting the elementary performances with regards to the involved interactions, we have chosen to let the aggregation operator take into account such interactions. The idea was to overcome the traditional principle of WAM, which considers that the overall performance is the weighted mean of the elementary performances. By considering the interactions, the overall

aggregated performance becomes expressed between the "min" and the "max" of the elementary performances. More generally, the fuzzy integrals family handles such computations, and particularly the Choquet Integral *CI* which can be seen as an extension of the WAM. Particularly, two kinds of parameters are involved in a 2-additive *CI*-based aggregation model [38]. The Shapley coefficient ϕ_i 's is the weight of each elementary performance $\left(\sum_{i=1}^{n}\phi_i = 1\right)$. The interaction I_{ij} of any pair of elementary performances p_i and p_j. I_{ij} ranges within $[-1,1]$. "−1" expresses a complete redundancy between and "1" means complete complementarity. The overall performance is:

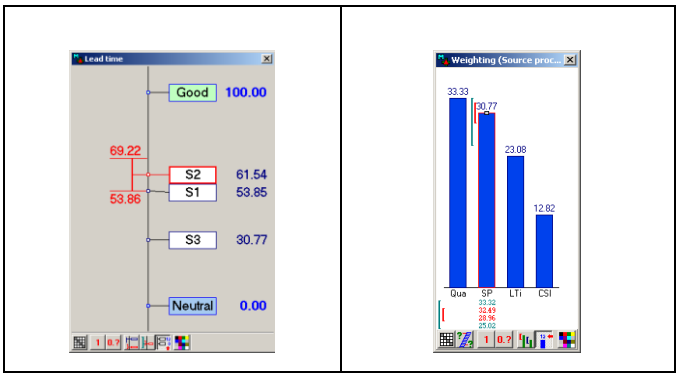
$$p_{Ov.} = CI(p_1 \ldots p_n) = \sum_{i=1}^{n}\phi_i - \frac{1}{2}\sum_{j\neq i} I_{ij}\left|p_i, p_j\right| \quad (1)$$

$$\text{with: } \left(\phi_i - \frac{1}{2}\sum_{j\neq i}\left|I_{ij}\right|\right) \geq 0, \ \forall\, i \in [1, n] \quad (2)$$

In order to elicit the CI parameters as well as the elementary performances, we have prospected the MCDA tools. In accordance with the measurement theory [39], such methods allow decision-makers to build a coherent aggregation model. We have particularly used the Multi Attractiveness Categorical Based Evaluation TecHnique MACBETH [40]. The building of the *CI* parameters as well as the elementary performances is done on the basis of expertise (Fig. 2). An additive assumption has been made, which concerns the link between the given expertise and the considered context. This expertise is given by preferences of the decision makers between considered "alternatives" with regards to the elementary performances.

An extension of the aggregated performance expression in the industry 4.0 context would concern the real-time decision. Indeed decision-maker is always time constraint and taking decision(s) from multiple performances expression can be risky. So a framework able to synthetise multiple performance expressions in a coherent way can be precious in order to take the right decision at the right time.

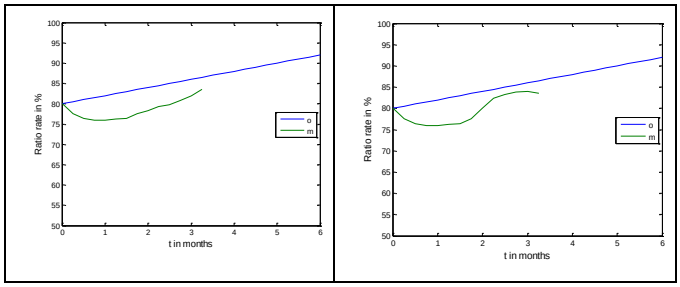
Fig. 2. MACBETH pieces of information for the aggregation.



IV. Temporal and visual management extensions

In the last 3-4 years, we have been interested in the temporal aspects related to the objective. We extend our vision by considering the objective no longer as a value to reach at a given milestone but as a trajectory to follow, throughout a temporal horizon, possibly generated from intermediate objectives, leading to the achievement of the expected value, at the end of the temporal horizon [41]. Our first idea was to compute the performance expression at any time, as long as measurements were available. Besides, we have been interested in "trend" objectives as in: "*Slightly increase the ratio rate throughout 6 months"*. Such declarations generate new questions about the representation of the trends and induced imprecision [42]. Moreover, the idea of comparing the trend of the measurement and the trend of the objective using a fuzzy linguistic approach has been emphasized, providing a performance in terms of trends. It has led us to distinguish between instantaneous performance and trend performance. The former gives information about the achievement of the objective, while the latter informs about the way this achievement is done. The fuzzy framework provides once again a natural tool to linguistically aggregate these pieces of information. The aggregated result makes it possible to differentiate a situation where the instantaneous performances are the same but the trend performances are different (Fig. 3).

Fig. 3. Instantaneous and trend performances.



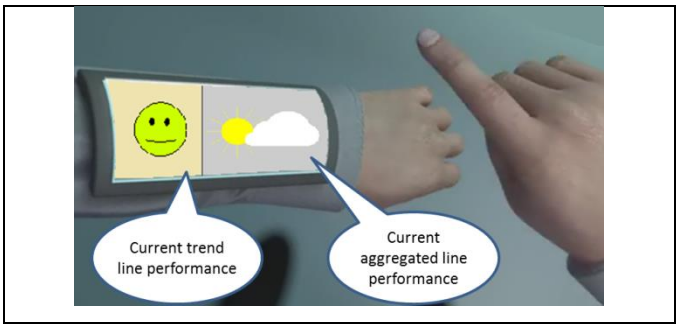
The temporal performance interest is significantly increasing in the industry 4.0. Indeed, due to its complexity and its quantity, temporal information leads to new issues in performance management. From this point of view, building user-friendly interfaces is a new challenge.

Recently, we have tackled this issue, in coherence with the visual management recommendations, in particular to provide an easy and real-time interpretation of the performance assessment. One interesting point concerning the fuzzy approach is the ease to convert numbers into linguistic terms, but another is the same ease to convert fuzzy linguistic terms into numbers. This conversion is known as defuzzification. For instance, when dealing with enriching conventional performance expressions by trend information, we proposed to represent an overall performance expression based on, on the one hand, a set of performances that are expressed on the set of terms L_1 = {*Very Good*, *Good*, *Medium*, *Bad*}, and, on the other hand, the trend performance expressions represented by the set of terms L_2 = {*Decreasing*, *Stable*, *Increasing*} [43]. We chose to express the result by a set of emoticons L_3 = {☺, 😐, ☹}. Let A be a fuzzy subset of L_1, for example $A = 0.85/VG + 0.15/G + 0/M + 0/B$, and B be a fuzzy subset of L_2, for example $B = 0.27/I + 0.73/S + 0/D$ where *VG* stands for *Very Good*, *G* for *Good*... Let *g* be the mapping associated with a rule base to perform the combination of both subsets as a global score. An example of rule is: "If performance is *Good* and performance trend is *Increasing* then score is 0.8/☺ + 0.2/😐". The global

score can be formally expressed as $C = g(A,B)$ where C is a fuzzy subset of L_3. One result is: C = 0.95/☺ + 0.05/😐 + 0/☹. The idea of friendly interfaces relies on the defuzzification of the previous fuzzy subset to generate parameters that can be used with graphical elements. For instance, it has been shown how to generate emoticons whose mouth openings and colours depends on these parameters. As an example, the emoticon represented in Fig. 4 is associated with the fuzzy subset given C = 0.95/☺ + 0.05/😐 + 0/☹. This fuzzy processing can be adapted to other symbolic representations such as weather symbols, traffic lights…

For the industry 4.0, such visual management could aid people to obtain a common understanding of the various performance expressions in an agile way. Indeed symbolic representation allows to communicate according to an universal meaning. Moreover the "IoT" makes it possible to select, refresh and trace the relevant performance in scorecards available everywhere in the company. Inspired by the Peugeot vision of the industry 4.0 [44] let us imagine such individual scorecard able to inform the employees everywhere and in real-time about the relevant performance for their activity. In the figure 4 a colored emoticon is associated with the line trend productivity, and a weather symbol represents the multicriteria (quality, lead time, safety, energy consumption…) aggregated performance of the line. Of course the displayed performance depends on the manpower objectives and the production events. Moreover it can be linked to the actions and decisions taken by the employees.

Fig. 4. Towards an autonomous visual scorecard 4.0 (inspired from [44]).



V. Conclusion

Performance measurement has been, over the last thirty years, a major field in the industrial control. In this paper, some aspects of the performance measurement have been considered and related to the industry 4.0 context. In fact the bases of the performance expression framework such as objective declaration, measurement acquisition, elementary comparison, overall aggregation will be maintained in the industry 4.0. This framework must be extended to better take into account the news performance expression characteristics due to the real-time decision, the numerous and various data available, the possibility of information displaying, the increasing employees autonomy, the IoT…

From these possible extensions, the visual management has been distinguished. This is way it can be applied to the various kinds of performance expressions (linguistic, numeric, temporal, aggregated); it can now be displayed all over the company through the IoT (mobile, watch, dedicated screens.) for the whole workforce; it can put into perspective the past, current and future performance using animated sequences. So, in this way developing connected interfaces based on visual management seems to be a promising PMS development.

References

[1] P. K. Paritala, S. Manchikatla, and P. K. Yarlagadda, "Digital manufacturing- applications past, current, and future trends," *Procedia Engineering*, vol. 174, pp. 982–991, 2017.
[2] "Collins on line dictionary," 1990. [Online]. Available: http://www.collinsdictionary.com/dictionary/english-french
[3] H. Johnson, "Management accounting in early integrated industry - e.i. dupont de nemours powder company 1903-1912," *Business History Review*, pp. 184–204, 1975.
[4] M. Lebas, H. Stolowy, and Y. Ding, *Financial Accounting and Reporting: A Global Perspective*. Cengage Learning EMEA, 2013.
[5] K. Cross and R. Lynch, "The "SMART" way to define and and sustain success," *National Productivity Review*, vol. 1, pp. 23–33, 1988.
[6] D. P. Keegan, R. G. Eiler, and C. R. Jones, "Are your performance measures obsolete," *Management Accounting*, pp. 45–50, 1989.
[7] S. Garcia, Y. Cintra, R. de Cássia S.R. Torres, and F. G. Lima, "Corporate sustainability management: a proposed multi-criteria model to support balanced decision-making," *Journal of Cleaner Production*, vol. 136, Part A, pp. 181–196, 2016.
[8] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 5, no. 15, pp. 2787–2805, 2010.
[9] L. Fortuin, "Performance indicators, why, where and how?" *European Journal of Operational Research*, vol. 34, pp. 1–9, 1988.
[10] C. Berliner and J. Brimson, *Cost Management for Today's Advanced Manufacturing: The CAM-I Conceptual Design*. Harvard Business School Press, 1988.
[11] J. Brimson, *Activity Accounting*, J. Wiley and N. Y. Sons, Eds., 1991.
[12] J. Browne, J. Devlin, A. Rolstadas, and B. Andersen, "Performance measurement: the ENAPS approach," *International Journal of Business Transformation*, vol. 69, no. 2, pp. 73–84, 1997.
[13] A. Gunasekaran, C. Patel, and R. E. McGaughey, "A framework for supply chain performance measurement," *International Journal of Production Economics*, vol. 87, no. 3, pp. 333–347, 2004.
[14] H. Balfaqih, Z. M. Nopiah, N. Saibani, and M. T. Al-Nory, "Review of supply chain performance measurement systems: 1998–2015," *Computers in Industry*, vol. 82, pp. 135–150, 2016.
[15] V. Veleva and M. Ellenbecker, "Indicators of sustainable production: framework and methodology," *Journal of Cleaner Production*, vol. 9, no. 6, pp. 519–549, 2001.
[16] M. Bitton, "Ecograi : méthode de conception et d'implantation de systèmes de mesure de performances pour organisations industrielles," Ph.D. dissertation, Thèse de doctorat en Automatique de l'Université de Bordeaux I, 1990.

[17] R. Kaplan and D. Norton, *The Balanced Scorecard: Measures that drive performances*. Harvard Business Review, 1992.
[18] A. Neely, M. Gregory, and K. Platts, "Performance measurement system design: a literature review and research agenda," *International Journal of Operations and Production Management*, vol. 48, no. 4, pp. 80–116, 1995.
[19] J. Wisner and S. Fawcett, "Linking firm strategy to operating decisions through performance measurement." *Production and Inventory Management Journal*, vol. 32, no. 3, pp. 5–11, 1991.
[20] A. Neely, C. Adams, and M. Kennerley, *The Performance Prism: The Scorecard for Measuring and Managing Business Success*. Financial Times Prentice Hall, London, 2002.
[21] A. M. Ghalayini, J. S. Noble, and T. J. Crowe, "An integrated dynamic performance measurement system for improving manufacturing competitiveness," *International Journal of Production Economics*, vol. 48, no. 3, pp. 207–225, 1997.
[22] F. Vernadat, L. Shah, A. Etienne, and A. Siadat, "Vr-pms: a new approach for performance measurement and management of industrial systems," *International Journal of Production Research*, vol. 51, no. 23-24, pp. 7420–7438, 2013.
[23] B. Sager, S. Hawer, and G. Reinhart, "A performance measurement system for global manufacturing networks," *Procedia CIRP*, vol. 57, pp. 61 – 66, 2016.
[24] Y. Zhang and F. Tao, *Optimization of Manufacturing Systems Using the Internet of Things*. Academic Press, 2017.
[25] A. Rymaszewska, P. Helo, and A. Gunasekaran, "Iot powered servitization of manufacturing – an exploratory case study," *International Journal of Production Economics*, In Press, Corrected Proof, 2017.
[26] U. S. Bititci, P. Suwignjo, and A. S. Carrie, "Strategy management through quantitative modelling of performance measurement systems," *International Journal of Production Economics*, vol. 69, no. 1, pp. 15–22, 2001.
[27] T. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, pp. 234–281, 1977.
[28] M. Bourne, A. Neely, J. Mills, and K. Platts, "Implementing performance measurement systems: a literature review," *International Journal of Business Performance Management*, vol. 5, no. 1, pp. 1–24, 2003.
[29] Y. Ducq, B. Vallespir, and G. Doumeingts, "Coherence analysis methods for production systems by performance aggregation," *International Journal of Production Economics*, vol. 1, pp. 23–37, 2001.
[30] V. Clivillé, L. Berrah, and G. Mauris, "Quantitative expression and aggregation of performance measurements based on the MACBETH multi-criteria method," *International Journal of Production Economics*, vol. 105, no. 1, pp. 171–189, 2007.
[31] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 38–353, 1965.
[32] ——, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.
[33] D. Dubois, L. Foulloy, G. Mauris, and H. Prade, "Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities," *Reliable Computing*, vol. 10, pp. 273–297, 2004.
[34] L. Zadeh, "A note on web intelligence, world knowledge and fuzzy logic," *Data & Knowledge Engineering*, vol. 50, pp. 291–304, 2004.
[35] L. Berrah, G. Mauris, A. Haurat, and L. Foulloy, "Global vision and performance indicators for an industrial improvement approach," *Computers in Industry*, vol. 43, no. 3, pp. 211–225, 2000.
[36] L. Berrah and L. Foulloy, "A fuzzy handling of the multi-criteria characteristic of manufacturing processes," in *Applications of Multi-Criteria and Game Theory Approaches*, ser. Springer Series in Advanced Manufacturing, L. Benyoucef, J.-C. Hennet, and M. K. Tiwari, Eds. Springer London, 2014, pp. 137–160.
[37] S. Globerson, "Issues in developing a performance criteria system for an organisation," *International Journal of Production Research*, vol. 23, no. 4, pp. 639–646, 1985.
[38] C. Labreuche and M. Grabisch, "The Choquet integral for the aggregation of interval scales in multicriteria decision making," *Fuzzy Sets and Systems*, vol. 137, no. 1, pp. 11–26, 2003.
[39] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement, Vol. I., Additive and Polynomial Representations*. Academic Press, New York, 1971.
[40] C. Bana e Costa and J. C. Vansnick, "Applications of the MACBETH approach in the framework of an additive aggregation model," *Journal of Multi-Criteria Decision Analysis*, vol. 6, pp. 107–114, 1997.
[41] L. Berrah and L. Foulloy, "Towards a unified descriptive framework for industrial objective declaration and performance measurement," *Computers in Industry*, vol. 64, no. 6, pp. 650 – 662, 2013.
[42] L. Foulloy and L. Berrah, "A fuzzy handling of trend objective declaration and trend performance expression," in *Proc. of the 2015 IFAC Symposium on Information Control in Manufacturing*, 11-13 mai 2015, Canada 2015.
[43] V. Clivillé, L. Berrah, and L. Foulloy, "Fuzzy symbolic handling of industrial instantaneous and trend performance expressions," in *Advances in Production Management Systems. Innovative and Knowledge-Based Production Management in a Global-Local World*, ser. IFIP Advances in Information and Communication Technology, B. Grabot, B. Vallespir, S. Gomes, A. Bouras, and D. Kiritsis, Eds. Springer Berlin Heidelberg, 2014, vol. 440, pp. 68–75.
[44] Factory of the future by the PSA group (in French). [Online]. Available: https://www.youtube.com/watch?v=igW-MyX7PkI

Molding 4.0 - The Economics of an Injection Molding As-a-Service Business Model

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Alessandro Charalambis*, Marco Alessandro Tonetti*, Guido Tosello*
*Department of Mechanical Engineering
Technical University of Denmark, Kgs. Lyngby, Denmark 2800
Email: alcha@mek.dtu.dk

***Abstract*—During the last few years, the term Industry 4.0 or The Fourth Industrial Revolution, made its appearance and spread across industries. While it is accepted that the term broadly refers to a set of recent innovations with potential to disrupt value and process chains, the heterogeneity of actors involved contributes to a hazy definition of the phenomenon. In this work, Industry 4.0 is analyzed by analyzing into its influence on the plastics industry, with a focus on the injection molding technology. A new business model for the plastic industry is proposed, which fosters closer cooperation with suppliers and Internet-of-Things integration of equipment on the manufacturing company shop-floor. Automated quality control, predictive maintenance, and energy efficiency are identified as main levers of improvement, and associated cost savings are quantified for a total of 4.13M€. This study shows how the paradigm of Industry 4.0 is able to disrupt the industry by decreasing machine downtime and offering remarkable improvements in machine up-time. The present research aims to highlighting some of the opportunities for the plastic industry enabled by the implementation of an Internet of Things architecture.**

I. Introduction

The increasing focus on digitization, and interconnection of product, machine and manufacturing process represent one of the main global trends in the industry today. The terms Industry 4.0, digital economy, intelligent production or Industrial Internet of Things (iIoT) have been used to describe the digital transformation of conventional manufacturing processes [1].

The main trends that capture the digitization of the manufacturing systems are the inevitable evolution of the way goods are produced, and the services that are offered. Small, medium and large organizations are continuously researching how to control and develop their processes by optimizing operational efficiency, improve quality, reduce scrap and waste. The rise and incremental use of digital manufacturing technologies such as additive manufacturing for the optimization of the conventional manufacturing process chain have proved to be able to considerably reduce cost and lead time [2]. In addition, the global macro trends that ask for an adaptive, fast, and flexible processes to meet the global needs for a cost efficient and reliable manufacturing unveiled the necessity to digitize, interconnect and link the production processes [3].

Four disruptive technologies drive the industrial advancement of digital manufacturing: *data and connectivity, analytics and intelligence, human-machine interaction, and digital-to-physical conversion"* [4]. In this sense, Industry 4.0 is a digital evolution of the manufacturing system that is personalized and tailored on the specific organization that is pursuing it.

Recent studies on the economics of IoT and Industry 4.0 suggest that the correct establishment of a digital-to-physical connection could generate *"up to $11.1 trillion a year in economic value by 2025"* [5]. A large share of benefits of IoT and Industry 4.0 will be obtained among others by improved operational efficiency, zero-defects manufacturing, optimized energy consumption, and reduction of costs across the entire manufacturing process chain [5].

To achieve the aforementioned benefits of Industry 4.0, it is a key issue to fully and correctly implement it. Industry 4.0 should not be interpreted as an IT implementation, but more as a business and technological decision strategy across the entire company that will inevitably create added value and change the value chain. New business models that are created upon the Industry 4.0 levers of value creation will bring added value and create profitable business opportunities.

II. Molding 4.0 - Business Model

Industry 4.0 and the digitization of manufacturing systems is not driven by the rise of new technologies but instead, more importantly, is the vision in which existing technologies are interconnected, communicate and work together to create operational excellence. In the Industry 4.0 scenario, the whole organization will be connected across the entire supply chain with its machines and products. However, a relationship between suppliers and manufacturers is an important step in the implementation of a successful digital infrastructure for manufacturers in IoT integration. Data connectivity is a key requirement in the future smart factories. Collection and elaboration of data opens new opportunities in terms of converting existing business models to the "as-a-service" business model. The advantages linked with the business model innovation will be presented in the next section, where the concept of molding as-a-service it is presented and analyzed.

An initial research on the profit and loss (P&L) statement of a leading European plastic manufacturer has shown that depreciation is burdening the financial structure of the company. This finding opens a broader discussion about the variabilization trend of fixed costs recently observed among corporations and, specifically, the lighter capital intensity enabled by Industry 4.0 [6]. High fixed costs are rightly linked to scale effects, cost advantages and profitability, and this proved to be a sound concept throughout the second half of the twentieth century. Variabilization offers the opportunity to lighten the requirements of financial resources on the demand-fulfillment side, in a situation where demand-creation will

already be extremely challenging. Variabilization of fixed costs can be enabled through *"as-a-service business models"* [4]: the idea is that technology and automation providers may move from selling machinery to a pay-by-usage model, in which the customers pay for the useful outcomes of the technology (up-time).

The service logic paradigm in which manufacturers commercialize their products by offering a service throughout time has been developed already in the first industrial revolution by James Watt (1736-1819) [7]:

> We let you have a steam engine cost free. We will install it and take over the customer service for five years. We guarantee that you will pay less for the engine's coal than you currently spend to feed the horses doing the same work. And all we are asking is that you give us one third of the money you will save. (James Watt, cited in [8])

A research on the direction for prioritizing digitization endeavors by identifying processes and parts of the value chain with the biggest opportunities of improvement identifies the main cost drivers as quality, repair & maintenance, and energy consumption. Such business model has the potential to reduce fixed operating cost, increase the manufacturer productivity by the use of machine that are in-line connected and continuously improved with predictive maintenance, efficiently energy consumption and high quality throughout the entire service. Throughout a continuous QC, high level of QA, and optimized maintenance service all offered by the supplier of the machine, we envision a business model where the manufacturer will pay only for the amount of goods that have been fabricated according to specialization meaning that in case the system detect a possible decrease in the molding parameters, which affect the cycle's quality then that batch will not be paid for.

Recent literature on QC and QA provide evidence of successful attempts to link part quality to machine process parameters through combination of hardware (sensors, actuators, ...) and different analytical methods. A study proposed in [9] shows the actual implementation of an IoT model to monitor real time part quality during the IM process. A trace management system collects and manage information stored by a sensor that analyze real-time the mould system temperature and pressure. A monitoring/analysis of trace system architecture detect abnormal status, analyze trace data, automatically detects defects, and optimize production by reducing scrap rates based on analysis of big data related to previous defective parts.

A similar approach to QC based on a predictive model, has been proposed in [10] in order to achieve a predictive maintenance approach for IM that uses process parameters to evaluate the conditions of the equipment. The system is able to indirectly predict malfunctioning of components through real-time monitoring of the identified significant process parameters.

Literature widely recognizes that one of the biggest improvements in energy efficiency, in the IM industry, comes from the introduction of all-electric machines, however one recurrent thread in Industry 4.0 is the possibility to optimize inputs of any process in real-time. This thread is also applicable to the minimization of energy consumption of IM machines. An energy management framework based on the IoT is proposed in [11], and it discuss the introduction of smart sensors and smart meters at machine and production line level for real-time energy consumption management.

III. "AS-A-SERVICE" MODEL POTENTIAL

A. Structure of the analysis & definition of initialized variables

The proposed IoT architecture of this analysis is outlined in figure 1, where the steps followed in the analysis are highlighted for savings related to QC, predictive maintenance and reduction of energy consumption.

B. Quantification of "as-a-service" model

The structure of the analysis presented in this work, starts with the analysis of the case company budget to identify that fixed and variable operating expenses are the main cost drivers accounting for 53% and 45% of total cost respectively. The proposed business model has the potential to impact variable and fixed operating expenses in two different ways. An analysis of the fixed operating expenses shows that depreciation constitutes the major part (55%) hence fixed operating expenses allocated to the investment in injection molding machines can be replaced due to the fact that machines can be now rented from the manufacturer and it will be instead paid the service acquired. Therefore, the first objective of this research is to establish possible cost reduction opportunities enabled by leasing injection molding machines instead of buying them. Then the indirect cost reduction opportunities enabled by an optimized QC, maintenance service, and energy consumption offered by molding as-a-service model will be discussed.

In the proposed model, suppliers would rent their machines to the manufacturer that would pay only for the up-time of the machine. The supplier would then receive a periodic cash-flows that depends on the performance of the equipment offered and the level of quality that the machine is able to achieve. As previously stated, we envision a business model where manufacturers would not pay for "non-conformed quality injection cycles".

Cost reduction in fixed operating cost have been estimated considering a company with around 52m€ stacked in ownership of IM machines with an average machine cost of 130.000€ (hence owning around 400 IM machines). Maintenance cost amounts to 2% of the purchasing price per year per machine ($2.600€/machine \times year$) and it is considered the average machine life time to be 20 years. This is a model that accounts only for the cost to purchase the machine and the maintenance cost quoted by the supplier. The cost associated with the manual operations involved in the use of the machine, electricity and other overheads are not directly accounted here but instead are considered as variable operating expenses. Therefore, it is chosen to depreciate the total cost for the machine and its maintenance contract over the amount of hour it actually works. Assuming a currently efficiency rate of 85% gives possible breakdowns and maintenance needed. Then the current operating hours per year amount to 5.100 hour/year. It follows that under the "as-a-service" model, the supplier charge a premium price (assumed to be a 10% markup) to achieve an efficiency rate of 95% or higher and the services offered (QC, QA, preventive maintenance,

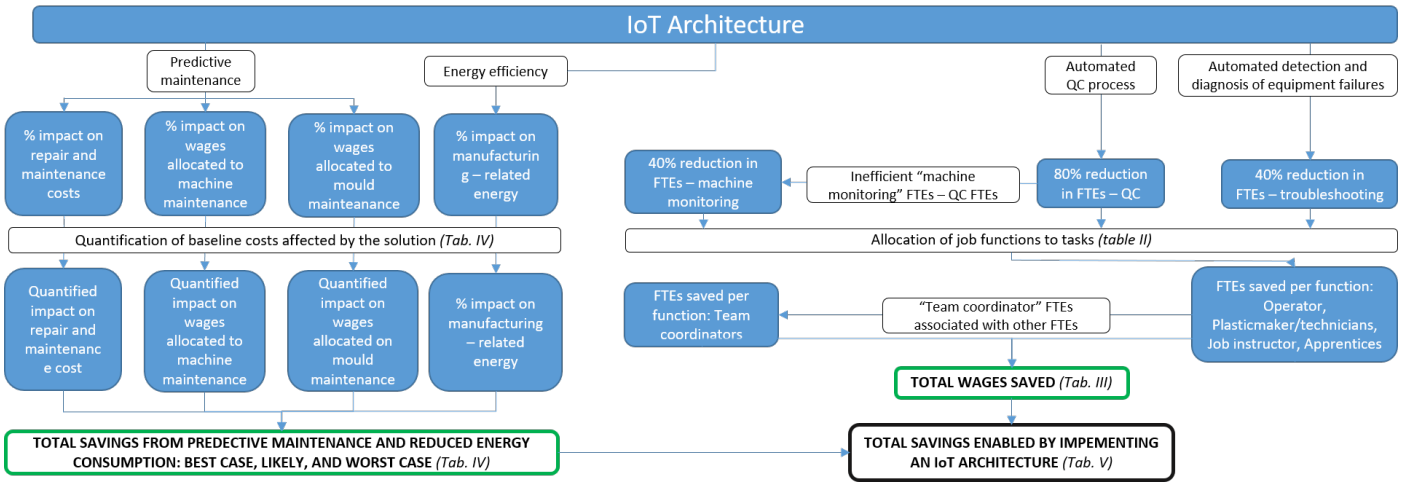


Fig. 1. Flowchart of the IoT architecture applied to injection moulding and summary of the structure of the analysis

efficient energy consumption) then a monthly rent for the machine could be equal to 3.600 €/$month$. Even though this monthly rent is translated in a higher cost per hour than the present, greater advantages and opportunities are linked with higher machine efficiency, lower breakdown and downtime, an optimized maintenance, efficient energy consumption, QA and QC practices as described in the next section.

C. Wages reduction enabled by automated QC

Variable operating expenses are investigated in the optimization of the above mentioned areas of improvement and are broke down as follows: *wages* - pertaining blue collar workers - dominate these expenses (54%) and will be further investigated below; *repair and maintenance* (12%) and *manufacturing-related energy* (11%) are also big enough to confirm their prioritization for improvement; *indirect materials* (20%) includes ware parts.

Table I allocates the budget for wages to the different departments. The next step consists in identifying the main drivers among the production department accounting for 45% of wages, and to determine what tasks consume most human resources.

TABLE I. ALLOCATION OF WAGES TO DEPARTMENTS

Department	M€
Production	7.52
Mould maintenance	1.96
High bay warehouse	1.42
Raw material	1.25
Machine maintenance	1.08
Changeover	1.02
Logistics maintenance	0.83
Mold maintenance CMS	0.62
Equipment	0.48
Internal transport	0.33
Molding process engineering	0.23
Molding management	0.00
Total	**16.74**

Table II shows a breakdown of job functions in the Production department employed in that function and their allocation to the task performed, with the percentage of total full-time equivalents (FTE)s, which represents 37,5 $men \times hours$, i.e. the human resources that are equivalent to one employee working full-time. It should be highlighted that four tasks account for almost 72% of total wage expenses: quality control, troubleshooting, monitor machines and replacing molds.

TABLE II. BREAKDOWN OF FTES

Job function	% Tot. FTEs	Task	Time	% FTEs
Operator	38,4%	Quality control	60%	23,0%
		Monitor machines	30%	11,5%
		Cleaning machines	10%	3,8%
Plastic maker /technician	36,1%	Troubleshooting	60%	21,7%
		Monitor output	10%	3,6%
		Replacing molds	15%	5,4%
		Train colleagues	5%	1,8%
		Other	10%	3,6%
Team coordinator	12,1%	Operational	50%	6,1%
		Task coordination	40%	4,8%
		Adm. work	10%	1,2%
Apprentice operator	5,3%	Quality control	60%	3,2%
		Monitor machines	30%	1,6%
		Cleaning machines	10%	0,5%
Apprentice plastic maker /technician	5,0%	Troubleshooting	60%	3,0%
		Monitor output	10%	0,5%
		Replacing molds	15%	0,8%
		Train colleagues	5%	0,3%
		Other	10%	0,5%
Job instructor /operator	1,6%	Training colleagues	35%	0,6%
		Quality control	39%	0,6%
		Monitor machines	20%	0,3%
		Cleaning machines	6%	0,1%
Job instructor /plastic maker /technician	1,5%	Training colleagues	35%	0,5%
		Troubleshooting	39%	0,6%
		Monitor output	7%	0,1%
		Replacing molds	10%	0,2%
		Other	9%	0,1%
TOTAL	100%			100%

The potential reduction in workforce that would result from the introduction of an iIoT architecture due to the "as-a-service" business model are now introduced together with the associated technological benefits, with particular focus on savings enabled by automated QC. It should be stressed that "cutting FTEs" does not entail layoffs and the associated social, psychological and reputation's downsides. The present analysis looks indeed at the long-term growth strategy of the company in which the free human resources are allocated to the new capacity required by the enterprise expansion.

The quality control task in the tables above refers to the activity of visual inspection that takes place in the pro-

duction modules of the case company. Employees perform visual inspection on parts produced by each machine every 8 hours (assessing superficial defects, geometry, color, and clutch power), relying solely on their own experience. This task could be almost entirely substituted by IoT-enabled QC offered by the machine suppliers on the long run, assuming to keep 20% of the quality control FTEs for complementary monitoring (at reduced frequency) and for performing ad-hoc checks when the machines identify potential quality issues.

The other main reduction concerns the troubleshooting task. This task, mostly handled by plastic makers and technicians, involves detecting root causes of failures/breakdowns of the injection molding equipment and production errors, and taking decisions on the countermeasures to undertake. It is expected that the frequency of breakdowns will be dramatically reduced in the first place (a recent study reports 70% to 75% elimination of breakdowns achieved from predictive maintenance programs [12]). Furthermore, when failures do happen, troubleshooting will be done instantaneously by the equipment intelligence (by recognizing patterns in machine data): the job of technicians and plastic makers will be reduced to interpreting and validating results of the automatic analysis.

Table III summarizes the results of this analysis concerning the amount of wages that can be saved. Column 2 in the top section of the table lists the number of FTEs allocated to QC, among the overall FTEs employed in the job function in column 1. Column 3 lists the assumptions regarding the percentage of FTEs in column 2 that can be cut (as a result of automatized QC) and column 4 lists the consequent reductions in term of FTEs. The indirect reductions in column 5 depend on the job descriptions of the three functions involved, and need some background explanation. Looking at table II, it is possible to see that, for operators and apprentice operators, 60% of time is allocated to quality control and 30% to monitor machines. In the case of an instructor - operator, 39% is allocated to quality control and 20% to monitoring machines; this means that - in all three functions - half an hour of monitoring machines is associated with each hour of quality control.

After an internal analysis of the case company it was understood that this task mostly represents idle time between other tasks (e.g. walking between machines, looking at the machines for general monitoring purposes). Assuming that all the machine monitoring task was idle time, then each saved hour of quality control would entail an extra half-hour saving of idle time in the first place. Furthermore - even if this task actually included useful monitoring-time - it can be assumed with a good degree of confidence that it could be easily performed automatically through the IoT architecture. Therefore, ideally, the number of FTEs directly saved per function (column 4) would come with an extra 50% savings. For the sake of conservativeness, this assumption has been reduced to 40% for all the three functions and is referred to as indirect reduction (column 5). Column 6 presents the total reduction in FTEs as sum of direct and indirect. Finally, column 7 lists the savings coming from reducing the number of FTEs in column 6. The central portion of the table quantifies, in an analogous fashion, the impact of automatizing troubleshooting, but no indirect reductions are considered in this case. The assumption regarding direct percentage reduction was also more conservative (40%) due to the more delicate nature of the troubleshooting task and the necessity to maintain technically educated employees, who would be able to interpret machine-generated reports and make the final decision on the most suitable course of action to undertake. In the lower part of the table, an additional reduction in FTEs is considered concerning team coordinators. Out of the overall FTEs, 12% are team coordinators and 88% are other functions (being supervised by team coordinators), with a consequent ratio of 7,23 additional FTEs per coordinator (or 0,14 coordinators per FTE). It was therefore assumed that a potential reduction in team coordinator FTEs was linearly associated with reductions in other types of FTEs with a ratio of 0,14 to 1 (column 2 of lower table III), and that this potential can be fully captured in this case (represented by the 100% assumption in column 3). The total savings in wages, under the assumptions stated, amounts to 3.53 M€ per year.

D. Savings from predictive maintenance and energy efficiency

This section, investigates potential cost reductions enabled by the IoT, but focuses on the impact of predictive maintenance and increased energy efficiency. Achieved savings from analogous actions (IoT-enabled improvements) reported by several different sources were gathered and achieved savings range between 8-12% [13] to 50% [14] for predictive maintenance. Two other studies [15] and [16] reported savings in the range between 5% and 31%. Table IV summarizes the results of calculations about cost reductions in the case company. Column 2 lists the cost areas affected by the lever of improvement indicated in column 1. Column 3 lists the baseline cost considered, related to each of the cost areas in column 2 (i.e. the cost base that can be affected by percentage reductions identified in the following columns). The baseline cost of Repair & Maintenance was calculated according to the following considerations: an average initial cost of an injection molding machine equal to 130.000 €/machine and a maintenance cost equal to 2%/year (of initial cost of the machine). Therefore, $130.000\ €/machine \times 2\% \times 400$ machines $= 1.04$ $M€$ per year for maintenance related to injection molding machines alone. This 1.04 M€belongs to the total Repair & Maintenance cost allocated to case company budget, among variable operating expenses, or 39%.

It was then assumed that this proportion would apply to the baseline cost of manufacturing-related energy as well: the 0.92 Me value is calculated as 29% of the total Manufacturing-related energy cost allocated among variable operating expenses. It should be stressed that this is a conservative hypothesis (presumably, injection molding machines constitute a bigger portion of manufacturing-related energy) which likely leads to underestimate total calculated savings. The baseline cost for Wages (mould maintenance) comes from the sum of wages allocated to Mould maintenance and Mould maintenance CMS departments (cf. table I). On the right part of table IV, three scenarios (best, likely, and worst case) are outlined for each of the cost areas in column 2, and apply to the respective baseline cost.

Percentages regarding best, likely, and worst case scenarios are based on cautious and conservative evaluation of achieved savings in previous research ([13]–[16]). For repair and maintenance, it seems that 12% is a realistic estimate of

TABLE III. POTENTIAL SAVINGS IN WAGES THAT CAN BE CAPTURED BY IMPLEMENTING IOT THROUGH "AS-A-SERVICE" BUSINESS MODEL

Job function	**QC-allocated FTEs**	**Direct reduction [%]**	**Direct reduction [FTEs]**	**Indirect reduction [%]**	**Total reduction in FTEs**	**Wages saved [M€/y]**
Operator	44,3%	80%	58,4%	40%	56,5%	1.92
Job instructor/operator	1,2%	80%	1,6%	40%	1,6%	0.05
Apprentice operator	6,1%	80%	8,1%	40%	7,8%	0.21
Job function	**Troubleshooting allocated FTEs**	**Direct reduction [%]**	**Direct reduction [FTEs]**		**Total reduction in FTEs**	**Wages saved [€/y]**
Plastic-maker/technician	41,5%	40%	27,3%		18,9%	0.74
Job instructor - pl. maker/tech.	1,1%	40%	0,8%		0,5%	0.02
Apprentice pl. maker/tech.	5,7%	40%	3,8%		2,6%	0.07
SUBTOTAL	**100%**		**100%**		**87,9%**	**3.01**
Job function	**Team coordinator FTEs needed per non-coordinator FTE**			**Reduction [%]**	**Total reduction in FTEs**	**Wages saved [M€/y]**
Team coordinator	0,14			100%	12,1%	0.52
TOTAL					**100%**	**3.53**

TABLE IV. OPTIMISTIC, LIKELY, AND PESSIMISTIC ESTIMATES OF SAVINGS FROM IOT BENEFITS RELATED TO PREDICTIVE MAINTENANCE AND ENERGY EFFICIENCY

Lever of improvement	**Cost area impacted**	**Baseline cost [M€/y]**	**Scenario** $\frac{[\%]}{[M€]/y}$		
			Best case	**Likely**	**Worst case**
Predictive maintenance	Repair & Maintenance	1.04	35% 0.36	12% 0.13	8% 0.08
	Wages (mold maint.)	2.74	20% 0.55	8% 0.22	4% 0.11
	Wages (mach. maint.)	1.14	20% 0.23	8% 0.09	4% 0.05
Energy efficiency	Manufacturing related energy	0.92	30% 0.28	15% 0.14	5% 0.05
TOTAL		**5.84**	**1.42** (24%)	**0.58** (10%)	**0.29** (5%)

savings achievable through predictive maintenance, when the base case is an effective and optimized preventive maintenance program (and even in that case there is large room for improvement). Indirect reduction on workforce (i.e. wages - machine maintenance) are expected to carry lower results than direct savings on the equipment, therefore the percentage savings have been almost cut in half for all scenarios. In the case of wages (mould maintenance), the relations between predictive maintenance on injection molding machines and maintenance on the molds, based on new findings regarding a predictive maintenance solution, it seems that 20% is a feasible range of savings achievable. Such solution is is based on a solution that could effectively keep track of critical mold parameters in real time and - based on ANNs and pattern recognition - determine the condition of moulds without needing inspection and shutting down the machine, consequently enabling optimized scheduling of maintenance activities [17].

As far as the energy efficiency is concerned, success case applications show that many different energy-saving initiatives can be triggered by the granular monitoring of consumption enabled by IoT enables. Taken individually, none of the reported initiatives brought less than 5% of energy savings. In fact when part of a comprehensive energy management endeavor, huge savings are possible (for example, BMW achieved a 31% reduction in 6 years and aims at 45% by 2020 [16]). Furthermore, improved maintenance is correlated to higher efficiency of the equipment (by 5-15%), which increases the total potential when joint effects of predictive maintenance are considered. Overall, energy improvements achieved in the injection molding industry are generally in the range of 25-35%. Even in this case, for the sake of conservativeness, the likely scenario is lowered to 15%. Overall, yearly savings for the case company under the assumptions stated range between 0.3 and 1.5 M€, and likely fall in the neighborhood of 0.7-1.0 M€.

IV. DISCUSSION

Table V summarize results of the previous analyses. The total savings achievable amount to 4.1 M€ under the assumptions stated, or 13% of total variable operating expenses. The largest portion of savings comes from the implementation of automated quality control, which revealed to be one of the most resource-consuming activities in the plant, and the associated reduction in workforce: reductions in operators, plastic-makers and technician FTEs, all directly imputable to the introduction of IoT-enabled QC that accounts in fact for 68% of the savings identified. Improvements enabled by predictive maintenance and energy efficiency account for at least 0.7 M€ - or 18% of the identified savings - and include both direct savings and further reduction of FTEs: it should be stressed that conservative assumptions used in this case might lead to underrated savings as far as predictive maintenance and energy efficiency are concerned.

TABLE V. SUMMARY OF SAVINGS ENABLED BY IMPLEMENTATION OF AN IOT ARCHITECTURE

Item		**Savings [M€]**
Wages	Operators	1.92
Wages	Plastic-maker/Technicians	0.74
Wages	Team Coordinators	0.52
Wages	Apprentices	0.29
Wages	Mold Maintenance	0.22
Manufacturing related energy		0.14
Repair & Maintenance		0.13
Wages	Machine Maintenance	0.09
Wages	Job instructor	0.08
TOTAL		**4.13**

As the implementation of the "as-a-service" business model would focus on the automation of human activities concerning QC and maintenance it should not come as a surprise that savings in wages dominate the total benefits identified. Other important issues that were not addressed include: reduction of cycle times/increased output due to self-optimizing settings, optimized scheduling of jobs, and faster batch changeovers; reduction of inventory levels, both thanks to better scheduling of jobs and to horizontal integration of supply chain data

(internally and externally); longer useful life of the machines (or higher end-of-life value) thanks to predictive maintenance; improved sustainability profile and reputation, thanks to higher energy efficiency and therefore lower pollution.

It was also proven that strong synergies exist among the three fields analyzed (quality assurance, predictive maintenance and energy efficiency), meaning that - once a networked environment is established - several benefits can be captured with relatively little extra effort.

In general, it is strongly recommended that companies interested in pursuing the digital evolution of their manufacturing systems cooperate with the suppliers of injection molding machinery. Success stories come from initiatives based on cooperation across the narrow boundaries imposed by the enterprise. In the present case, success can be fostered by leveraging the superior technological knowledge of suppliers and the massive amounts of production data generated on the shop floors. It is also recommended that manufacturing companies consider investigating feasibility of the "as-a-service" business model, applied to injection molding machines. Through the acquisition of injection molding machine as a service, the manufacturer has the potential to free itself from the large investment in buying injection molding equipment. It has the opportunity to optimize its production setup through a more effective and continuous and in-line/interconnected quality control, preventive machine maintenance, and energy consumption control of the machines.

V. Conclusion

Industry 4.0 is a complex phenomenon, which is hard to map according to simplified frameworks without losing some of its most important aspects. A case study has been conducted in the present research on a large injection molding factory aimed at quantifying the impact of networked machinery. Major improvements enabled by Industry 4.0 in the molding process can be obtained by: automatizing quality control; optimizing maintenance; and maximizing energy efficiency. The IoT, not limited to the plastic industry, simplifies the adoption of industrial-machine-as-a-service business models, and extends their applicability to lower capital intensive assets. These business models create boundary conditions that inherently generate superior value for any given supply chain. In the present case study, the implementation of an IoT architecture is estimated to reduce total variable operating expenses of the plant considered by 13%. This cost reduction comes to a great extent from a reduction of the wages connected to maintenance and QC activities which is mostly a result of automatizing the QC process. At the same time, this research proposes new perspectives on a business model innovation for injection molding value chain, which shifts from its conventional paradigm (i.e. direct sale of machines) to the new idea that machine suppliers will move towards leasing their machines, offering a new value proposition that capitalize on optimized maintenance programs, automated QC and efficient energy consumption, gaining a share of the increased value offered to their customer. It is essential to stress once again that with this research we are not proposing to lay off employees but instead to use actual manpower for value adding activities, since labor intensive tasks such as quality control, monitor machine, troubleshooting, can be automatized through Industry 4.0. To conclude, it is expected that a possible cost model for molding as-a-service envisages a control of the amount of successful injection cycles and the real-time computation of a premium charged to the manufacturer for the higher performance.

Acknowledgment

The authors would like to thank the Marie og M. B. Richter's Fond (Denmark) for supporting the presented research and a large manufacturing company for the the development of the case study.

References

[1] S. I. Shafiq, C. Sanin, C. Toro, and E. Szczerbicki, "Virtual engineering object (veo): toward experience-based design and manufacturing for industry 4.0," *Cybernetics and Systems*, vol. 46, no. 1-2, pp. 35–50, 2015.

[2] A. Charalambis, L. Kerbache, G. Tosello, D. B. Pedersen, and M. Mischkot, "Economic analysis of additive manufacturing integration in injection molding process chain," in *7th IESM Conference, October 11–13, 2017, Saarbrucken, Germany*, 2017.

[3] D. Schweer and J. C. Sahl, "The digital transformation of industry–the benefit for germany," in *The Drivers of Digital Transformation*. Springer, 2017, pp. 23–31.

[4] D. Wee, R. Kelly, J. Cattel, and M. Breunig, "Industry 4.0 how to navigate digitization of the manufacturing sector," *McKinsey & Company*, 2015.

[5] J. Manyika, M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin, and D. Aharon, "Unlocking the potential of the internet of things," *McKinsey Global Institute*, 2015.

[6] M. Blanchet, T. Rinn, and A. Dujin, "The industrie 4.0 transition quantified," *Roland Berger*, p. 19, 2016.

[7] M. Ehret and J. Wirtz, "Unlocking value from machines: business models and the industrial internet of things," *Journal of Marketing Management*, pp. 1–20, 2016.

[8] E. Hofmann, D. Maucher, J. Hornstein, and R. den Ouden, *Capital equipment purchasing: Optimizing the total cost of CapEx sourcing*. Springer Science & Business Media, 2012, vol. 2.

[9] H. Park, "Iot-based trace management system for upgrading quality on production of plastic injection molding," *International Information Institute (Tokyo). Information*, vol. 19, no. 1, p. 207, 2016.

[10] C. Park, D. Moon, N. Do, and S. M. Bae, "A predictive maintenance approach based on real-time internal parameter monitoring," *The International Journal of Advanced Manufacturing Technology*, vol. 85, no. 1-4, pp. 623–632, 2016.

[11] F. Shrouf and G. Miragliotta, "Energy management based on internet of things: practices and framework for adoption in production management," *Journal of Cleaner Production*, vol. 100, pp. 235–246, 2015.

[12] R. Thomson, M. Edwards, E. Britton, and B. Rabenau, "Predictive maintenance - is the timing right for predictive maintenance in the manufacturing sector?" *Roland Berger*, 2014.

[13] G. Sullivan, R. Pugh, A. Melendez, and W. Hunt, "Operations & maintenance best practices," *A guide to achieving operational efficiency, Release*, vol. 2, 2004.

[14] Emerson Process Management, "Reducing operations & maintenance cost," 2003. [Online]. Available: http://www.controlglobal.com/assets/Media/MediaManager/ReducingOperationsAndMaintenanceCosts.pdf

[15] M. H. Naitove, "Sustainable injection molding: Amcor looks at the big picture," 2013.

[16] BMW Group, "Industry 4.0: intelligent energy data managemet for sustainable production," 2014.

[17] Applied Market Information Ltd., "Creating a "smart factory" - exploring potential and identifying key challenges," in *Industry 4.0 for plastic processor*, 2017.

Economic Analysis of Additive Manufacturing Integration in Injection Molding Process Chain

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Alessandro Charalambis*, Laoucine Kerbache†, Guido Tosello*, David Bue Pedersen* and Michael Mischkot*
*Department of Mechanical Engineering
Technical University of Denmark, Kgs. Lyngby, Denmark 2800
Email: alcha@mek.dtu.dk, guto@mek.dtu.dk, dbpe@mek.dtu.dk, micmi@mek.dtu.dk
†Department of Informations Systems and Operations Management
HEC Paris, Jouy-en-Josas, France 78350
Email: kerbache@hec.fr

***Abstract*—The purpose of this research is to analyze how additive manufacturing can create value when it is utilized as a supportive technology to injection molding by quantifying the cost advantages that can be obtained. Tooling for the product development phase is investigated as pilot integration area of additive manufacturing with injection molding. Cost considerations are discussed through the development of a cost estimation model. The study shows that integration of additive manufacturing in the product development phase for fabrication of soft tooling is economically convenient with a cost reduction of 79,8% and 89,9%. The cost models on additive manufacturing have been built so far on the idea of substituting injection molding with additive manufacturing. In response to this literature gap, this research addresses the advantages of additive manufacturing utilized in a synergistic rather than disruptive way to create value in the injection molding process chain.**

I. Introduction

The increasing focus on digitization and interconnection of products, machines and manufacturing processes represent one of the current main trends in industry globally. This phenomenon is referred as Industry 4.0 or fourth industrial revolution, and many companies are looking towards integration of products, machines, and people using digital technologies. Four disruptive technologies drive the industrial advancement: *data and connectivity, analytics and intelligence, human-machine interaction, and digital-to-physical conversion* [1].

Additive Manufacturing (AM) as a digital technology is capable of creating a linkage between the digital and physical value chains and fits in this scenario within the global trend of the digitization of manufacturing process chains and the rise of Industry 4.0. For this reason, the increasing adoption of AM within conventional manufacturing systems is pushing companies to research new ways of adapting their business models and optimize their value chains by integrating the new manufacturing technology and move towards a digital-to-physical conversion.

AM refers to a series of processes that have in common the use of a 3D CAD design to manufacture an object layer by layer. The origin of AM dates back to 1987 when it was first introduced the process of stereolithography (SL) or stereolithography apparatus (SLA) commercialized by 3D Systems for prototyping purposes [2], [3].

The series of processes manufacturing an object layer-by-layer have been called under the common name of "*3D Printing*". To distinguish the use of layer based manufacturing technologies between their use for prototyping and end-use production, the ASTM International committee F2792 defined AM as the "*process of joining materials to make parts from 3D model data, usually layer upon layer, as opposed to subtractive manufacturing and formative manufacturing methodologies*" [4].

AM technologies have been characterized by an impressive double digit growth, with an average compound annual growth rate (CAGR) over the past 27 years of 26.2% [2]. A large market for AM has been created in the past 20 years and a higher industrial adoption is foreseen in the near future. The market growth and the increasing use of AM have created a surrounding hype on the technology of being able to replace conventional manufacturing processes.

This paper intends to present how synergies between AM and conventional manufacturing processes can be established and how is possible to foster the adoption of AM. The focus of the research is then to identify how AM is able to add value if adopted to support the formative manufacturing process of injection molding. Following the continuous technological development of AM, its use for fabrication of injection mold inserts is increasingly becoming a possibility. Many firms have already started integrating polyjet printed molds in their product development phase, as presented e.g. in [5] and by a selection of Stratasys customers' application stories [6]–[10].

Two types of inserts were tested and considered; more specifically an insert for a two cavities mold and the micro insert discussed in [11], [12] to fit a four cavity mold. A drawing of each insert is shown in figure 1 and 2 respectively. The tested inserts are two cuboids with dimension of $80 \times 60 \times 10mm^3$ and $20 \times 20 \times 2.7mm^3$ respectively. The two inserts type are referred to in this paper as medal and micro insert due to their surface design.

II. Problem Statement and Methodology

Three main trends characterize the digitization in injection molding process chain: *Zero-defect production by digital molding platform, Additive manufacturing, and Integrated Modeling of molding process chain* [13]. Companies in order to capture

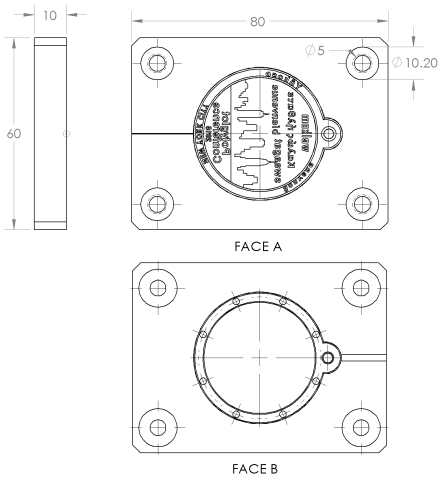


Fig. 1. Drawing of the medal insert (dimensions in mm) based on [12]

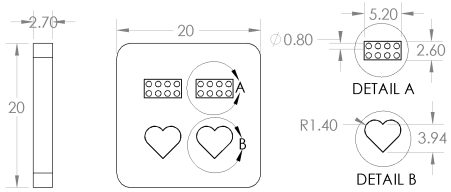


Fig. 2. Drawing of the micro insert (dimensions in mm) based on [11]

the benefit of Industry 4.0, have to conquer the digitization of the production processes, the increased demand for digital technologies and an interconnected network of machines and product. Those benefits reside within the ability to achieve operational effectiveness, adapt the business model to capture additional value offered by new technologies and work towards the implementation of the digital manufacturing transformation. The role that AM has within the technological revolution is to increase the operational effectiveness and efficiency, and create additional value for the company.

Much effort has been spent to discuss the potential impact that AM might have on the traditional supply chain and how the technology will revolutionize the entire production process. Researches and consulting firms have foreseen a future where conventional manufacturing processes are going to be replaced by AM [14]. Indeed the technology have great advantages and potential to change the traditional supply chain, but a gap in the literature is represented by the idea of integrating AM to create synergies and support conventional manufacturing processes. The next step would be to evaluate the value created by implementing AM within the value chain of a company that is pursuing the digitization of their value chain.

To create a thorough overview of AM and its adoption in the industry, is important to evaluate the technology from an economic perspective. Understanding the cost components that in the past mostly affected the price to fabricate an object can reveal opportunities for the future. The cost to manufacture an object is one of the factors that influence managers in deciding whether or not to implement a new technology. Two cost models have discussed the economics of AM. However, both have evaluated the value of AM under the assumption of replacing injection molding [15], [16].

Before presenting the economic model developed in this research it is important to draw some considerations on the economics that characterize the technology. The aim of this research is to thoroughly analyze the creation of synergies between AM and conventional manufacturing. Therefore, it is essential to create an understanding of what would be the added value of AM as supportive technology rather than a replacement for injection molding.

Review of literature, cost models, application cases in companies belonging to different firms, all reveal that when AM is used to replace injection molding it is cost effective but only up to low production volumes. Only few parts can be fabricated before AM becomes economically disadvantageous and manufacturers are forced to switch to conventional manufacturing technologies. Figure 3 summarizes the behaviour of AM through a curve when used to replace injection molding, and when is used through a synergistic approach.

AM for production of single units requires a low initial investment but the cost raises with an increase in the production quantities. On the contrary, the injection molding as fabrication method is characterized by a conspicuous initial investment which is then followed by a relatively low cost to manufacture a part. Therefore, the break-even-point in this scenario is equal to few units manufactured. It follows that manufacturers have to switch quickly to the injection molding in order to reach a cost effective solution for mass production.

The graph in figure 3 also demonstrate that the transition curve of AM when applied to create synergies with conventional manufacturing is gradual compared to the steep transition curve of AM when it replace injection molding. Moreover, the integration of AM in the injection molding process chain for fabrication of tooling, as it will be further analyzed in this paper, point out that a much larger quantity can be manufactured. Application of AM for tooling production makes the technology cost effectiveness for much larger production quantities even though fabrication of complex and time lavish mold inserts require a higher initial investment. The discussion about the economics that characterize AM continues in the definition of the AM cost estimation model of this research.

A. Additive Manufacturing

A cost model for the vat-photopolymerization technology, Digital Light Processing (DLP) [4], has not at this time been developed or discussed in literature. In this section, it is introduced a cost estimation model for an insert manufactured with DLP technology.

DLP is one of the most accurate AM technology available on the market [2], [3]. A DLP machine is characterized by a vat filled with a photopolymer liquid resin and an horizontal platform that moves vertically on which the part is attached and lifted horizontally from the bottom to the top from the vat. A digital light projector, project the object geometry layer-by-layer in the resin vat, which absorb the light and solidify the material following a photopolymerization process. Fabrication on a DLP machine involves three main steps starting from (1)

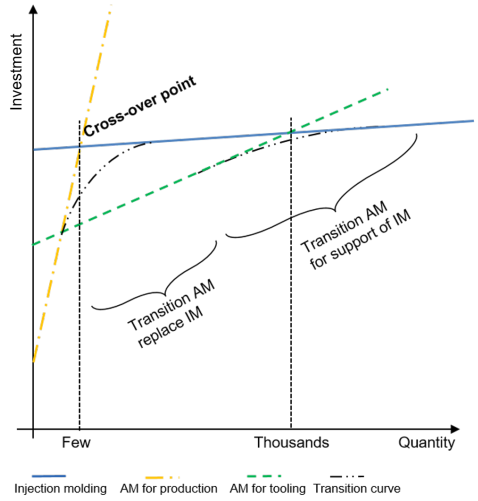


Fig. 3. Cost curve of AM when either replace or support conventional manufacturing

pre-processing that includes setting the building environment by determining and choosing the appropriate material, layer thickness, and build style; (2) verify the correctness of the STL file and resolution of any errors; (3) setting the appropriate part orientation in the build platform by considering the trade-off between surface quality and building speed; (4) generation of necessary support structures for the part; (5) finally, the production step is carried out, specifically a top-down layer-by-layer manufacturing process. (6) Last step, post-processing includes cleaning the part from liquid resin; removing support material and drying the part; post curing the part into a UV light oven; sanding, grinding or machining to achieve the desired surface finish.

The cost of an AM part, manufactured with DLP technology is made up of five main cost elements: pre-processing cost; build cost; material cost; post-processing cost, and overhead cost, as shown in Eq. (1):

$$C_{DLP} = C_{\mathrm{P_p}} + C_{\mathrm{Pr}} + C_{\mathrm{M}} + C_{\mathrm{Pp}} + C_{\mathrm{O}} \tag{1}$$

where C_{DLP} is the cost of a part fabricated on a DLP machine ($€/part$), $C_{\mathrm{P_p}}$ is the total cost for preparatory activities, C_{Pr} is the production cost including investment in machinery, C_{M} is the material cost, C_{Pp} is the total cost for post-processing activities, and C_{O} is the sum of all other expenses. Cost components for AM vary depending on the part size, level of surface finish, and material selection. Equation (2) is used to calculate the cost associated with pre-processing activities:

$$C_{\mathrm{P_p}} = \frac{C_{\mathrm{Op}}}{\mathrm{N}} \times T_{\mathrm{S}} \tag{2}$$

where C_{Op} is the hourly cost for an operator ($€/hour$) including overheads. T_{S} is the time ($hour$) spent to set up the print job and N is the number of parts that can be printed in the same build. The cost to run the DLP machine and print an insert is estimated as shown in Eq. (3).

$$C_{\mathrm{Pr}} = C_{\mathrm{Machine_{Hour}}} \times \frac{T_{\mathrm{B}}}{\mathrm{N}} \tag{3}$$

where $C_{\mathrm{Machine_{Hour}}}$ is the hourly cost of the machine ($€/hour$) and T_{B} the time ($hour$) to build a part. The error factor, δ, is attributed to the build time in order to consider possible failures in the printing process of a part and it is accounted as follows: $T_{\mathrm{B}} = T_{\mathrm{B}}/(1-\delta)$. Errors during the production process depends mainly on the level of experience the operator. Based on observation and collection of data within the case company, the error is within a range of 5% to 20%. Since N parts can be built simultaneously, the production cost is distributed over the N parts fabricated per build. Equation (4) present the hourly machine cost, $C_{\mathrm{MachineHour}}$:

$$C_{\mathrm{Machine_{Hour}}} = \frac{C_{\mathrm{I}}}{\mathrm{Y}} \times \frac{1}{\sigma \times \mathrm{H}} \tag{4}$$

where C_{I} is the cost to purchase the AM machine (€), and Y is the machine lifetime ($year$). The factor σ is used to account for the machine utilization, while H is the total number of hours the machine can work in a year ($hour/year$). Equation (5) present the estimation for the cost of material.

$$C_{\mathrm{M}} = C_{\mathrm{M_{kg}}} \times \mathrm{V} \tag{5}$$

where $C_{\mathrm{M_{kg}}}$ is the cost per kilogram of material ($€/kg$), and V is the total volume used in the build ($liter$). Based on observation the volume filled by the part is around the 70% of the resin volume used which includes the use of support structure and the amount of resin wasted.The simple equation for the volume $V = V_{\mathrm{Part}} + V_{\mathrm{Support}} + V_{\mathrm{Waste}}$ can be summarized as $V = (V_{\mathrm{Part}}/0.7)$. Equation (6) introduce the calculation for the cost of post-processing.

$$C_{\mathrm{Pp}} = \frac{C_{\mathrm{Op}} \times T_{\mathrm{Pp}}}{\mathrm{N}} \tag{6}$$

where the time for post-process a part, T_{Pp} is the sum of the time to remove support material T_{SR}, the time to post curing T_{Pc}, the time required to dry a part $T_{\mathrm{P_d}}$, and the time required for machining T_{M} as shown in Eq. (7):

$$T_{\mathrm{Pp}} = T_{\mathrm{SR}} + \alpha \times (T_{\mathrm{Pc}} + T_{\mathrm{P_d}} + T_{\mathrm{M}}) \tag{7}$$

Since post-processing activities can be performed over a number N of parts, the cost can be distributed over the number of parts that are built simultaneously. The multiplicative factor α is used to consider that the actual amount of time an operator is required to perform post-processing activities.

The last cost component considered in this model is the overhead cost. In the cost model of [15] overheads are neglected because contributed less than 1% of total costs, while in [16] model it is discussed how overheads contribute for a 10% of total costs. The importance of considering overhead costs rely on the fact that the increasing level of automation in manufacturing decrease manual labor but inversely raise the importance of overheads. Therefore, overheads cost are simply

TABLE I. DETAILED COST BREAKDOWN FOR DLP

	Insert type	
	Medal	Micro
Number of parts per build	1	[1; 3]
Operator cost (€/$hours$)	68	68
Set-up time (min)	30	30
Pre-processing cost (€/$part$)	34	[11,33; 34]
Purchase machine cost (€)	[50k; 100k]	[50k; 100k]
Depreciation time ($years$)	5	5
Utilization (%)	90	90
Hours per year ($hours$)	8760	8760
Machine cost per hour (€/$hour$)	[1,27; 2,54]	[1,27; 2,54]
Time to build	[10; 11][1]	[2; 3][1]
Printing failure (%)	[5; 20]	[5; 20]
Production cost (€/$part$)	[13,35; 34,88]	[2,67; 3,17]
Cost of material per kg (€/kg)	400	400
Part volume (lt)	0,044	0,0099
Volume of the build (lt)	0,063	[0,001; 0,0004]
Material cost (€/$part$)	22,54	0,57
Support removal time (min)	[30; 60]	[30; 60]
Post curing time (min)	[15; 30]	[15; 30]
Drying time (min)	[15; 30]	[15; 30]
Machining time (min)	[15; 30]	[15; 30]
Manual labour in Post-processing (%)	[10; 20]	[10; 20]
Post-processing time ($hours$)	[1,25; 2,50]	[1,25; 2,50]
Post-processing cost (€/$part$)	[39,10; 88,40]	[29,47; 39,10]
Yearly maintenance cost (€/$year$)	[6250; 12500]	[6250; 12500]
Hourly maintenance cost (€/$hour$)	[0,79; 1,59]	[0,79; 1,59]
Maintenance cost (€/$part$)	[0,08; 0,14]	[0,40; 1,59]
Electricity consumption (KWh)	[0,35; 0,60]	[0,35; 0,60]
Cost per KW (€/KW)	0,091	0,091
Cost of electricity (€/$part$)	[0,32; 0,60]	[0,05; 0,06]
Cost per square meter (€/m^2)	[15; 20]	[15; 20]
Machine space (m^2)	12	12
Cost of space rental (€/$part$)	[0,002; 0,003]	[0,010; 0,025]
Overhead cost (€/$part$)	[0,40; 0,75]	[0,47; 1,67]
Cost per insert (€)	**[109,39; 180,57]**	**[28,84; 76,81]**

the sum of three components as illustrated in the following equation:

$$C_{\mathrm{O}} = \frac{C_{\mathrm{El}} + C_{\mathrm{Sr}} + C_{\mathrm{Me}}}{\mathrm{N}} \tag{8}$$

where C_{El} is the cost for electricity usage (€/$part$), can be estimated as $C_{\mathrm{El}} = (E \times C_{\mathrm{E}})/\mathrm{P}$ where E is the amount of electricity consumed by the machine (KW/hour), C_{E} is the hourly cost of kilowatt consumed (€/KW), and P is the hourly productivity of the machine (part/hour).

The cost for space rental C_{Sr} (€/$part$) can be obtained from a simple relation between the space occupied by the machine S (m^2), the cost per square meter utilized C_s (€/hour), the amount of hour the machine work in a year and the hourly productivity: $C_{Sr} = (S \times C_s)/(\sigma \times H) \times P$.

Lastly, the maintenance cost per part can be estimated as $C_{\mathrm{Me}} = (C^y_{Me} \times P)/(\sigma \times P)$. The yearly maintenance cost C^y_{Me} (€/year of an EnvisionTEC P3 MML is linked to a support contract equal to the 12% of the purchase machine cost per year. Table I presents a detailed cost breakdown for the medal and micro-inserts manufactured with DLP technology.

B. Injection Molding

The cost components of injection molding were investigated to understand what are the main sources of costs to manufacture plastic molded parts and quantify the impact of AM to lower the total cost. In the cost model it is assumed that 100 to 200 new parts are designed and tested each year in the product development phase. For each new developed part, around 100 injection shots are performed to evaluate each part before final production starts. This is an important figure to keep in mind when the cost per molded plastic part with CNC machine mold inserts is compared with AM inserts.

There are three main cost elements that largely contribute to the manufacturing cost of an injection molded plastic part: mold cost; material cost; and production cost, as shown in Eq. (9).

$$C_{\mathrm{IM}} = C_{\mathrm{T}} + C_{\mathrm{M}} + C_{\mathrm{Pr}} \tag{9}$$

where C_{IM} is the unit cost of a molded part (€/$part$), C_{T} is the total tooling manufacturing cost per part (€/$part$), C_{M} is the material cost (€/$part$), and C_{Pr} is the processing cost (€/$part$).

The three cost components for injection molding are influenced by the part size, part geometry and level of complexity, and material selection. Moreover, cost per part is highly influenced by the production volume since the depreciation of expensive investment in tooling is distributed over the production volume. At low production volumes, the proportion of cost due to the use of CNC machining to fabricate molds in brass or steel is extremely high while inversely at larger production volumes, the ratio of cost due to tooling decreases and material and production cost became the major cost components.

Mold cost is mainly influenced by the cost of the mold base and the cost to machine the inserts. Based on the number of cavities in the mold, an equivalent number of inserts are machined. Inserts are dependent on the part size and geometry, and on the complexity of fabricating cavities and cores and the level of surface finish required. The mold cost is calculated as shown in Eq. (10):

$$C_{\mathrm{T}} = \frac{C_{\mathrm{M_B}}}{\mathrm{LC}} + \frac{C_{\mathrm{In}} \times n}{B} \tag{10}$$

where $C_{\mathrm{M_B}}$ is the cost of a mold base (€), C_{In} is the cost to manufacture one insert (€/$insert$), n is the number of cavities, LC is the life cycle of the mold base ($parts$) and B is the batch size or number of injection shots for which the inserts is used for. Two are the possible strategies to depreciate the investment in tooling: depreciate the insert cost over the inserts technical life which for the CNC machined brass or steel insert is between 100,000 and 500,000 injection shots; depreciate the insert cost over a fixed number or the production volume for which the insert is used. Since the analysis is carried out within the pilot production, it is chosen to depreciate the mold base cost over its lifetime equal to around 2 million injection shots. Instead, the inserts are amortized over the average number of injection shots for which they are used.

Material cost is a major factor to take into consideration. Molded part costs also depend on the choice of materials. Part weight is calculated considering the sum of the part, runners, and sprue weights to account for the material used to mold the part, but also wasted for the runners and sprues as calculated in Eq. (11):

[1]Obtained considering the nominal machine building speed equal to $\frac{1\ \mathrm{min}}{layer}$ that print layers down to 16 μm.

$$C_{\mathrm{M}} = C_{\mathrm{M_{kg}}} \times W \tag{11}$$

where $C_{\mathrm{M_{kg}}}$ is material cost per kilogram of material used ($€/kg$), and W is the total weight in kilogram. Part processing cost is the sum of the expense related to the amount of time the operator spend to run the injection molding machine and the cost of the machine per part. Equation (12) determines the processing cost per part:

$$C_{\mathrm{Pr}} = C_{\mathrm{h}} + \frac{C_{\mathrm{m}}}{Cap_h} + \frac{\alpha \times C_{\mathrm{O}}}{Cap_{\mathrm{h}}} \tag{12}$$

where C_{h} is the cost to set-up and run the injection molding machine ($€/part$) and is calculated as the production of hourly operator cost and the sum of time to setup a mold and the time the machine is not running because the mold has been extracted, distributed over the batch size B. Cap_{h} is the hourly capacity of the machine ($part$) considering that a mold fabricated with n cavities will produce n parts over the cycle time. C_{O} is the hourly cost of an operator ($€/hour$). The analysis carried out in the case company have shown an average utilization (α) of the workers between 10% to 20%. Equation (13) and (14) introduce the calculation of C_{h} and Cap_{h}:

$$C_{\mathrm{h}} = \frac{(T_{\mathrm{S}} + T_{\mathrm{D}}) \times C_{\mathrm{O}}}{B} \tag{13}$$

$$Cap_{\mathrm{hr}} = \frac{3600 \times n}{CT} \tag{14}$$

where T_S is the time to setup the machine, T_D the time the mold is down, and CT the cycle time.

Table II presents a detailed cost breakdown for injection molding when the medal and micro inserts are CNC machined or additively manufactured. Some additional consideration must be drawn before proceeding with the discussion of AM cost advantages. Mold cost are obtained from industry average and inserts are depreciated over the batch size since the analysis is bench-marked within the pilot production. To obtain the insert cost for AM is necessary to examine the insert lifetime which is set to 50 successful shots for the medal and 100 for the micro-insert. The 12 times longer cycle time for the medal inserts reflect the need for a longer cooling time. Accounting for a longer cooling time helps to protect the the medal AM insert against thermal deterioration.

III. Results

The cost model proves the potential cost savings in applying AM inserts for pilot production when relatively small batches of parts are molded. For small batch size of 100 shots, for each new part that is designed and tested before final production the larger cost component is the investment in tooling. Figure 4 shows how the three cost component affect the total cost per part when inserts are machined or fabricated with AM. For a part molded with the medal insert manufactured in brass, the cost for tooling is the 97% of the total cost per part, while with AM inserts is equal to 78% of the total cost. Moreover, the tool cost decreases by 80% when AM is used instead of CNC machining, while the processing cost increases by 69%. The increase in processing cost with AM inserts is due to the longer cycle time required to efficiently cool down the insert. Material cost in both cases account for a slight percentage, 0,002 and 0,09 % of total cost respectively.

TABLE II. Detailed cost breakdown for injection molding.

	Insert type			
	CNC		AM	
	Medal	Micro	Medal	Micro
Mold base cost (€)	23522	23522	23522	23522
Insert cost (€)	4369	1331[2]	[437,58; 722,28]	[115,36; 307,24][2]
Number of cavities	2	4	2	4
Batch size	100	100	100	100
Tool cost ($€/part$)	43,70	13,32	[4,39; 7,23]	[1,17; 3,08]
Material cost per kg ($€/kg$)	1,74	1,74	1,74	1,74
Weight (gr)	5,5	1,1	5,5	1,1
Material cost ($€/part$)	0,01	0,002	0,01	0,002
Setup time (hours)	1	1	1	1
Down time (hours)	1	1	1	1
Operator cost ($€/hour$)	68	68	68	68
Handle cost ($€/part$)	1,36	1,36	1,36	1,36
Cycle time ($second$)	25	20	300	50
Hourly capacity ($part/hour$)	288	720	24	288
Machine cost ($€/hour$)	12,60	12,60	12,60	12,60
Production cost ($€/part$)	1,45	1,40	2,45	1,45
Cost per part ($€/part$)	**45,16**	**14,72**	**[6,85; 9,70]**	**[2,62; 4,5]**

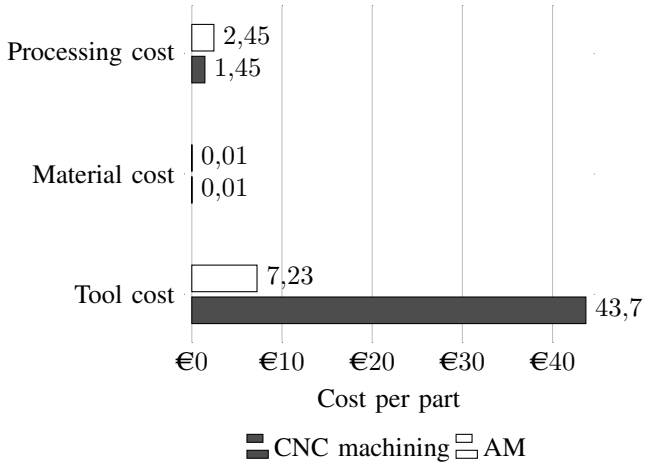


Fig. 4. Cost per part comparison with CNC machined and AM inserts

To assess the integration of AM for inserts production, it is performed a break-even-analysis that is helpful to identify how an increment in the batch size would affect the cost per part in the pilot production and the cost effectiveness of AM. Figure 5 shows the break-even point for molded parts when inserts are fabricated with AM and CNC machining. The chart of figure 5 shows that with a batch size up to 810 parts and a unit cost of €5,65, it is cost effective to mold parts with photo-polymer inserts. After that point, then it is cost effective to machine brass inserts. It is possible to notice that for the micro-insert the break-even point is much higher, 3900 units, due to several factors including the possibility to fabricate n

[2]Cost obtained from industrial standards and equal to €665,32 to manufacture 2 inserts with a lead time of 6 $hours$.

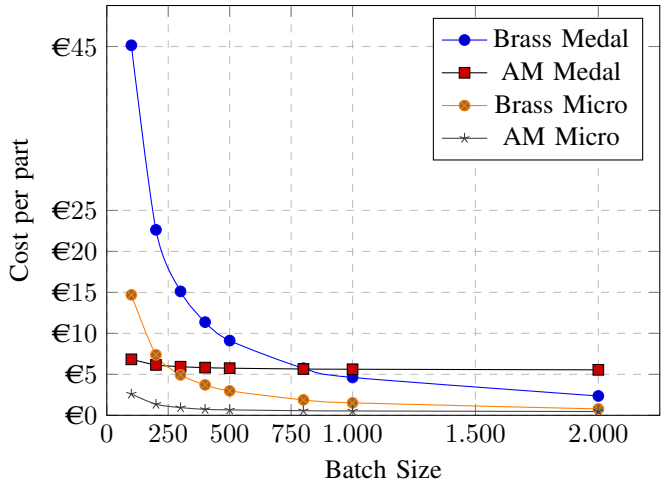


Fig. 5. Break-Even-Analysis of AM inserts for the product development phase.

parts simultaneously, and the lower production time required.

IV. Conclusion

This research contributes to the analysis and development of additive manufacturing adoption within the conventional manufacturing process of injection molding. It is of a particular interest in order to facilitate companies to research how to exploit the advantages that AM can offer. This article elaborates on the possible synergies between AM and conventional manufacturing to create added value by decreasing operational cost in the pilot production.

The impact of layer manufacturing technologies has been a largely discussed topic in the literature, where it has been argued that the way products have been manufactured until now it is meant to be overturned by AM. However, it is interesting to evaluate another possible scenario in which AM is not conceived as a disruptive technology but instead as a set of processes that have the potential to improve a production process instead of replacing it.

One of the most important points that emerged from the analysis is the indisputable power of AM to decrease substantially the manufacturing cost. The main topic researched within this article is the integration of AM for soft tooling. The work of this research demonstrates the possibility to reduce manufacturing cost from €4369 to [€437,58; €722,28] for the medal batch and from €1331 to [€115,36; €307,24] for the micro batch ([-83%; -90%] and [-77%; -91%]). On-demand production of tooling is now a reality and the investment in tooling is remarkably reduced as cost for manufacturing mold inserts are substantially decreased.

Even though AM has the potential to optimize the operational efficiency of a firm, it is important to highlight that the current technological limitations can still slow down its adoption in some department and areas of the job floor. A reflection on the current restrictions of the technology is necessary to correctly establish where and how layer manufacturing techniques can be applied, for example when more restrictive production requirements arise. Consequently, the main areas where the technology has to improve are identified and discussed throughout the development of a cost estimation model, which highlights the key cost components. The cost model represents a source of novelty in the literature since there were not other resources available discussing the cost to fabricate parts with DLP technology and especially integration of AM with injection molding.

Acknowledgment

The authors thank the HEC Paris Business School (France) for the partial financial support used to conduct this research and an international manufacturing company that allowed to conduct the case study on additive manufacturing. The Marie og M. B. Richter's Fond (Denmark) is acknowledged for supporting the present research.

References

[1] D. Wee, R. Kelly, J. Cattel, and M. Breunig, "Industry 4.0 - How to navigate digitization of the manufacturing sector," *McKinsey & Company*, pp. 1–62, 2015.

[2] T. Wohlers, T. Caffrey, W. Associates, and R. Campbell, *Wohlers Report 2016: 3D Printing and Additive Manufacturing State of the Industry : Annual Worldwide Progress Report*. Wohlers Associates, 2016.

[3] I. Gibson, D. Rosen, and B. Stucker, *Additive manufacturing technologies: 3D printing, rapid prototyping, and direct digital manufacturing*. Springer, 2014.

[4] ASTM International, "Standard Terminology for Additive Manufacturing Technologies General Principles Terminology," *Iso/Astm 52900*, vol. i, pp. 1–9, 2015.

[5] M. Hall and M. Krystofik, "Conformal Cooling," Rochester Institute of Technology, Tech. Rep., 2015.

[6] Stratasys, "Polyjet for injection molding," 2015, AB_PJ_InjectionMolding_EN_0915.

[7] ——, "PolyJet Enables Functional Testing of Injection Molded Prototypes Faster and Less Expensively Than Ever Before," 2015, ACS-Injection Molding-1-14.

[8] ——, "Medical Prototyping - Custom Injection Molder Embraces 3D Printing," 2015, CS_PJ_Med_DiversifiedPlastics_EN_0915.

[9] L. Zonder and N. Sella, "Precision Prototyping - The Role of 3D Printed Molds in the Injection Molding Industry," 2015, WP_PJ_InjectionMolding_EN_1014.

[10] Stratasys, "Durable Prototyping - 3D Printing Builds Tough Injection Molds Fast," 2015, cS_PJ_CM_Turck_0616a.

[11] M. Mischkot, H. N. Hansen, and D. B. Pedersen, "Additive manufacturing for the production of inserts for micro injection moulding," in *Proceedings of Euspen's 15th International Conference & Exhibition*, 2015.

[12] M. Mischkot, G. Tosello, D. K. Y. Nielsen, D. B. Pedersen, Y. Zhang, T. Hofstätter, L. Herbin, and H. N. Hansen, "Injection moulding pilot production: Performance assessment of tooling process chains based on tool inserts made from brass and a 3d printed photopolymer," in *Annual Technical Conference (ANTEC2017) of the Society of Plastic Engineers*, Anaheim, CA (USA), May 2017.

[13] G. Tosello, "Digital manufacturing technologies for precision injection moulding," *Manufacturing Academy of Denmark (MADE) Innovation Conference, Flexible and Rapid Production with Digital Production Technologies, Aarhus (Denmark)*, Feb. 2016.

[14] B. Grynol and Deloitte (Firm), "Disruptive manufacturing: The effects of 3d printing," 2014.

[15] N. Hopkinson and P. Dicknes, "Analysis of rapid manufacturing using layer manufacturing processes for production," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 217, no. 1, pp. 31–39, 2003.

[16] M. Ruffo, C. Tuck, and R. Hague, "Cost estimation for rapid manufacturing-laser sintering production for low to medium volumes," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 220, no. 9, pp. 1417–1427, 2006.

3D Printing Technology: Are we ready for the next step?

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Prodromos D. Chatzoglou [*1,2]
[*1]Production & Management Engineering Department
Democritus University of Thrace
Xanthi, Greece
pchatzog@pme.duth.gr

Vassiliki N. Michailidou [*2]
[*2]Business Administration Department
Hellenic Open University
Patras, Greece
vassmich@hotmail.com

Abstract—3D printing technology redesigns the production chain and contributes to the production of customized products, or products with specific characteristics. The evolution of this production technology and its various applications improve our daily life and lead to important changes in economic and social level. The aim of this dissertation is to examine the attitude of people towards 3D printing technological innovation and their intention to use it at their workplace. A new conceptual model has been designed and empirically tested using primary data collected using a structured questionnaire. The sample consists of 258 respondents from a number of European countries. The findings highlight the importance of perceived usefulness, output usability, compatibility and attitude towards understanding respondents' intention to use 3D printing technology.

Keywords—3D Printing Technology; Technology Acceptance Models; Attitude; Output Usability; Perceived Ease of Use; Perceived Usefulness.

I. Introduction

Human history has always been affected by technology probably much more than any other field. New technologies have improved life in different ways, opening up new possibilities in many different areas. However, it must be realized that time is required, years or even decades, before the technology becomes apparent and useful.

For most of the people, the definition of printing involves the process of putting ink on a piece of paper. By contrast, 3D printing is the process of using a machine to create layer by layer in three dimensions the desired object. More specifically, a 3D printer is a "printing" machine that can turn a design created using a computer program into a physical solid object [1] in three dimensions; width, height and depth [2] and involves creating products layer by layer using a variety of different materials [3].

In the mid-1980s, Charles Hull and Scott Crump, both pioneers, started developing a range of technologies now known as 3D printing [4]. 3D Systems and Stratasys were the two major companies in the 3D printing field and rapid prototyping [5]. Within the next thirty years, 3D printing technology evolved and developed very quietly.

Although 3D printing might lead the industry in a few years to a new revolution its current impact is rather limited. Some of the most promising areas for 3D printing implementation include medical applications, custom parts replacement and customized consumer products. 3D printing, as an additive manufacturing process could transform the manufacturing process in many critical ways. The nature of 3D technology permits new and different ways of thinking in term of economic, environmental and social implications. This could lead to better products that competitors will not be able to match without also adopting the new design and manufacturing process [6]. Moreover, 3D printing is widely believed that has an enormous potential to become one of these technologies that affect everyday life since it will bring production closer to the customer. Consumers could also have 3D printers at home to manufacture spare parts for household items [7]. However, the crucial question is "Are we ready for the next step?"

The aim of this dissertation is to examine the attitude of people towards 3D printing technological innovation and their intention to use it at their workplace.

II. Research Model and Hypotheses Development

A. The Proposed Research Model

The technology acceptance concept has been examined by different theories in an attempt to explain users' acceptance of various technologies. Three of the most cited of these theories are the Technology Acceptance Model, known as TAM [8], [9], the Theory of Planned Behavior, also known as TPB [10], [11] and the Diffusion of Innovation Theory, known as DOI [12].

This research incorporates different components used by TAM, TPB and DOI, into a new and unique theoretical framework that is empirically tested. The proposed theoretical framework is aiming to explain the acceptance of 3D printers and the effect of the factors that influence the acceptance of the particular technology.

The dependent variable is the behavioural intention to use 3D printers. The first set of constructs considered to influence behavioural intention to use 3D printers consists of attitude and behavioural beliefs. It also includes compatibility, relative advantage, job relevance, output quality and output usability. The reason for using compatibility, as well as relative advantage, is to examine an individual's perception of the benefits of using a 3D printer and, consequently, his/her attitude towards using a 3D printer. High compatibility will

adversely affect potential adopters and will increase the likelihood of adoption. In the context of 3D printers, knowledge compatibility is the extent to which adopters believe 3D printers can be integrated into their daily lives. Relative advantage refers to the degree to which an innovation is perceived to be better than earlier innovations of similar nature i.e. 2D printers. This relative advantage can be expressed in terms of economic factors, efficiency, or other types of benefits. Job relevance, output quality of the product produced, as well as output usability are all directly connected with perceived ease of usefulness of using a 3D printer and would have a strong positive impact on the particular framework. Job relevance would determine if individuals believe that 3D printing technology would be important for their job, while output quality and usability would determine if individuals believe the 3D printing technology would help them in their job.

The second set of constructs (considered to influence behavioural intention) includes experience regarding 3D printers, attitude towards using 3D printers, as well as the usefulness and use of a 3D printer. Prior knowledge and experience of 3D printers and technology would have a positive influence on the attitude of individuals towards adopting using 3D printers. An innovation, such as 3D printers, is anticipated to be of high usefulness when the potential adopters believe that there is a direct relationship between use, on the one hand, and productivity, performance, effectiveness or satisfaction, on the other. Usefulness recognition would be important because there is a strong direct effect on the intention of individuals to use the innovation. As far as perceived ease of use is concerned, it represents the degree to which individuals associate usage difficulty with the use of 3D printers in everyday usage. If the use would be threatening to individuals, that would have a negative influence on the intention to use 3D printers.

B. Hypotheses Development

The first hypothesis refers to the link between compatibility and attitude [13]. In the context of using a 3D printer, if an individual considers that this will offer at least the same level of benefits as other current technologies, then the individual is likely to have a more positive attitude towards using 3D printers. However, if the individual consider using a 3D printer not to be on par with other technologies, then the individual will have a more negative attitude towards using 3D printers.

Hypothesis 1: The compatibility of 3D printers with current printing technologies has a positive direct effect on the attitude towards using 3D printers.

When an innovation is perceived to offer higher benefits, an individual's perception of positive outcomes from using that innovation will increase. The degree to which an innovation is perceived to offer better benefits than earlier innovations of a similar nature is referred to as relative advantage [14]. The second proposed hypothesis would be that relative advantage has a strong positive effect on attitude. A potential adopter's positive or negative evaluation of using 3D printers is directly affected by the perceived improvements in benefits gained from using 3D printers.

Hypothesis 2: The relative advantage offered by using 3D printers over other technologies has a significant positive direct effect on the attitude towards using 3D printing technology.

Job relevance determines if an individual thinks that an innovation makes sense and would be relevant for his job [15]. In other words, job relevance would have a moderate effect on an individual's perceived use of usefulness. Usefulness of a technology would be essential if it would increase an adopter's productivity by being relevant to his job [16].

Hypothesis 3: Job relevance has a direct positive effect on perceived usefulness.

Output quality determines the degree that an individual thinks the technology helps them do their job well [15]. When an individual would believe that the use of the particular technology would help him to proceed with his job in a more efficient way that would have a strong positive effect on perceived usefulness [17].

Hypothesis 4: Output quality has a direct positive effect on perceived usefulness.

Usability output involves a comparison of the new innovation technology used versus other already existing technologies based on the actual level of effort required to complete specific tasks [18]. If the effort is relatively low, or acceptable according to the result, there will be a positive effect on perceived ease of use [17].

Hypothesis 5: Usability output has a direct positive effect on perceived ease of use.

Venkatesh and Davis [19] have shown that even over a wide variety of jobs, the factors that affect technology acceptance vary as a function of experience with the system. Specifically, they found that more variance in perceived ease of use is explained at higher degrees of experience (60%) than at lower experience levels (40%). Thus, it is normal to expect that the experience of an individual would have a positive direct effect on his behavioural intention to use an innovation [16].

Hypothesis 6: Experience has a direct effect on the behavioural intention to use 3D printers.

In the technology acceptance literature, attitude is defined as a person's positive or negative evaluation of performing a type of behaviour [20]. If an individual feels that there will be positive consequences from using 3D printers then the individual will have more intention to adopt 3D printers. On the other hand, feelings of negative or not important consequences lower adoption intention. In this study, the proposed hypothesis draws a link between attitude and behavioural intention in the context of 3D printers (Davis et al., 1989).

Hypothesis 7: Attitude towards the use of 3D printers has a positive direct effect on the intention to use 3D printers.

Perceived usefulness is the perception that a given technology will help an individual achieve his or her work goals. Within the context of adopting and using a new technology at the workplace, [21] provide evidence that the most important determinant of an employee's attitude toward

adopting and using a new technology is his perception of the usefulness of the technology. Employees are much more likely to adopt a system that they believe will help them achieve their work goals. If the technology is perceived to be useful it is associated with increased behavioural intention to use the technology [9].

Hypothesis 8: Perceived usefulness has a direct positive effect on the intention to use 3D printers.

The final hypothesis involves that ease of use will have a positive direct effect on perceived behavioural control in the 3D printers context. The amount of effort that is required to use this innovation will have an effect on the adopter's sense of self-efficacy with regards to using the innovation. If individuals find that they can comprehend fairly easily how to use an innovation, they are more likely to feel confident about using the innovation effectively [14]. This implies, in the context of 3D printers, the degree of simplicity or understanding that use of 3D printers provides will significantly influence a person's confidence about his or her ability to use 3D printers.

Hypothesis 9: 3D printers' ease of use has a significant positive direct effect on the intention to use 3D printers.

III. Research Methodology

In order to study the acceptance of 3D printers, a survey was conducted where a number of people were asked to complete an online structured questionnaire. In some cases, face-to-face interviews took also place to ensure that people who are not very familiar with the specific technology, but do understand and use the results of technological innovations such as 3D printers, are also included.

A. Questionnaire Design

To ensure content validity, the questionnaire used in this study was adapted from the original measurement scales used in TAM and other models found in literature (TPB,DOI) with some modifications, the necessary wording changes and validation to fit the context of 3D printers' usage. The questionnaire was also translated into Greek because most of the participants are native Greek speakers.

The research instrument consists of two main sections. The first section contains specific and simple demographic questions mainly using the nominal scale. It contains 13 items such as gender, age, working experience, educational background and current employment situation.

TABLE 1: Construct Measurement

Constructs	Items	Sources
Compatibility (COM)	3	[14], [22], [13]
Relative Advantage (REL)	4	[14]
Job Relevance (JR)	5	[16]
Output Quality (OQ)	3	[23] [17]
Output Usability (OU)	3	[23], [17]
Experience (EXP)	4	[16]
Attitude (ATT)	4	[8], [24], [25], [26], [27], [22]
Perceived Usefulness (PU)	6	[8], [9]
Perceived Ease of Use (PEU)	5	[8], [9],[14]
Intention to Use (BIU)	3	[22], [9], [19]

The second section includes questions about the main factors incorporated into the proposed theoretical framework (research model), which are measured using a 5-point Likert scale (1: Strongly disagree, 2: Slightly disagree, 3: Neither agree nor disagree, 4: Slightly agree, and 5: Strongly agree). As shown in Table (1), the total number of questions used to measure these factors is 40.

B. Sampling Approach and Sample Characteristics

The sample of this survey is consisted of people living in European countries (mainly in the UK, Finland, Germany and Greece), who are well educated and/or qualified people in in the use of technology. The sampling techniques used in the present study are the non-probability convenience and snowball sampling. Non-probability convenience sampling is probably the most common of all sampling techniques. On the other hand, the snowball approach is mainly used when online research instruments for collecting primary data are used via social media.

The final actual sample consists of 258 valid responses (15 more were discarded because they were incomplete). From those, 52,3% belongs to male participants, in the 30-45 or 46-55 years old age groups (53,5% and 32,6% respectively), who are very well educated since they are either university graduates (39,5%) or hold a postgraduate degree (41,9%). Almost all of them are full time employees (90,7%), with more than 10 years of working experience (74,4%) at a managerial position (45,3%) or as an employee (34,9%), in various departments and business sectors. The size of the companies they are working for are small (52,3%), medium (11,6%), large (16,3%) and very large (19,8%).

It should be stressed that most of the participants (83,7%) do not use 3D printers in their job, while 90,7% do not use 3D printers at all. This indicates that, although 7% do have 3D printer in the company they work for, they do not use it at all, probably because the 3D printer is in another department or they do not know how to use it. Although these numbers are low, it should be taken into consideration that the new technology of 3D printers is indeed very innovative and still in a very limited use, at least for most of the industrial sectors. This is why the 16,3% of the business that use 3D printers is a rather good number, since the main aim of this research is to examine the behavior of all the people towards the use of 3D printers.

IV. Results

The content validity test of the questionnaire was conducted prior to the beginning of the survey and included: (a) consultation with academics of the field, (b) consultation with experienced practitioners, and (c) pilot testing. The construct validity test was conducted in two steps. Each of the research factors was evaluated (a) for its unidimensionality and reliability, (b) for the goodness of fit to the proposed research model.

A confirmatory factor analysis has been initially conducted to assess the validity of the constructs included in the research model [28]. The Kaiser–Meyer–Oklin (KMO) measure of sampling adequacy and Barlett's test of sphericity are the two measures that are recommended for measuring construct

validity [29], while [30] point out that the Cronbach's alpha reliability test can be used for assessing the internal consistency of the measurements. Total variance explained (TVE) is also used to measure the percentage of the common variance that is explained by all the factors.

The results of the confirmatory factor and reliability analyses (Table 2) show that KMO is above the 0.5 threshold [29], while the TVE score for all the factors is very satisfactory, above 0.5 [31] and the factor loadings for all the items are within acceptable levels (above the 0.5 threshold; [29]). Finally, the Cronbach's alpha is also above the 0.6 threshold [32] for all the factors. Overall the results of this analysis indicate that the factors included in the theoretical model are valid and reliable and, therefore, can be used to test the overall model fit (using the SEM approach).

Next, correlation analysis (Spearmans' r_s), was performed to investigate not only the correlation between all the factors proposed in this research, but also the correlation between these factors and some of the most important demographic variables. Thus, initially it is found that there is a moderate or even a strong correlation between all the constructs incorporated in the proposed research model. The only exception is "Experience" which is only related (weak) with "Attitude" and "Perceived ease of Use". This is a very interesting result indicating that it is irrelevant whether people are experienced or not with the intention to use the new technology. Then, it is found that gender, family status, job experience, job status and position are not related with any of the factors included in the proposed research model (Table 3). On the other hand, the age of the respondents, the department they are working in, the industrial sector their firms belong in and its size are related with many of the main factors examined. However, as it was expected, there is a strong relationship between these factors and the ownership of a 3D printer as well as its actual use. Actually, the results of this analysis are similar with the results of the mean test where it was found that the scores of the factors significantly differ between respondents who own or use a 3D printer or not.

TABLE 2: Factor and Reliability Analysis

	Mean	St.D.	Items	Factor Loading	K.M.O	T.V.E	Cronbach α-value
COM	3.22	1.147	1-3	0.903 - 0.950	0.739	87.303	0.926
REL	3.46	1.180	1-4	0.882 - 0.944	0.837	85.035	0.941
JR	3.60	1.085	2-5, (1)	0.864 - 0.924	0.844	70.539	0.927
OQ	3.41	1.147	1-3	0.916 - 0.923	0.757	84.372	0.906
OU	3.08	1.121	1-3	0.822 - 0.893	0.694	73.342	0.817
EXP	3.34	0.924	2-4 (1)	0.849 - 0.930	0.703	79.903	0.873
ATT	3.78	1.012	1-4	0.894 - 0.950	0.825	83.784	0.934
PU	3.42	1.027	1-6	0.739 - 0.934	0.880	76.469	0.938
PEU	3.62	0.737	1-3, 5 (4)	0.681 - 0.880	0.786	68.355	0.828
BIU	3.77	1.054	1-3	0.913 - 0.961	0.721	87.338	0.927

TABLE 3: Correlation Analysis

	Age	Educa-tion	Depar-tment	Indust-rial Sector	No. of Empl	Sales	Hold 3D printer	Use 3D printer
COM				,302			,473	,412
REL	,195			,288	,193		,513	,347
JR	,227		,275				,473	,265
OQ				,231	,228		,428	,304
OU			,184	,181	,192		,484	,285
EXP								-,187
ATT				,179		-,340	,459	,196
PU			,186	,225			,489	,278
PEU	,204	,186	,343	,288			,452	,283
BIU			,206	,289			,460	,286

All the above relationships are statistically significant at .05 level

Further, the examination of hypotheses was conducted with the use of the Structural Equation Modelling technique [33], [34]. The estimation of the structural model was conducted with the Maximum Likelihood Estimation method [29], [35].

According to [36] (p. 3) "overall fit is evaluated by how well the model explained all of the data in the entire analysis. Fit indices evaluate model fit for the data being examined". In this research the following five best known and commonly used overall model fit indices are used: the X^2/DF score (1.346), which is within the accepted limits, below 5 [37], the Goodness of Fit Index (GFI) score (.915), which is above the 0.90 minimum threshold [38], the Comparative Fit Index (CFI) score (.987), which is above the 0.9 threshold [36], the Tucker-Lewis Index (TLI) (.951), which is close to the 0.9 threshold [29], the RMSEA score (.054), which is below the 0.1 threshold and the Root Mean square Residual (RMR) value (.023), which is below the 0.1 threshold [39], [29]. All the aforementioned fit indices suggest that the overall fit of the extracted model is statistically significant and acceptable.

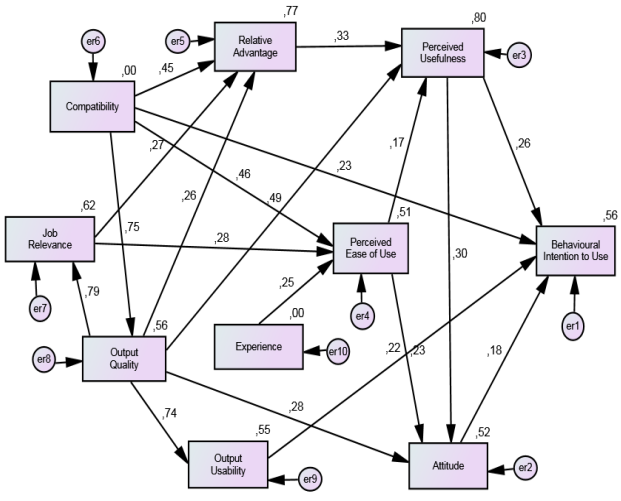


Figure 1: SEM Model

Figure 1 demonstrates the overall structural model, along with the extracted path coefficients and the adjusted R^2 scores. Overall, the final model can explain 56% of the variance of Behavioural Intention to Use, 80% of the variance of Perceived Usefulness, 52% of the variance of Attitude, and 51% of the variance of Perceived Ease of Use.

The results of the SEM analysis indicate that hypotheses H4, H7 and H8 are fully accepted, while hypotheses H1, H2, H3, H6 and H9 are partially accepted (since there is an indirect relationship between the factors examined), and only H5 is rejected. More specifically, H4, which refers to the Output Quality that determines the degree that an individual thinks the technology helps them do their job well [15] (Perceived Usefulness), is supported by the results of the SEM analysis. Individuals believe that the 3D printing technology would help to proceed with the job in a more efficient way which has a strong positive effect on Perceived Usefulness. Also, Output Quality is found to be very important, having a direct positive effect on Relative Advantage, Output Usability and Attitude towards using 3D printers. The next accepted hypotheses (H7 & H8) refer to the Attitude and the Perceived Usefulness with the Behavioral Intention to Use 3D printers. As Attitude is defined as a person's positive or negative evaluation of performing a type of behavior [20], when an individual feels that there will be positive consequences from using 3D printers, then, the individual will have stronger intention to adopt 3D printers. Also, it is found that Perceived Usefulness is associated with increased Behavioral Intention to Use.

Hypothesis H1 (link between Compatibility and Attitude) is partially accepted. It appears that, although Compatibility of 3D printers with current technologies has no direct effect on the Attitude towards using a 3D printer, it has an indirect effect through Perceived Ease of Use. On the other hand, it is directly related with Behavioral Intention to Use 3D printers, as well as with Relative Advantage, Output Quality and Perceived Ease of Use. The second partially accepted hypothesis (H2) refers to the link between the Relative Advantage and Attitude (through Perceived Usefulness). It is also found that it is directly related to Perceived Usefulness. Further, H3, which explores the possible effect of Job Relevance on Perceived Usefulness, is also partially accepted (through Relative Advantage). Job relevance is found to be directly linked to both Relative Advantage and Perceived Ease of Use. Moreover, H6 (effect of Experience on the Behavioral Intention to Use) is also partially accepted (through Perceived Ease of Use, Perceived Usefulness and Attitude). It is found that Experience is directly linked only to Perceived Ease of Use, which was expected since [19] have shown that, even over a wide variety of jobs, the factors that affect technology acceptance vary as a function of experience with the system. Specifically, they found that more variance in perceived ease of use was explained at higher degrees of experience (60%) than at lower experience levels (40%).

Finally, Output Usability (new innovation technology used versus other already existing technologies) is based on the actual level of effort required to complete specific tasks [18], seems to have no effect on Perceived Ease of Use (H5), while, on the other hand, it has a direct effect on the Behavioral Intention to Use 3D printers.

V. Summary and Conclusions

The technology acceptance concept has been examined by different theories in an attempt to explain users' acceptance of various technologies. We aimed to examine the 3D acceptance through a number of participants and the main and stronger outcomes are summarized below.

The majority of the participants (83,7%) do not have 3D printers in their job and a 90,7% do not use 3D printers at all, so our main aim to examine the behavior of all the people towards the use of 3D printers involves mainly people who do not use 3D printers yet.

Users' intention to use 3D technology is of paramount importance for both researchers as well as practitioners. A deeper insight into theory-based research is required to better understand the underlying motivators and barriers that will lead users to, or inhibit them from, adopting 3D printing technology. In this research we have explored and reviewed existing technology acceptance theories. Relevant constructs of existing models have been discussed in the light of the evolving 3D printing technology and, then, incorporated into an extended acceptance model, taking into account several perspectives including attitude, compatibility, perceived ease of use and perceived ease of usefulness . The proposed model attempts to view behavioral intention of use 3D printing. As technological advances are not slowing, technology acceptance will continue to be important. This study confirms some of the previous findings in the area of technology acceptance and proposes some new aspects that is worth investigating in future.

A very important finding was that the demographics are independent of the constructs and, consequently, the behavioral intention of a person to use 3D printers. Correlation analysis has shown that all factors have a moderate or strong correlation. A strong and positive effect was found between compatibility and relative advantage. People believe that 3D printers can be integrated into their daily lives also find them to have a relative advantage as compared with other printing techniques. Moreover, the output quality of the products produced by using 3D printers was found to have a strong and positive correlation with the relative advantage and the job relevance as well. With a more precise investigation to these constructs, managers may find some interesting correlation between compatibility, output quality and relative advantage with different business sectors regarding job relevance. Another very important find for managers, refers to the lack of correlation between experience and behavioral intention to use 3D printers. Managers may introduce 3D printing technology regardless the experience of the potential users.

This study has some limitations. First of all, the sample size is relatively small and comes from many European countries with differences in the technology acceptance readiness and, therefore, the findings may not be fully representative. Even though the survey was carried out both on medium and large companies in different business sectors, the part of people using the 3D technology was very small, so we should take into account that the results of this work are mainly derived by examining people who are not using 3D printing technology.

References

[1] Rideout, B., (2011). Printing the Impossible Triangle: the Copyright implications of Three-Dimensional Printing. Journal of Business. *Entrepreneurship & Law*. 5 (1): 163.

[2] Jones, S., McMorrow, M. and Stokes, G., (2011). Clone wars:3D printing and intellectual property, DLA Pipers, April 11, 1-8 [files.dlapiper.com/files/Uploads/Documents]

[3] Sissons, A. and Thompson, S., (2012). Three Dimensional Policy: Why Britain needs a policy framework for 3D printing. The Big Innovation Centre, 8.

[4] Srinivasan, V. and J. Bassan, J., (2012). 3D Printing and the Future of Manufacturing; Remaking Manufacturing. CSC Leading Edge Forum, 5.

[5] GridLogics Technologies Pvt Ltd. (2014). 3D Printing Technology Insight Report: An analysis of patenting activities from 1990-current. GridLogics Technologies Ltd.

[6] Berman, B., (2012). 3D Printing: The new Industrial Revolution. Business Horizons, 55: 155.

[7] Pannett, L., (2014). 3D: The future of printing. *Supply Management*. 19 (1), 34-37.

[8] Davis, F.D., Bagozzi, R.P., and Warshaw, P.R., (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, 35 (8), 982-1002.

[9] Davis, F., (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*. 13: 319.

[10] Ajzen, I., (1985). From intentions to actions: A theory of planned behaviour, In J.B.Kuhland (ed), Action-control: From cognitions to behavior(11-39), Heidelberg: Springer

[11] Ajzen, I. & Madden, T.J., (1986). Prediction of goal-directed behavior: The role of intention, perceived control and prior behavior. *Journal of Experimental Social Psychology*, 22, 453-474.

[12] Rogers, E.M., (1983), Diffusion of innovations, The Free Press, New York, NY.

[13] Karahanna, E., and Straub, D. W., (1999). The Psychological Origins of Perceived Usefulness and Ease-of-use. *Information and Management*. 35(4): 237.

[14] Moore, G. C., and Benbasat, I., (1991). Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation. *Information Systems Research*. 2(3): 192.

[15] Venkatesh, V. & Bala, H., (2008). Technology acceptance model 3 and a research agenda for interventions, *Decision Sciences*, 39 (2), 273-315

[16] Thompson, R. L., Higgins, C. A., and J.M. Howell (1991). Personal Computing Toward a Conceptual Model of Utilization. *MIS Quarterly*, 15(1): 125.

[17] Compeau, D.R., Higgins, C.A., and Huff, S., (1999). Social Cognitive Theory and Individual Reactions to Computing Technology: A Longitudinal Study. *MIS Quarterly*, 23 (2), 1999, pp. 145- 158.

[18] Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information Systems Research*, 11 (4), 342-365.

[19] Venkatesh, V., and Davis, F.D., (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Managem. Science, 46(2): 186.

[20] Chau, P.Y.K., and Hu, P.J. (2001). Information technology acceptance by professionals: A model comparison approach. *Decision Sciences*, 23(4)699 – 719.

[21] Venkatesh, V., and Morris, M. G., (2000). Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *MIS Quarterly*, 24(1): 115.

[22] Benham, H., and Raymond, B., (1996). Information technology adoption: Evidence from a voice mail introduction. *Computer Personnel*, 3: 25.

[23] Compeau, D., and Higgins, C., (1995). Computer Self-Efficacy: Development of a Measure and Initial Test. *MIS Quarterly*. 19(2): 189.

[24] Fishbein, M., and Ajzen, I., (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research, Addison-Wesley, Massachusetts.

[25] Taylor, S., and Todd, P.A., (1995a). Understanding Information Technology Usage: A Test of Competing Models. *Information Systems Research*. 6(2): 144.

[26] Taylor, S., and Todd, P.A., (1995b). Assessing IT Usage: The Role of Prior Experience. *MIS Quarterly*, 19(4): 561.

[27] Mathieson, K. (1991). Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. *Inform. Systems Research*, 2(2),173–191.

[28] Chang, M. K. (1998). Predicting unethical behavior: a comparison of the theory of reasoned action and the theory of planned behavior. *Journal of Business Ethics*, 17, 1825–1834.

[29] Hair, J. F., Anderson, R. E., Tatham, R. L., & Black W. C. (1998). *Multivariate Data Analysis with Readings*. Englewood Cliffs, NJ: Prentice Hall.

[30] Straub, D. W., Boudreau, M. C., and Gefen D. (2004). Validation guidelines for IS positivist research. *Communications of AIS*, 13, 380-427.

[31] Straub, D.W. (1989). Validating Instrument in MIS Research. *MIS Quarterly*, 12 (2): 47 170.

[32] Malhotra, N., (1999), Marketing research: An applied orientation, 3rd Ed., Prentice Hall, Inc., New Jersey.

[33] Blunch, N. (2012), Introduction to Structural Equation Modeling Using IBM SPSS Statistics and Amos, SAGE Publications Ltd, USA.

[34] Kline, R.B. (2010), Principles and Practice of Structural Equation Modeling (Methodology in the Social Sciences), Guilford Press, UK.

[35] Kelloway, K. (1998), Using LISREL for Structural Equation Modeling: A Researcher's Guide, SAGE Publications, London.

[36] Smith, T. D., & McMillan, B. F. (2001). A primer of model fit Indices in Structural Equation Model. Annual Meeting of the Southwest Educational Research Association. (New Orleans, LA, February 1-3).

[37] Harrison, A. W., & Rainer, R. K. (1996). A general measure of user computing satisfaction. *Computers in Human Behavior*, 12, 79-92.

[38] Bollen, K. A., & Long J. S. (1993). Introduction. Testing structural equation models, Newbury Park. CA: Sage Publications

[39] Bollen, K. A. (1989). *Structural Equations with Latent Variables*, NY: Wiley.

The Software and Hardware Complex for Automatic Feed Control of Steam Generators: Modeling and Optimization

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Liudmila Denisova
Omsk State Technical University
Omsk, Russia
e-mail: denisova@asoiu.com

Vitalii Meshcheryakov
Siberian State Automobile and Highway University
Omsk, Russia
e-mail: meshcheryakov_va@sibadi.org

Abstract—This paper presents the software and hardware complex for an automatic feed control system of steam generators in the power unit of a nuclear power plant on the basis of digital local controllers. The mathematical model of automatic feed control system of steam generators in the power unit is considered. As initial data for model design, archival information about industrial process in the power unit is used. We obtain the transfer function of a shaping filter for random signal generation with a required autocovariance function. The results of simulation studies and optimization of automatic control system's parameters using a genetic algorithm are produced.

Keywords—software and hardware complex; feed control system of steam generators; mathematical model; random process; shaping filter; quality criteria; genetic algorithm

I. Introduction

For updating the automatic control system (ACS) of steam generators (SGs) feed used in the power units of Kol'skaya nuclear power plant (NPP), the specialists of ZAO Avtomatika-E (Omsk, Russia) and the scientists of Omsk State Technical University developed the software and hardware complex (SHC) on the basis of the remote local controllers VLR-2.1 [1]. The previous feed control system was updated due to its incompatibility with the modern safety and reliability requirements for an NPP. Note that the digital controller VLR-2.1 represents a freely programmable controller and realizes not only control functions, but also the functions of automatic protection and blocking of processing equipment. According to the requirements of Kol'skaya NPP, for each SG the technical specialists replaced the electronic regulating units and some components of the relay circuit responsible for technological blockings by the controllers VLR-2.1.

A steam generator feed system is one of the most important elements in a power unit of an NPP with a water-water energetic reactor (VVER), therefore, the SHC for the ACS of SG feed with reliable and cost-saving operation of a steam generator feed system affects the performance of the whole power unit. In this paper, we present the mathematical model of the ACS of SG feed in the power unit of a NPP taking into account the actions of random disturbances on a controlled object. The autocovariance functions of random changes for technological parameters are estimated on the basis of archival information on industrial processes (IPs) in the power unit.

A shaping filter is synthesized for random signals generation with required autocovariance functions. Simulation studies and optimization of the ACS of SG are carried out using the developed mathematical model and a genetic algorithm (GA).

II. Description of the SHC equipment

The SHC for the ACS of SG feed based on the controllers VLR-2.1 serves for maintaining feed water flows in SGs in all modes of power unit operation according to generated steam flows and maintaining a water level setpoint in an SG.

The software and hardware complex for the automatic control system of SG feed forms control actions to the electric drives of the main and start-stop feed control valves (FCVs). These valves perform water feed from a discharge header into a steam generator heated by a primary-circuit coolant. Strict requirements are imposed on water level stabilization in an SG. For the SG with VVER-440, the rated water level is 1900 mm. The precision of maintaining the specified level in stationary modes makes up ±50 mm. The admissible dynamic overshoot is ±75 mm under load variation. Water level increase (decrease) with respect to the rated level is prohibited due to possible steam carryover into a turbine (heating surface exposure, respectively).

In addition to control functions, the SHC for the ACS of SG feed performs informational functions, namely, acquisition, processing and transmission of data on control parameters and process variables for further display to the operating staff. The SHC for the ACS of SG feed implements the following service functions: operability monitoring for the hardware and software means of the complex; configuring and tuning of control loops from an engineering workstation (EW). Besides, an EW realizes the functions of display, registration and archiving of all data on the status of processing equipment used in the ACS for SG feed and the status of the controllers.

The SHC for the ACS of SG feed consists of six main controllers and six start-stop water level controllers used in SGs and realized via VLR-2.1 according to the number of SGs in a power unit. The main controller of each SG operates in the normal mode of the power unit, whereas the start-stop controller serves launching and terminating functioning of the unit. The scheme of level control channels using VLR-2.1 for a single SG (see Fig. 1) represents the structure of data exchange between VLR-2.1, signal sources and users (i.e., sensors of process variables, an engineering workstation, FCV control technical tools, protection and blocking means).

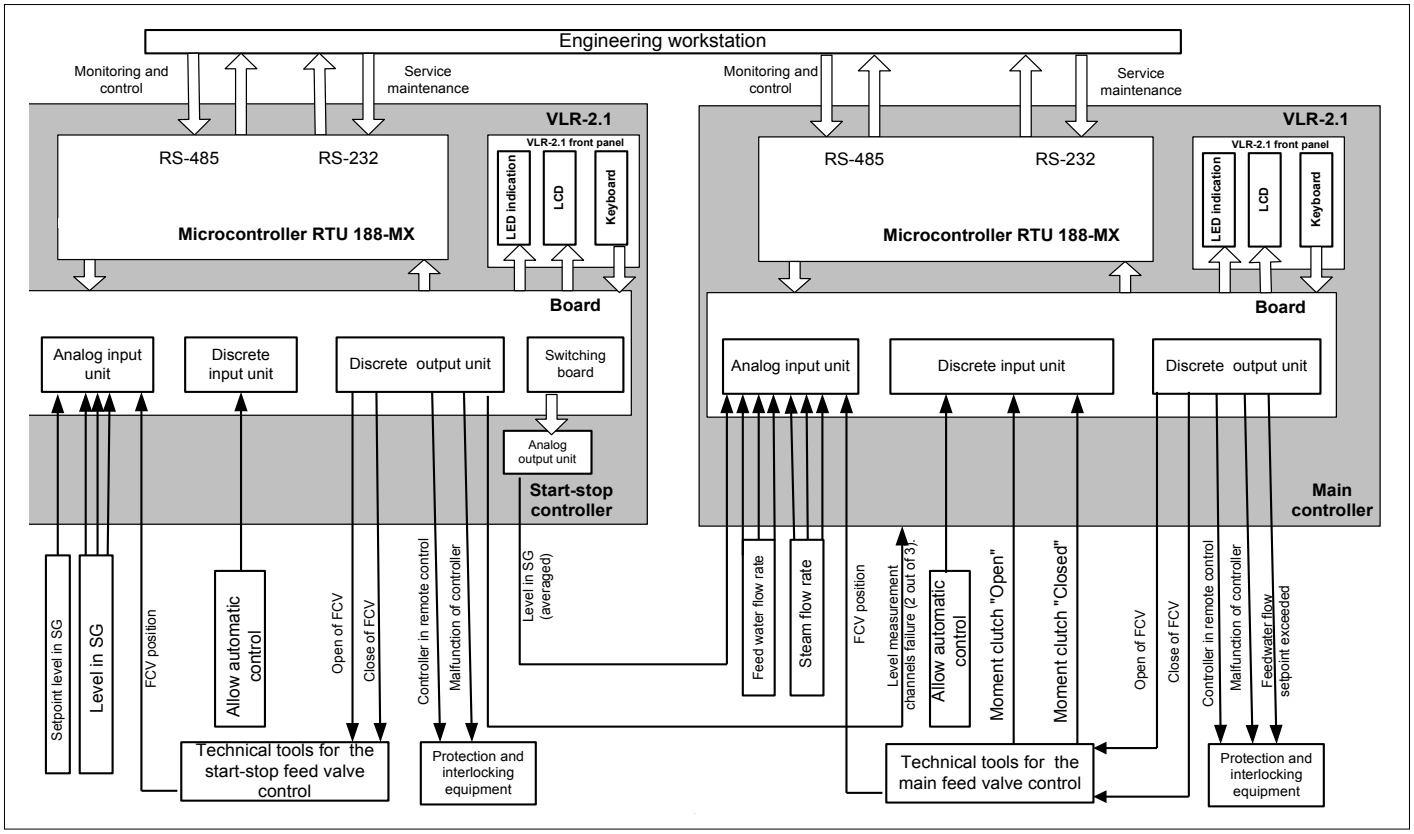


Fig. 1. Realization of water level control channels in the SG using VLR-2.1.

Each of the main controllers of the SHC for the ACS of SG feed receives analog signals from three water flow sensors and three steam flow sensors, as well as from the actuating mechanism (AM) of the FCV position. Besides, the main controller receives the SG level signal from the start-stop controller (this signal is averaged over the indications of three sensors) and the discrete signal on an existing failure in two of three level measuring channels. In addition, the discrete signals from a block-type control panel (BCP) on activating the automatic operation mode of the controller and on the current status of the moment clutch of the FCV ("open"/"closed") are supplied to the inputs of the main controller. According to the proportional-integral (PI) control law (formed jointly with the FCV AM), each of the main controllers in the SHC for the ACS of SG feed produces the control discrete signals for opening/closing of the main FCV, as well as the signal to the blocking circuit on exceeding the admissible value by water flow.

Each start-stop controller in the SHC for the ACS of SG feed receives the analog signals from three sensors and the SG level setter, from the FCV actuator, as well as the discrete signal from the mode switcher of the BCP on activating the automatic operation mode of the controller. According to the proportional control law (formed jointly with the FCV AM and the signal from the FCV position indicator), each start-stop controller in the SHC for the ACS of SG feed produces the control discrete signals for opening/closing of the start-stop FCV. Both the main and start-stop controllers form the discrete informational signals on the controller's operation mode and failures in the circuits of analog and discrete channels of VLR-2.1. Besides, the signals on the controller's operation mode and input commands, process variables and failures in the processing equipment of control channels are outputted to an LCD and LED indicators on the face panel of VLR-2.1 and an engineering workstation. The SHC for the ACS of SG feed with reliable and cost-saving operation of a steam generator feed system affects the performance of the whole power unit. An adequate mathematical model is necessary to test the application software of the SHC for the ACS of SG feed for its compliance with technical requirements and to perform optimization of control channels' settings. Since information on the behavior of basic industrial processes in active ACSs of SG is registered and archived by an available data-processing system, it can be used as initial data for constructing a mathematical model for a controlled object. Both data about SG dynamic characteristics and data about the statistical characteristics of active random disturbances can be obtained on the basis of archival information.

We will consider the instantiations of optimization for the ACS with the main controller. The main controllers of the SHC realize a three-pulse control system: the signals of measured water level (the basic controlled variable), feed water flow (the auxiliary variable) and superheated steam flow (the disturbance) are supplied to the controller's input. The ACS of SG forms control actions for the electric drive of the feed control valve which performs water feed from a discharge header into a steam generator heated by a primary-circuit coolant. Additionally supplied signals allow decreasing the

influence of disturbances from the FCV and steam load variations on SG level.

The control action of the ACS of SG is feed water flow, and the disturbance is turbine steam flow which varies simultaneously with turbine power. Moreover, reactor power, heat supply from the primary circuit and other factors [2, 3] affect water level, too.

III. Estimation of Disturbance Statistical Characteristics and Synthesis of the Shaping Filter

The changes of SG water level, steam and feed water flows represent random processes. Fig. 2 a–c shows the dynamics of system technological parameters according to the archival information of Kol'skaya NPP. From the available data arrays, we have chosen almost stationary realizations in the sense of mathematical expectation constancy and have estimated the statistical characteristics of fluctuations by them. Fig. 2 a–c demonstrates the derived estimates of the normalized autocovariance functions of these random processes for different samples of size n = 100.

To model the random changes of technological parameters using the obtained initial data, it is necessary to synthesize a shaping filter for random signal generation with a required autocovariance function. The transfer function of a shaping discrete filter can be obtained via the spectral decomposition method [4]. The following should be taken into consideration here. As is well-known, supplying the discrete normalized white noise to a discrete linear filter with the transfer function $K(z)=\dfrac{A(z)}{B(z)}$ yields on its output a discrete random process with the spectral density equal to the squared modulus of the transfer function:

$$F(z)=\left|\frac{A(z)}{B(z)}\right|^2=\frac{A^{'}(z)}{B^{'}(z)}\frac{A^{'}(z^{-1})}{B^{'}(z^{-1})}=\left|K^{*}(z)\right|^2=\left|K(z^{-1})\right|^2, \quad (1)$$

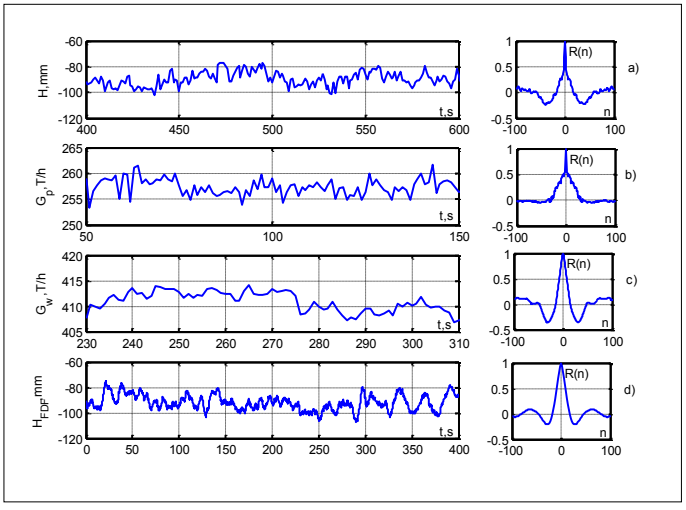


Fig. 2. The random processes and the corresponding autocovariance functions: (a) SG level variation; (b) steam flow variation; (c) feed water flow variation; (d) SG modeled level variation with the shaping filler.

where $K^{*}(z)$ and $K(z^{-1})$ indicate the discrete transfer function of the shaping filter for the direct and inverse shift operators, respectively; $A^{'}(z)$ ($A^{'}(z^{-1})$), $B^{'}(z)$ ($B^{'}(z^{-1})$) denote the numerator and denominator polynomials of $K^{*}(z)$ ($K(z^{-1})$), respectively. The input signal of the shaping filter is the normally distributed discrete white noise $\xi(n)$, i.e., an uncorrelated sequence of random numbers with the mean $M_{\xi}=0$ and the variance $\sigma_{\xi}^2=1$. The experimentally obtained autocovariance functions of signal fluctuations for a continuous process are approximated by the expression $R(\tau)=\sigma^2 e^{-\tilde{\alpha}|\tau|}\cos\tilde{\beta}\tau$, where the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ characterize the decay of the autocovariance function and the frequency of the periodic component of the process, respectively; σ designates the standard deviation of the signal; and finally, τ represents a time lag.

A stationary random signal with bandpass frequency spectrum realizes the signal pulsations with the above-mentioned autocovariance function. This allows forming pulsations in the required frequency range through varying the parameters $\tilde{\alpha}$ and $\tilde{\beta}$.

The normalized autocovariance function of the corresponding discrete process is $R(n)=e^{-\alpha n}\cos\beta n$, where $\alpha=\tilde{\alpha}t_0$; $\beta=\tilde{\beta}t_0$, and t_0 stands for time sampling interval.

The spectral density of a discrete random process is the discrete-time bilateral Laplace transform of its autocovariance function: $F(z)=\sum_{n=-\infty}^{+\infty}R(n)e^{-j\omega n}=\sum_{n=-\infty}^{+\infty}R(n)z^n$.

For obtaining the discrete-time bilateral Laplace transform, take advantage of the formula

$$F(z)=F^{+}(z)+F^{+}(z^{-1})-R(0), \quad (2)$$

where $F^{+}(z)=\sum_{n=o}^{+\infty}R(n)z^n$ is the unilateral z-transform of the autocovariance function. Then

$$F^{+}(z)=\sum_{n=o}^{+\infty}R(n)z^n=\sum_{n=0}^{+\infty}(e^{-\alpha n}\cos\beta n)z^n$$

The table of z-transforms [4, p. 764] gives

$$F^{+}(z)=\frac{1-ze^{-\alpha}\cos\beta}{1-2ze^{-\alpha}\cos\beta+z^2e^{-2\alpha}}=\frac{1-z\gamma\cos\beta}{1-2z\gamma\cos\beta+z^2\gamma^2};\ \gamma=e^{-\alpha}.$$

Hence, according to the expression (2) the spectral function $F(z)$ takes the form

$$F(z)=\frac{1-z\gamma\cos\beta}{1-2z\gamma\cos\beta+z^2\gamma^2}+\frac{1-z^{-1}\gamma\cos\beta}{1-2z^{-1}\gamma\cos\beta+z^{-2}\gamma^2}-1.$$

By reducing to a common denominator, collecting terms and introducing new notation, we obtain that

$$F(z) = \frac{-z^{-1}(A_0 + A_1 z + A_0 z^2)}{\left|1 + B_1 z + B_2 z^2\right|^2}, \qquad (3)$$

where

$A_0 = (1-\gamma^2)\gamma\cos\beta;\ A_1 = -(1-\gamma^4);\ B_1 = -2\gamma\cos\beta;\ B_2 = \gamma^2$.

Perform factorization of the spectral function $F(z)$, i.e., divide the expression in the righthand side of (3) into two factors according to (1). The denominator of the expression (3) already represents the product of two factors. For numerator factorization, find its roots:

$$v_{1,2} = \frac{-A_1 \pm \sqrt{A_1^2 - 4A_0^2}}{2A_0} = v_0 \pm \sqrt{v_0^2 - 1};\quad v_0 = \frac{1+\gamma_0}{2\gamma\cos\beta}.$$

The equation $A_0 + A_1 z + A_0 z^2 = 0$ enjoys symmetry and any of its roots $v_{1,2}$ can be used in the expansion, e.g., v_1. Then the numerator can be represented as

$$-z^{-1}(A_0 + A_1 z + A_0 z^2) = \rho(z - v_1)(z^{-1} - v_1);\ \rho = A_0 / v_1 .$$

Thus, the discrete transfer function of the shaping filter for realization of a random process with the autocovariance function $R(\tau) = \sigma^2 e^{-\tilde{\alpha}|\tau|}\cos\tilde{\beta}\tau$ acquires the form

$$K(z^{-1}) = \frac{a_0 + a_1 z^{-1}}{b_0 + b_1 z^{-1} + b_2 z^{-2}},$$

where $a_0 = -\sigma v_1\sqrt{\rho};\ a_1 = \sigma\sqrt{\rho};\ b_1 = B_1;\ b_2 = B_2$

To implement the random signal model, define the shaping filter by the recurrent relationship

$$\tilde{\xi}(n) = a_0\xi(n) + a_1\xi(n-1) - b_1\tilde{\xi}(n-1) - b_2\tilde{\xi}(n-2), \quad (4)$$

where ξ, $\tilde{\xi}$ are the input and output signals of the shaping filter, respectively; a_0, a_1, b_1, b_2 denote the shaping filter parameters dependent on α and β. Fig. 1d shows the approximated autocovariance function and the corresponding random process realization obtained through modeling of the random signal of SG level using the shaping filter (4).

IV. The Mathematical Model of SG Feed

In MATLAB/Simulink [5] the mathematical model of the ACS of SG has been developed (see Fig. 3) which includes the following submodels*: controller*, *control valve* (that's FCV) *with AM* and *control object*. The digital impulse controller of the ACS of SG whose parameters are necessary to optimize realizes a traditional control method adopted in ACSs with constant-speed actuating mechanisms. The algorithm bases on a relay-impulse proportional-plus-derivative (PD) converter representing two bi-position relay sections with a dead zone and hysteresis in a common feedback loop.

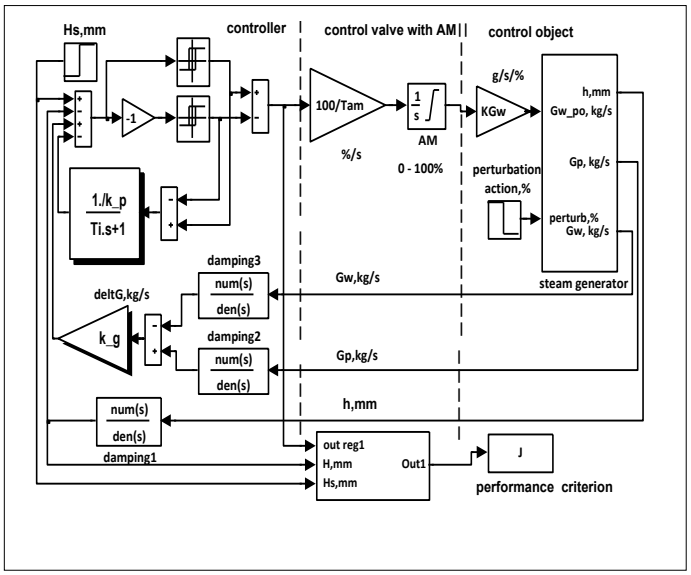


Fig. 3. The model of the ACS of SG feed.

The inertial section in the feedback loop has the transfer function $W_{fb}(s) = \frac{1/k_p}{T_i s + 1}$, where k_p is the controller's gain; T_i denotes the integration time constant. The error signal at the controller's input is formed by the weighted sum of the signals of a given and current levels and also steam and feed water flows.

The material balance signal between steam and feed water flows multiplied by a gain k_g serves as local feedback in the system, guaranteeing high accuracy of required level maintenance. In the described relay-impulse controller, the pulse ratio of output impulses demonstrates the PD relationship with error signal variations. The controller jointly with the AM approximately realizes PI-control law. In addition, the model incorporates inertial sections for input signals damping. The controller's tuning parameters are the gain k_p, the integration time constant T_i and the feedback gain k_g. The sections with these parameters are shaded in Fig. 3.

Being generated by the controller, the output pulse sequence comes to the AM of the FCV, which is represented by the amplifying and integration sections. From the amplifying section output whose gain takes into account the FCV opening time T_{am}, the FCV moving speed signal comes to the integrator with limitations (0...100%). The FCV position signal comes to the controlled object, where the feed water flow signal is formed as the product of the FCV position signal and the corresponding gain.

The subsystem *steam generator* is used for SG simulation (see Fig. 4). This subsystem has two inputs, namely, the CV feed water flow and the parameter perturbation for forming deep disturbances of steam flow variations. The output variables are SG water level, steam and feed water flows with proper consideration of their mutual influence.

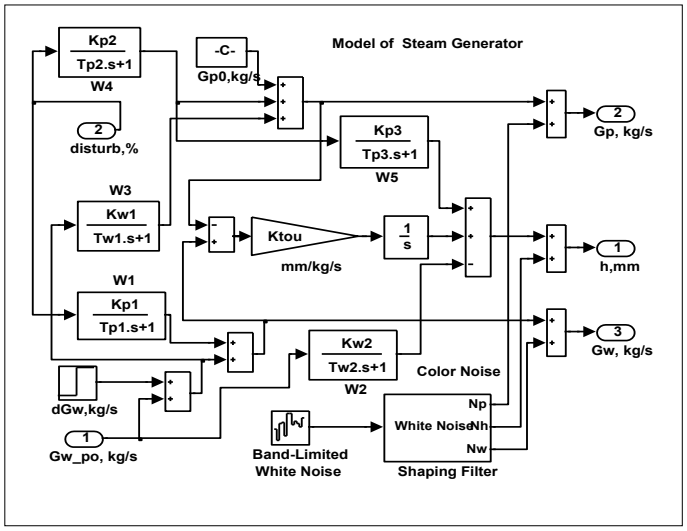


Fig. 4. The SG model considering random disturbances.

The parameters of the above linear SG model have been evaluated on the basis of design-technological parameters of industrial equipment and also by approximation of SG dynamic characteristics computed using the nonlinear power unit model [3].

SG water level is the integral from the material imbalance between steam and feed water flows. For modeling of SG dynamic properties by water level, each of influence channels is described by the sum of an integrator and an aperiodic section (note that the former is common for the both channels). In addition, the SG model includes aperiodical sections necessary for describing the mutual influence of disturbances for steam and water flows. The fluctuations of random disturbances of water level, steam and feed water flows are realized by the shaping filter *Shaping Filter* using the experimentally obtained autocovariance functions of signal fluctuations according to the expression (4) and are further summed up with the corresponding outputs of the SG model. At the same time, the block *Band-Limited White Noise* generates the normally distributed white noise supplied to the input.

For optimization, the model computes an integral control performance criterion (the *performance criterion* subsystem in Fig. 3): $J = \int_0^T (\varepsilon^2 + \lambda u)^2 dt$, where ε is the controlled parameter deviation from the setpoint value; u denotes the controller's output signal supplied to the AM; λ represents a weighting coefficient; t indicates the current time; T means the upper limit of integration chosen not smaller than the response time. Minimization of the performance criterion J guarantees high response and the absence of overshoot, as well as decreases the number of AM actuations.

The developed mathematical model of the ACS of SG feed allows carrying out simulation studies in different operation modes and defining the parameters leading to minimization of the chosen performance criterion of ACS operation.

V. Experimental Studies and Parameter Optimization of the ACS of SG

Experimental studies have been performed using *m*-files, i.e., MATLAB scripts repeatedly invoking the Simulink model of the ACS of SG. The following aspect has been taken into consideration. Computations with complete enumeration of adjustable parameters require a considerable time, whereas wider range of parameters or their smaller increments for higher solution accuracy actually further increases time costs and makes the task unfeasible.

ACS optimization based on the genetic algorithm [6–9] from MATLAB Global Optimization Toolbox eliminates this problem. Using the GA, we have found the optimal parameters of the ACS of SG and obtained the corresponding transient responses under for insignificant time costs.

The GA represents an optimal solution search method based on genetic inheritance and natural selection [6]. In comparison with traditional optimization methods, the GA enjoys an advantage that solution search involves a set (population) of points. The algorithm considers only most promising search regions in the sense of optimal value approximation owing to the crossover of the most adapted individuals (application of the crossover operator). Next, the random mutation operator yields individuals with new properties. In addition, the best individuals (elite) pass to a next generation without any changes; therefore, a new generation contains a refined ratio of characteristics inherent in the best individuals of a previous generation, which is further extended throughout the whole generation. As a result, the population has evolutionary convergence to the optimal solution. During ACS optimization, we search for the parameters k_p, T_i, and k_g, minimizing the performance criterion $J = f(k_p; T_i; k_g)$. Hence, in the genetic algorithm the role of an individual belongs to the parameter vector $X = (k_p; T_i; k_g)$. At each step, the GA considers the parameter vector X and the corresponding value of J. The initial values of the ACS parameters are chosen randomly from the accepted ranges: $k_p \in [0{,}01;1]$, $T_i \in [2; 20]$, $k_p \in [1;100]$. Fig. 5 illustrates GA operation in one of the performed instantiations during ACS optimization.

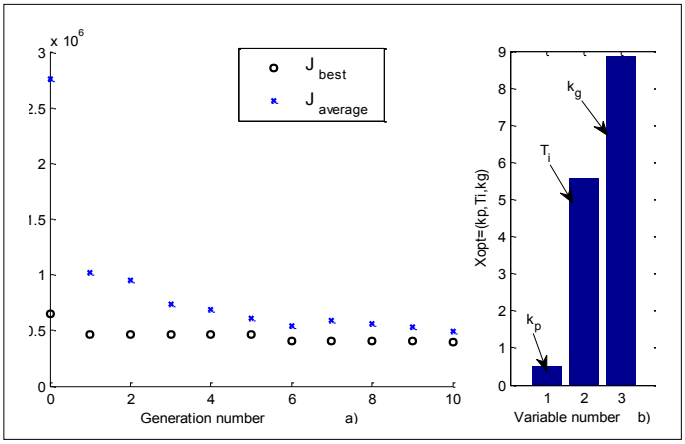


Fig. 5. ACS optimization using the genetic algorithm.

Particularly, Fig. 5a shows the dynamics of the performance criterion for the best individuals and its average value across the whole population. Merely 10 generations have been required for obtaining optimal parameters: clearly, the algorithm converges. And Fig. 5b represents the elements of the optimal parameter vector $X_{opt} = (k_p; T_i; k_g)$ in the form of bar charts.

Since a random-number generator specifies the initial population in the GA, we have carried out several realizations of computations and selected the best ones among the obtained solutions. The optimization result for one of the realizations is demonstrated by Fig. 6. Here SG runs in the disturbed mode: steam flow goes down by 10% from the nominal value at the 150th second. The graphics show that the steam flow decrease is well processed by the regulator controlling SG water feed within one pulse only. Feed water flow is adjusted in accordance with steam flow, and the door position in operation range varies by less than 5 %. SG level (i.e. H_{FDF}) is maintained within the required accuracy ± 50 mm.

Note that the disturbance has been successfully processed by a small number of controller's pulses in the majority of realizations. The conducted studies lead to a conclusion that the suggested optimization method based on the genetic algorithm is efficient for computer-aided design of ACS optimization with proper consideration of active random disturbances.

VI. Conclusion

The developed mathematical model of SG feed allows performing control channels adjustment in the SHC for the ACS of SG feed and conducting simulation experiments for its dynamic characteristics in transient modes with different load variations under active random disturbances. The dynamic tests of the developed model have proved that the SHC for the ACS of SG feed maintains the SG parameters in the normal operation ranges of the power unit.

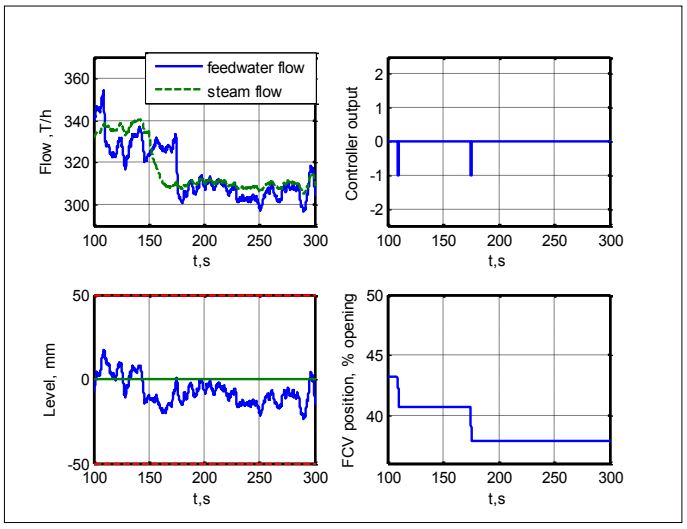


Fig. 6. The transient responses for the ACS of SG under steam flow disturbance.

Note that the obtained transfer function of the shaping filter can be employed to model other industrial processes incorporating random disturbances under strict requirements imposed on control quality.

The mathematical model of SG feed with active random disturbances realized on the testing rig of ZAO Avtomatika-E makes it possible to carry out a prompt control system tests in different operation modes and to elaborate recommendations on parameter optimization of the SHC for the ACS of SG feed.

The updated ACSs of SG feed based on the digital controllers VLR-2.1 were put into industrial operation at power units No. 3 and No. 4 of Kol'skaya NPP.

References

[1] E.M. Raskin, L.A. Denisova, and P.N. Nadtochiy, "The Software and Hardware Complex for Automatic Feed Control of Steam Generators in the Power Unit of a Nuclear Power Plant", Automation and Remote Control, vol. 76, no. 12, pp. 2241-2248, 2015.

[2] K.A. Pupkov, and N.D. Egupov, "Metody klassicheskoi i sovremennoi teorii avtomaticheskogo upravleniya, tom 3: Sintez regulyatorov sistem avtomaticheskogo upravleniya" ("The Methods of Classical and Modern Automatic Control Theory, vol. 3: Controller Design in Automatic Control Systems"), Moscow: Gos. Tekhn. Univ. im. Baumana, 2004.

[3] E.M. Raskin, L.A. Denisova, V.P. Sinitsyn, and Yu.V. Nesterov, "A Mathematical Model of Steam Generator Feed System at Power Unit of a Nuclear Plant", Automation and Remote Control, vol. 72, no. 5, pp. 1118-1126, 2011.

[4] R.C. Dorf, R.H. and Bishop, "Modern Control Systems", New York: Addison-Wesley, 1967. Translated under the title "Sovremennye sistemy upravleniya", Moscow: Laboratoriya Bazovykh Znanii, 2002.

[5] D.Xue and Y. Q. Chen, "System simulation techniques with MATLAB and Simulink", John Wiley & Sons, 2013.

[6] J.H. Holland, "Adaptation in natural and artificial systems. An introductory analysis with application to biology, control and artificial intelligence". London: Bradford book edition, 1994.

[7] G. N. Purohit, A. M. Sherry, and M. Saraswat, "Optimization of function by using a new MATLAB based genetic algorithm procedure", International journal of computer applications, vol. 61(15), pp. 1–5, 2013.

[8] L.A. Denisova and V.A. Meshcheryakov, "Automatic Parametric Synthesis of a Control System Using the Genetic Algorithm," Automation and Remote Control, vol. 76, no. 1, pp. 149-156, 2015.

[9] L. Denisova and V. Meshcheryakov, "Synthesis of a Control System Using the Genetic Algorithms / IFAC-PapersOnLine, vol. 49, issue: 12, pp: 156-161, 2016.

A User-centric Process Management for System and Software Engineering Projects

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Mojtaba Hajmoosaei
IRIT, University of Toulouse,
Toulouse, France
mojtaba.hajmoosaei@irit.fr

Hanh Nhi Tran
IRIT, University of Toulouse,
Toulouse, France
hanh-nhi.tran@irit.fr

Christian Percebois
IRIT, University of Toulouse,
Toulouse, France
christian.percebois@irit.fr

***Abstract*—In traditional process environments, process modeling is performed by process designers and process enacting is performed by process actors. Due to this separation, there is often a gap between process models and their real enactments. As a consequence, the operational level of process environments has stayed low, especially in system and software industry, because they are not directly relevant to process actors' needs. In order to facilitate the usage of process environments, this paper presents a solution that enables process actors to perform both the modeling and enacting of their real processes. To this end, first, an end-user process modeling approach was proposed to allow each process actor to easily describe the process fragment containing the activities carried out by his role. Second, an artifact-centric process engine was developed to enact activities coming from different process fragments. Our process engine does not require predefined work-sequence relations among these activities to synchronize them, but deduces such dependencies from their exchanged artifacts. As a result, the process engine can enact even a partially defined process where some fragments are missing.**

I. Introduction

Modern System and Software Engineering (SSE) processes are often extremely complex, involve multi-disciplines and can be realized by different teams. Such complicated processes can be modeled and then applied to different projects. Process's execution should be managed by a process environment that monitors and controls the project's activities.

Most process environments adopt a top-down approach where the whole process is first modeled by process designer and then deployed in a project where real enactment happens by process actors, i.e., persons operating the process. However, in general it is almost impossible for process designers to model accurately the entire process as they don't have a thorough knowledge about the process's application domain. Thus, several details of the process may be left initially unspecified; some execution scenarios may be unseen and the process model does not describe exactly what will be done by process actors. Consequently, deploying a process model in a process environment to manage a project requires an important refinement [4] that is not evident for process actors. First, they are not trained to work with complicated process modeling languages. Second, they don't have a global view on the process; each process actor only knows his own activities. For this reason traditional process environments have been still weakly adopted by end-users in system and software industry.

To promote the practical use of process environments, we put the main emphasis on process actors and aim at providing them a friendly environment allowing each process actor to model himself his working process so that the obtained model is closer to his real activities and thus executable after a simple deployment. In contrast to the top-down approach that requires the whole process model to enable its enactment and synchronization, we propose: (1) modeling the global process in separate fragments belonging to different roles and then (2) constructing dynamically the global view of the system at enactment-time to synchronize activities belonging to different process fragments that are progressively instantiated during the process's execution. In order to achieve these objectives, we developed:

- *A Structural Process Modeling Language* enabling each process actor to specify his process fragment independently of other process fragments. To be more relevant to process actor's vision about process, our language suppresses the work-sequence relations in process models but strengthens the description of artifact exchanges among process activities. This simple language will be described in Section III.
- *An Artifact-centric Process Engine* enacting process fragments performed by different process actors. The process engine analyzes the data exchanged among the running activities to deduce their inter-dependencies and then synchronize them. Section IV presents the proposed process engine.

In Section V, we report the experiments carried out by using our prototype process environment with participation of our industrial partners.

II. Motivating Example

Figure 1 presents an extract of a process given by our industrial partners in the project FUI[1] ACOVAS (Agile tools for COnception and VAlidation of Systems) to reconfigure the wiring system of a testbench due to an evolution of the system under test. The process *Modify Testbench Wiring* concerns several expertise domains but we only focus on the *electrical* and *instrumentation* ones to simplify the description. The process is performed by five roles: *Analyst, Electrical Designer, Instrumentation Team, Wiring Team* and *Bench Coordinator*.

In practice, each role describes his fragment of process as a list of activities that he performs, with their input and output artifacts in a specific state. For instance, an *Analyst* performs the activity *Detail Change Requirement*, which receives the artifact *Change Demand* in the state *defined* and produces three output artifacts: a *Component Purchase List* (*defined*), an *Electrical*

[1]French academic industry joint research program.

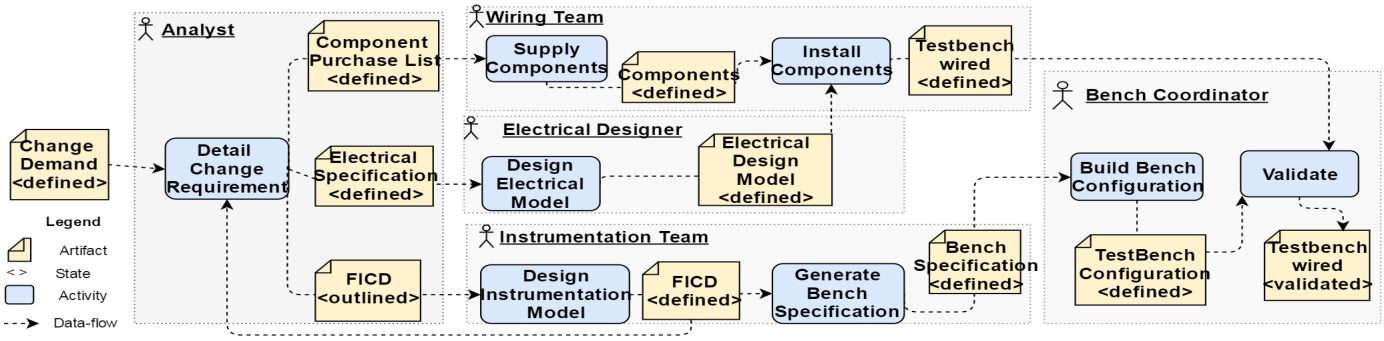


Fig. 1: Process Modify Testbench Wiring

Specification (*defined*) and a *Functional Interface Control Document (FICD)* (*outlined*). Often, activities are described without an explicit execution order; their inter-dependencies however can be deduced by analyzing the artifacts that they exchange. For example, *Design Electrical Model* can only start after the termination of *Detail Change Requirement* because it requires the *Electrical Specification* produced by *Detail Change Requirement*.

In a top-down approach, to obtain an executable process, the process designer must refine the process in Figure 1 to model the enactable tasks of all roles as well as their execution order in all possible circumstances that may happen at enactment-time. Figure 2 shows one possible modeling of the process fragments concerning the activities of *Analyst* and *Instrumentation Team*.

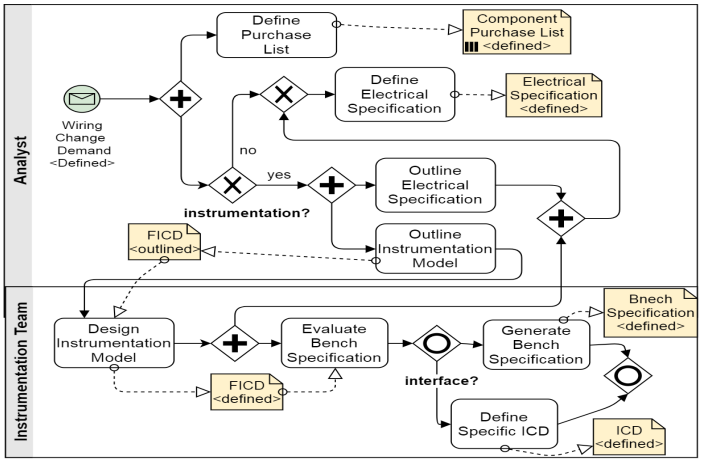


Fig. 2: Process Modify Testbench Wiring in BPMN

For complex processes, this modeling is challenging, particularly when tasks' sequence is non linear because different modeling constructs as loops or choices are needed to express execution orders. For example, inside the activity *Detail Change Requirement* while the tasks, *Define Purchase List* and *Define Electrical Specification* are always performed, the task *Outline Instrumentation Model* will be enacted only if the option *instrumentation* is chosen. Often, the process designer has to spend several interviews with involved process actors to obtain the formal process model which can become super complex. Moreover, it requires a lot of process designer's effort in order to deploy the process by establishing the data mapping concerning information flow from the process context to a particular task and from the task to the process.

Top-down process engines enact and synchronize all process activities in an established way as described in the process model. This strict solution has two disadvantages. Firstly, it does not allow executing a process partially defined, i.e. some process fragments are missing when the process starts to run. Unluckily, it is difficult to fully model SSE processes because they are generally performed by various teams and also by external subsidiaries. Secondly, it cannot let process actors handle emerging events that are not described in the model. However, it is hard, especially for process designers, to anticipate in the process model all process's execution possibilities.

Being aware of the above drawbacks of top-down based process environments, we are interested in an opposite approach. We argue that by separating the performers of process modeling and process enacting, a process is often inadequately described and therefore not faithful to its real enactment. In contrast, our end-user approach that puts the main emphasis on process actors by letting them carry out both modeling and enacting can eliminate the inconvenience of the top-down approach and reduce process gap. We present in the following sections the solutions to implement this idea.

III. End-user Process Modeling

This section explains how our approach enables process actors to model their working process. For this purpose, the standard process modeling languages as BPMN [10] and SPEM [9], which are developed for process designer and thus very complex for process actors to use, are not suitable for our intention. We aim at an end-user friendly process modeling language which includes only the concepts known by process actors in their daily work and avoids introducing concepts requiring process-special knowledge.

Our process modeling language, named SPML (Structural Process Modeling Language), has two key features: (1) removing the work-sequence relations among activities from the process model; (2) describing in detail the artifacts entering and leaving the activities to provide sufficient information to deduce the behavioral aspect of process at enactment-time.

In our language, a *Process Fragment* encapsulates the activities belonging to a *Role*. An *Activity* describes the functional view of a work performed by a specific role and is

composed of tasks corresponding to different sub-objectives. A *Task*, *mandatory* or *optional*, represents an enactable and manageable action taken to achieve one specific activity's sub-objective. An *Artifact* is a tangible work product describing the system to be delivered; at a moment in its lifecycle, an artifact is defined in a specific *State*. A task can use an artifact as input, output or input/output parameter via the relation *TaskParameter*. The property *UsageKind* of the relation *TaskParameter* indicates if the concerned artifact is necessary for a task's pre-condition (*ToStart*) or for a task's post-condition (*ToFinish*). Finally, a *Process* is comprised of several process fragments.

Our language introduces a concept *Option* to allow process actors specifying alternative ways to realize an activity in different execution contexts of a process. An option represents a specific choice that activates the use of some optional tasks and artifacts.

Using the proposed process modeling language, a process actor can model the process fragment corresponding to the role that he plays in the project. Each process fragment can be modeled independently of the other fragments. However all fragments are described based on common company assets including the definitions of participating roles, working artifacts together with their possible states, used resources, and development options. In practice, such company assets always exist even if the process is not formally modeled.

Each role models simply the structure of his process fragment without specifying its behavioral aspect, i.e. the sequence flows between activities and tasks. Concretely, he defines a list of activities which are decomposed further into tasks. Each task is specified with the input and output artifacts in a specific required state.

Figure 3 illustrates the *Analyst*'s process fragment modeled with SPML. This process fragment includes an activity comprising three tasks without specifying any work-sequence flow between them. The task *Outline Instrumentation Model* is optional and activated under option *instrumentation*.

Process Modify Testbench Wiring
options instrumentation, interface
  ProcessFragment Analyst
    Activity Detail Change Requirement **options** instrumentation
      MandatoryTask Define Purchase List
        Artifact Wiring Change Demand(**in**, defined)
        Artifact Component Purchase List(**out**, defined)
      MandatoryTask Define Electrical Specification
        Artifact Wiring Change Demand(**in**, defined)
        Artifact FICD(**in**, defined, **toFinish**)
          ifOption instrumentation
        Artifact Electrical Specification(**out**, defined)
      OptionalTask Outline Instrumentation Model
      ifOption instrumentation
        Artifact Wiring Change Demand(**in**, defined)
        Artifact FICD(**out**, outlined)

Fig. 3: Process Fragment of Analyst

By suppressing the work-sequence flows, an activity can invoke freely and separately its inner tasks. This feature facilitates the reworks and enables variable ways to perform an activity. Another advantage of our end-user modeling approach is that it is not necessary to define the whole process from the beginning of the project. The process fragments can be progressively defined during the project execution. Moreover, if some process fragments are not provided, the rest of the process can be specified and enacted independently to the missing parts.

Thanks to this fine-grained modeling, the deployment of a process in a project becomes a simple instantiation of the process fragments models, i.e., creating concrete instances of activities and tasks, then providing these tasks the required effective parameters selected from the real artifacts managed in the project.

IV. Artifact-Centric Enactment

This section presents how our process engine instantiates process fragment models and systematically controls the enactment of the running processes. As tasks are real actions that change artifacts' state and progress the process's execution, the main function of the process engine is managing tasks. The lack of work-sequence relations in our process models causes a new issue for the process engine in synchronizing tasks because it must deduce the dependencies among tasks from the real enactment circumstances.

The conventional process engines cannot satisfy our special requirement on tasks synchronizing. Hence, we developed a new process engine that can establish dynamically the work-sequence dependencies among running tasks by analyzing the artifacts that they exchange. To do so, we proposed: (1) a *Process Dependency Graph (PDG)* as a run-time storage keeping the information of process element instances coming from different process fragments and therefore represent the global view of the system; (2) An artifact-centric process synchronization mechanism to manage tasks' life-cycle. We implemented these solutions in the *Bottom-up Artifact-centric Process Environment (BAPE)*.

A. Process Environment BAPE

Figure 4 shows our Bottom-up Artifact-centric Process Environment (BAPE) comprising the following components:

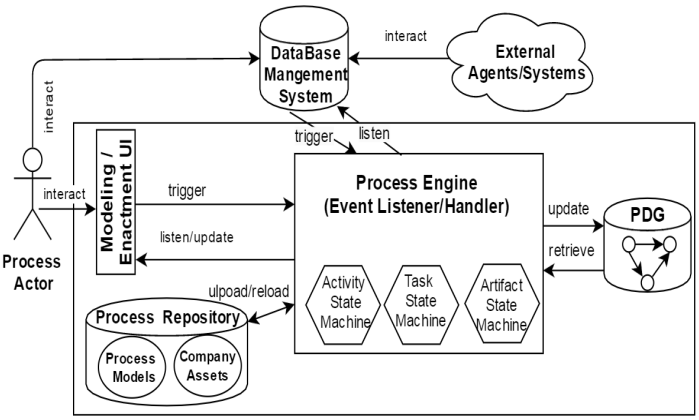


Fig. 4: Architecture of BAPE

1) Modeling/Enactment UI: This component provides interfaces that allow process actors to interact with the process environment during the modeling and enacting phases. The modeling interface enables process actors to define company assets and model the fragments of their processes then store

them into the *Process Repository*. The enactment interface enables process actors to manage their activities and tasks (e.g., create activities, start and complete tasks, etc.).

2) Database Management System (DBMS): We suppose that resources (e.g., process actors, tools, etc.) and working artifacts are externally managed by each company proper tools. However, those external tools use a central DBMS which is connected to BAPE. This connection makes the process environment being aware of any events changing the state of artifacts and resources in the external systems.

3) Process Dependency Graph (PDG): This component stores the information of running processes managed by BAPE. PDG is progressively updated by the process engine to reflect always the current state and the global view of the system at run-time. Hence, PDG is a source of information for the process engine to coordinate and synchronize processes.

4) Process Engine: This component is in charge of enacting and synchronizing the processes managed by BAPE. The behavior of process engine is based on state machines specifying the life-cycles of activities, tasks and artifacts. These state machines pilot the engine by defining the list of listened events triggered by users or external systems along with engine's actions to handle these events to update the PDG and make process enactment progress.

In the following, we present how the PDG and the artifact-centric process engine enable process execution and synchronization in a flexible way.

B. Process Dependency Graph (PDG)

PDG is our solution to monitor process elements' instances as well as the relations among them [13]. As process models are fragmented in our system, the role of PDG is not only storing runtime information but also establishing the global view of the system from separate process fragments. For this purpose, PDG is defined as a directed and typed graph where each PDG element refers to its defining process element in our SPML language. In this way, PDG nodes represent process element instances and PDG edges represent the relations between them. Representing the behavioral aspect of the process, each PDG node has the property describing its current state in the system. The instantiation of a process fragment is then a creation of PDG's nodes and edges conforming to the elements specified in the fragment's model. The execution of process concerns updating the PDG progressively.

Figure 5 illustrates an extraction of the PDG representing a snapshot of the system corresponding to the execution of the *Modify Testbench Wiring* process in the project $Proj$. This snapshot includes only the fragments of *Analyst* and *Instrumentation Team* performed respectively by the actors A_1 and IT_1. The PDG shows the current state of process elements and the relations among them. The dependencies among tasks, both inside or among process fragments, can be established based on their exchanged artifacts. For example, in the activity *Detail Change Requirement* DCR_1, the task *Outline Instrumentation Model* OIM_1 produced the artifact $FICD_1$ in the state *outlined*; the task *Define Instrumentation Model* DIM_1 is modifying an *outlined* $FICD_1$ and will produce the *defined* $FICD_1$ which will be used by the task *Define Electrical Specification* DES_1. Thus, OIM_1 has to finish so that DIM_1 can start and DIM_1 has to finish so that DES_1 can finish.

Thanks to the information provided by the external systems which manages company's resources and data, the PDG's *ActorNodes* and *ArtifactNodes* can be created and updated. When process actors enact process's activities and tasks, the corresponding *ActivityNodes* and *TaskNodes* will be added progressively into the PDG and the edges among the concerned nodes will be established or updated.

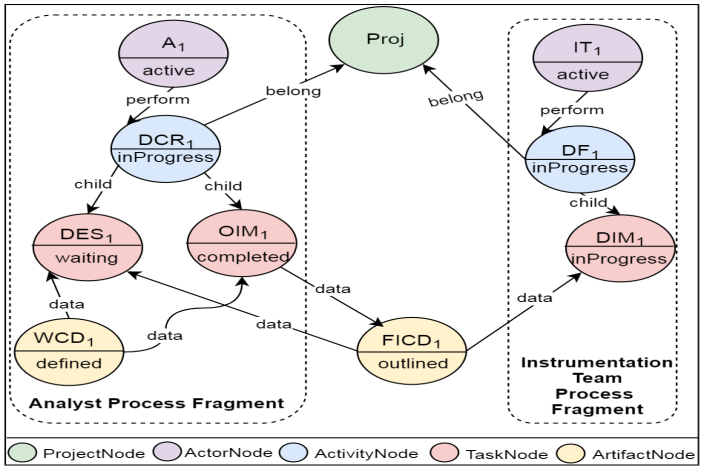


Fig. 5: A Snapshot of PDG

C. Artifact-centric Task Synchronization

To manage the life-cycle of tasks, we adopt a state machine inspired from the standard WS-HumanTask Specification [8] as illustrated in Figure 6.

This state machine pilots the process engine in creating and updating the *TaskNode*s of the PDG when a task event is triggered by process actors through their *Enactment UI*. For each task event, the process engine verifies the condition of the required transition to enable or refuse the transition. If the transition is accepted, the process engine updates the concerned PDG *TaskNode* and its parameters by using the predefined actions for the transition and then allows the actor to perform task. If the transaction is refused, the demanding task has to wait until another event validates the transition condition. In this way, the process engine synchronizes the execution of different process actors' tasks.

Note that in conventional process engines, the condition of a transition between task's states is principally based on work-sequence relations defined among tasks. This solution cannot be applied to our process engine as we do not describe the work-sequence relations among tasks in process models. Instead of work-sequence relations, our process engine uses artifacts as the main source to enable the transitions between task states.

The conditions of two most important transitions to start and to finish a task are defined in the task's model as pre- and post-conditions that specify respectively the required state of the task's input and output artifacts. The process engine verifies the pre/post conditions of a task instance in the PDG by checking if the states of the concerned *ArtifactNodes* are the required ones. If the pre-condition is satisfied, the process engine updates the *TaskNodes*'s state from *created* to *inProgress*. If not, the *TaskNode* will be put in *waiting* state to wait until the required input *ArtifactNode* is available. The post-conditions are verified similarly. For instance, Figure 5

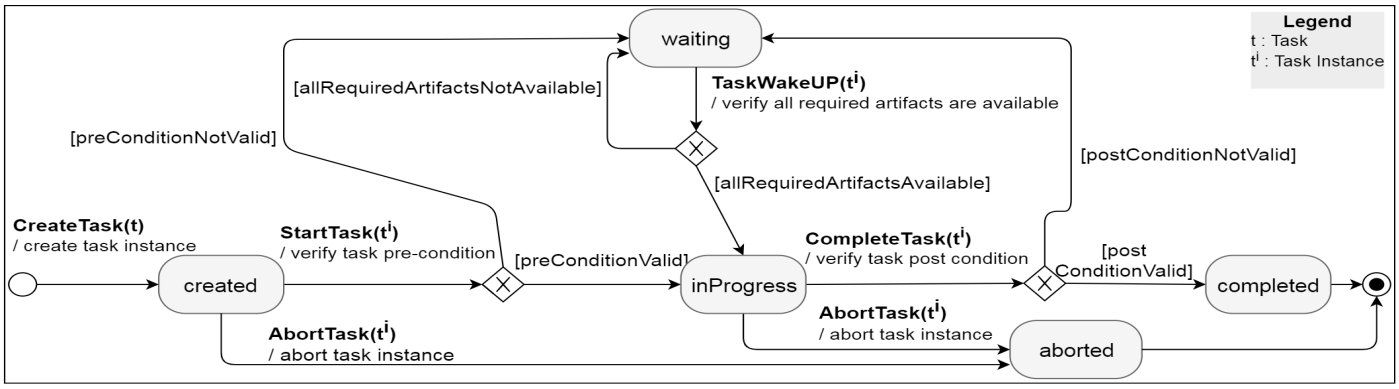


Fig. 6: Task State Machine

illustrates the situation where the task DES_1 is waiting for the artifact $FICD_1$ in a state $defined$ which is required to produce the artifact *Electrical Specification*.

V. Industrial Case Study

The experiment presented in this paper is a simplified version of a real case study conducted in the context of the project ACOVAS. We carried out our evaluation in two phases of modeling and enactment for the process described in Figure 1. We applied a simple setup of the scenario presented in Section II for a project including five process actors as one actor in each team. Our evaluation has been carried out using top-down process environments jBPM [3] and AristaFlow [1], which use BPMN [10] as their modeling language, and our prototype BAPE.

In order to efficiently manipulate the PDG, we used the graph database system Neo4j [7], which excels at managing highly connected, rapidly changing data thanks to its powerful traversal framework. Our framework is based on the assumptions that a central graph database, a central resource repository and an application server (e.g. jBOSS) are provided to support process management in a distributed environment. Figure 7 shows the BAPE *Enactment UI* of role *Analyst*.

Our assessment on modeling phase is based on the required roles to participate into two sub-phases process modeling and process deployment. Our evaluation on enactment phase is based on the required degree of process completeness to enable the synchronization of activities and tasks.

The result of our modeling and enactment evaluations is presented in the Table I.

TABLE I: Modeling and Enactment Phases Evaluations

	jBPM	AristaFlow	BAPE
Modeling	designer process actor	designer process actor	process actor
Deployment	designer process actor	designer process actor	process actor
Synchronization	complete-defined	complete-defined	partial-defined

As learned from our experience, jBPM and ArtistaFlow require the participation of both process actors and process designers to model and deploy the process. By proposing SPML with a user-friendly modeling interface developed in BAPE, process actors can conveniently model their executable processes.

Regarding process synchronization, both jBPM and ArtistaFlow require that the whole process is modeled before enactment so that the process engine can know how to synchronize and coordinate the activities. As a main advantage, BAPE can enact and synchronize a partially defined process where some process fragments are missing.

VI. Related Work

The lack of adoptation by end-users has been recognized as a major problem of existing process environments in SSE [6]. To deal with this issue, many researches aim at integrating end-users into process management life-cycle in order to suit process actors needs [11], [2], [12].

Front et al. [2] propose a participative modeling approach named ISEA that allows process actors to collaborate in a centralized way to improve the process elicitation. Oppl [11] shares the same idea as ours on eliminating the work-sequence flows among activities and fragmentating of process models. However, both works lack enactment support in their solution. They propose the transformation of process models into BPMN [10] that requires the efforts of process actors and process designers to model the behavioral aspect of their processes and then validate and deploy the process into target process environment.

Stoitsev et al. [12] develop a collaborative task management tool CTM (as add-in to Microsoft Outlook) to enable end-users creating hierarchical to-do task lists. Tasks can be delegated over email exchange. Tracking of email exchanges on task delegation integrates the end-user's personal to-do list to overall Task Delegation Graphs (TDG). TDGs represent weakly-structured process models that are captured snapshots of the actual process enactment. Like us, they apply a graph structure for the re-composition of the global model, but the work-sequence relations among tasks are defined as a suggestion based on the time of tasks processing changes.

We share the same vision as PHILharmonicFlows [5] and PET [4] about the significant role of artifacts in process modeling and enacting. Kuhrmann et al. [4] propose a solution to ease the process deployment for process actors by transforming the process models from SPEM into the artifact format of the project tools (e.g., Team Foundation server, SharePoint and OfficeWord). Their solution eases the deployment of processes into some specific development tools

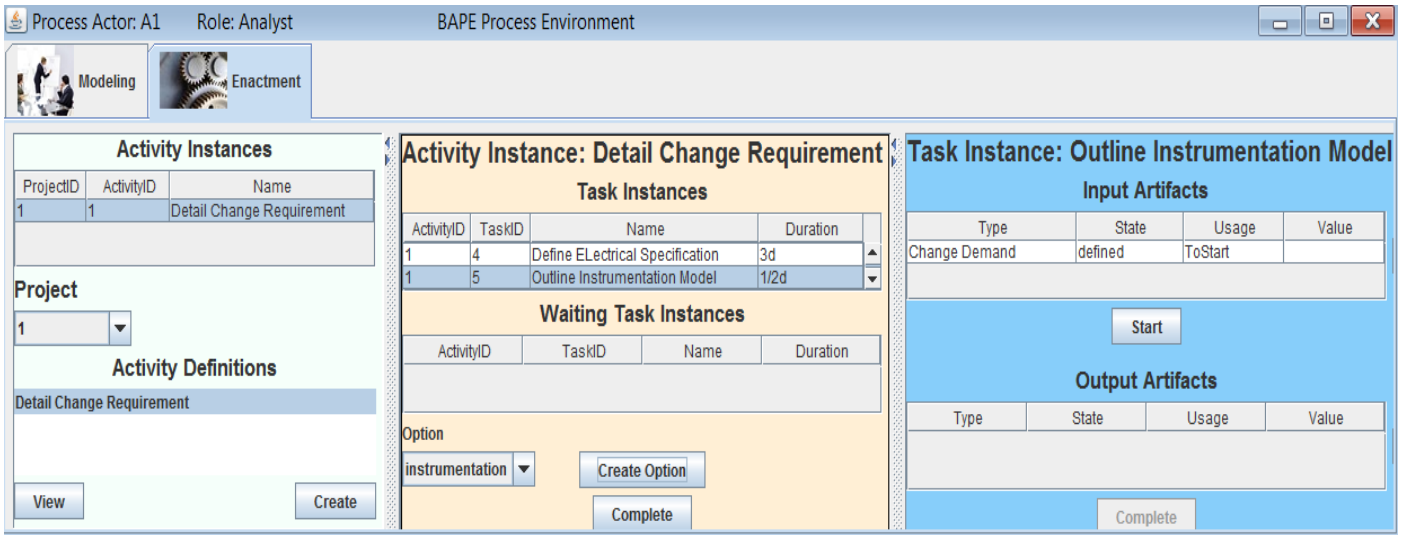


Fig. 7: Enactment UI of Analyst

but does not consider the process synchronization. Kunzle et al. [5] emphasize on the integration of process and data to offer more flexible and efficient process enactment through the PHILharmonicFlows framework. They address modeling and enactment of object-aware processes. Object types and object relations are defined in a data model, while object behavior are expressed in terms of a process whose execution is driven by object attribute changes. However, their work adopts the top-down approach that requires a global view of process model before enactment.

VII. Conclusions

To avoid the gap between a process model at high-abstraction level and its execution in real projects, we proposed an end-user approach that allows process actors to model and enact the process that they perform in practice.

We developed a prototype for our process environment BAPE that supports process actors in modeling and enacting their process fragments. The first contribution of this work is the simple process modeling language SPML which offers structural and fragmented process modeling to process actors. The second contribution is an artifact-centric process engine that enables enacting and synchronizing of processes which are defined, even partially, by several fragments. Our solution makes the process modeling simpler and offers more flexibilities to enactment by giving more freedom to process actors in controlling their activities.

The first feedback of our industrial partners is rather positive thanks to the end-user friendly characteristic of our approach. As future work, we investigate a solution to integrate process actor's working tools to BAPE in order to offer them a consolidated environment to manage their daily activities.

VIII. Acknowledgement

Part of this research has been supported by the French research project FUI ACOVAS which associates AIRBUS, GFI, IRIT, LIEBHERR, NEXEYA, PROMETIL, S2C and ZODIAC. We are grateful to our industrial partners who have helped us on developing the case studies for our research and on validating our prototype.

References

[1] AristaFlow. http://www.aristaflow.com/.

[2] A. Front, D. Rieu, M. Santorum, and F. Movahedian. A participative end-user method for multi-perspective business process elicitation and improvement. *Software & Systems Modeling*, pages 1–24, 2015.

[3] JBoss. jbpm, http://www.jbpm.org/.

[4] M. Kuhrmann, G. Kalus, and M. Then. The process enactment tool framework-transformation of software process models to prepare enactment. *Sci. Comput. Program.*, 79:172–188, Jan. 2014.

[5] V. Künzle and M. Reichert. Philharmonicflows: towards a framework for object-aware process management. *Journal of Software Maintenance and Evolution: Research and Practice*, 23(4):205–244, June 2011.

[6] R. Matinnejad and R. Ramsin. An analytical review of process-centered software engineering environments. In *IEEE 19th International Conference and Workshops on Engineering of Computer-Based Systems, ECBS 2012, Novi Sad, Serbia, April 11-13,2012*, pages 64–73, 2012.

[7] Neo4j. Neo4j website : http://www.neo4j.com.

[8] OASIS. Web Services Human Task (WS-HumanTask) Specification Version 1.1, August 2010.

[9] OMG. Software & Systems Process Engineering Metamodel Specification (SPEM) Version 2.0, Apr. 2008.

[10] OMG. Business Process Model and Notation (BPMN) Version 2.0.2, Dec. 2013.

[11] S. Oppl. Articulation of work process models for organizational alignment and informed information system design. *Information & Management*, 53(5):591 – 608, 2016.

[12] T. Stoitsev, S. Scheidl, F. Flentge, and M. Mühlhäuser. From personal task management to end-user driven business process modeling. In *Proceedings of the 6th International Conference on Business Process Management*, BPM '08, pages 84–99, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] H. N. Tran, M. Hajmoosaei, C. Percebois, A. Front, and C. Roncancio. Integrating run-time changes into system and software process enactment. *Journal of Software: Evolution and Process*, 28(9):762–782, 2016.

Development of an Industrie 4.0 Maturity Index for Small and Medium-Sized Enterprises

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Sebastian Häberer
Fraunhofer Institute for Factory Operation and Automation IFF
Magdeburg, Germany
sebastian.haeberer@iff.fraunhofer.de

Lina Katrin Lau
Fraunhofer Institute for Factory Operation and Automation IFF
Magdeburg, Germany
lina.katrin.lau@iff.fraunhofer.de

Dr. Fabian Behrendt
Fraunhofer Institute for Factory Operation and Automation IFF and Otto von Guericke University
Magdeburg, Germany
fabian.behrendt@iff.fraunhofer.de

Abstract—
The importance of digitization for manufacturing and service operations is growing rapidly. Uncertainties about the actual capabilities expected and the structural impact on value chains are, however, currently impeding the implementation of Industrie 4.0 in SMEs. An Industrie 4.0 maturity assessment can be performed to easily overcome hurdles on the way to the factory of the future. Although some products that determine a company's Industrie 4.0 maturity exist, requirements analyses reveal that none of the existing methods meet the needs of SMEs. Two existing methods have therefore been selected and elucidated and the needs of SMEs have been analyzed. This paper examines the main results of this analysis. Finally, the Industrie 4.0 Quick CheckUp,, a new method developed based on the needs of SMEs, is presented. The paper concludes with a discussion of future prospects for this method.

Keywords— Industrie 4.0, digitization, maturity index, potential analysis, performance evaluation, innovation management

I. Introduction

Digitization is rapidly sweeping German industry. The German Federal Ministry of Education and Research has identified a future industrial manufacturing scenario designated Industrie 4.0. It is intended to provide German industry a means to boost its international competitiveness in the medium and long term. [1] Research and industry concur on the aim of Industrie 4.0. Yet contrary to what this implies, the variety of problems and related challenges render a revolution at the push of a button challenging. The virtual and real world are increasingly converging. Actual products, their components and parts, machines, plants and infrastructural entities exchange relevant and situational information and thus have a direct impact on manufacturing and supply chain operations.

The concept of digitization has to be viewed from a variety of angles. Digitization generally denotes the conversion of analog data for the purpose of electronic processing, storage and transmission [2]. This paper examines digitization from the point of view of industry. Buzzwords such as big data, Internet of Things and, above all, Industrie 4.0 are connected with this. Digitization of processes and information as well as active use of digital technologies are the basic prerequisites for the introduction of Industrie 4.0 [3]. The central component is the linkage of digital worlds with physical systems with so-called cyber-physical systems [4]. Yet contrary to what this implies, there will not be any revolution at the push of a button. Companies' individual orientation as well as the interconnection of the various technologies in different stages of development make Industrie 4.0 a complex undertaking. An objective performance assessment of a company's current capabilities is needed to optimize business processes. To this end, capability maturity models are employed to derive a systematic plan to progressively enhance performance based on the analysis of a company.

This paper examines the importance of digitization to German industry. It additionally presents various capability maturity models designed to help companies define their own path toward digitization. A requirements analysis performed for SMEs revealed that none of the existing methods fully meets their needs. A new method has therefore been developed and is presented briefly in this paper, which closes with a discussion of future prospects for this method.

II. Digitization Requirements

Industrial management is focusing on factories and their infrastructures in the fourth industrial revolution. In the 20th century, ongoing globalization and the increasing scarcity of resources gave rise to intense competition stemming from pressure to be cost effective and innovative. Current conventional value added paradigms primarily apply economies of scale to minimize costs. New digital business models will alter basic conditions faster than many companies suspect, though. Foreign competition has also recognized this potential and operates under other basic legal conditions (e.g. in the realm of data privacy). In Germany, established and disparate value chains are making the alignment of individual companies and the interconnection of the various technologies at different stages of development a complex undertaking in which employees absolutely have to be integrated as well [5]. Industrial manufacturing accounts for nearly one quarter of all jobs in Germany [6]. Manufacturing and the service sector will have to use the opportunities of digitization constructively

so as not to jeopardize the backbone of German prosperity in the future.

The increasing digitization of manufacturing processes will change work in industrial manufacturing significantly in the coming years [7]. People, machines and products will be complementary elements in cyber-physical production systems (CPPS) exchanging digital information with each other and interconnecting physical and virtual objects and processes. “It would appear obvious that the same technologies also harbor risks and can lead to more monitoring, elimination of constraints and rationalization. Concrete approaches to doing ‘good work’ in this complex field hardly exist though” [8]. This is also corroborated by current studies: “The increasing digitization of work processes is making output more transparent. Half of the interviewees expect that employees will play an active role in this development but reject aspects of individualized performance assessment” [9]. Employees will therefore assume a special role in the context of digitization in the future. They are situated among demographic change, changing demands for qualifications and dynamic migratory trends in this complex field. Demographic change, a buzzword of current economic, social and environmental developments, denotes the growing scarcity of qualified labor on the German job market, which is indirectly causing the average retirement age to rise. The resultant longer periods of employment in the working world underscore the need for action because employees’ ability to meet the demands of their jobs diminishes as they age [10, 11]. Manufacturing in particular is increasingly being confronted by the challenge of the rising average age. One out of every three employees in the machinery manufacturing industry is between forty-five and fifty-five years old and performs manual labor over 95% percent of the time, e.g. in final automotive assembly. Apart from psychological stresses, high physical stresses resulting from excessively heavy lifting and poor posture are the other main cause of employee absence. One-third of all absences are attributable to musculoskeletal and psychological ailments. Production losses of € 21 billion were incurred by absences in Germany in 2013 alone [12-14].

Since sweeping change of demands in people’s working world goes hand in hand with change in society, manufacturing workers are increasingly moving into the focus of operational measures. Their position in the factory of the future and the categorization of their skills constitute a major realm of activity. Yet the variety of capabilities of manufacturing and information technology when Industrie 4.0 is implemented and the disparate company profiles as jobs become more complex do not allow any general recommendations or trends. Individual “location tracking” of companies on the path to the factory of the future, i.e. a maturity index assessment, is needed. Their classification in stages of Industrie 4.0 integration is the basis for identifying and placing the "right" measures in a managerially expedient sequence. The Industrie 4.0 CheckUp is intended to deliver information on individual companies and to identify the next necessary steps expediently for decision makers. An abundance of established methods for assessing a company’s maturity index are commercially available.

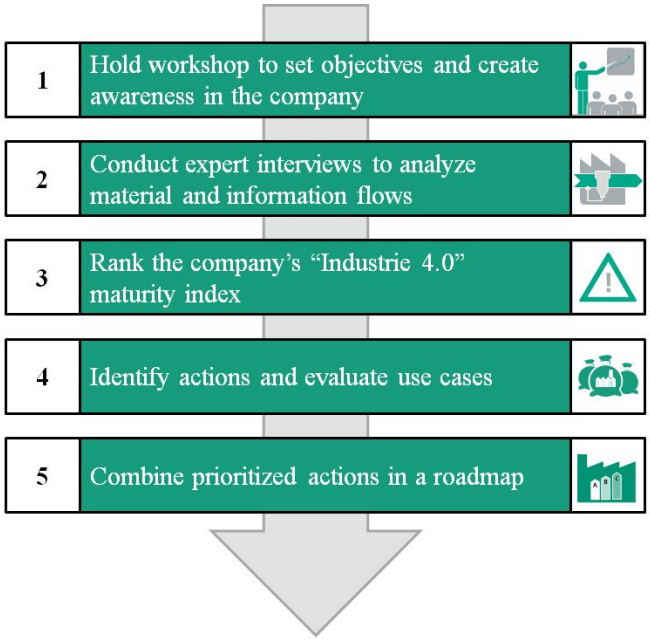
III. Capability Maturity Models

Two of the numerous existing Industrie 4.0 capability maturity models are presented here.

The Fraunhofer IFF Industrie 4.0 CheckUp

The Industrie 4.0 CheckUp entails five steps, each of which is adapted to the object of analysis. The general procedure is presented in Fig. 1

Fig. 1: The Fraunhofer IFF Industrie 4.0 CheckUp

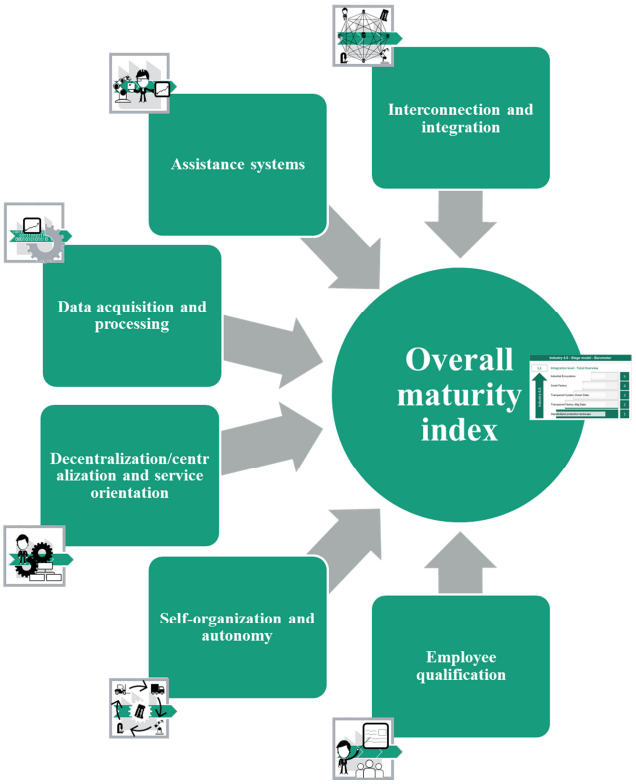


An Industrie 4.0 CheckUp is started by conducting a kickoff workshop with the company in order to make the capabilities and attendant opportunities of digitization tangible. At the workshop, visions of Industrie 4.0 are presented, concrete digitization actions are discussed, and basics of digitization and the desired interconnectivity of the entire value chain are imparted. Expert interviews with select company representatives to compile a common base of data and information constitute the base of knowledge underlying the Industrie 4.0 CheckUp. Along with executives and employees’ longstanding experience, engineers’ planning know-how and technology assessments are also incorporated. This top-down management approach to implementing Industrie 4.0 in combination with a bottom-up improvement is quite promising since the solutions are developed with the participation of value-adding staff and largely accepted as a result. This participatory planning approach has repeatedly proven effective, especially when implementing digitization and automated solutions. The capability to introduce and implement interconnected manufacturing technologies is analyzed by examining:

- the location and corporate environment,
- customers and suppliers,
- the organizational and operational structure, planning and control,
- product and performance development,
- manufacturing, logistics, information and communications,
- workspace design and ergonomics, IT management, service and security, and
- energy and resource management.

A questionnaire that is based on the principles of Industrie 4.0 and completely analyzes fields of activity allows ranking a cross-section of the company in the stages of Industrie 4.0, integration defined by the Fraunhofer IFF. This ranking situates the company in the demands of Industrie 4.0 constitutes the basis for analysis of every structural unit in interconnected manufacturing. Following the principles of Industrie 4.0 presented in Agiplan's study, the Fraunhofer IFF has added an employee qualification module [4]. All of the organizational units included in the Industrie 4.0 CheckUp are presented in Fig. 2

Fig. 2: Overall maturity index assessment derived from an evaluation of the individual functional units of Industrie 4.0



This assessment delivers information on a company's ranking in terms of the requirements of Industrie 4.0 and serves as the basis for analyzing every structural unit of interconnected manufacturing. Since the capability maturity model is modified for each individual company's requirements, objectives and the organization, the companies assessed cannot be benchmarked.

Industrie 4.0 CheckUps performed in companies have revealed that some business units complete more activities and projects intuitively and iteratively than other units. This can result in the coexistence of different stages of Industrie 4.0 integration in one company. Format changes and interface problems among different generations of technology manifest themselves as obstacles to improvement. Since interdisciplinary and process-driven plans are often lacking, the task is to advance every unit to the same stage of integration based on:

- identifying and weighting innovation drivers,
- identifying concrete measures for each unit and placing them in the overall focus,
- analyzing different options for action and likelihoods of success,
- creating a capability maturity model and performing a cost-benefit analysis, and
- providing decision support for potential capital projects.

Concrete measures for each business unit can subsequently be identified and placed in the company's overall focus, always under the premise of avoiding local optima by using interdisciplinary and process-driven plans. Measures for employee awareness creation and training are as much a part of this as changes to and modifications of processes and technology.

The Fraunhofer IFF's experts use assessment models specially modified for digitization to evaluate measures qualitatively or quantitatively as the client desires. These assessments ultimately establish the basis for drawing up a strategy roadmap. This provides a company a digitization strategy with potential migration paths, thus revealing a concrete evolutionary path.

The VDMA Guideline Industrie 4.0

Together with the Technische Universität Darmstadt's Department Computer Integrated Design of the and the University of Karlsruhe's wbk Institute of Production Science, the German Engineering Federation VDMA has compiled a Guideline Industrie 4.0 for the implementation of Industrie 4.0 in German medium SMEs,. Available to VDMA members, it is intended to be a tool that helps SMEs generate new Industrie 4.0 business models and ideas. The area of analysis is the products and manufacturing itself, which are used in a systematic process to identify a particular company's Industrie 4.0 capabilities. It supplies tools and methods for continued individual development of particular strengths and skills. Detailed strategies for the implementation of Industrie 4.0 are dispensed with [15].

The guideline primarily provides instructions for conducting an intra-disciplinary workshop. The analysis phase

and a so-called Toolbox Industrie 4.0 serve as the basis for this. The toolbox, divided into the components of “products” and “production”, is used during the analysis phase and the actual workshop and is intended to be a “continuously evolving impulse generator” [15]. It divides the visions and technologies of Industrie 4.0 into five stages of development. The rows in the toolbox display the individual areas’ levels of application, the columns their stages of development. User-friendly pictograms indicate the different levels of development [15].

Products are distributed among the six application levels (rows) of "Sensors and actuators", "Communication and connectivity", "Data storage and information exchange", "Monitoring", "Product-related IT services" and "Business models". The toolbox can be used for products and their subcomponents. The production component is also divided into six application levels (rows) to optimize manufacturing processes and reduce costs. The application levels are “data processing in manufacturing”, “machine-to-machine communication”, “company-wide interconnectivity with manufacturing”, “ICT infrastructure in manufacturing”, “human-machine interfaces” and “efficiency for small lot sizes”. The five different stages (columns) of Industrie 4.0 in the VDMA toolbox are untitled.

The toolbox is employed in four steps:

1. Select an example with Industrie 4.0 capabilities from the existing product portfolio or manufacturing.
2. Select the promising application level(s).
3. Identify the company’s current stage of development.
4. Generate ideas for expediently improving the stage (s) of development.

There are nevertheless other capability maturity models for self-assessment, which are evaluated below in terms of the needs of SMEs.

IV. Requirements of an SME Performance Assessment Based on Maturity

A prevalent feature of company ranking is size based on the quantitative characteristics of number of employees and annual sales turnover. The Institut für Mittelstandsforschung (IfM) Bonn distinguishes between micro, small and medium-sized businesses and generally defines SMEs as businesses with a maximum size of 500 employees and an annual sales turnover of € 50 million or less [16].

SMEs generate around 37 % of German companies’ total sales turnover and around 57 % of their value added [17]. This underscores the important role played by German SMEs. A great deal of uncertainty about which visions of Industrie 4.0 will actually enable individual companies to continue evolving still remains, though. German SMEs will especially have to maintain and expand their business success in global competition. Implementing another off-the-shelf IT system or purchasing multi-sensor automated solutions such as pseudo-cyber-physical systems promises little success and will not result in the desired improvement. According to McKinsey & Company’s study “Industrie 4.0 - How to navigate digitization of the manufacturing sector”, only six out of ten German companies feel well prepared for Industrie 4.0. Individual “location tracking”, i.e. a maturity assessment, is essential on the path to the factory of the future [7]. Ranking in the stages of Industrie 4.0 integration serves as the basis for identifying proper actions and placing them in a sequence expedient for business. Commercially available capability maturity models are either very time-consuming and expensive or so general that they are incapable of analyzing companies individually.

A workshop catalog was compiled and validated together with six representatives from SMEs in order to develop a method that fully meet the needs of SMEs. The list of criteria for capability maturity models for SMEs is based on the requirements of systematized maturity assessment defined by Bensik [18].

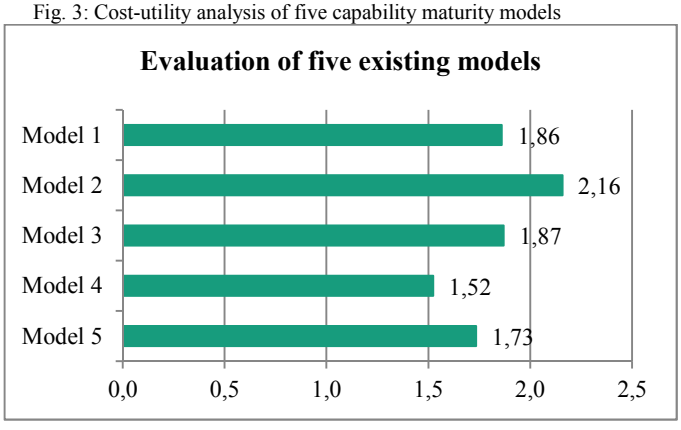
The following requirements for an Industrie 4.0 capability maturity model for SMEs were identified:

- reduced time and costs,
- incorporation of individual company objectives,
- significant results (positioning in relation to Industrie 4.0),
- support to interpret the results,
- independent development of actions,
- no need for specialist knowledge (easily understandable methodology), and
- a software and Internet-based application.

These requirements were used to evaluate five representative Industrie 4.0 capability maturity models.

To this end, the requirements were analyzed in a pairwise comparison and the five models were subsequently evaluated based on the requirements. An anonymized overview of the evaluation is presented in Fig. 3.

Fig. 3: Cost-utility analysis of five capability maturity models



The analysis revealed that none of the existing models meets the needs of SMEs: One independent Industrie 4.0

capability maturity model is inexpensive but its results do not provide an accurate assessment of the implementation of Industrie 4.0. A sound, resource-efficient, independent method of Industrie 4.0 maturity assessment is needed for SMEs.

V. Industrie 4.0 Quick CheckUp

In response, a method has been developed, which includes a capability maturity model that fully meets the needs of SMEs.

Since offering a free and independent version would greatly advance the Fraunhofer IFF's Industrie 4.0 CheckUp, it is being refined by incorporating the new requirements. The Industrie 4.0 Quick CheckUp is a version of the Fraunhofer IFF's Industrie 4.0 CheckUp modified for SMEs.

The method is used for an initial self-assessment. It identifies potentials for improvement in a company, which can be used to take advantage of the economic, environmental and social benefits of Industrie 4.0.

The method is based on the Fraunhofer IFF's Industrie 4.0 stage model of the incremental development required to integrate Industrie 4.0 fully, itself based on the evolutionary nature of Industrie 4.0. The model consists of five stages of Industrie 4.0 integration:

- standards (stage 1),
- big data (stage 2),
- smart data (stage 3),
- dark factory (stage 4), and
- industrial ecosystem (stage 5).

The Quick CheckUp is structured to factor in the complete product life cycle in order to identify potentials for improvement in company. This results in five company units:

- business management,
- development,
- manufacturing,
- supply chain, and
- service.

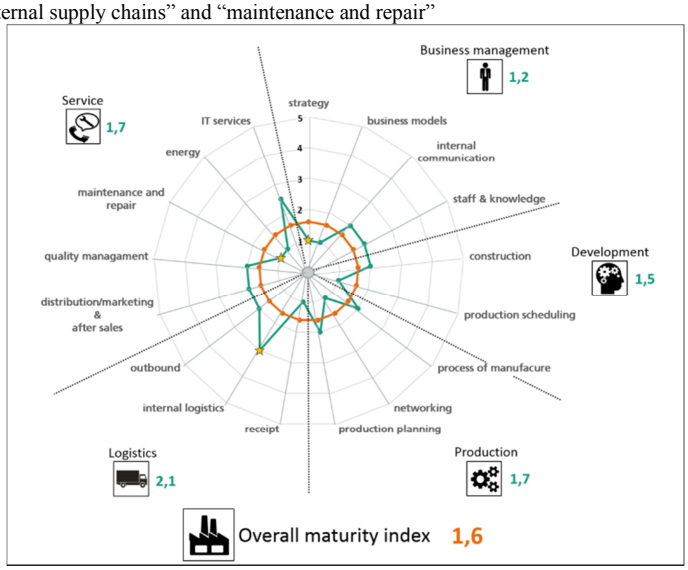
These units are assigned to thematic areas based on the key themes.

Data is collected by a survey covering every topic. The questions were selected to address every company unit and functional unit of Industrie 4.0 (Fig. 2). All of a company's units can be assessed objectively in conjunction with the stages of Industrie 4.0 integration.

An option for independent weighting is included to allow for a company's individual objectives. Weighting can be redistributed and users can weight up to three units individually.

The results of the Industrie 4.0 maturity assessment are visualized in a radar chart. Sample results are presented in Fig. 4. The visualization enables users to recognize which topics exhibit potential for improvement. The topic maturity indices, company unit maturity indices and the overall maturity index are visualized. In addition, the current and next stage of Industrie 4.0 integration is elucidated. Concrete actions for the implementation of Industrie 4.0 are not shown because they have to be formulated individually for each company. Users can develop actions independently or together with researchers from the Fraunhofer IFF.

Fig. 4: The result of the Industrie 4.0 Quick CheckUp's Industrie 4.0 maturity assessment with independent weighting of the topics "strategy", internal supply chains" and "maintenance and repair"



A second area is added to the Industrie 4.0 Quick CheckUp web application to increase user-friendliness and acquire knowledge. Along with a definition of Industrie 4.0 and its capabilities, it provides information on the individual functional units of Industrie 4.0, company units and related topics.

Following successful implementation, users can:

- acquire knowledge independently by using the second subject area of the Industrie 4.0 Quick CheckUp website,
- develop independent actions to integrate Industrie 4.0,
- compare industries and regions as a result of contacting the Fraunhofer IFF,
- have the Fraunhofer IFF evaluate the self-assessment, and
- have the Fraunhofer IFF assist with the interpretation of results and the formulation of actions.

A combined approach is recommended to avoid unprofitable capital expenditures. Knowledge should be acquired independently to develop an Industrie 4.0 strategy independently, which is used to continue developing a company. An evaluation of the self-assessment by the Fraunhofer IFF's Industrie 4.0 experts is useful. Furthermore, actions for the implementation of Industrie 4.0 can be

formulated together with the researchers. Additionally performing an industry and regional comparison is recommended.

VI. Conclusion

The CheckUp spanning every level is a comprehensive set of tools and methods for developing incremental migration to Industrie 4.0. It will become crucially important for companies to be aware of their own position relative to future versions of Industrie 4.0 and to play an active role in change by developing disruptive business models. Digitizing all processes and interconnecting local solutions in one integrated infrastructure ultimately boosts manufacturers' value added.

The goal of the analyses is invariably to deliver sound and assessable recommendations for concrete options for action. The team of researchers is steadily expanding and deepening the Industrie 4.0 CheckUp to keep pace with developments. New manufacturing technologies in conjunction with new (composite) materials such as carbon are continually changing basic conditions, all and any changes having to be factored into the decision-making process. Factories and their infrastructures, jobs and their qualification requirements will change in the future, too, as a function of and in keeping with ongoing development of supply chain and control processes supported by IT.

In all likelihood, these changes will occur even more rapidly and diversely as Industrie 4.0 advances. Plans for the physical arrangement of entire sequences of machines and production lines, including their supply and disposal processes, are priorities in intensive responsiveness analyses. Production systems tuned for quality, cycle time and cost effectiveness generate all sorts of capabilities for volatile manufacturing. They can operate demand-driven and profit-driven in future markets for manufactured end products, while simultaneously conserving resources and energy and generating disruptive business models as well.

References

[1] Federal Ministry of Education and Research, https://www.bmbf.de/de/zukunftsprojekt-industrie-4-0-848.html [access: 04.02.2017], 2016.

[2] Schröder, H.-D., "Digitization", "Digitalisierung" in (Hans-Bredow-Institut Hrsg.), "Medien von A bis Z", VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH Wiesbaden, Wiesbaden, 2006; p. 95–97

[3] Schuh, G, Anderl, R., Gausemeier, J., ten Hompel, M., Wahlster, W., "Industrie 4.0 Maturity Index. Managing the Digital Transformation of Companies" (acatech STUDY), "Industrie 4.0 Maturity Index. Die digitale Transformation von Unternehmen gestalten", Herbert Utz Verlag, 2017

[4] agiplan GmbH; Fraunhofer IML; ZENIT GmbH, "Conclude the potential of the application of ‚Industrie 4.0' in SMEs", "Erschliessen der Potenziale der Anwendung von Industrie 4.0' im Mittelstand", Mühlheim an der Ruhr, 2015

[5] Federal Ministry of Labour and Social Affairs, "Work 4.0. Keep working", "Arbeiten 4.0. Arbeit weiter denken", Berlin, 2015

[6] Federal Office of Statistics (Dstasia), https://www.destatis.de/DE/ZahlenFakten/Indikatoren/LangeReihen/Arbeitsmarkt/lrerw013.html [access: 29.03.2017], Wiesbaden, 2017

[7] McKinsy & Company, "Industry 4.0 - How to navigate digitization often the manufacturing sector", https://www.mckinsey.de/files/mck_industry_40_report.pdf [access: 24.03.2017], McKinsy Digital, 2015.

[8] W. Apt, M. Bovenschulte, E.A. et al., foresight study "Digital working world", "Digitale Arbeitswelt" for the Federal Ministry of Labour and Social Affair., Germany, 2016

[9] W. Bauer, ingenics study "Industry 4.0 - A revolution in work design. How automation and digitization will change our production", "Industrie 4.0 – Eine Revolution der Arbeitsgestaltung. Wie Automatisierung und Digitalisierung unsere Produktion verändern werden", 2014

[10] M. Schenk, S. Wirth, E. Müller, "Factory planning and factory operation. Methods for the transformable, networked and resource efficient factory", „Fabrikplanung und Fabrikbetrieb. Methoden für die wandlungsfähige, vernetzte und ressourceneffiziente Fabrik", 2nd edition, Springer, Berlin Heidelberg, 2014

[11] Federal Ministry of Labour and Social Affairs, report by the Federal Government, "Departure into the age-appropriate working world", "Aufbruch in die altersgerechte Arbeitswelt", Germany, 2010

[12] M. Holweg, F. K. Pil, "The Second Century - Reconnecting Customer and Value Chain through Build-to-Ord", Vambridge: MIT, Press 2004

[13] VDMA, Volkswirtschaft und "Statistik, Statistical manual for mechanical engineering", "Statistisches Handbuch für den Maschinenbau", 2012

[14] Federal Institute for Occupational Safety and Health, „Security and healthiness at work", "Sicherheit und Gesundheit bei der Arbeit, 2013

[15] VDMA, „Industry 4.0 guide", "Leitfaden Industrie 4.0", Frakfurt am Main, VDMA-Verlag, 2015

[16] Institut für Mittelstandsforschung IfM, "SME definition of IfM Bonn", "KMU-Definition des IfM Bonn", http://www.ifm-bonn.org/definitionen/kmu-definition-des-ifm-bonn/ [access: 08.02.2017], 2016

[17] Institut für Mittelstandsforschung IfM, "SMEs - Key figures for SMEs as defined by the IfM Bonn", "Mittelstand im Überblick – Kennzahlen der KMU nach Definition des IfM Bonn", www.ifm-bonn.org/statistiken/mittelstand-im-ueberblick/#accordion=0&tab=1 [access: 24.03.2017], December 2016

[18] T. Bensiek, "System for the maturity-based performance assessment and increase of business processes in the middle class", "Systematik zur reifegradbasierten Leistungsbewertung und -steigerung von Geschäftsprozessen im Mittelstand", Paderborn: University Paderborn, Dissertation, 2013

Survey of the Renovation Sector in France: A Necessary Step Towards Implementing BIM in Microenterprises

(Presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

L. Joblot* **, D. Deneux**,
*Arts et Métiers ParisTech
151 boulevard de l'Hôpital, 75013 Paris, France
Laurent.joblot@ensam.eu

T. Paviot*, S. Lamouri* **
**LAMIH-UMR CNRS 8201
UVHC, Le Mont-Houy
59313 Valenciennes Cedex 9, France

Abstract—In recent years, Building Information Modeling (BIM) has spread throughout many Architecture, Engineering and Construction (AEC) companies and contributes to "digitizing" these sectors. Are there any limits to this expansion? Opinions seem to diverge on this matter. What are some of the conditions enabling it to be implemented in what might be called the "average model company" of the French construction sector; that is, a Microenterprise (µEnt) that primarily does renovations? Implementation templates or studies have little to do with µEnt and renovation. In order to provide a better definition for the activities, needs and limitations of this average business model and for a better understanding of the mechanisms required for implementing BIM in eligible companies, a survey has been carried out. The main outcomes of this survey are reported in this article.

Keywords— Building Information Modeling (BIM), Survey, Renovation, Microenterprise

I. Introduction

"Industry4.0" was publicly promoted in 2011 in Germany [1] and spread throughout Europe in 2013, after the "Final report of Industry4.0 Working Group" [2] was released. Since then, in many sectors keywords and concepts such as the Internet of Things and Services, smart machines autonomously exchanging information, facilitating optimized decision-making, dynamic business and engineering processes, novel business models, and more, are used to describe future industrial developments. In this expected future World, people, machines, logistic systems and products will communicate and cooperate with each other. This development also concerns the AEC Industry. Management consulting firms use the same vocabulary to describe the transition that awaits this sector and present the transformations imposed by the implementation of BIM [3], [4]. They tend to predict that *"companies that continue to ignore the digital wave will struggle to survive"*. The mere definition of BIM by the "worldwide authority" (http://buildingsmart.org) reflects the parallel that can be made with Industry4.0. BIM is defined as "*a digital representation of physical and functional characteristics of a facility. A BIM is a shared knowledge resource for information about a facility forming a reliable basis for decisions during its life-cycle; defined as existing from earliest conception to demolition*". It is important for AEC companies to become aware of their digital capabilities and to widely modify their behaviors, practices and organization. The French renovation sector will be impacted, even though its specific characteristics will likely make the evolution a long and complex process. This sector is of particular interest for several different reasons: it represents 57% of the French construction market. The major part of this market (55%) is shared by 650,000 companies that employ fewer than 20 individuals, and represents an annual turnover close to 42 billion Euros [5]. This results in a very diverse and stimulating case study. The typology of companies defined by the European Commission [6] was considered as a reference framework: Microenterprise (µEnt) stands for enterprises that have fewer than 10 employees and an annual turnover of below €2 million. There are around 635,000 of these enterprises and they represent an annual turnover close to 33.5 billion Euros of renovation work in France. Small enterprises (SmEnt) are those with fewer than 50 employees and an annual turnover below €10 million, while small and medium-sized enterprises (SMEs) employ fewer than 250 individuals. It is difficult to provide appropriate support to µEnt in their process of adopting the BIM technology without characterizing their current situation, understanding their needs and identifying their potential step-by-step improvements. Almost everything has to be done in this area. As proof, in the sector of energy-driven retrofits of existing buildings, the most frequently studied panel from a scientific point of view (about 60% of scientific publications [7] despite a relative significance of only 1/3 of renovation works [8]), BIM is still considered immature "*because of technical, informational and organizational complications*" [9]. The great diversity of businesses and trades [10], added to the lack of available, significant and ready-to-use scientific data [7] by small companies with tight budgets, are reasons stimulating and justifying this work.

This article includes three main sections. The first aims at detailing the different possible scenarios for implanting BIM in SmEnt, which mostly constitutes the French renovation sector. The decision to carry out a survey with 34 companies to better identify and map the environment as a whole will then be justified. The methodology used for this latter and the main results are detailed in the third section. The conclusion will introduce future works, i.e., how the collected information will be used and will enable the development of a BIM maturity referential, adapted to this particular and important renovation sector.

II. Implementation scenarios

A first approach to accelerate BIM implementation in France could be to transpose what leading countries in Europe, such as the UK, have achieved in the past 5 years. Similarities exist with the UK, where µEnt represent 93% of the construction sector [11]. There is also a resemblance between the 2012 BIM posture of the UK's AEC industry [12] and the French industry in early 2017 [13]. According to these studies, only 25% of SMEs were using 3D digital models. 81% of SMEs' contractors were only vaguely familiar with the levels of BIM while 32% of 400 French SmEnts' executives had

heard about it. At least 3% of SmEnt surveyed planned to integrate BIM in the upcoming months. Despite this, however, a cautious approach is advisable. The French and UK construction market are different. Repair and Maintenance of buildings only account for 38% in the UK's construction outcomes (vs. 57% in France) [14]. Another study in 2013 is also of interest to, but not representative of, the UK's present situation [15]. In this "*survey of current status of and perceived changes required for BIM adoption in the UK"*, only 13% of answers were from enterprises with fewer than 20 employees, 90% of organizations that responded were using BIM tools, but 3 years later, in a representative sample of UK construction SMEs [16], it was stipulated that out of the surveyed SMEs, "*mostly SME Companies (75%) in UK are at BIM infant level, as their business processes involve 2D workflows, generally using AutoCAD, have some knowledge of BIM and have yet to start their BIM journey….Around 75% of construction SMEs which form the backbone of construction market are non BIM"*. Works that present strategies to implement BIM in a practical context were also considered [17] [18]. The input model (project cost of several million Euro, size and sector of the industry, systematic use of accepted construction project management practices, use of a formal technical path scheduling techniques to estimate the duration or structure project, Work Breakdown Structures of the physical project work) revealed to be out of the scope and the transposition of it to the French renovation sector was found to be too risky.

Previous studies have shown that the needs of stakeholders and the diversity of renovation fields are extremely varied [7][10]. Implementing BIM requires developments with and for users with a logical step-by-step evolution, on-site, by gradually capitalizing [19]. BIM has to adapt to these realities, and the environment should also adapt to BIM. Important gaps have to be addressed and a more in-depth analysis of this French area is necessary to map business processes and uses in renovation. For example, 3D scanning is often mentioned as a key lever and many solutions are imagined by the scientific community [20]. However, the ROI of the process, change effort and the added-value of involving a capable surveyor remain unclear. Should it become systematic? It is also difficult for a µEnt to identify affordable partners that master the scanning technology. Architects are more likely to use it, but they have very few relationships with the renovation shareholders. So as to characterize BIM implementation scenarios, it is necessary to map the situations that coexist today and those that could coexist tomorrow. Practically, this means accounting for the French renovation sector according to the point of view of diversity in renovation, which is mostly performed by µEnt and SmEnt. Renovation processes in public projects are formally prescribed in French law. Depending on European thresholds, public customers in France can choose between 4 regulatory process [21]; **Adapted procedure, Competition, Competitive Procedure with negotiation** or **Competitive dialogue tender**. All this is part of the "MOP law" published by the French Government [22], which governs public project contracts and their relationship with private project management. The aim of the current study is to better understand the way of working outside these later cases. Indeed, 75% of the French market is constituted by individuals and private clients [23]. The remainder is state and local authorities. The regulatory process that was previously defined impacts not more than 15% of the enterprises with fewer than 20 employees [5].

In this context, best practices from abroad cannot be directly transferred to the French renovation sector without further examining what the current situation of this sector is. This is why a survey and interviews with companies mainly engaged in renovation was undertaken and is presented in the next section.

III. Objective of the survey

A. Representative nature of the survey

According to the Trades Construction Observatory[10], construction covers 2 major sectors: "building companies" and "Civil Engineering Industry". This latter, representing only 7% of the total, is considered out of the scope of this work.

34 companies were interviewed. 85% of the panel consisted of individuals in place for 16 years (with a standard deviation (SD) of 9 years). The panel was split into 2 complementary classes:

- Group A (GrA) comprised of 23 "technical companies". 70% only have one laying activity; the others are both a manufacturer and layer.
- Group B (GrB) includes Architects, Project Managers, Construction Consultants and Designers, Quantity surveyors, Geometrician expert, Commercial and Social Landlords. Their primary missions are to advise and organize projects.

In accordance with the major trends described above, 70% of GrA realize more than 80% of their projects' turnover in renovation. 65% work is done exclusively for individuals. For 90% of the group, tenders and processes related to the MOP law represent less than 10% of the activity. The average turnover of these companies is between 250 and 500 k€. The objective was to be as representative as possible, so the survey was carried out as long as necessary to obtain the mapping of a sufficient number of projects, specifically here 3200/year for the whole GrA, mainly carried out by Microenterprises (average of 7,65 employees SD of 7,5). This panel was formed from networking, advice from the French Building Federation and snowball sampling [24]. The survey was mainly conducted in the Bourgogne-Franche-Comté (BFC) Region. BFC has an enterprise size distribution very close to national figures as shown in "TABLE 1", synthesized from [25]. From a "technological or cultural" point of view, it is equally possible to justify this restrictive choice by two studies that tend to cancel each other out: [13] highlights that, in France, the BIM integration and knowledge level are quite steady over different regions, estimated between 23% and 37%.

TABLE 1 French Construction size distribution of enterprises

size / total	0	1 to 9 employees	10 to 19 employees	20 to 49 employees	More than 50
Enterprises in France	666 739	469 267	166 446	17 887	9 764	3 375
distribution	100%	70,4%	25%	2,7%	1,5%	0,5%
Enterprises in BFC	24 271	16 014	6 948	775	431	103
distribution	100%	66%	28,6%	3,2%	1,8%	0,4%

This corresponds to the percentage of SMEs that have only heard about BIM. Unexpectedly, the Paris region is just under the average, with 30%. At the same time, the engineering (GrB) is well documented in a recent study [26], showing that BIM projects are concentrated around Paris. Therefore, a survey focused on a French region is capable of representing the overall situation in France.

B. Methodology

Data were collected through 1 to 3 hours of face-to-face interviews, between Dec. 2016 and Feb. 2017. They were semi-structured, with open and closed ended questions. This approach allows the researcher to discover the way people work, visiting offices, identifying software, computer means, and sometimes machines or inventories to better understand the type of company it is. As described in [27], the intention here is "descriptive," "explanatory," and certainly in works that will follow, "prescriptive". As also noted in this book, the objectives of these interviews are mid-way between consulting (*obtaining results that enrich knowledge and the corpus of existing hypotheses*) and research-intervention (*no preconceived outcome*) for the continued work. In this sense, the approach adopted is quite similar to the one used in [28] for the UK construction industry. The aim is to identify a "*true picture of the situation*", the needs and uses of the companies to which the BIM might respond and lastly the optimal way to achieve it. So a representative population of the renovation sector (GrA) was chosen so as not to distort the photography and then to understand and build models on a strong basis, making them as realistic as possible. Individuals in the main trades and professions involved in renovation, as defined by the Observatory of the Construction Industry [10], were interviewed (electricity, plumbing, roofer, mason, etc.). Companies apply to several work areas, so the particularities of each of the 18 main trades and professions were addressed between 2 and 7 times during the survey. The feedback did not stop with just their main trades and profession, since each stakeholder is in contact with many others; therefore, the working habits of a much broader panel of companies and professions were also addressed.

The objective transcription of the data collected following the combination of qualitative and quantitative questions to identify the uses of µEnt was not easy [26]; however, the results of this survey are presented in the following section.

IV. Major results

A. Technical actors (GrA)

Of the around one hundred points covered during the interviews, only the useful parts for decrypting the functions of the SmEnt are presented here. Some elements or details of the survey will, however, be useful in the continuation of the work, which is detailed in part V.

1) General information

Among the key observations, raising some doubts about the introduction and use of the digital mock-up or BIM, we can point out, for example, that 60% of GrA work without quote software or simply use MS EXCEL. More than 95% have no procedures and 2/3 work without indicators. They practically have no relationship (90%) with surveyors who can perform 2D or 3D studies. For 50%, 2D drawings are provided less than 20% of the time, and only 15% of them are confident in using 100% exploitable and up-to-date drawings. 33% estimate that 2D drawings are useless in renovation activities. When they realize a drawing or diagram, they do it by hand or without dedicated software (50%). Less than 40% are equipped with computer solutions for this, yet it is necessary in 44% of cases (SD of 41). In this context, it is understood that only 13% have ever used electronic data interchange systems for their 2D drawings (very rarely) and favor the exchange of information by phone, email or on-site meetings.

Other elements that can be helpful for future works are presented "Fig 1 & 2". In these representations, the bubble size represents the proportion of turnover obtained from an invitation to tender for each company. "Fig 1" clearly indicates that a correlation between the size of the enterprise and the relationship it weaves with Architects exists. In the same way, the need for companies to respond to calls for tender and to seek to work on public projects tends to grow with its size. "Fig 2" shows the involvement of companies in new constructions according to the firm size. These two elements seem to be linked and this was explained during the interviews by the fact that companies with a relatively high payroll need to identify and smooth their load, and also to fill their order backlog.

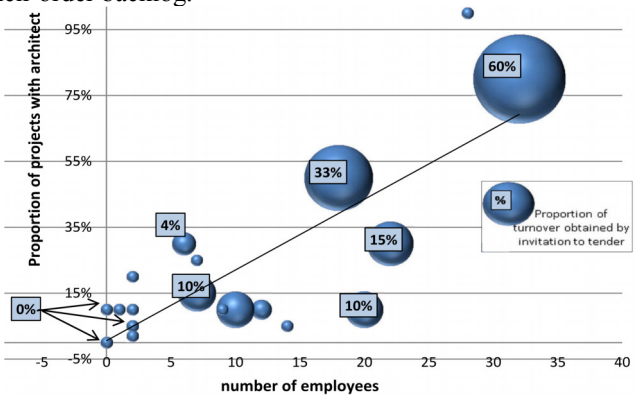


Fig. 1 Responding to calls for tender and proportion of projects with architects according to the number of employees.

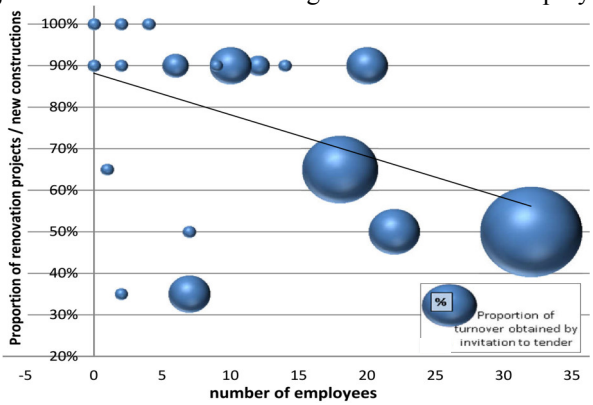


Fig. 2 Proportion of renovation and new construction according to the number of employees.

This is then made possible by new constructions, which require more advance planning, and also by invitation to tender (proportional to bubble size).

2) Technology, 3D and BIM knowledge

The aim of this survey is to understand, among other items, what BIM could provide and help to resolve. It is important to note that major difficulties include unexpected events and their consequences in terms of planning, workload and costs (56% of the answers), but also relationships, increasing workload and red tape year after year at once for 1/3 of those surveyed. During projects, partners use tablets in fewer than 6% of the cases and they are absent in 7 cases out of 10.

After the general principles of BIM were presented to participants (only 26% had previously heard of it at all), a third were not concerned by BIM. Time and money investments and complexity are fears raised and revealed by more than 50% of those surveyed. Almost 20% of the projects or interventions are carried out on sites where the company has already intervened. This point emphasizes the information and management capacities proposed by BIM. According to our interviews, changes have already taken place in the construction sector since 80% of orders are already made online (existence of some business software for visualization, order components via WEB portals and QR and barcodes scanning on Smartphone). Despite the fact that construction actors are attached to their benchmark and their working methodology, some expectations are still persistent and could be bridged by BIM and 3D. It is possible to mention, for example, the improvement of exchanges and interoperability, or reducing conflicts for 26% of the participants, and better apprehension, visualization and anticipation of difficulties by 44%. Finally, this could facilitate an understanding and decision-making on the part of customers, 57%.

Unfortunately, the path is still long, since 65% of renovation projects are done without 3D and in more than 8 times out of 10, 3D is used for fewer than 10% of the projects. 3D models are read via acrobat viewer, only one person uses Autodesk AUTOCAD. In other rare cases, 3D uses correspond to visualization via computer-aided manufacturing software (for manufacturers) or reseller applications (kitchen specialists).

B. Group B

GrB is composed of a construction economist, two members of design office (one digitization specialist for deconstruction, and one thermal), 5 architectural firms ranging in size from 1 to 6 people, a sales rep in custom joinery, the chairman of a company of 25 geometrician experts and a project manager in a BFC Social Landlords, which rehabilitate 800 housing units per year (http://www.opacsaoneetloire.fr). There are therefore 8 µEnt, 1SmEnt and 2SMEs, which allows for an extension of the field of analysis. The panel respects the scope, since for 90% of businesses, more than 50% of their activity is in renovation. Fairly logically, half of the activity of this group is realized through tenders and are subject to the MOP law. Difficulties encountered are contingencies in 63% of cases, but the latter was said to be reduced by scanning after minor demolitions. Relationships with clients and employees are other noted challenges (36%), particularly in occupied sites, and finally, technology and administrative (27%), especially for requesting and obtaining planning permission in classified areas. There are stronger relationships between these actors and the geometrician experts that intervene in 3 projects out of 10 for topographical surveys and two times less often for internal surveys. 91% of this panel is equipped with 2D and 3D software. 3D is used in 53% of renovation projects. The strengths of this approach (with the same response rate of 36%) are the reduction of conflicts, the possibility of storing and visualizing more information, assistance for client appropriation and a better understanding of the difficulties. Realistic renderings are also appreciated for insertions built in the landscape, an obligatory stage for obtaining permission.

Due to the specific nature of the professions combined with a "higher level" of studies (high school diploma plus 2 years for 90% of GrB versus a maximum of high school diploma level for 60% of GrA), GrB members logically have fewer difficulties in projecting themselves in BIM, as far as is known for 72% of them. Only one representative of this group works in BIM. Autodesk REVIT and Abvent ARCHICAD are present in 45% of companies, but only to be able to open files. Except for the digitization and demolition specialist, no collaborative work is set up even for the social landlords that realize more than 1000 interventions each year, of which 75% are in renovation. There is a growing awareness and it is possible to point out that the people interviewed feel that their working methods are about to change (more than 5 cases out of 10). They expect BIM to be able to improve interoperability (72%) and reduce conflicts by facilitating anticipation (45%). This point will also be important for future works because these kinds of actors are essential links in the transformation of working methodologies of the renovation profession as a whole. They are likely to be pioneering or privileged users of the digital model. They will certainly take the lead on BIM procedures or conventions in effect on projects. This awareness and this will to evolve is therefore a plus on which we can rely in the future. Lastly, profitability and sharing of investments are the main concern for 72% of GrB. All architects highlight that the additional time and fees for them must be taken in consideration if they have to play the role of BIM Manager.

C. Fundamental contrasts

To conclude this analysis, exposing some remarkable results, “Fig 3” shows few differences found (Proportion of positive responses) within the two groups that impact the companies’ capacities and investment needs. A third of GrA do not feel concerned with BIM and changing in the near future. It is, however, interesting to contrast this feeling with a recent survey [29] showing that only about a third of owners judged that craftsmen were well coordinated, according to 1500 people interviewed after renovation works. It is important to add that current and future constructions built with BIM will one day need to be renovated, or that whatever happens, new business start-ups are emerging with BIM. They are logically less conservative and will not have to implement BIM since they are innate users.

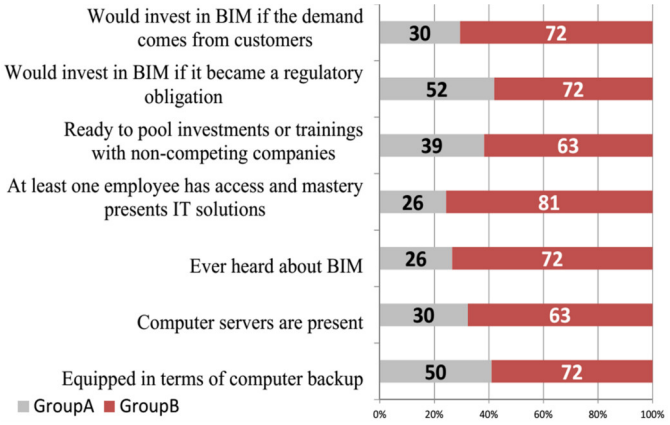


Fig. 3 Fundamental contrasts

These competitors, this latter point, and several difficulties and issues revealed in the survey, are sufficient to justify that changes are coming.

Data collected during the interviews corroborate well with recent studies, highlighting the persistence of a traditional 2D paper-based in AEC [30]. This is often explained by the history and organization of the profession itself. It was therefore necessary to organize exchanges, interviews and observation of the current practices. This has allowed for a better understanding of "*new insights into how to make the transition easier and more functional*". In conclusion, the following section presents how this is planned to work in the future, what will be our core target group, and subsequently explains why the creation of a BIM maturity framework adapted to the sector seems to be relevant.

V. Lessons learned and implications for future works

In preliminary works, the authors had planned to carry out this survey [7] "*to gain a better overview of relevant renovation processes and actors…to establish a maturity measurement tool for the renovation sector as flexible as possible to fit to the majority of scenarios observed in the field.*". The present study confirms that the task will be significant, alongside AEC professionals, to deploy BIM and a fortiori in companies performing renovation. Both computing and organizational means have to be provided to achieve this goal. This survey clearly reinforces recent studies carried out with 400 SmEnt that point out that only 1% of them actually have a BIM approach [13]. The current study has attempted to bring a description, as accurate as possible, of the French companies that realize the bulk of the repair and renovation projects of buildings. Presently, it seems complicated, as a first step, to respond and contribute to the transformation of all of the business models encountered in the field. As indicated in Part IV, some professions (GrB) are more ready to invest and more aware of this urgency and the opportunities that BIM can provide. The imminent decision of the French government to impose a certain level of BIM during tenders, the emergence of digital models in new constructions coupled with interviews and analysis of figures 1 and 2, lead to a conclusion that companies with fewer than 5 employees may be considered only in exceptional cases, and without high priority, in future works. Priority targets for future development are also the ones that will be most quickly impacted by BIM and are firms with the following features:

- Between 5 and 50 employees
- Making (but not exclusively) renovation work
- Taking part in few public procurement procedures
- In contact with the professions targeted by GrB.

Thanks to this survey, their workflows, the means at their disposal, exchange and communication vectors that are usually set up, the "concrete" needs and challenges they face, and more, are now characterized. The specificities of this sector, as well as some habits or state of mind differences have emerged and appear to depend on the membership of GrA or GrB, but also depend on the activity; whether they are a layer and/or manufacturer. Although only partially presented before, some professions are evolving and the emergence of new skills or requirements, such as property or facility management, house automation, etc., were addressed during the interviews and will be taken into consideration in models to be constructed. SmEnt have to be supported during their evolution. For this, it still seems appropriate to establish a maturity measurement tool and associated metrics for the renovation sector [7]. In full agreement with other works, these developments should help adopt a *collaborative mindset contributes towards some of the aims of BIM* [31]. To enhance this, some collaborative models discussed and implemented by audited companies (but not presented here) would be in-depth and at the heart of the reflection. Many BIM maturity matrices already presented [32] [33], will be analyzed in greater detail, but other authors [16] share this sense of a lack of an adapted tool, and highlight that for a BIM novice (or BIM infant) SMEs, *using an inappropriate BIM assessment & adoption strategy might result into waste of valuable time & resources*. So as not to overthrow the codes and uses of the companies encountered and lead to an evolution within this matrix, the mapping and the processes identified here will be a skeleton of the future maturity measurement tools designed. To increase the strength of these tools, we will also synthesize and take into account key factors of the success of Implementations of BIM and ERP in SMEs, as well as process and methodology advised in BIM conventions.

Our ambition is to contribute to progress from an equivalent of the level 0 defined in the UK, highlighted in [15], and widely encountered among the 34 companies surveyed, to a specific French renovation sector at Level 3, corresponding more precisely to the industry 4.0 that was initially presented. The number of steps and the levers to achieve this remain to be defined and will be subsequently submitted.

References

[1] H. Kagermann, W.-D. Lukas, & W. Wahlster, "Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution" *VDI Nachrichten*, vol.13, 2011.

[2] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, Recommendations for Implementing the strategic initiative INDUSTRIE 4.0. Forschungsunion, 2013.

[3] P. Gerbert, S. Castagnino, C. Rothballer, A. Renz, and R. Filitz, "The Transformative Power of Building Information Modeling," bcgperspectives.com, accessed Mar. 1. 2017.

[4] McKinsey, “Accélérer la mutation numérique des entreprises : un gisement de croissance et de compétitivité pour la France,” mckinsey.com/, accessed Jun. 29. 2016. .
[5] CAPEB, “artisanat du bâtiment Les chiffres clés 2015,” capeb.fr/chiffres-cles-segmentation-region-metiers-poids-et-evolution-secteur/, accessed Feb. 8. 2016. .
[6] Europeen commission, “Micro-, small- and medium-sized enterprises: definition and scope,” http://eur-lex.europa.eu/eli/reco/2003/361/oj, accessed Sep. 7. 2016.
[7] L. Joblot, T. Paviot, D. Deneux, and S. Lamouri, “Analyse du BIM appliqué à la rénovation,” 11th International Conference on Modeling, Optimization and SIMulation (MOSIM 2016), Montréal, Canada, Aug. 2016.
[8] ademe, “Bâtiment : chiffres clés 2013,” ademe.fr/sites/default/files/assets/documents/chiffres-cles-batiment-edition-2013-8123.pdf. .
[9] M. Khaddaj and I. Srour, “Using BIM to Retrofit Existing Buildings,” *Procedia Eng.*, vol.145, pp.1526–1533, 2016.
[10] Observatoire des métiers du BTP, “Chiffres clés des entreprises,” metiers-btp.fr/les-chiffres/les-entreprises.html, accessed Nov. 9. 2016. .
[11] Office for National Statistics, “UKBC Enterprise/local units by Industry, Employment size band and Legal status,” http://web.ons.gov.uk/, accessed Mar. 8. 2017. .
[12] National Federation of Builders, “building information modelling BIM: ready or not?,” http://builders.org.uk/eblock/services/resources.ashx/000/318/333/NFB_BIM_Survey_BIM-ready_or_not.pdf, accessed Jan. 3. 2017.
[13] M. Giroud, “Baromètre des entreprises du bâtiment / KPMG,” home.kpmg.com/fr/fr/home/media/press-releases/2017/01/barometre-des-entreprises-du-batiment.html, accessed Mar. 1. 2017.
[14] Experian, “UK Construction Forecast A report by the Forecasting Committee for the Construction Industries,” experian.co.uk/assets/economics/samples/uk-construction-forecast-sample.pdf, accessed Mar. 8. 2017.
[15] R. Eadie, M. Browne, H. Odeyinka, C. McKeown, and S. McNiff, “A survey of current status of and perceived changes required for BIM adoption in the UK,” *Built Environ. Proj. Asset Manag.*, vol.5, no. 1, pp.4–21, 2015.
[16] A. Ghaffarianhoseini, D. T. Doan, T. Zhang, A. Ghaffarianhoseini, N. Naismith, and J. Tookey, “A BIM Readiness & Implementation Strategy for SME Construction Companies in the UK,” Proc. of the 33rd CIB W78 Conference 2016, Oct. 31st – Nov. 2nd 2016, Brisbane, Australia.
[17] T. Hartmann, H. van Meerveld, N. Vossebeld, and A. Adriaanse, “Aligning building information model tools and construction management methods,” *Autom. Constr.*, vol.22, pp.605–613, 2012.
[18] B. Gledson, D. Henry, and P. Bleanch, “Does size matter? Experiences and perspectives of BIM implementation from large and SME construction contractors,” 1st UK Academic Conference on Building Information Management (BIM) 2012, Northumbria University, Newcastle upon Tyne, UK, Sep. 2012.
[19] R. Miettinen and S. Paavola, “Beyond the BIM utopia: Approaches to the development and implementation of building information modeling,” *Autom. Constr.*, vol.43, pp.84–91, Jul. 2014.
[20] R. Volk, J. Stengel, and F. Schultmann, “Building Information Modeling (BIM) for existing buildings — Literature review and future needs,” *Autom. Constr.*, vol.38, pp.109–127, Mar. 2014.
[21] MIQCP, “La réforme de la commande publique de 2016 applicable à la commande de projets de construction : quelles évolutions pour choisir les concepteurs ?,” miqcp.gouv.fr/, accessed Jan. 5. 2017. .
[22] Ministère français de l’industrie, des postes et télécommunications et du commerce extérieur, “modalités techniques d’exécution des éléments de mission de maîtrise d’oeuvre confiés par des maîtres d’ouvrage publics à des prestataires de droit privé,” legifrance.gouv.fr/, accessed Feb. 5. 2016. .
[23] CONSTRUCTIF, “Grands chiffres et figures - Le Bâtiment : regards, enjeux, défis,” constructif.fr/bibliotheque/, accessed Feb. 6. 2017
[24] L. A. Goodman, “Snowball Sampling,” *Ann. Math. Stat.*, vol.32, no. 1, pp.148–170, 1961.
[25] Insee, “Caractéristiques des établissements au 31/12/2014,” insee.fr/fr/statistiques/2021289, accessed Nov. 8. 2016. .
[26] KYU Lab pour OPIIEC –, “Etude sur l’évolution de l’ingénierie française de la construction liée au BIM,” eduscol.education.fr/sti/sites/eduscol.education.fr.sti/files/ressources/techniques/8118/8118-la-synthese.pdf. .
[27] H. Savall and V. Zardet, Recherche en Sciences de Gestion : Approche Qualimétrique, observer l’objet complexe. Economica, 2004.
[28] F. Khosrowshahi & Y. Arayici, “Roadmap for implementation of BIM in the UK construction industry” *Eng. Constr. Archit. Mng.*, vol.19, no. 6, pp.610–635, 2012.
[29] ademe, “OPEN : Observatoire permanent de l’amélioration énergétique du logement,” ademe.fr/open-observatoire-permanent-lamelioration-energetique-logement-campagne-2015, accessed Feb. 6. 2017. .
[30] Cuš Babic and Danijel Rebolj, “Culture change in construction industry: from 2D toward BIM based construction,” itcon.org/cgi-bin/works/Show?2016_6, accessed Aug. 9. 2016. .
[31] S. Mellon and T. Kouider, “SMEs and BIM in preparation for 2016: a case study.,” Architectural Technology, Towards Innovative Professional Practice: Conference Proceedings of the 5th International Congress of Architectural Technology, Aberdeen 2014, 2014.
[32] A. Copping, P. Shepherd, and A. Azzouz, “Emergence of Building Information Modelling Assessment Methods,” *Integr. Des. 50 Build. Our Future*, 2016.
[33] B. Succar, W. Sher, and A. Williams, “Measuring BIM performance: Five metrics,” *Archit. Eng. Des. Manag.*, vol.8, no. 2, pp.120–142, May 2012.

Industry 4.0 and the SME: a technology-focused review of the empirical literature

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Alexandre MOEUF
ISMEP-Supmeca
QUARTZ
Saint-Ouen, France
alexandre.moeuf@exxelia.com

Romain EBURDY
Mines Paristech
CAOR
Paris, France
romain.eburdy@mines-paristech.fr

Samir LAMOURI
Arts et Métiers Paristech
LAMIH
Paris, France
samir.lamouri@ensam.eu

Simon TAMAYO
Mines Paristech
CAOR
Paris, France
simon.tamayo@mines-paristech.fr

Robert PELLERIN
Polytechnique Montréal
Department of Mathematical and Industrial Engineering
Montréal, Canada
Robert.pellerin@polymtl.ca

Abstract— The concept of Industry 4.0 has many advantages for Small and Medium-sized Enterprises (SME) in relation to other industrial management methods such as Just-In-Time or MRPII. Its adoption can be achieved through the use of many technologies. However, these technologies are not well mastered by SMEs. In order to identify the difficulties encountered by SMEs, this article presents an analysis of the exploitation of the different means of realization of the industry 4.0. From a scientific literature review, we show disparities in the exploitation of different technological group. Among the important elements, we note a sub-consideration of the data generated as a source of added value for SMEs, an under-exploitation of certain means of realization and a lack of expertise in SMEs that slows penetration of certain technological groups. The exploitation of the different technologies is often approached individually and targeted, which leads us to conclude that the concept of industry 4.0 is not approached from the angle of industrial management strategy.

Keywords— SME, Industry 4.0, Production Planning and Control.

I. Introduction

As major industrial players, SMEs have to meet ever more complex customer needs. Therefore, they are working on improving operations management: production planning and execution, production control, operational performance measurement and evaluation [1].

Just-In-Time (JIT), which stems from the Toyota Production System [2], [3], is one of several existing industrial management methods. It consists in standardizing tasks and processes, creating continuous process flows, and adopting a market-driven pull production strategy. Though it has produced some excellent results [4], [5], it has also proved difficult to implement in SMEs, for lack of leadership and expertise [6], [7]. Likewise, the MRP and MRPII methods, which are supported by the deployment of computer software such as ERPs [8], are complex methods and costly to implement in SMEs. They quite often proved too rigid a system once installed [9].

The concept of industry 4.0 emerged more recently as a new approach for controlling production processes by providing real time synchronization of flows and by enabling a make-to-order, customized, fabrication of products [10].

This concept relies on emerging technology such as cloud computing, Internet of Things, cyber-physical systems, and Big Data. These technologies foster information sharing all along the managed system; this allow for streamlined processes as well as real time synchronization to meet fluctuation in demand. This concept appears more flexible and less expensive, but relies on many technological means that SMEs struggle to deploy.

This article aims to identify and analyze the different means of Industry 4.0 deployment for production planning and control in SMEs. We will first describe more formally what Industry 4.0 is as a concept in the next section, and proceed to describe the means of realization that it relies on. We will the present a literature review of production planning and control in the Industry 4.0 era. This article ends with a discussion of the results obtained.

II. Industry 4.0

Groups of technologies such as the Internet of Things [11], [12], Big Data [13], as well as Cloud Computing and networks improvement [14], [15], recently entered the industrial landscape fostering the emergence of new concepts for process management. Consequently, several strategic initiatives emerged across the world, including "Industry 4.0" in Germany, "Industry of the future" in France, "Smart Manufacturing" in the United States, or "Internet +" in China. The concept of Industry 4.0 appears to be dominant and we will use this term in the remainder of this paper [16]. The concept of Industry 4.0 arose in the German Hanover Fair from a discussion between representatives of industry, academic institutions, trade unions, and state authorities. The term of Industry 4.0 refers to the fourth industrial revolution, fallowing mechanization (1st revolution), the creation of production lines taking advantage of new energy sources such as electricity (2nd revolution), and PLCs in manufacturing (3rd revolution). Industry 4.0 aims to all objects and stakeholders within tomorrow's factory. Communication between objects, machines and people promotes the integration and synchronization of all resources along the value chain. Industry 4.0 allows 3 types of integration:

- Horizontal integration along the value chain, fostering collaboration among enterprises and connecting remote units.
- Vertical and hierarchical integration helps in the management of subsystems through the flexibility and reconfiguration of manufacturing networks.
- Temporal integration throughout the product life-cycle allow for customization and make-to-order manufacturing while capitalizing data from design, manufacturing and product use.

To achieve the aforementioned forms on integration, Industry 4.0 relies on several characteristics [17]:

- the interoperability of objects, machines, people and computers systems which communicate with one another;
- the virtualization of the physical world through a copy in the virtual world based on data collected by sensors;
- the decentralization of decision-making which could happen directly on in-the-field cyber-physical systems;
- the real-time management of load and capacity through data collection and analysis;
- a service-focused approach to enhance customer offers; and
- the modularity of production systems allowing for increased.

Underlying this concept and its characteristics are the emergence of new technologies that we group together under the term of means of Industry 4.0 realization [18]. These technologies require significant investments and strong expertise for their deployment, but seem to respond to the need for flexibility in SMEs. We present the technologies in the next section and proceed in section 5 to analyze real deployment cases o these means of realization within SMEs.

III. Groups of technologies in Industry 4.0

The concept of Industry 4.0 is based on the emergence of a set of several groups of technologies. We adapted the list of the Boston Consulting Group© which includes 9 means of realization [18]:

- Big Data: a large amount of data becomes available for use. Real-time or delayed analysis of large data sets is a source of added value both commercially and industrially. Processing this massive amount of data is one of the major issues of industry 4.0 [13], [19]–[21];
- Simulation: modeling products, production lines and multi-site networks allows to evaluate different scenarios using computer software [22]–[25]. Thus, designers and decision-makers can stimulate decision outcome and evaluate its performance [26]–[28];
- Internet of Things: new technologies allow for physical objects to have communication capabilities [29], [30]. This approach makes it possible to capture the state of the system in real time while decentralizing decision-making. These data can be used in massive data analysis to predict system behaviors;
- Cyber Physical Systems: these are in-system mechanisms to algorithmically control and monitor systems and their surrounding users. It allows objects to communicate with their environment and reconfigure in real time according to needs [15];
- Cloud computing: network connectivity, reactions times of a few milliseconds, and increased bandwidth allow for real-time information sharing. Thus, data and software are available anywhere, at any time and from any terminal [31], [32];
- Virtual reality: using systems such as Google's smart glasses, data become available directly into the field of view of employees [33];
- Cyber security: an increase in data transfer between businesses and services must not be to the detriment of industrial information safety. Cyber security aims to protect physical and intangible assets of companies as well as networks [34];
- Collaborative Robots: robots are becoming more flexible, better able to communicate and cooperate [35]. Product connectivity reduces batch sizes, and embedded sensor technologies allow for better collaboration with humans on the most meticulous tasks [36]; and

- Machine-to-Machine communication: with the increase in number of autonomous machines, communication machines are developing. This communication based on standard protocols and technologies allows the autonomous management of machines fleets.

Before assessing each group of technologies, we briefly describe in the following section the review strategy we used to select empirical Industry 4.0 cases in SMEs.

IV. Literature review strategy

The objective of our research is to identify the articles describing the use of the means of Industry 4.0 realization for production planning and control in SMEs. To do this, we applied Tranfield's systematic review method [37]. The selection of this method is justified on the grounds of reproducibility and formality, and this method has already shown its relevance in other literature reviews regarding SMEs [38].

The fallowing databases have been used:

- Elsevier (http://www.sciencedirect.com/)
- Emerald (http://www.emeraldinsight.com/)
- Springer (http://link.springer.com/)
- Taylor and Francis (http://tandfonline.com/)

We performed the requests on April 4, 2016 using the fallowing queries:

- "Industry 4.0" AND "small and medium";
- "digital production" AND "small and medium";
- "digital manufacturing" AND "small and medium";
- "internet of things" AND "small and medium";
- "cyber physical systems" AND "small and medium";
- "cyber factory" AND "small and medium"; and
- "production planning and control" AND "small and medium".

We obtained 2081 articles from the different queries. We excluded non-English language literature and the duplicate paper. We selected articles describing at least one empirical case of Industry 4.0 or one of its means of realization, for production planning and control in SMEs.

Among the 80 articles describing empirical cases, 54% do not specify the size of the company and 17% concern large companies. We found 23 items that match our criteria, which we analyze in the next section.

V. Literature analysis

The reviewed publications show that not all the groups of technology related to Industry 4.0 are present in the SMEs setting. As shown in Table 1, Cloud Computing and the Internet of Things are the most used technologies to implement Industry 4.0 initiatives. A more detailed analysis is provided in this section.

A. Big Data & Analytics

The use of data in the sense of Big Data is discussed in the paper of Ren et al. [39] without giving case of application. Through the design of a cloud computing platform for SMEs, they propose to exploit data coming from the Internet of Things via MapReduce algorithms and the Hadoop platform [40], [41].

The lack of research in this area confirms the observation of Bi and Cochran [42] that showed the weakness of SMEs regarding research and development activities and their difficulties in managing complex computer solutions. They emphasized that Cloud Computing is a viable solution for SMEs for providing analytical services and the data structuring means for using Big Data.

B. Simulation

Barenji et al. [43] observed that the planning of SMEs had several limits: these systems are centralized and not distributed, nor do they allow parallel queries or respond to the need to reconfigure production lines. Masood et al. [44] observed that the planning of SMEs is complex due to the many processes and competences involved in response to a great variety of products. Finally, several Cloud Computing simulation platforms have emerged to support distributed production on several sites and the optimization of multi-site resources [45].

According to Chalal et al. [46], some SMEs want their commercial offering by adding services in addition to their current products. They proposed two connected simulation models to replicate each offering. The first subsystem modeled demand and the second control the production plan in response to the demand of the first generated by the first subsystem.

Barenji et al. [43] proposed using the PROMETHEUS method to develop a planning simulation software application. The author observed that usual methods consisted in taking into account only the dynamic demand of customers or production variations. The author presented a method that considered both at the same time to better suit the needs of SMEs.

Peng et al. [47] reported the deployment of a Cloud Manufacturing platform. They presented an algorithm designed to optimize distributed resource management in a collaborative situation. They use a hybrid algorithm that combines VNS (Variable Neighborhood search) and PSO (Particle Swarm optimization). This combination permits optimizing multi-objective problems for flexible planning in a job shop working in a collaborative network of SMEs.

Givehchi et al. [48] presented a simulation to optimize the machining of a part. In their approach, the authors enhanced the data from numerical design of the part by defining new data recorded by CPS.

Table 1 : Articles dealing with different means of realizations

	Big Data & Analytics	Simulation		IoT		CPS	Cloud Computing					Virtual reality	Cyber security	Machine to machine	Collaborative Robot
		Scheduling optimization	Scenario-based simulation	RFID	Smart glasses		Sharing documents	Servitization	Collaboration	Distributed production	Optimization resources				
[49]				X			X	X	X						
[39]	X			X			X	X	X						
[50]								X							
[51]				X			X	X	X						
[52]															
[53]				X											
[43]		X													
[47]		X					X	X	X	X	X				
[44]			X												
[54]				X			X		X						
[55]								X							
[46]		X													
[56]							X	X	X	X					
[57]				X									X		
[58]							X	X	X	X					
[59]				X						X					
[48]		X				X					X				
[60]		X									X				
[61]			X												
[12]					X						X	X			
[62]				X											
[63]		X									X				
[64]							X	X	X				X		

Masood et al. [44] presented the concept of DPU (Dynamic Producer Unit) intended to model the "role" of a resource in order to carry out a coherent breakdown between employees, machines and information systems. The DPU concept can be used to facilitate the modification of system models by changing only the "roles" involved in the simulation.

Lastly, Dombrowski and Ernst [61] presented an approach to perform a simulation of potential production scenarios in 6 steps: the development of different scenarios, the design of possible changes of production lines, modeling alternatives, experimentation and optimization, the evaluation of alternatives, the implementation of the alternative chosen.

Among all papers identified, only Dombrowski and Ernst [61] presented a practical guide for SMEs for implementing their approach. In other cases, research teams used action research as their research methodology.

C. Internet of Things

Several research use the IoT coupled with RFID technology to obtain production feedback in real time [39], [51], [57], [54], [49], [52].

In this regards, Sena Ferreira et al. [54] proposed several indicators to measure and validate the performance of the collaborative system. They suggested indicators focusing on individual operational performance criteria as well as global performance indicators for measuring the success of the partner's network.

Denkena et al. [19] observed that most SMEs do not have reliable data, the data not being considered as a source of added value by the manager of this kind of company. They proposed using the IoT associated with RFID technology to manage flows and to facilitate implementation of Lean Manufacturing. This system made the data flow reliable and made possible to target improvement initiatives more quickly than with the use of classical Value Stream Mapping.

Xia et al. [59] used the IoT to recover data form a production machine and to analyze its performance and variance. The approach was also associated with a continuous improvement program. Segura Velandia et al. [62] use a similar approach to gather data from the produced parts. In both cases, the aim was to use IoT to acquire data and evaluate the performance of the production system.

Constantinescu et al. [60], [63] observed that the IoT gives too much data for humans to process. They developed the concept of JITIR (Just-In-Time Information Retrieval) consisting of three steps: the analysis of need through interviews with the employees, the recovery of information, and the periodic review of the employee's environment to track any need for change to improve the quality of decision making.

Hao and Helo [12] observed that most research focused on using the IoT to improve automation and flexibility in organizations. Their work uses a different approach by focusing on the man –machine link through connected objects. Used in parallel with Cloud Computing and virtual reality, their approach connects employees with other to optimize access to expert functions.

The main research approaches in these papers followed an action research method [65].

D. Cyber physical systems

The works of Givehchi et al. [48] are the only found case reporting the use of Cyber Physical System applied in SMEs for production planning and control. They showed that adding connectivity to a production machine and defining a new data format for parts produced makes it possible for the machine to control and optimize its operations. As CPS are complex systems that incorporate processing algorithms, it is not surprising to note the lack of in house competences in SMEs is a major barrier for implementing CPS [66].

E. Cloud Computing

Cloud Computing is the most used means of realization of Industry 4.0 practices in SMEs as we found that 65% or our selected publications reported its use.

Several works used Cloud Computing with the goal of building *Virtual Enterprises* between SMEs [39], [47], [51], [49], [52], [56], [64], [58]. Based on the observation that SMEs do not possess all the knowledge and capacities to satisfy complex clients' needs, the proposed models favor the development of industrial collaboration between several partners. Cloud Computing platforms allow the *servitization* of the products, i.e. a change the vision of "What I have" to "What I can achieved" and how it can be shared in the network of partners [39].

The creation of such a network does not only depend on the availability of a Cloud Computing platform. Hao et al. [67] outlined the first steps of building a Virtual Enterprise: find partners and then contractualize the commitments and risks. Once these steps have been taken, it is possible to progress to collaboration and operational optimization.

Shamsuzzoha et al. [49] presented a concept of a Cloud Computing platform based on the Net-Challenge Framework responding to make-to-order and engineering-to-order strategies. Collaboration between partners is achieved for the specific needs of each customer. Once the need has been satisfied, the virtual organization is dismantled.

However, Herdon et al. [55] observed that SMEs have internal information systems that do not permit direct connection with Cloud Computing systems. They proposed transferring ERP data to the Cloud Computing platform free of charge to promote the appropriation of their solution by enterprises.

Bonfanti et al. [50] showed that Cloud Computing allows Italian artisanal enterprises to offer products and services online. Creating a new product or service via Web interfaces and Cloud Computing platforms strengthens client loyalty, while providing access to new markets.

In conclusion, we note that Cloud Computing platforms favor the planning and utilization of shared resources, control over processes and evaluation of performance.

F. Virtual reality

Hao and Helo [12] showed the advantage of using IoT, Cloud Computing and Virtual Reality simultaneously. The use of smart glasses allows for information to be displayed directly in the users' field of vision in real time. Disturbing events appear more visible, which causes employees to be more reactive. Likewise, maintenance of production resources is facilitated by the availability of the data required to restore faulty equipment to operational condition.

G. Cyber Security

MacKerron et al. [57] and Holtewert et al. [64] studied cyber security in the production environment where the Internet of Things and Cloud Computing are both used. However, readers should note that cyber security was not central to their works.

H. Collaborative robots and Machine-to-Machine communication

Collaborative robots and communication between machines permit decentralized-decision making at the heart of the production processes. Unfortunately, we could not report any cases on the implementation of these technologies in SMEs.

VI. Discussion

SMEs are companies recognized for their flexibility and proximity to customers [68]. Their size ease communication and adjustments between employees, thereby enabling quicker reconfiguration in the event of change in demand [69], [70]. However, SMEs are weak on investment capacity and operational performance, with high costs and subpar on-time delivery compared to large companies. Finally, SMEs manly have short-term strategies, which do not favor long-term investments.

From our analysis, three groups of means of realization are emerging: collaborative tools, production tools, and optimization tools.

The first group, the collaborative tools, is strongly present in the identified articles; it is cyber security and Cloud Computing that promote exchanges and their security between companies through the creation of *virtual enterprise*. Cloud Computing is in the form of PaaS (platform as a service), which does not require high skill and investment. However, this mean of realization push SMEs towards concentrating on their core business activities, at the risk of being increasingly dependent on their external value chain. A more detailed analysis is needed to estimate the optimal exploitation strategy of Cloud Computing.

The second group, the production tools, concerns the cyber physical systems, the virtual reality, collaborative robot and the communication between machines. These means of implementation bring a capacity to the physical systems to become more flexible. SMEs do not need to improve their flexibility for they are already recognized for this. Moreover, these means require heavy investment, which may hamper the reconfiguration capacity of SMEs and does not favor the deployment of these means of realization. This group of technologies thus appears ill suited to SMEs and unable to meet their specific needs.

The third group, the optimization tools, combines the Internet of Things, Big Data analysis and simulation. These tools are inexpensive compared with collaborative robot for example. The Internet of Things ease the acquisition of data from the whole production system; some authors relied on this mean to realize a VSM in a Lean Manufacturing implementation project. The Big Data analysis can be exploited to consider predictive maintenance or predictive planning with ERP [71]. Likewise, all of these data can be used to design a simulation model and thus make a decision between scenarios. However, these means of realization require a strong expertise, which is usually not present in SMEs. While the cost of these solutions seems to be compatible with SMEs investment capacity, the required expertise is not. Research should focus on the definition of methods / tools available to SMEs, taking into account the specific characteristics of these enterprises.

In our analyses, the cases of application are described in terms of the use of one or more means of Industry 4.0 realization. However, we have not identified an article discussing industry 4.0 as a concept supporting the company's strategy. Geographic growth, growth throughout the value chain or growth through diversification of activity, Industry 4.0 provide help according to the strategy chosen, the research must extend the study of industry 4.0 in this sense.

Our literature review has some limitations. First, we found only 23 applied cases to analyze, which involve eight different means of realization. In most cases, the implementation processes of Industry 4.0 technologies within SMEs are not clearly described. Indeed, in order to take an exhaustive look at the real cases and their success, it would be pertinent to supplement this type of approach with field surveys. Second, there is a subjective bias in the reading and selection of papers in a review. In our case, we have ignored commercial types of journals and focused on scientific journals only. Delays between the acceptance and publication of such papers may certainly impact the number of selected papers and underestimate the real level of Industry 4.0 adoptions within SMEs. Third, case studies sometimes describe qualitative results and not quantitative results. It is therefore difficult to judge the real benefits achieved by SMEs in exploiting new technologies and practices.

VII. Conclusion

The aim of this paper was to analyze the exploitation of the means of realization of Industry 4.0 within SMEs for production planning and control.

Our study has been limited to the analysis of empirical cases concerning SMEs since 2011. The concept of Industry 4.0 is recent, but the penetration of new technologies within companies has existed for many years. In order to

complement our research, we should analyze the articles with the same objectives using ICT (Information and Communication Technology) or AMS (Advanced Manufacturing System) [72], [73]. Moreover, an extension to industrial documents (reports, white papers) describing concrete cases could enrich the analysis.

SMEs are recognized for their flexibility and proximity to customers, and are also recognized for their low investment capacity and short-term strategy. In this sense, some means of realization cannot support the strategy of SMEs. Production tools require too much investment, while collaboration tools and optimization tools seem less expensive. Research should focus on adapting optimization tools to the characteristics of SMEs, mainly exploiting massive data to improve operational performance.

VIII. References

[1] Apics, *Apics Dictionary*. The American Production and Inventory Control Society, 2005.

[2] T. Ohno, *The Toyota Production System — Beyond Large-Scale Production*. 1988.

[3] J. K. Liker, *The Toyota way: 14 management principles from the world's greatest manufacturer*, vol. 4, no. 1. 2007.

[4] R. Shah and P. T. Ward, 'Lean manufacturing: Context, practice bundles, and performance', *J. Oper. Manag.*, vol. 21, no. 2, pp. 129–149, 2003.

[5] J. P. Womack and T. J. Daniel, *Lean thinking*, First Free. Simon & Schuster, Inc., 2003.

[6] P. Achanga, E. Shehab, R. Roy, and G. Nelder, 'Critical success factors for lean implementation within SMEs', *J. Manuf. Technol. Manag.*, vol. 17, no. 4, pp. 460–471, 2006.

[7] A. Moeuf, S. Tamayo, S. Lamouri, R. Pellerin, and A. Lelievre, 'Strengths and weaknesses of small and medium sized enterprises regarding the implementation of lean manufacturing', *IFAC-PapersOnLine*, vol. 49, no. 12, pp. 71–76, 2016.

[8] A. Deep, P. Guttridge, S. Dani, and N. Burns, 'Investigating factors affecting ERP selection in made-to-order SME sector', *J. Manuf. Technol. Manag.*, vol. 19, no. 4, pp. 430–446, 2008.

[9] C. Cullinan, S. G. Sutton, and V. Arnold, 'Technology monoculture: ERP systems, "techno-process diversity" and the threat to the information technology ecosystem', *Adv. Account. Behav. Res.*, vol. 13, pp. 13–30, 2010.

[10] D. Kohler and J.-D. Weisz, *Industrie : Les défis de la transformation numérique du modèle industriel allemand*. Paris, 2016.

[11] R. M. Dijkman, B. Sprenkels, T. Peeters, and A. Janssen, 'Business models for the Internet of Things', *Int. J. Inf. Manage.*, vol. 35, no. 6, pp. 672–678, Dec. 2015.

[12] Y. Hao and P. Helo, 'The role of wearable devices in meeting the needs of cloud manufacturing: A case study', *Robot. Comput. Integr. Manuf.*, p. , Oct. 2015.

[13] R. F. Babiceanu and R. Seker, 'Big Data and virtualization for manufacturing cyber-physical systems: A survey of the current status and future outlook', *Comput. Ind.*, vol. 81, no. 2015, pp. 128–137, Sep. 2016.

[14] X. Xu, 'From cloud computing to cloud manufacturing', *Robot. Comput. Integr. Manuf.*, vol. 28, no. 1, pp. 75–86, Feb. 2012.

[15] X. Yue, H. Cai, H. Yan, C. Zou, and K. Zhou, 'Cloud-assisted industrial cyber-physical systems: An insight', *Microprocess. Microsyst.*, vol. 39, no. 8, pp. 1262–1270, Nov. 2015.

[16] S. Wang, J. Wan, D. Li, and C. Zhang, 'Implementing Smart Factory of Industrie 4.0: An Outlook', *Int. J. Distrib. Sens. Networks*, vol. 2016, pp. 1–10, 2016.

[17] J. Smit, S. Kreutzer, C. Moeller, and M. Carlberg, 'Industry 4.0 - Study for the ITRE Committee', 2016.

[18] M. Ruessmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel, and M. Harnisch, 'Industry 4.0: the future of productivity and growth in manufacturing industries', 2015.

[19] B. Denkena, J. Schmidt, and M. Krüger, 'Data Mining Approach for Knowledge-based Process Planning', *Procedia Technol.*, vol. 15, pp. 406–415, 2014.

[20] N. Kushiro, S. Matsuda, and K. Takahara, 'Model Oriented System Design on Big-data', *Procedia Comput. Sci.*, vol. 35, no. C, pp. 961–968, 2014.

[21] F. Rago, 'A Smart Adaptable Architecture Based on Contexts for Cyber Physical Systems', *Procedia Comput. Sci.*, vol. 61, pp. 301–306, 2015.

[22] A. Azevedo and A. Almeida, 'Factory Templates for Digital Factories Framework', *Robot. Comput. Integr. Manuf.*, vol. 27, no. 4, pp. 755–771, Aug. 2011.

[23] M. Matsuda, Y. Sudo, and F. Kimura, 'A Multi-agent Based Construction of the Digital Eco-factory for a Printed-circuit Assembly Line', *Procedia CIRP*, vol. 41, pp. 218–223, 2016.

[24] A. Mehrsai, B. Henriksen, C. C. Røstad, K. A. Hribernik, and K.-D. Thoben, 'Make-to-XGrade for the Design and Manufacturing of Flexible, Adaptive, and Reactive Products', *Procedia CIRP*, vol. 21, pp. 199–205, 2014.

[25] I. Veza, M. Mladineo, and N. Gjeldum, 'Managing Innovative Production Network of Smart Factories', *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 555–560, 2015.

[26] A. Caggiano, F. Caiazzo, and R. Teti, 'Digital Factory Approach for Flexible and Efficient Manufacturing Systems in the Aerospace Industry', *Procedia CIRP*, vol. 37, pp. 122–127, 2015.

[27] A. Caggiano and R. Teti, 'Modelling, Analysis and Improvement of Mass and Small Batch Production through Advanced Simulation Tools', *Procedia*

CIRP, vol. 12, pp. 426–431, 2013.

[28] X. Luo, W. Li, Y. Tu, D. Xue, and J. Tang, 'Operator allocation planning for reconfigurable production line in one-of-a-kind production', *Int. J. Prod. Res.*, vol. 49, no. 3, pp. 689–705, Feb. 2011.

[29] P. Kopacek, 'Automation and TECIS', *IFAC-PapersOnLine*, vol. 48, no. 24, pp. 21–27, 2015.

[30] L. Wang, M. Törngren, and M. Onori, 'Current status and advancement of cyber-physical systems in manufacturing', *J. Manuf. Syst.*, vol. 37, pp. 517–527, Oct. 2015.

[31] P. Gupta, A. Seetharaman, and J. R. Raj, 'The usage and adoption of cloud computing by small and medium businesses', *Int. J. Inf. Manage.*, vol. 33, no. 5, pp. 861–874, Oct. 2013.

[32] Z. Wu, Z. Gao, Y. Cao, X. Ye, and J. Yang, 'Tolerance Design and Adjustment of Complex Customized Product Based on Cloud Manufacturing', *Procedia CIRP*, vol. 27, pp. 169–175, 2015.

[33] J. Lee, S. Han, and J. Yang, 'Construction of a computer-simulated mixed reality environment for virtual factory layout planning', *Comput. Ind.*, vol. 62, no. 1, pp. 86–98, Jan. 2011.

[34] D. Airehrour, J. Gutierrez, and S. K. Ray, 'Secure routing for internet of things: A survey', *J. Netw. Comput. Appl.*, vol. 66, pp. 198–213, May 2016.

[35] J. Michniewicz and G. Reinhart, 'Cyber-physical Robotics – Automated Analysis, Programming and Configuration of Robot Cells based on Cyber-physical-systems', *Procedia Technol.*, vol. 15, pp. 566–575, 2014.

[36] C. Constantinescu, D. Popescu, P.-C. Muresan, and S.-I. Stana, 'Exoskeleton-centered Process Optimization in Advanced Factory Environments', *Procedia CIRP*, vol. 41, pp. 740–745, 2016.

[37] D. Tranfield, D. Denyer, and P. Smart, 'Towards a methodology for developing evidence-informed management knowledge by means of systematic review *', *Br. J. Manag.*, vol. 14, pp. 207–222, 2003.

[38] P. Garengo, S. Biazzo, and U. S. Bititci, 'Performance measurement systems in SMEs: A review for a research agenda', *Int. J. Manag. Rev.*, vol. 7, no. 1, pp. 25–47, Mar. 2005.

[39] L. Ren, L. Zhang, F. Tao, C. Zhao, X. Chai, and X. Zhao, 'Cloud manufacturing: from concept to practice', *Enterp. Inf. Syst.*, vol. 9, no. 2, pp. 186–209, Feb. 2015.

[40] J. Dean and S. Ghemawat, 'MapReduce', *Commun. ACM*, vol. 51, no. 1, p. 107, Jan. 2008.

[41] Y. Bao, L. Ren, L. Zhang, X. Zhang, and Y. Luo, 'Massive sensor data management framework in Cloud manufacturing based on Hadoop', in *IEEE 10th International Conference on Industrial Informatics*, 2012, pp. 397–401.

[42] Z. Bi and D. Cochran, 'Big data analytics with applications', *J. Manag. Anal.*, vol. 1, no. 4, pp. 249–265, Oct. 2014.

[43] A. V. Barenji, R. V. Barenji, D. Roudi, and M. Hashemipour, 'A dynamic multi-agent-based scheduling approach for SMEs', *Int. J. Adv. Manuf. Technol.*, Aug. 2016.

[44] T. Masood, R. Weston, and A. Rahimifard, 'A model-driven approach to enabling change capability in SMEs', *Int. J. Adv. Manuf. Technol.*, vol. 69, no. 1–4, pp. 805–821, Oct. 2013.

[45] J. S. Srai, M. Kumar, G. Graham, W. Phillips, J. Tooze, S. Ford, P. Beecher, B. Raj, M. Gregory, M. K. Tiwari, B. Ravi, A. Neely, R. Shankar, F. Charnley, and A. Tiwari, 'Distributed manufacturing: scope, challenges and opportunities', *Int. J. Prod. Res.*, vol. 54, no. 23, pp. 6917–6935, Dec. 2016.

[46] M. Chalal, X. Boucher, and G. Marques, 'Decision support system for servitization of industrial SMEs: a modelling and simulation approach', *J. Decis. Syst.*, vol. 24, no. 4, pp. 355–382, Oct. 2015.

[47] G. Peng, Y. Jiang, J. Xu, and X. Li, 'A collaborative manufacturing execution platform for space product development', *Int. J. Adv. Manuf. Technol.*, vol. 62, no. 5–8, pp. 443–455, Sep. 2012.

[48] M. Givehchi, A. Haghighi, and L. Wang, 'Generic machining process sequencing through a revised enriched machining feature concept', *J. Manuf. Syst.*, vol. 37, pp. 564–575, Oct. 2015.

[49] A. Shamsuzzoha, C. Toscano, L. M. Carneiro, V. Kumar, and P. Helo, 'ICT-based solution approach for collaborative delivery of customised products', *Prod. Plan. Control*, vol. 27, no. 4, pp. 280–298, Mar. 2016.

[50] A. Bonfanti, M. Del Giudice, and A. Papa, 'Italian Craft Firms Between Digital Manufacturing, Open Innovation, and Servitization', *J. Knowl. Econ.*, Nov. 2015.

[51] B. Huang, C. Li, C. Yin, and X. Zhao, 'Cloud manufacturing service platform for small- and medium-sized enterprises', *Int. J. Adv. Manuf. Technol.*, vol. 65, no. 9–12, pp. 1261–1272, Apr. 2013.

[52] T. Song, H. Liu, C. Wei, and C. Zhang, 'Common engines of cloud manufacturing service platform for SMEs', *Int. J. Adv. Manuf. Technol.*, vol. 73, no. 1–4, pp. 557–569, Jul. 2014.

[53] B. Denkena, B. Dengler, K. Doreth, C. Krull, and G. Horton, 'Interpretation and optimization of material flow via system behavior reconstruction', *Prod. Eng.*, vol. 8, no. 5, pp. 659–668, Oct. 2014.

[54] P. Sena Ferreira, A. H. M. Shamsuzzoha, C. Toscano, and P. Cunha, 'Framework for performance measurement and management in a collaborative business environment', *Int. J. Product. Perform. Manag.*, vol. 61, no. 6, pp. 672–690, Jul. 2012.

[55] M. Herdon, L. Várallyai, and Á. Péntek, 'Digital business ecosystem prototyping for SMEs', *J. Syst. Inf. Technol.*, vol. 14, no. 4, pp. 286–301, Nov. 2012.

[56] R. Fornasiero and A. Zangiacomi, 'A structured approach for customised production in SME collaborative networks', *Int. J. Prod. Res.*, vol. 51, no. 7, pp. 2110–2122, Apr. 2013.

[57] G. MacKerron, M. Kumar, V. Kumar, and A. Esain, 'Supplier replenishment policy using e-Kanban: A framework for successful implementation', *Prod. Plan. Control*, vol. 25, no. 2, pp. 161–175, Jan. 2014.

[58] X. V. Wang and X. W. Xu, 'An interoperable solution for Cloud manufacturing', *Robot. Comput. Integr. Manuf.*, vol. 29, no. 4, pp. 232–247, 2013.

[59] M. Xia, T. Li, Y. Zhang, and C. W. de Silva, 'Closed-loop design evolution of engineering system using condition monitoring through internet of things and cloud computing', *Comput. Networks*, vol. 101, pp. 5–18, Jun. 2016.

[60] C. L. Constantinescu, E. Francalanza, D. Matarazzo, and O. Balkan, 'Information Support and Interactive Planning in the Digital Factory: Approach and Industry-driven Evaluation', *Procedia CIRP*, vol. 25, no. C, pp. 269–275, 2014.

[61] U. Dombrowski and S. Ernst, 'Scenario-based Simulation Approach for Layout Planning', *Procedia CIRP*, vol. 12, pp. 354–359, 2013.

[62] D. M. Segura Velandia, N. Kaur, W. G. Whittow, P. P. Conway, and A. A. West, 'Towards industrial internet of things: Crankshaft monitoring, traceability and tracking using RFID', *Robot. Comput. Integr. Manuf.*, vol. 41, pp. 66–77, Oct. 2016.

[63] C. L. Constantinescu, E. Francalanza, and D. Matarazzo, 'Towards Knowledge Capturing and Innovative Human-system Interface in an Open-source Factory Modelling and Simulation Environment', *Procedia CIRP*, vol. 33, pp. 23–28, 2015.

[64] P. Holtewert, R. Wutzke, J. Seidelmann, and T. Bauernhansl, 'Virtual Fort Knox Federative, Secure and Cloud-based Platform for Manufacturing', *Procedia CIRP*, vol. 7, pp. 527–532, 2013.

[65] L. Cappelletti, 'La Recherche-Intervention : Quels Usages En Controle De Gestion ?', *Congrès l'Association Francoph. Comptab.*, pp. 1–25, 2010.

[66] P. Evangelista, A. Mckinnon, and E. Sweeney, 'Technology adoption in small and medium-sized logistics providers', vol. 113, no. 7, pp. 967–989, 2013.

[67] Y. Hao, P. Helo, and A. Shamsuzzoha, 'Virtual factory system design and implementation: integrated sustainable manufacturing', *Int. J. Syst. Sci. Oper. Logist.*, vol. 2674, no. November, pp. 1–17, Oct. 2016.

[68] O. Torres, *Les PME*. 1999.

[69] H. Mintzberg, *Structure et Dynamique des organisations*, Vingt-Troi. Eyrolles, 1982.

[70] A. Ates, P. Garengo, P. Cocca, and U. Bititci, 'The development of SME managerial practice for effective performance management', *J. Small Bus. Enterp. Dev.*, vol. 20, no. 1, pp. 28–54, Feb. 2013.

[71] H. Mousannif, H. Sabah, Y. Douiji, and Y. Oulad Sayad, 'Big data projects: just jump right in!', *Int. J. Pervasive Comput. Commun.*, vol. 12, no. 2, pp. 260–288, Jun. 2016.

[72] J. P. T. Mo, 'The role of lean in the application of information technology to manufacturing', *Comput. Ind.*, vol. 60, no. 4, pp. 266–276, 2009.

[73] A. Tarutė and R. Gatautis, 'ICT Impact on SMEs Performance', *Procedia - Soc. Behav. Sci.*, vol. 110, pp. 1218–1225, Jan. 2014.

Quantification of Industry 4.0 Manufacturing Complexity based on Information Theory

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mourtzis Dimitris, Vlachou Ekaterini, Fotia Sophia
Department of Mechanical Engineering and Aeronautics, University of Patras
Laboratory for Manufacturing Systems and Automation (LMS)
Patras, Greece
mourtzis@lms.mech.upatras.gr

Abstract — The new industrial revolution, called Industry 4.0, imposes new challenges to the research community. One of the main issues in Industry 4.0, is the management of the huge amount of information exchanged among its different integrated systems. The present work demonstrates how the traditional manufacturing system is transformed to the Industry 4.0 manufacturing system, and proposes a quantification approach which includes metrics from the Information Theory, estimating the complexity and the capacity of the Industry 4.0, from the perspective of the communication among the systems. The application impact of the proposed Industry 4.0 system and the comparison of these metrics before and after shifting to Industry 4.0 is also analyzed and demonstrated in a bicycle manufacturing company.

Keywords — Industry 4.0, Complexity, Information Theory, Entropy

I. Introduction

The transition towards the fourth Industrial revolution will transform the design, manufacture, operation, as well as the service of products and systems. The advent of modern technologies such as Cyber-physical systems (CPS) and Internet of Things (IoT), opens new horizons towards the industrial digitalization by enabling automated procedures and communication by means that were not attainable in the past. The increased connectivity and interaction among systems, humans, and machines support the integration of various automated or semi-automated systems, increasing flexibility and productivity [1]**.** This will also lead to interconnected manufacturing systems and supply chains, that constitute an integrated whole which provides the needed products at the right place, at the right time.

More specifically, the integration of smart sensor systems and mobile devices, together with the use of industrial communication protocols and standards, will support industries to achieve the desired awareness on the actual condition and plan efficiently and accurately. The adoption of CPS by applying IoT, results in the creation of vast amounts of data [2]. Once this generated data is manipulated and analyzed in order to perform meaningful reasoning, they provide added value information to the systems, transforming them into adaptive and flexible systems.

Although the adoption of the Industry 4.0 paradigm by the manufacturing sector is continuously increasing, a main question/concern that is raised by the manufacturing companies is how and in what level digitalization and the adoption of enabling technologies will change their companies and the way they are working.

The present work introduces a modeling of Industry 4.0, including all the core aspects that are incorporated in this new manufacturing paradigm. After that, the structural complexity and the capacity of a manufacturing system shifted to Industry 4.0 compared to a traditional one, are quantified.

II. State of the Art

The transformation of today's manufacturing systems into Industry 4.0 manufacturing systems is mainly based on the integration of CPS with production, logistics, as well as services [3]. This transformation is enabled by the higher availability and affordability of sensors, data acquisition systems and computer networks [4]. The adoption of smart sensors and the integration with industrial communication protocols and standards will generate a great amount of data. The smart objects, enabled by IoT, can be dynamically reconfigured to achieve high flexibility, whereas big data analytics can provide global feedback and coordination to achieve high efficiency [5].

Several approaches have been developed in the field of CPS from the design phase to manufacturing, assembly and final products provision. An approach for product configuration using augmented reality technology, integrating the customer easily and retrieving all the necessary data for designing or re-designing products is presented by [6]. In the field of process planning, an approach for adaptive process planning considering the availability and capabilities of the machine tools is presented by [7]. Adaptive or integrating scheduling approaches and methods are also presented, taken into account the generated data from the different smart objects and systems. In [8], an integrated approach for production and maintenance planning is presented, considering the condition of the machines.

One of the most important outcomes of the emergence of IoT is the generation of data in increasing volume, variety and velocity, also called as Big Data. The analysis of this data constitutes the basis for the modern scope of mass customization, which implies the fulfilling of the needs of individualized customer markets [9]. The analysis of large sets of data can enhance the knowledge repositories and improve decision making in different manufacturing stages, such as assembly operations [10].

Although the adoption of Industry 4.0 and IoT paradigms in manufacturing reveal positive consequences, the increased complexity that may occur is a main challenge. Towards identifying how and in what level manufacturing systems are changed, several approaches have been reported in literature.

For the quantification of complexity in engineering design, several approaches have been developed, mainly based on statistics, probabilities, and heuristics [11]. One of the most important approaches, which will be used in the present study, is Shannon's Entropy of Information theory. These approaches have been widely used for measuring structural complexity of manufacturing systems [11]-[13].

The literature review makes apparent that several approaches have been developed in the field of Industry 4.0. Moreover, several approaches that utilize entropic measures to identify the transformation of manufacturing systems are also reported. However, quantification approaches that investigate how the adoption of the Industry 4.0 paradigm will transform manufacturing are limited. Towards that end, the present work provides how the traditional manufacturing system is transformed to the Industry 4.0 manufacturing system and proposes a quantification approach for the complexity of Industry 4.0 from the perspective of the communication among the systems. The application impact of the proposed Industry 4.0 system and the comparison of the entropy measurements before and after shifting to the Industry 4.0 system, are also analyzed and demonstrated.

III. Manufacturing Systems transformation through Industry 4.0 paradigm

Industry is a part of the economy where different paradigms have been introduced over the years, from mechanization (1st industrial revolution), to the intensive use of electrical energy (2nd industrial revolution), to the widespread digitalization (3rd industrial revolution), and finally to cyber-physical systems and "smart" objects (4th Industrial revolution).

The Industry 4.0 revolution is constituted by a number of disruptive technologies that enable the digitalization of the manufacturing sector. The internet of Things, together with Big Data analytics, virtual and augmented reality and cloud technologies, aim to integrate and analyze data across sources and companies, sharing outcomes across the value chain, ensuring the integration with physical production assets, and rethinking the design of classical production systems. The Industry 4.0 paradigm is identified in every manufacturing phase, from the design phase to final products.

The increased use of internet-connected mobile devices and social media supports manufacturing companies to gather the pulse of market and customers' needs easily and effectively, offering high added value products and solutions to their customers. In the design phase, the use of virtual reality technology as well as 3D printing have revealed great advantages for the designers, supporting them to identify possible issues in the early stages and resolve them. The use of smart sensing devices, mobile devices as well as the advanced Human-Machines Interfaces (HMIs), during the manufacturing phase, enables companies to sense the actual condition of their production and retrieve valuable feedback. This great amount of data that is produced by different sources, can be integrated and analyzed providing useful information and insights to the planning systems, transforming them to adaptive, autonomous, as well as self-learning systems. The integration of data from the different sources is mainly performed through the adoption of industrial communication protocols (e.g. OPC-UA, ROS, etc.). These protocols enable the integration of the data and the efficient transfer of data among the various systems. One of the main challenges occurring by the great amount of data is its analysis. For that reason, big data analysis techniques have been introduced, capable of analyzing data of high volume, velocity, variety, and veracity [14]. This analyzed data can be further used by advanced algorithms for prediction and integrated planning, increasing systems' efficiency and productivity.

Aiming to increase interoperability and communication among the various systems, store the generated data and support ubiquitous data access, cloud technology is also used in the context of the Industry 4.0 paradigm. Cloud manufacturing enables the development of different services and the use of these services following the needs and the business model of each industry.

Summarizing the above, industry 4.0 technologies aim to unlock new value potential through new types of business models. The low cost of the IoT devices and apps enables companies and especially SMEs to become digital and empower their position in a global value chain. In Figure 1 below, the transformation of a traditional system (with blue color) to the industry 4.0 system (with red color) is presented, showing how Industry 4.0 technologies and frameworks enhance the existing system in each of its phases.

IV. Entropy measurement for traditional and industry 4.0 manufacturing systems

As it has been analysed in the previous section, the model that represents a typical industry 4.0 manufacturing system is illustrated in Figure 1. Following this modelling, the quantification of the complexity will be performed by measuring the entropy of each system that comprises the Industry 4.0 paradigm.

The Entropy measurement of a system, is based on the analogous of the Shannon's communication system [14]. According to this, the basic components of a communication system are the information source, the message, the transmitter, the signal, the noise source, the receiver, and the destination. More specifically, the information source selects a desired *message* out of a set of possible messages (this is a particularly important remark, which requires considerable explanation later). The *transmitter* transforms this *message* into the *signal* which is actually sent over the *communication channel* from the transmitter to the *receiver*. The *receiver* changes the transmitted signal back into a message, and handing this message to the destination. During the transmission, distortions which were not intended by the information source, are added to the signal. These unwanted additions comprise the noise of the channel. Based on the Shannon's theory about an information source, if $x_1,\ldots,x_n$ are the messages sent by a selected source, and $P(x_i)$ is the probability of the transmission of the x_i message to the receiver without any distortions, where $i = 1,..,n$, the average amount of information per message of the information source is given by the equation below:

$$H(x) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \quad \textbf{(1)}$$

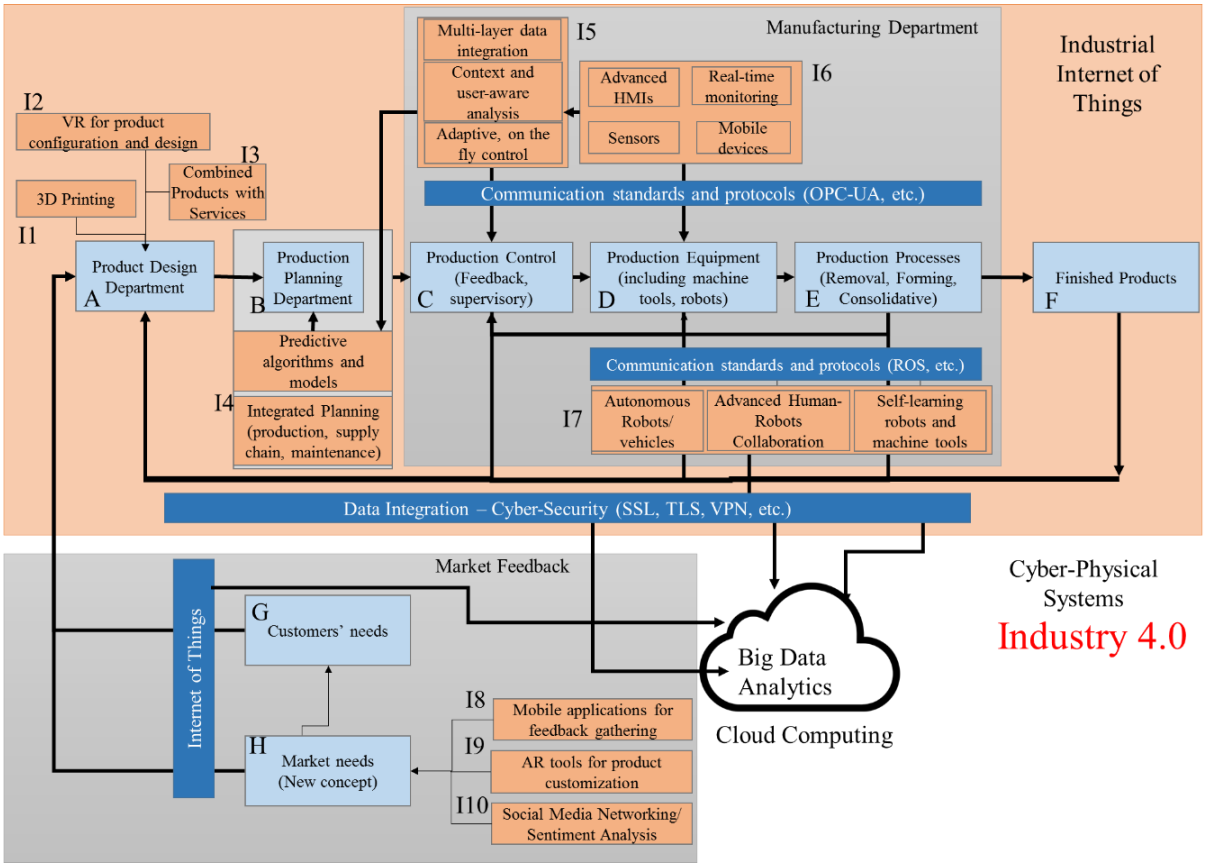


Figure 1 Industry 4.0 model

- If $P(x_1) \approx \cdots \approx P(x_n)$, then H is maximum
- If one of the $P(x_i)$ is much greater than the other ones, then $H<1$
- If one of the $P(x_i)$ is equal to one, and the others are zero, then $H=0$, meaning no uncertainty, no freedom of choice, no information

When a source generates (N) messages, these can be encoded and transferred using $N*H$ bits. This value represents the information context transmitted between the departments.

Entropy is a measure of unpredictability of the state, or, of its average information content. The meaning of the events observed (the meaning of messages) does not matter in the definition of entropy. Entropy only takes into account the probability of observing a specific event, so the information it encapsulates is the information about the underlying probability distribution, not the meaning of the events themselves [14].

In the present study, it is assumed that the Industry 4.0 manufacturing system is probabilistic, since the Shannon entropy is calculated only in the context of a probabilistic model. Moreover, it is assumed that the events are equiprobable since it is the worst-case scenario, ones the H takes its maximum value when the probabilities inserted in the logarithm are equal. The equiprobability of the worst case can become more understandable with the following example: if it is considered that in a fair coins toss the $H= - \log_2 (0.5^2) = 2$, while if the toss is not fair and one of the two states is more probable, $H= - \log_2(1) = 0$, meaning no uncertainty. The more choices there are, the larger the H will be.

As it has been mentioned previously, the Entropy defines the average uncertainty concerning the received signal when the sent message is known. The uncertainty is zero when there is no noise. If noise is introduced to the signal, then the received message contains certain distortions, certain errors, and then additional material that would certainly lead to a received signal of higher information content but of low-quality.

Another important value in the Information Theory is the channel capacity. The channel capacity is defined as the ability of the channel to transfer information, and is given by the following expression:

$$C = \frac{N \cdot H}{T} \; bits/\sec \; or \; bps \qquad (2)$$

where,
N: the number of messages that a source generates
H: Shannon entropy
T: the duration for transmission N messages

Following the presentation of the main points of the Entropy of the Information theory, Table IV-1 and Table IV-2 comprise the actual modelling of a manufacturing system. Table IV-1 includes the information transferred between the systems of the traditional manufacturing systems. These systems are also included in the new manufacturing paradigm of Industry 4.0, but

are enhanced with new systems that exhange information, which are presented in Table IV-2.

Table IV-1 The modelling parameters of the traditional system

Systems	Information exchanged
A-B	k_{A-B} = number of variables required for the coding of the designed product p_{A-B} = the number of parameters needed to specify the geometry variables
B-CDE	k_{B-CDE} = the number of digits that can describe each geometric feature of the component. p_{B-CDE} = the number of variables that the process plan considers for each part.
D-E	k_{DE} = The number of variables included in the source depending on the number of operations, which can be performed by the machine tools. p_{DE} = the number of parameters needed to specify each operation. The number of messages, generated by each process plan depends both on the numbers of variables and of the attached parameters.
E-F	k_{EF} = The number of codes of the assembly processes which are considered to be the variables in the source p_{EF} = The attached parameter provides information on the degree of difficulty of the assembly operations as well as on the use of auxiliary instruments.
GH-A	k_{GH-A} = The number messages relevant to the market feedback, associated with a specific product feature p_{GH-A} = The number of the product features for which feedback is received.

Table IV-2 The additional modelling parameters comprised in the Industry 4.0

Systems	Information exchanged
I1- A	k_{I1-A} = The number of variables required for the coding of 3D object p_{I1-A} = The minimum number of parameters needed to specify the 3D object's variables
I2- A	k_{I2-A} = The number of variables required for the coding for VR visualization p_{I2-A} = The minimum number of parameters needed to specify the variables
I3- A	k_{I3-A} = The number of product features connected with services p_{I3-A} = The average number of services that are associated with the above product features.
I4-I5I6	$k_{I4-I5I6}$= The number of measurements from the different integrated systems, that are associated with a specific variable on the production plan $p_{I4-I5I6}$= The number variables of the production plan algorithm that affected by external feedback
I5I6-CD	$k_{I5I6-CD}$ = Number of messages/measurements from ERP systems, MES, SCADA (high level systems), that are associated with a specific variable on the production plan $p_{I5I6-CD}$ = Number of variables that used in the control, which are affected by the external feedback
I7 - ED	k_{I7-ED}= Number of code lines that have been developed for human-robot collaboration, and self-learning and autonomous robots p_{I7-ED}=The number of tasks that request these codes
I8 – GH	kI8 – GH = The number of characters that are included in a set of messages associated with a specific feature of the product. pI8 – GH = The number of product-service features for which the customers provide their opinion.
I9 – GH	k_{I9-GH} = The number of characters that are included in a set of messages associated with a specific feature of the product. p_{I9-GH} = The number of product-service features which the customers can customize through the AR tool
I10 – GH	k_{I0-GH} = The number of the characters that are included in a set of posts per product or service features which commented by the customers. p_{I0-GH} = The number of product-service features which were commented by the customers.

The logarithm of the probability distribution is useful as a measure of entropy because it is additive for independent sources. It is assumed that the probability of all the messages transferred from the sender to the receiver are equal. Moreover, the communication channel is assumed to be noiseless. Based on the above, and considering Table IV-1, Table IV-2, two metrics are defined for quantifying of the complexity and the capacity of an Industry 4.0 manufacturing system, based on Eq. (1). The ***complexity metric of an Industry 4.0 manufacturing system***, is defined as the sum of the individual information content of the traditional and the enhanced manufacturing systems, as presented below:

$$CI^{I4.0} = CI^{traditional} + CI^{enhanced} \quad \textbf{(3)}$$

For the traditional system, the sum of entropies is:

$$\begin{aligned} CI^{traditional} = & -p_{A-B}\log_2\frac{1}{k_{A-B}} - p_{B-CDE}\log_2\frac{1}{k_{B-CDE}} \\ & - p_{D-E}\log_2\frac{1}{k_{D-E}} - p_{E-F}\log_2\frac{1}{k_{E-F}} \\ & - p_{GH-A}\log_2\frac{1}{k_{GH-A}} \end{aligned} \quad \textbf{(4)}$$

Each component of the Eq. (4) represents the average amount of information per message which is sent between source and receiver, meaning the defined systems (A, B, …etc.).

For each additional component included in the Industry 4.0 manufacturing system the sum of entropies is:

$$CI^{enhanced} = -p_{I1-A}\log_2\frac{1}{k_{I1-A}} - p_{I2-A}\log_2\frac{1}{k_{I2-A}} - p_{I3-A}\log_2\frac{1}{k_{I3-A}} - p_{I4-I5I6}\log_2\frac{1}{k_{I4-I5I6}} - p_{I5I6-CD}\log_2\frac{1}{k_{I5I6-CD}} - p_{I6I7-D}\log_2\frac{1}{k_{I6I7-D}} - p_{I7-ED}\log_2\frac{1}{k_{I7-ED}} - p_{I8-GH}\log_2\frac{1}{k_{I8-GH}} - p_{I9-GH}\log_2\frac{1}{k_{I9-GH}} - p_{I10-GH}\log_2\frac{1}{k_{I10-GH}} \quad (5)$$

It is clear that $CI^{I4.0} > CI^{traditional}$. However, it is reasonable to estimate how much larger the entropies of Industry 4.0 are, compared to those of the traditional manufacturing. Aiming to address that, the capacity of the channel should be calculated.

Following the equation (2) of the capacity of a communication channel, the ***capacity metric of an Industry 4.0 manufacturing system*** is defined as the sum of the individual capacities of all the components of the system as presented below:

$$CC^{I4.0} = CC^{traditional} + CC^{enhanced} \quad (6)$$

Where:

$$CC^{traditional} = -\frac{p_{A-B}}{T_{A-B}}\log_2\frac{1}{k_{A-B}} - \frac{p_{B-CDE}}{T_{B-CDE}}\log_2\frac{1}{k_{B-CDE}} - \frac{p_{D-E}}{T_{D-E}}\log_2\frac{1}{k_{D-E}} - \frac{p_{E-F}}{T_{E-F}}\log_2\frac{1}{k_{E-F}} - \frac{p_{GH-A}}{T_{GH-A}}\log_2\frac{1}{k_{GH-A}} \quad (7)$$

$$CC^{enhanced} = -\frac{p_{I1-A}}{T_{I1-A}}\log_2\frac{1}{k_{I1-A}} - \frac{p_{I2-A}}{T_{I2-A}}\log_2\frac{1}{k_{I2-A}} = -\frac{p_{I3-A}}{T_{I3-A}}\log_2\frac{1}{k_{I3-A}} - \frac{p_{I4-I5I6}}{T_{I4-I5I6}}\log_2\frac{1}{k_{I4-I5I6}} = -\frac{p_{I5I6-CD}}{T_{I5I6-CD}}\log_2\frac{1}{k_{I5I6-CD}} - \frac{p_{I6I7-D}}{T_{I6I7-D}}\log_2\frac{1}{k_{I6I7-D}} = -\frac{p_{I7-ED}}{T_{I7-ED}}\log_2\frac{1}{k_{I7-ED}} - \frac{p_{I8-GH}}{T_{I8-GH}}\log_2\frac{1}{k_{I8-GH}} = -\frac{p_{I9-GH}}{T_{I9-GH}}\log_2\frac{1}{k_{I9-GH}} - \frac{p_{I10-GH}}{T_{I10-GH}}\log_2\frac{1}{k_{I10-GH}} \quad (8)$$

The systems that are comprised in the Industry 4.0 system, permit the faster communication between different systems. As a result, based on the equations (7) and (8), it is valid to say that since $T^{enhanced} << T^{traditional}$, the capacity of the enhanced system is greater than that of the traditional one.

The estimation of the channels' capacity is important when the noise of the channel is considered, since the noise in a channel increases the amount of transmitted information with spurious content, and given a specific capacity of a channel, it may fail to transfer the information. From the manufacturing perspective, the potential errors in the design drawings of parts, can be considered as noise.

V. Case study and Results

The proposed approach is investigated in a bicycle manufacturing company. In a first level of analysis, comparison of i) the entropy, ii) the information content, and iii) the capacity, is performed between the components that comprise the manufacturing systems. Subsequently, the two-defined metrics namely complexity and capacity, are calculated before and after the shifting of this industry to the Industry 4.0 paradigm.

Based on a previous study [12], for a traditional manufacturing system, the variables k which represent the transmitted messages from the selected sources, take values in the interval $k\epsilon$[8-26]. Additionally, the variables p which represent the number of messages that are required to be sent by the selected source, take values $p\epsilon$[2-48]. For example, the designs of the parts for a typical bicycle, use k = 26 different variables, for which the total number of messages needed to describe these design variable is 48. Following the same research work [12], the time required for the transmission of a message between the systems is $T\epsilon$[30-700] seconds.

For the enhanced systems of Industry 4.0, it is assumed that the variables $k'\epsilon[10\text{-}10^6]$. This high range of values which could be taken by the variables k' reflects the large amount of feedback from the control systems, social media, etc., that all together comprise the Big Data and should be managed in order to obtain useful information for the design and the development of products. Contrarily to the previous, the product features which are changing when the feedback is received, as well as the control variables that are implying adaptation are too small, $p'\epsilon$[1-20]. Moreover, the time required for the transmission of a message between the enhanced systems of Industry 4.0 is assumed to be $T'\epsilon$[0.5-120] seconds. After selecting random values between these intervals to match the variables of Table IV-1 and Table IV-2, the results can be visualized in Figure 2 and Figure 3.

Figure 2 depicts the Information Content and the Entropy of each system of Figure 1 calculated with the random data of known intervals, which are provided in the previous paragraphs. From A-B to GH-A are the components of a traditional manufacturing system, while from I1-A to I10-GH, the components of Industry 4.0 system are depicted. It is observed that in the transition from the traditional to Industry 4.0 system there is a jump in the entropy value from 3.5 bits/ message to 19.8 bits/ message. On the other hand, the information content of the traditional versus the industry 4.0 systems, is increasing in a low rate, and all the values converge in the average.

Based on the information theory fundamentals, the rate of information, or the signal which can communicate over the channel should be maximized, since it is assumed that the greater the capacity of the channel, the more possible for the channel to transmit information with arbitrarily small block error during the procedure of coding and recoding. The capacity calculation of the manufacturing systems is depicted in Figure 3, where a big

jump is observed for the systems I1-A, meaning that the industry 4.0 systems performed better in the existence of noise.

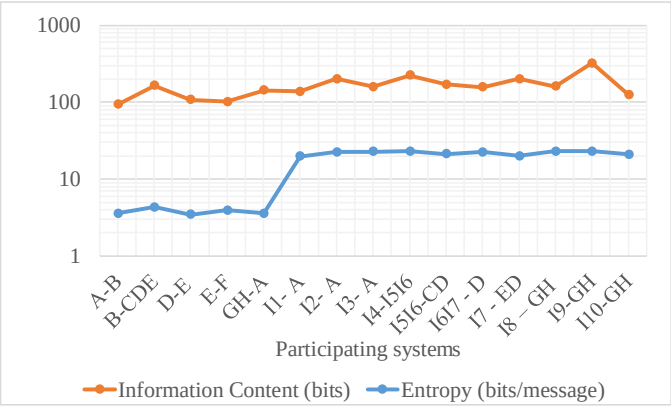


Figure 2 Information content and Entropy calculation for all the systems of a traditional and an Industry 4.0 manufacturing

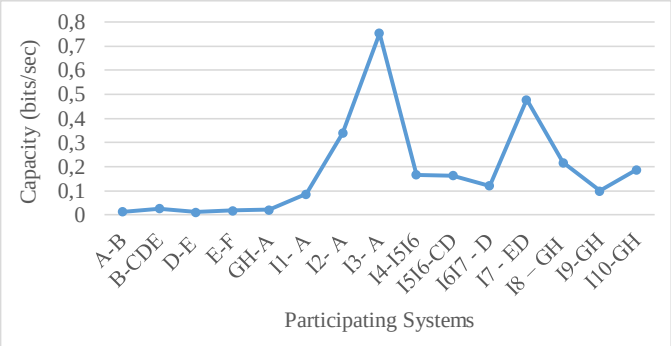


Figure 3 Capacity calculation for all the systems of a traditional and an Industry 4.0 manufacturing

Finally, for the defined values of *k* and *p*, the Complexity and the Capacity metrics of the company before and after shifting to industry 4.0 are calculated. Concluding the above, it is shown that the complexity of an industry 4.0 system is 70% higher than the complexity of a traditional system. Also, the capacity of the industry 4.0 manufacturing system was observed as 96% higher than the one of the traditional system, something which denotes that the Industry 4.0 manufacturing system, apart from its higher complexity, performs better in noisy communication channels and enables better exchange of information among the systems.

VI. Conclusions

The present work aims to bridge the literature gaps regarding the quantification approaches for the new manufacturing paradigm of Industry 4.0. To this end, in the present work a modelling of an Industry 4.0 manufacturing system based on Information Theory is introduced, and an approach for the quantitative evaluation of such systems by defining metrics for complexity and capacity of Industry 4.0 manufacturing systems. The proposed approach was applied in a bicycle manufacturing company, and the results prove an increase in the complexity of manufacturing system by shifting to Industry 4.0, with a concurrently larger increase on the capacity of the system.

In future work, the proposed approach will be enhanced by revoking some of the assumptions. Particularly, the noise of the communication channels will be further investigated. In addition to that the messages transferred from the senders to the receivers will not be considered as equiprobable.

References

[1] G. Chryssolouris, Manufacturing Systems: Theory and Practice. 2nd New York: Springer-Verlag, 2006.

[2] D. Mourtzis, E. Vlachou, “Cloud-based Cyber-Physical Systems and Quality of Services”, TQM Emerald Journal, vol. 28, Issue 5, 2016, pp.704-733.

[3] L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, K. Ueda, “Cyber-physical systems in manufacturing”, CIRP Annals - Manufacturing Technology, Volume 65, Issue 2, 2016, pp. 621–641.

[4] J. Lee, B. Bagheri, H.-A. Kao, “A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems”, Manufacturing Letters, vol. 3, 2015, pp 18–23.

[5] S. Wang, J. Wan, D. Zhang, D. Li, C. Zhang, “Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination”, Computer Networks, vol. 101, 2016, pp 158–168.

[6] D. Mourtzis, E. Vlachou, C. Giannoulis, E. Siganakis, V. Zogopoulos, "Applications for Frugal Product Customization and Design of Manufacturing Networks", Procedia CIRP, 6th International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2016), vol. 52, 2016, pp.228-233.

[7] D. Mourtzis, E. Vlachou, N. Xanthopoulos, M. Givehchi, L. Wang, "Cloud based- adaptive process planning considering availability and capabilities of machine tools", JMSY, vol. 39, 2016, pp.1-8.

[8] M. A. Bajestani, D. Banjevic, J.C. Beck, “Integrated maintenance planning and production scheduling with Markovian deteriorating machine conditions”. IJPR. 52/24, 2014, pp. 7377-7400.

[9] J.M. Tien, “The next industrial revolution: Integrated services and goods, Journal of System Sciences and Systems Engineering”, vol 21, 2012, pp. 257-296.

[10] R.S. Renu, G. Mocko, A. Koneru, “Use of Big Data and Knowledge Discovery to Create Data Backbones for Decision Support Systems, Procedia Computer Science, 2013; 20:446-453.

[11] W. ElMaraghy, H. ElMaraghy, T. Tomiyama, L. Monostorid, CIRP Annals - Manufacturing Technology, vol. 61, Issue 2, 2012, pp 793–814

[12] G. Chryssolouris, E. Vassiliou, D. Mavrikios, “Application of Information Theory to the Quantification of Concurrent Engineering Processes”, 13th International Conference on Concurrent Engineering (ISPE), France, 2006, pp.679-695.

[13] K. Efthymiou, A. Pagoropoulos, N. Papakostas, D. Mourtzis, G. Chryssolouris, “Manufacturing Systems Complexity Analysis Methods Review”, International Journal of Computer Integrated Manufacturing, vol. 29, issue 9, 2016, pp 1025-1044

[14] ICTC, Big Data & The Intelligence Economy, Accessed by http://www.ictcctic.ca/wp-content/uploads/2015/12/BIG-DATA-2015.pdf

[15] C. E. Shannon, (1948), A mathematical theory of communication, Bell Sys. Tech. f. 27, 379-432, 623-6.

IoT for All: Architectural Design of an Extensible and Lightweight IoT Analytics Platform

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Stephan Müller, Patrick Wiener, Adrian Bürger and Jens Nimis
Karlsruhe University of Applied Sciences, Karlsruhe, Germany
mail@stephanmueller.eu, {patrick.wiener, adrian.buerger, jens.nimis}@hs-karlsruhe.de

***Abstract*—Realizing Internet of Things (IoT) is still a huge challenge for Small and Medium-sized Enterprises (SMEs) caused by heavy upfront capital investments and inadequate infrastructure. An opportunity to counteract the concern is provided in this paper by introducing an extensible and lightweight IoT analytics platform. Therefore, we present an architectural design offering the possibility of connecting arbitrary IoT devices to the analytics platform which includes data processing and analytics, data storage as well as data delivery. Regardless of the integration of already purchased non IoT-ready devices or acquisition of new IoT capable devices, the design allows to gradually build an IoT environment in enterprises provoked by its modular structure and relying on only open source software-based tools.**

I. Introduction

A lot has been written over the past years on the huge potential of Internet of Things (IoT) in industries and more specifically of industry 4.0 in the secondary sector. The comprehensive digitalization of enterprises introduces additional sensors into any area of production and organisation. On the one hand, respective data allows for competitive advantages due to better classical optimization of processes [1] and/or resource consumption [2]. More advanced techniques like time series-based predictive analytics lead to promising applications such as anomaly detection and especially predictive maintenance [3], [4] which increase efficiency and availability of production plants [1].

On the other hand, the vision of IoT-supported data driven applications is much broader. It ranges from improved general decision making [3], [1] to the deduction of new, formerly hidden knowledge [5], [6], thus facilitating new revenue streams and business models [1].

However, a great deal of additional value still remains to be captured especially by Small and Medium-sized Enterprises (SMEs), as they face numerous challenges on their way into IoT. A lack of predominant comprehensive standards bears the risk of incompatibilities and especially a lack of interoperability with existing systems. This doubts sustainability of the generally heavy upfront capital investment required to start IoT activities, e.g. due to the high cost of sensors and infrastructure [7]. Other technological and organizational concerns comprise issues like stability, complexity, data security, or a lack of staff's expertise.

To lower some of these entry barriers, we propose *IoT for All*, an architectural design of an extensible and lightweight IoT analytics platform, which is sketched in Figure 1.

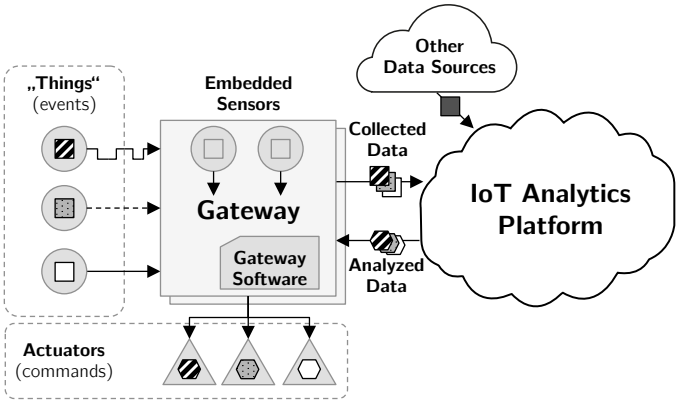


Fig. 1. High-level architectural design of the IoT for All analytics platform interacting with the physical world.

In the platform, sensors measure or detect physical properties of "things", e.g., temperature, humidity, or flow volume, while actuators control system components' settings, e.g., servo motor angles or valve positions. Sensor readings are collected by a gateway, that itself may contain various embedded sensors, and sent to an IoT platform where the received data are transformed, aggregated, enriched and analyzed. Relevant results are emitted back to a dedicated gateway to control the actuators.

The outline of the paper is as following: in Section II we present a motivating use case for the IoT for All platform and derive general architectural design requirements accordingly. The IoT analytics platform design is elaborated in Section III and an exemplary implementation is given in Section IV. Section V discusses the results wrt. related work and Section VI concludes the paper.

II. Motivating use case

A. Solar Thermal Climate System

At the Faculty of Management Science and Engineering at Karlsruhe University of Applied Sciences, a Solar Thermal Climate System (STCS) is installed as depicted in Figure 2 and Figure 3. Central component is an Adsorption Cooling Machine (ACM) that uses solar heat as driving energy. The cooling energy produced by the ACM is used to cool down parts of the faculty building during summer. During winter, the solar heat is used directly to support the buildings conventional heating system. Shortfalls in supplies of heating and cooling

energy can be balanced through a high temperature and a low temperature water storage [8].



Fig. 2. Central STCS components: cold water storage (left), ACM (center left), hot water storages (center right), pumps and heat exchangers (right) [8].

In addition to its use as a climate system, the STCS is utilized within several research projects. One project focuses on the development of software for efficient, predictive control of renewable energy systems, wherein the machines current, conventional controller is replaced by an optimization-based, model predictive controller [8]. These elaborate control strategies are in need of additional measurement devices which have not been required by the conventional controller and which now must be installed. Respective data must be accessible alongside with those of present devices. Table I shows the volume of data produced by the several measurement devices, where the symbol "(+)" marks subsequent installations.

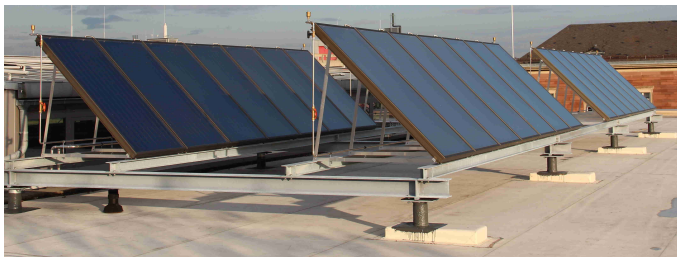


Fig. 3. Solar panels on the roof of the faculty building collect thermal energy for either (1) heating of the building's interior, or (2) operating the ACM [8].

The main purpose of the collected sensor data is to monitor and control the system's state optimally, while the peripheral actuators must be controllable in a simple and unified way to facilitate automated tests and control. Additionally, the collected data is provided to various consumers for further processing and analyses.

B. Requirements

The presented use case is one example to motivate a wide field of applications within the range of the IoT domain. Thus defining specific functional and non-functional requirements that shape our systems' design approach is not feasible, which is why we refer to the more general term of requirements in the following.

1) Extensibility: Due to the ongoing research work on the STCS, a flexible design with modular structure is needed to allow further integration of (1) new IoT devices (hardware side), as well as (2) additional analytics jobs (software side).

TABLE I. OVERVIEW OF ACCUMULATING SENSOR DATA

Devices	Request rate	Values / day	MB / day
40 temperature sensors	1 Hz	3,456,000	13.183
36 temperature sensors (+)	0.05 Hz	155,520	0.593
3 volume flow sensors	1 Hz	518,400	1.978
3 volume flow sensors (+)	1 Hz	518,400	1.978
3 control signals (double)	1 Hz	259,200	0.989
15 control signals (boolean)	1 Hz	1,296,000	1.236
4 solar radiation sensors (+)	0.2 Hz	17,280	0.07
2 weather forecasts (+)	4 / day	1,296[1)]	0.005
Total amount		6,222,096	20.023

[1)] 10 hourly forecast quantities for 144 hours

2) Lightweight design: All involved software components should be realizable with only open source software-based tools in combination with low-cost hardware to minimize investment costs.

3) Connectivity: External applications should be able to communicate with the system's IoT devices and to retrieve collected data.

4) Real-time processing capability: In general, the architectural design must account for the amount of data that is produced at a high frequency, as shown in Table I. Thus, to derive meaningful insights from the sensor readings and define measurements accordingly, it is necessary to process and analyze these data streams in real-time.

5) Data persistency: Raw data as well as processed data should be stored for future data analysis, for example evaluation of historical observations or training of machine learning models.

6) Fault data-aware: Real-world data collected by IoT devices can be errorprone and inaccurate due to system outages, damaged sensors or erroneous communication. Therefore, faulty data should be tackled at an early stage of the IoT analytics platform to prevent the spread of it.

7) Scalability: Scalability is desired for handling future system extensions that might result in a higher data throughput.

III. AN ARCHITECTURAL DESIGN OF AN EXTENSIBLE AND LIGHTWEIGHT IOT ANALYTICS PLATFORM

An architectural design for an IoT analytics platform, which accounts for the above requirements, involves decomposing the overall systems' design into various architectural layers and corresponding elements as shown in Figure 4. In this respect, the Kappa Architecture [9] is used as the theoretical foundation of our design. It evolves from the generic, scalable and fault-tolerant Lambda Architecture [10] in the area of big data processing. In contrast to the Lambda Architecture, which is comprised of three layers (batch layer, speed layer, serving layer) to hybridly combine batch and streaming analytics, the Kappa Architecture neglects the batch layer to simplify the overall systems' design, assuming that batch processing is only a subset of stream processing on a finite set of input data.

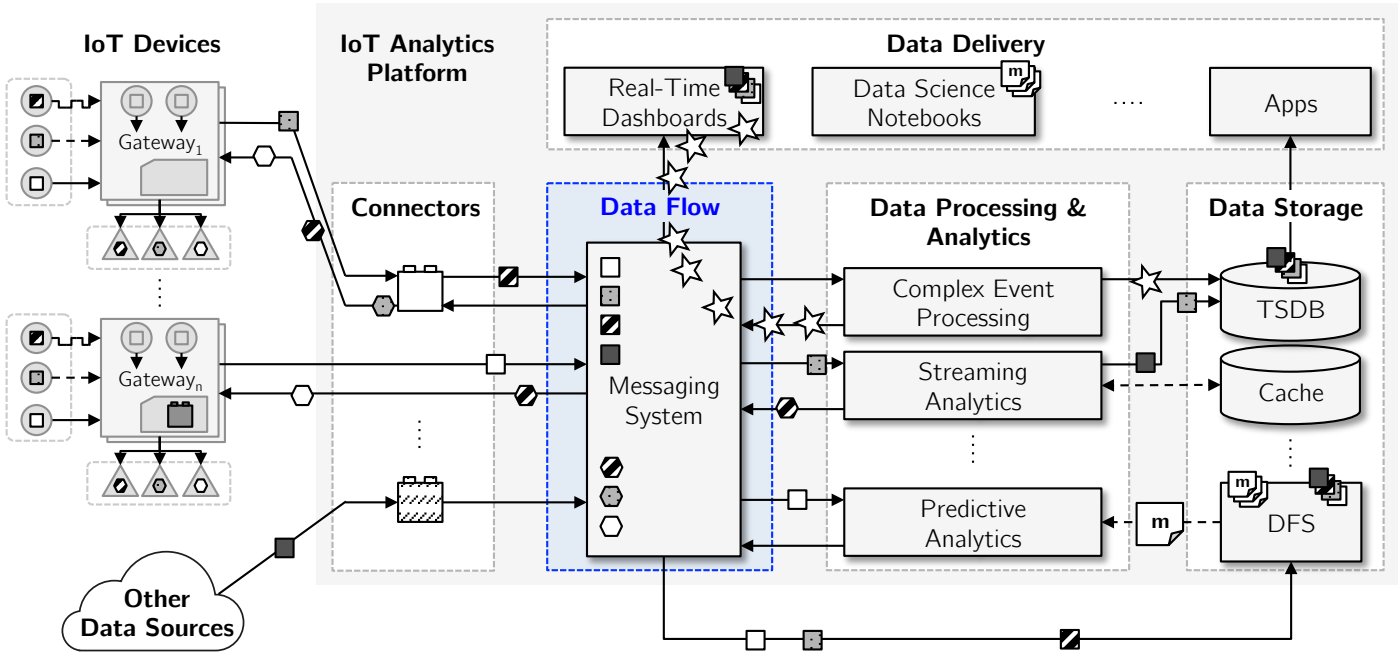


Fig. 4. Scalable and modular architectural design of an IoT analytics platform for processing and analyzing data streams from a variety of existing IoT devices and other useful data sources.

A. Connectors

A typical system configuration in the IoT domain consists of a huge variety of different devices, each providing access to a manifold of sensors and actuators. Additionally, these devices often operate in a constrained environment due to limited resources such as processing power, amount of memory or battery consumption. Furthermore, integrating other (external) data sources, e.g., to enrich the sensor readings in the course of data exploitation in the downstream pipeline elements, can help to generate valuable insights. Thus, it is crucial to provide connectors to allow (1) IoT devices and other data sources to be easily integrated in the analytics platform, and, if needed, (2) to send commands to dedicated actuators accordingly. As shown in Figure 4, connectors can either run as stand-alone interfaces on the IoT analytics platform or operate integrated on the device if feasible.

B. Data Flow

Loosely coupling arbitrary data producers and data consumers through a unified message broker is crucial for any type of IoT application. Therefore, the data flow layer, which comprises a messaging system, is key to our systems' design approach. Thereby, a topic-based publish-subscribe approach paired with a suitable data processing semantics, e.g., at-least-once semantics, enables multiple consumers to process and analyze the subscribed IoT data streams in order to meet different business requirements, thus allowing the producer and consumer side to be scaled independently.

C. Data Processing and Analytics

The data processing and analytics layer consists of various elements, that each fulfill different computational tasks by processing incoming sensor data streams in a linear and scalable manner, e.g., transforming, aggregating, enriching the consumed IoT data, and storing the results into persistent data storages such as a Time Series Database (TSDB), and continuously updating a cache storage. In addition to these streaming analytics tasks, business rules embedded in complex event processing jobs as well as pre-trained machine learning models, e.g., to classify faulty system states, complement the analytics layer and publish their analysis results back to the messaging system of the data flow layer.

D. Data Storage

The data storage layer holds space for a variety of different storage backends to (1) persistently store newly processed and analyzed IoT data results, e.g., in a dedicated TSDB and/or cache storage to allow for fast ingest and retrieval of recent analysis results, as well as to (2) provide cheap long-term historical data storages, e.g., a Distributed Filesystem (DFS) such as the Hadoop Distributed Filesystem (HDFS)[1] that allows for profound historical analysis or training of machine learning models.

E. Data Delivery

Numerous elements in the data delivery layer provide access to either data in motion, such as monitoring the raw IoT data and presenting triggered complex events in adequate real-time dashboards, or to data at rest by retrieving it from the storage backends, e.g., to allow data scientists to train and evaluate machine learning models or feed other applications with information extracted by the IoT analytics platform. As a consequence, various user roles such as domain experts or data scientists interacting with the platform, can easily be provided with personalized information relevant for their specific processes.

[1] http://hadoop.apache.org/

IV. IMPLEMENTATION

In this section, we give an overview of the implementation of the proposed architectural design, including the involved technologies as shown in Figure 5. In addition, we demonstrate how further IoT devices can be integrated in the analytics platform.

A. Implemented Elements of the IoT Analytics Platform

A key factor for ensuring flexibility and modularity (see Section II-B1) is the utilization of software container technology, with the additional advantages of hardware independency and opportunities of scalable applications (Section II-B7). Hence, all elements of the different layers are packed into software containers, realized with Docker[2].

1) IoT Devices: Several measurement devices like temperature sensors (via 1Wire protocol), volume flow sensors (via voltage signals) and solar irradiation sensors (via RS232 protocol) are connected, e.g., to Arduino microcontrollers and Raspberry Pi single board computers, both low-cost and low-energy hardware (Section II-B2) supported by a large community. Other devices are accessed via a variety of different communication protocols, like the standard system controller and its connected measurement devices by the controller's ModbusTCP interface, or like weather forecasts over HTTP.

2) Connectors: Connectors as the entrypoint for sensor data to the analytics platform have to support a variety of messaging protocols like illustrated in the previous section. In order to reduce programming heterogeneity, Python has been selected as a platform independent modular programming language supported by a large community and applicable as generalist with support of any relevant communication protocol. During the research project, a dedicated Python module, based on both a Kafka client and an Avro serialization module, has been developed. It simplifies building connectors that allow for communication with the messaging system and thereby enables fast integration of new IoT devices.
Furthermore, connectors are also responsible for controlling actuators, which is currently not implemented but already prepared for integration.

3) Data Flow: Providing a message broker as explained in Section III-B is a part of the data flow layer. For fulfilling the needs, Apache Kafka[3] is used, a highly scalable and distributed messaging system, capable for big data (Section II-B4). Standardization, a fundamental precondition for providing data over public interfaces, is realized with Apache Avro[4], a compact, binary and fast data format and its according toolset. Therefore, data schemas have to be defined, containing data types for each sensor value. Schema management and versioning is provided by Confluent's Schema Registry. On this basis, Avro serializers, which are part of the data producing elements, detect and reject incorrect data types and structures during the serialization process, i.e. before messages can enter the messaging system. Nevertheless, faulty messages could be delivered to consumers, e.g. data with wrong schema versions or missing schema. However, they can immediately be detected during the deserialization process on consumer side.

4) Data Processing and Analytics: A complex event processing framework, realized with Apache Flink[5], is integrated at an early stage that catches events, stores them in the database and informs other layer elements by sending occurred events back to the message system.
Real-time classification of time series is an example for a predictive analytics task. The classification step is the precursor for a current research project of a forecasting unit to predict future STCS states.
Depending on their computational complexity, stream processing tasks are either realized with Flink or Python. Whereas Python is sufficient for simple jobs, Flink should be preferred for scalable jobs. In our case, Python is used for calculating real-time indicators depending exclusively on the latest sensor values, while Flink is used for calculation of aggregations over longer intervals as input for real-time classification.

5) Data Storage: Because the amount of data is not excessive in our case, the relational database management system Postgres[6] sufficiently provides a storage back-end. On growing data stream, optimized Time Series Databases such as InfluxDB should preferably be chosen. Currently, the data storage includes (1) collection of trained predictive models, and (2) data logging of all messages ever arrived the messaging system and it therefore fulfills the requirement of a long-term data storage (Section II-B5). Logging of data is done automatically as soon as a new message topic is created after deriving the database schema from the Avro schema.

6) Data Delivery: The user's interaction with the system is enabled through Jupyter Notebook[7], a powerful data science notebook application as extension of the Python functionalities. Training predictive models can easily be performed, ensured by the existing Python predictive modeling libraries.
In addition, a real-time web monitor is realized to show the current state of all system components. Since native web applications cannot directly communicate with Kafka, a Websocket server as a requirement for real-time native web apps, which forwards all incoming device messages, acts as an intermediate stage to allow dynamic processing of data (Section II-B3).

B. Demonstration of Modularity: Integrating New IoT Devices

Integration of further IoT devices as additional data providers leads to efforts in nearly all layers of the analytics platform. While a few components such as data persistence are provided automatically, others need manual modifications. In this subsection, the focus is on the installation and integration of IoT devices to demonstrate modularity. After a general description of the integration process, it is exemplified by the integration of a solar irradiation sensor.

1) Description of the integration process: Mounting a new device consists of several steps that include the physical installation of the device, configuration of the device according to its intended usage and connecting the device to an already existing or new IoT gateway, followed by creating a new connector application that propagates the data to the messaging system for further distribution.

[2]https://www.docker.com/
[3]https://kafka.apache.org/
[4]https://avro.apache.org/
[5]https://flink.apache.org/
[6]https://www.postgresql.org/
[7]https://jupyter.org/

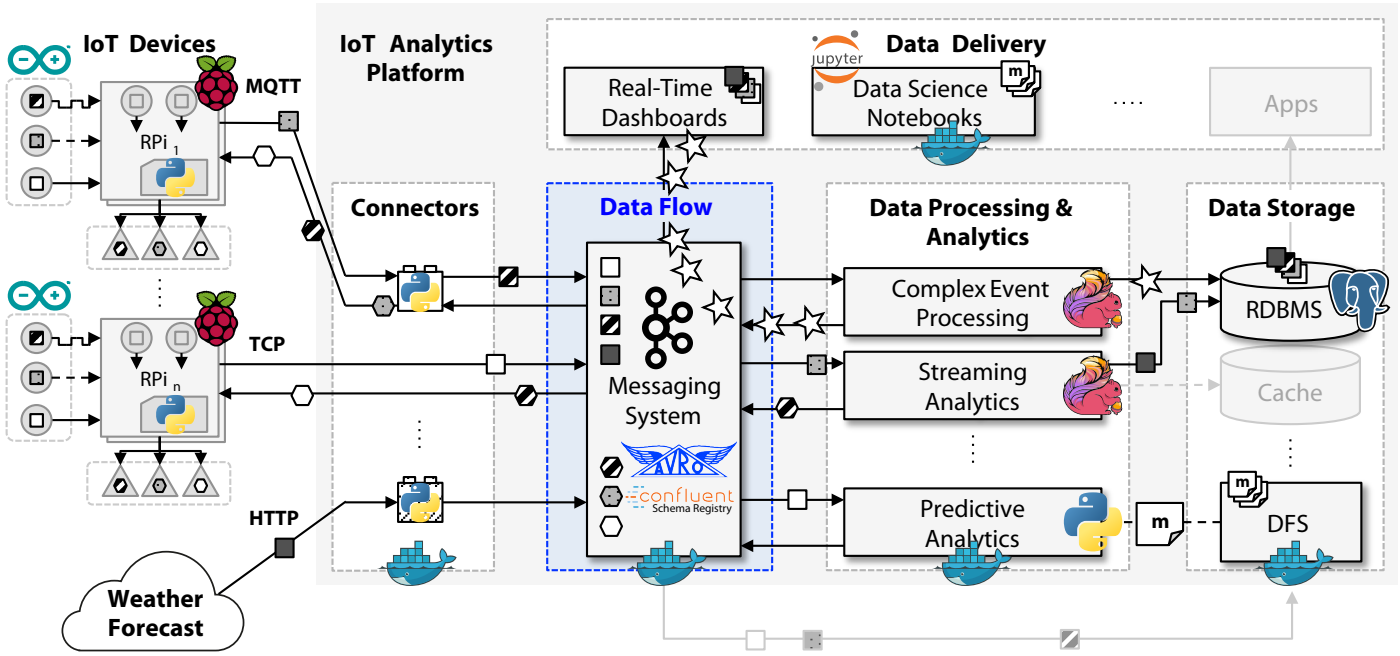


Fig. 5. Specific implementation of the proposed IoT analytics platform architectural design, highlighting used software and hardware tools.

In our case, mostly Raspberry Pi boards are used as gateways to which the device has to be connected with. After configuration, a new Python producer script has to be created with (1) a Kafka client implementation, (2) a function to read IoT device values and (3) a loop in which the reading function and the produce function of the Kafka client are called recurrently. Creating such a script is supported by the specific developed Python module as mentioned in Section IV-A2.

Since we use Kafka in our implementation, a topic has to be defined to separate the messages of different device or sensor groups. The usage of Avro requires a definition of a schema file including a mandatory timestamp field and at least one additional data field.

After the setup, the collector script can be executed, while continuous execution of the sensor readings can be ensured by using a process control system such as Supervisor[8]. Incoming data then are immediately available to any subscribing consumer via the Kafka messaging system.

2) Example: Integration of a solar irradiation sensor: Within the system described in Subsection II-A, a solar irradiation sensor of type SPN1 from Delta-T Devices is integrated. In addition to analogous outputs, measurements of the SPN1 can be read via an RS232 interface which is designed for connection to a PC or similar devices [11].

In our case, the RS232 interface is connected to a Raspberry Pi using an RS232 adapter that can be connected to the Raspberry Pi's GPIO pins, which is simultaneously used to provide electric power to the measurement unit of the SPN1.

The command sequences to be sent to the SPN1 for retrieving measurement data via serial console are given in the SPN1 manual [11]. For automation of the communication process, the Python serial communication module pySerial [12] is used for sending commands to the SPN1 and collecting its responses. In the sporadic case of communication errors, unexpected messages are detected and the communication process is repeated.

On successful communication, the measurement data obtained from the serial connection is parsed, and the relevant values and a corresponding timestamp are matched to the fields defined in the according Avro schema for the measurement device's Kafka topic and communicated to Kafka using the Python module mentioned in Section IV-A2.

The procedure is repeated within a loop on a defined time interval. The process control system Supervisor is used to ensure that the process is restarted also after occurrence of unexpected errors.

V. Discussion and Related Work

Our approach to lower some of SMEs' entry barriers into IoT is an architectural design of an extensible and lightweight IoT analytics platform as presented in the sections above. Its focus was to provide a generally applicable design, which addresses the general requirements that can be found in most IoT applications, while reducing upfront capital investments. The implementation example has shown how this can be achieved by the use of open source software-based tools and low priced commodity hardware.

The most similar approach to IoT for All has been presented by Patil and Reddy [13]. While their selection of software and hardware tools is close to ours, the focus of their work is rather on implementation than on design side. Thus their approach does not cover all requirements and the complete design addressed by our approach.

IoTOne [14] provides an organizational approach to interoperability of an infrastructure for heterogeneous IoT devices

[8]http://supervisord.org/

and as such has its focus more on devices than on data processing and analytics. The vendor-driven platform aims to be user friendly especially towards homeowners, allows robust control of devices, and supports secure communications.

Cloud4IoT [15] is a heterogeneous, distributed and autonomic Cloud platform for IoT and provides seamless integration of IoT and Cloud Computing. Therefore, architectural patterns typical of Cloud Computing (e.g., Infrastructure as Code) are combined with patterns from IoT. It covers features typical of both approaches and addresses the requirements of flexibility, scalability, fault tolerance, and distribution of IoT and data-intensive applications. The automatic deployment and orchestration of IoT support applications and data logic and processing applications is performed using containerization solutions based on open source technologies. In contrast to Cloud4IoT, our approach is deployment model agnostic and focuses more on simplification, as we believe that the introduction of IoT and Cloud Computing at the same time might overburden SMEs at IoT entry level.

The IoT big data Analytics Framework (IBDA) [16] has three major integrated technology components: IoT sensors, big data management and analytics. As the used infrastructure is restricted to Apache's HDFS it is missing the generality and openness of IoT for All.

For the aforementioned approaches, the fulfillment of the presented requirements (Section II-B) is shown in Table II.

TABLE II. OVERVIEW OF RELATED WORK WITH RESPECT TO THE PREDEFINED REQUIREMENTS.

Legend:
○ Requirement not fulfilled
◐ Requirement partially fulfilled
● Requirement fulfilled

Requirements:	Patil and Reddy [13]	IoTOne [14]	Cloud4IoT [15]	IBDA [16]
Extensibility	●	●	●	●
Lightweight design	●	●	●	●
Connectivity	●	◐	●	●
Real-time processing capability	●	○ [1)]	●	◐
Data persistency	●	○	●	●
Fault data-aware	○	○	○	○
Scalability	●	○	●	●

[1)] Real-time capable for home automation but not for high data volumes.

Besides the more comprehensive approaches presented as related work in this section, there exist a number of offerings that highlight selected aspects of IoT for All, like e.g., IoT gateways provided by industrial suppliers as Beckhoff or Harting MICA, or data management tools like Confluent's Contol Center.

VI. CONCLUSION AND FUTURE WORK

IoT for All shows how to connect arbitrary IoT devices to an analytics platform including data processing and analytics, data storage as well as data delivery in a lightweight fashion. It is suitable for pre-existing non IoT-ready devices and infrastructure components as well as for gradual introduction and extension of comprehensive IoT applications. The STCS implementation of our approach is currently used in a number of research projects and further elaborated in this context.

Within industrial contexts, IoT for All could be used within monitoring of production processes and quality management as well as for supervision of energy consumptions of processes and plants, while especially SMEs could benefit from low preliminary investment costs and simplified integration of heterogeneous data sources.

As a future activity on IoT for All, a better integration of data security is planned with secure communication, authentication and authorization based on SSL and OAuth. After this extension, we intend to provide the core software packages as Docker containers together with its documentation as open source project.

REFERENCES

[1] S. Haller, S. Karnouskos, and C. Schroth, *Future Internet — FIS 2008*. Berlin, Heidelberg: Springer-Verlag, 2009, ch. The Internet of Things in an Enterprise Context, pp. 14–28.

[2] L. Coetzee and J. Eksteen, "The Internet of Things - Promise for the Future? An Introduction," in *2011 IST-Africa Conf. Proc.*, May 2011, pp. 1–9.

[3] G. Kecskemeti, G. Casale, D. N. Jha, J. Lyon, and R. Ranjan, "Modelling and Simulation Challenges in Internet of Things," *IEEE Cloud Computing*, vol. 4, no. 1, pp. 62–69, Jan. 2017.

[4] Y. Yamato, H. Kumazaki, and Y. Fukumoto, "Proposal of Lambda Architecture Adoption for Real Time Predictive Maintenance," in *2016 4th Int. Symposium on Computing and Networking (CANDAR)*, Nov. 2016, pp. 713–715.

[5] N. Lelwala, "Ensemble Inference Based Framework for Creating Knowledge from Big Data in IoT," in *2016 16th Int. Conf on Advances in ICT for Emerging Regions (ICTer)*, Sep. 2016, pp. 245–245.

[6] A. Gyrard, P. Patel, A. Sheth, and M. Serrano, "Building the Web of Knowledge with Smart IoT Applications," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 83–88, Sep. 2016.

[7] D. O'Halloran and E. Kvochko, "Industrial Internet of Things: Unleashing the Potential of Connected Products and Services," http://www3.weforum.org/docs/WEFUSA_IndustrialInternet_Report2015.pdf, Jan. 2015.

[8] A. Bürger, A. Altmann-Dieses, M. Diehl, and M. Braun, "Software und Methoden zur optimierungsbasierten Regelung nachhaltiger Energiesysteme," in *Forschung aktuell 2017*. Karlsruhe University of Applied Sciences, 2017.

[9] J. Kreps, "Questioning the Lambda Architecture," https://www.oreilly.com/ideas/questioning-the-lambda-architecture, 2014.

[10] N. Marz and J. Warren, *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Manning Publications, 2013.

[11] Delta T Devices Ltd., "User Manual for the Sunshine Pyranometer Type SPN1," http://www.dynamax.com/images/uploads/papers/SPN1_Manual.pdf, visited on August 3, 2017.

[12] C. Liechti, "pySerial Documentation," https://pythonhosted.org/pyserial/, visited on August 3, 2017.

[13] S. Patil and S. Reddy, "Designing and Implementing your IOT Solutions with Open Source," https://de.slideshare.net/HadoopSummit/designing-and-implementing-your-iot-solutions-with-open-source, Jul. 2016.

[14] N. Gyory and M. Chuah, "IoTOne: Integrated platform for heterogeneous IoT devices," in *2017 Int. Conf. on Computing, Networking and Communications (ICNC)*, Jan. 2017, pp. 783–787.

[15] D. Pizzolli, G. Cossu, D. Santoro, L. Capra, C. Dupont, D. Charalampos, F. D. Pellegrini, F. Antonelli, and S. Cretti, "Cloud4IoT: A Heterogeneous, Distributed and Autonomic Cloud Platform for the IoT," in *2016 IEEE Int. Conf. on Cloud Computing Technology and Science (CloudCom)*, Dec. 2016, pp. 476–479.

[16] M. R. Bashir and A. Q. Gill, "Towards an IoT Big Data Analytics Framework: Smart Buildings Systems," in *2016 IEEE 18th Int. Conf. on High Performance Computing and Communications; IEEE 14th Int. Conf. on Smart City; IEEE 2nd Int. Conf. on Data Science and Systems (HPCC/SmartCity/DSS)*, Dec. 2016, pp. 1325–1332.

Improving visibility in the supply chain with digital technologies

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Karine Evrard Samuel
Univ. Grenoble Alpes, CNRS,
CERAG,
F-38000 Grenoble, France

karine.samuel@grenoble-inp.fr

Abstract— It has become a major objective of many firms to increase their supply chain visibility because this leads to better control of flows within an uncertain and complex environment. The rapid development of digital technologies would appear to suggest that their use in industrial contexts will lead to major changes in the design and the management of supply chains. The ability to generate, store and manipulate large quantities of data, together with the possibility of using advanced analysis capabilities allow new ways of increasing supply chain visibility to be envisaged. The aim of this article is to understand how visibility can be improved with digital technologies. The paper provides a complete literature review on supply chain visibility, and highlights basics requirements to develop the ability to identify, capture and share heterogeneous information and events by capturing and analyzing data as assets move through business processes.

Keywords—visibility, supply chain, digital technologies, big data

I. Introduction

Organizations today operate within a dynamic world characterized by four main factors known as VUCA challenges[1]: Volatility, which is expressed by the nature, speed, volume, extent and dynamic of change; Uncertainty, characterized by the lack of predictability of events; Complexity, which means the interconnected nature of products and services; and the Ambiguity of situations and their interpretations in a global context.

In this context, the combination of large quantities of physical flows and varied data, together with rapid, large-scale changes and digital technologies (social media networks, mobile devices, large-scale data analysis and cloud storage) mark a turning point in Supply Chain Management (SCM). Supply Chain (SC) processes can now be improved thanks to the ability to generate, store and handle large quantities of data (*Big Data, Cloud*) as a result of new treatment and analysis capacities (*Analytics*) [1]. These developments will make it possible to have significant impact on the increase of visibility across the entire SC. The concept of visibility, which is often linked to Business Intelligence (BI) [2], is not a recent one, but it assumes greater magnitude when assessed in the light of the digital transformation of SCs [3].

Recent research shows that the lack of visibility in SCs increases organizations' vulnerability and the risks of stock shortages, both upstream and downstream [4]. Furthermore, while customer satisfaction remains a key objective which requires efficient supply chains, consumer habits and customer communication in particular have developed very rapidly over the last fifteen years as a result of the massive development of digital trade (e-commerce, m-commerce). Structural visibility across the entire network has therefore emerged as one of the ways in which these risks can be identified and contained, but it requires a paradigm change, which transcends dyadic approaches. Such approaches no longer reflect the full complexity of global supply chains that are made up of multi-party networks and interconnected by multi-system information architectures. In this context that is evolving rapidly in both organizational and technological terms, the professional literature notes the rise of new challenges related to an absence of visibility over the entire network, that means that companies express difficulties to identify, capture and share milestone events as assets move through a business process. However, until now, the academic literature dealing with the concept of visibility has concentrated more closely on the management of risks in SC, without truly specifying the stages, tools and methods which enable SC visibility to be increased [5]. Some more recent research does emphasize the benefits related to the increase of SC visibility due to an improvement in agility [6], but only a limited number of studies in this area have been conducted so far to highlight the basic requirements.

With this contribution, we seek to structure the knowledge relating to the concept of visibility by considering multi-party and multi-system interdependencies, and by integrating digital technologies that give rise to new applications and make it possible to rethink the ways how SCs are controlled. Based on a literature review on the concept of visibility (section II), we review digital technologies that play a role in transforming a firm's ability to render its entire SC network visible (section III). Finally, we propose a framework establishing basics foundations to analyze visibility issues within supply chains (section IV).

[1] The notion of VUCA was introduced by the U.S. Army War College to describe the more volatile, uncertain, complex and ambiguous multilateral world, which resulted from the end of the Cold War.

II. Defining Supply Chain Visibility

The concept of visibility in SCM is based on the *Bullwhip effect* demonstrated in 1958 by Forrester who explains that the amplification of demand variability in SCs results from a lack of communication between the different links in the chain [7]. The capacity of the entire SC network to react to the variation in demand is part of the Resource Based View (RBV) theoretical framework, and particularly the dynamic capabilities approach. Teece and Pisano (1994) define dynamic capabilities as the characteristics or capacities of an organization, which enable it to integrate, construct and reconfigure competencies in order to achieve greater congruency in a changing environment [8]. From this perspective, a firm's performance is achieved by the combination of two types of organizational capability: dynamic capabilities and operational capabilities [9].

The latter are defined as the means, methods, practices and procedures by which a firm operates in order to survive in the present. Some authors consider visibility as a 'low level' capability that facilitates information sharing between members of the same chain and improves coordination between its members [10]. However, given that visibility makes it possible to detect market developments or/and supplier's failures and to cope with the consequent disruptions, it makes a direct contribution to an organization's resilience capability [11]. Recent technological developments, and particularly the rapid rise of digital technologies, lead us to reconsider this view. Indeed, the concept of visibility has evolved from an approach focused mostly on the benefits of information-sharing [12], to a solution-based logic involving the implementation of actions [13], and then to an attempt to view the SC as a concept based on information system (IS) architectures [14].

The definitions proposed for the concept of visibility rarely achieve a satisfactory consensus because the concept tends to be confused with the notions of connectivity and traceability. Nevertheless, the capacity to identify, capture, and share information and providing the access to it in order to enhance decision-making does meet with approval from the majority of researchers studying this concept [15] [16] [6]. Two levels of analysis are proposed to explain the concept of visibility as a support mechanism for the control of SCs: firstly, the SC design and then, the operations connecting the actors within that SC [17].

SC design involves making strategic long-term decisions for a company, especially when choosing suppliers and markets. At the organizational level, the need for visibility depends on the configuration of the SC analyzed by Caridi *et al.* from two dimensions [18]. The first is the SC's virtuality, i.e. the extent to which the value chain is integrated or externalized (the more a SC is 'disintegrated', the greater the focal firm's need for visibility will be because the company does not then have direct control over its SC). Secondly, the SC's complexity, which is generally related to the number of customer-supplier relationships involved (number of suppliers and third parties, number of distributors and retailors) and to the degree of differentiation between these suppliers in terms of size, technologies, products etc.

To better control SC operations upstream and downstream, the organizational commitment of the SC members is a *sine qua non* condition for visibility within the SC [19] and is characterized by a commitment to physical, human and informational resources [20]. A desire on behalf of the participating organizations to analyze and share information between the various links in the SC is also a key requirement.

From the demand side, collaboration occurs between manufacturers and retailers. True demand management isn't just about sales forecasting; it is also about having a global vision of demand signals from points of sale and distribution centers (information held by retailers) and from warehouses and plants (manufacturers and 3rd party logistics). From the supply side, the ability to catch these demand signals is a key success factor for suppliers' agility. The difficulty, however, is to overcome the breaking point in the upstream information flow due to the weakness of demand signals exchange between retailers, manufacturers and suppliers. Better supply-demand alignment could be obtained by sharing volumes and variability demand, but many works on that topic have shown that the knowledge-sharing stage of collaboration is rarely reached because retailer and manufacturer don't think the same way about sharing data [21]. At the operational level, increasing SC visibility implies access to information about flows and the knots between these flows. Upstream, this information mainly concerns location and volumes of inventories of parts, assemblies and finished products, available capacities, the reception or dispatch of orders and the availability of raw materials. Downstream, visibility focuses on feedback on quality, real rather than expected sales, stocks in the distribution network (available to promise) and real demand at points of sale. The data for this "big picture" is critical to visualize the SC network and reach efficient control and visibility. However, this big picture doesn't exist in companies.

To effectively develop an end-to-end visibility, technologies and processes can be seen as supports of information sharing as they facilitate the data transmission-reception process, i.e. a reciprocal information exchange that is essential to avoid SC disruptions. Such reciprocity enables information to be analyzed in a rapid and, above all, collective manner. In addition, it facilitates the ability to take decisions at the right time.

Several stages have been identified in the process aimed at increasing SC visibility: data collection, data integration, data intelligence and decision-making [17]. Some researchers focus on the properties of information exchanges in terms of accuracy, reliability, utility and the format in which they are used [20] [22]. The quality of data is an essential pre-requisite to the data integration phase and to the whole process in general. Data collection is conducted via multiple mechanisms and communication methods ranging from manual techniques which rely on human intervention to trigger the act of communication (reporting to and from suppliers), to semi-automatic means which require partial human intervention to format the information (collaborative platforms, web portals and interfaces), and finally to automatic means such as Electronic Data Interchange (EDI), or 'Track and Trace' technologies which enable information to be shared in predefined formats with little or no human intervention.

However, the lack of standardized norms renders the automating of information exchanges complex and costly when interface applications have to be developed between the various information systems used by the different SC partners. Tracking packages is not new technology but tracking events as assets move requires new tools and methods to be implemented as digital technologies are transforming industries.

This state of art on SC visibility reveals a consciousness regarding the need for a better anticipation of disruptions thanks to SC visualization systems, but no existing research really allows to characterize such systems, integrating recent technological evolutions. This is the object of the next section.

III. Digital technologies as a means of increasing supply chain visibility

In the recent literature, a large number of terms are used to describe new communications systems that help to reach supply chain goals: industrial internet [23], physical internet [24], smart factory [25], e-supply chain [26]. However, it appears that the informational architectures currently implemented within firms are a barrier to the development of SC vertical and horizontal collaborations, and therefore constitute a brake on their control of supply chains [27] [28]. In fact, IS architectures are mainly aligned with one particular firm's strategy and not with the strategy envisaged by a network of firms. In 2004, Gunasekaran and Ngai examined how ISs could better facilitate the integration of the various SC entities [29]. Their research perspectives already emphasized the need to examine how the IS architecture and the SC strategy and objectives could be aligned. Research on this topic was rare at that time and continues to be so today. A recently published study shows that the ISs currently used to control SC (ERP, APS, TMS and WMS) mainly facilitate transactional exchanges in dyadic relationships, but that this is not sufficient to create and gain better management control by connecting events and hard goods through disparate data sources. Furthermore, the professional SCM press has been discussing for many years now the notion of "control tower" whose purpose is to collect, centralize and control a range of information within one single tool using new data collection and processing technologies [30]. This concept revisits an idea which appeared in Supply Chain Event Management at the start of the 2000s but which failed to be adopted by firms at that time probably in spite of its early innovative nature [31]. The question is therefore whether new digital technologies that have been developed in the last past years will enable supply chain visibility to be increased by combining three factors related to their development:

1) ever-increasing amounts of data which are accessible in real time thanks to the development of Internet of Things [32], and building a network of hardware, software, devices, databases, objects, sensors, and systems [33];

2) analysis and optimization tools which are more accessible and efficient (solutions for storing, calculating and analyzing data and for predictive analysis) [34];

3) constantly progressing calculation capabilities and network developments which enable data from heterogeneous and geolocated sources to be integrated [35].

These factors are accompanying industry in its mutation towards the so-called 'Industry 4.0' and open a path toward a deeper integration of the digital world with the physical world. This lead to rethink the control of supply chains by envisaging a network of interconnected firms with access to extensive information sources, which supports more comprehensive intelligent services. Full interoperability between various interconnected devices will allow the network to be reconfigured when unexpected events occur and which is not only capable of meeting customer demand rapidly, but also of anticipating changes in trends. However, despite the consensus on the great potential of this new way to integrate supply chains through advanced technology, the main questions with regard to this mutation focus on how to realize the system on the current IS basis which relies on management control tools whose 'rear view mirror' effect is not always helpful in developing anticipatory scenarios.

From this perspective, it is necessary to draw the path toward a SC visualization system integrating the latest advances in digital technologies. Our aim in the next section is to highlight basics requirements to develop such a system.

IV. Basic foundations for SC visibility

To deal with VUCA challenges, supply chains must become smarter [36]. The visibility capacity enables a synchronized, closed loop between customer orders, production scheduling and manufacturing execution through intelligent, cost-effective devices and global communication networks. However, even if many firms have identified visibility as a top Supply Chain issue, requirements to make it real have not been adequately addressed. To clarify the visibility needs in order to improve the SC decision-making capability, we propose to work on three key components, which are information management, processes monitoring and technologies enabling visibility.

A. Information management

Data can be considered as "raw facts" of an event. Big data has been defined as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation"[2]. An interconnected SC network can generate more real-time information thanks to machine-generated sensor data but the difficulty is now to produce better information from this big data [3]. The need for good data management has increased as the volume of data has grown over the last few years and the use advanced analytics that draw on the immense data sets created by supply chain activities is becoming a major stake for today's organizations. For example, using predictive analytics allow firms to pull in many data sources to anticipate changes

[2] Gartner IT Glossary, http://gartner.com/fr-glossary/big data/

in the economy or how competitive environment may affect demand for their offerings.

Better manage information in a SC means having the right type of information at the right time, and also improving information quality, even if it comes from various sources. As supply chains have become more interconnected and global, they have also become more vulnerable, with more potential points of failure and less margin of error. Information is held by firms belonging to a large multi-tier supply chain network and risks emanate both from within this network and from outside (macro-environment risks). Regarding the extended value chain, the supply chain visibility lies in the most granular view of the process, that is the physical location, status, and condition of the individual items themselves and not the truck, ship, plane or train they are supposed to be on. For that reason, the first requirement that underpin a SC visibility framework is to identify **what** information should be visible and shared by SC partners regarding the occurrence of some events.

The table below gathers information needs from the demand side to the supply side, including 3rd party services. Information is structured into three main sources of identified risks in the extended value chain that are the level of inventories, lead times and quality issues (see table 1).

Table 1. Informational framework for end-to-end visibility

	Supply side **Tier N Tier 1 ←**	**3rd Party Services**	**Demand side** **→ Distributors End Users**
Inventories	Raw materials, parts and assemblies availability Suppliers' location Marketing plans	Replenishment status of inventories Inventory levels	Available To Promise in DCs Orders and sales (real-time consumption) Promotions
Lead times	Manufacturing capacities and production scheduling	Delays and shipping time	Availability of products at customer's required date
Quality	Part/assemblies quality management Supplier compliance Quality control of finished goods	Tracking data	Traceability Safety issues

B. Process monitoring

Understanding the processes and problems specific to SCM in the current business environment is the second step that will help to plug information flows to strategic stages of product development. The process mapping will help to answer to the questions **how and when** to access to information and use it. The IS architectures which underpin activity within the SC may be an impediment in that they do not allow for the collection and analysis of data beyond the dyadic relationship. Downstream, *Advanced Planning Systems* (APS) which are sophisticated planning systems coupled with ERP software make it possible to draw up demand forecasts by combining sales history with planned operations such as promotions and product launches [37]. However, these systems are unable to give visibility when demand is highly volatile, and their planning function does not therefore permit a rapid reconfiguration of the network when the SC design has to be adapted due to rapid changes. These systems have to be coupled with what-if analysis and scenarios that simulate several market trends in order to give the possibility to the suppliers to rapidly change their manufacturing plans.

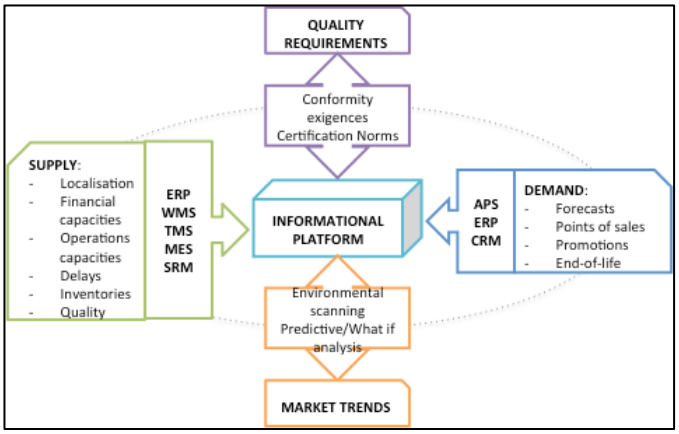
Upstream, *Manufacturing Resource Planning* (MRPII), integrated with ERP softwares, which calculates needs for materials and components does not enable information to be provided on capabilities or inventories held by suppliers because the system does not report on production operations in real time. None of the currently utilized IS make it possible to meet the need to collect, aggregate, store, process, analyze or compile data in real time so that decisions can be taken rapidly across an entire SC. The enabling contribution of digital technologies can counteract some of these limitations, but several bars still need to be lifted with regard to the integration of such technologies with current SC management systems. RFID technology can provide better visibility by improving manufacturing efficiency and effectiveness [38]. The major effects of using RFID are automational effects on operational processes followed by informational effects on managerial processes [39]. Other technologies such as wireless networking technology also allow extensive interconnection but their adoption by companies takes time and these systems are up-to-now unequally deployed and not completely interoperable [40]. Achieving interoperability via standards, protocols and connectivity remains a major challenge [41]. The web-enabled technologies are a good alternative for data standardization. They provide more flexible and scalable software systems that can standardize data, make it accessible, and help effectively trans that aggregated data into performance-improving information.

Sharing information raises the issue of confidentiality and security of these exchanges [42]. Indeed, no visibility in SC is possible if large-scale data is not coded and stored in a standardized format, which is secure and exchangeable between all partners. It is therefore essential to undertake an examination of data property and handling rights between the SC members, and to reflect on the redistribution of this data with the aim of increasing the warning capability and enhancing decision-making abilities.

The strategies for reconfiguring a network are necessarily based on reliable systems, which make it possible to interpret the weak warning signs that precede a potential risk. The technological data approach (Big Data, predictive analysis) must not overshadow the organizational approach that enables the complexity of SC relationships. It is the combination of these two approaches that enables visibility to be increased [18].

To take into consideration these elements, we propose to structure an informational platform that organizes the information feedback from both supply and demand processes, integrating quality requirements and market trends which can have an impact on any portion of the supply chain, or across the entire supply chain (see figure 1).

Fig. 1. Framework for an IT platform for SC visibility



C. Technologies enabling SC visibility

At least, there is a third question related to **which** technologies will enable to develop SC visualization systems.

First, the digitization of manufacturing thanks to the Industrial Internet of Things (IIoT) offers the possibility to capture real-time shared data across robotics and smart machines, sensors, systems, from the whole SC network (customers, suppliers and 3rd party services). This manufacturing data sources can be transformed into libraries of common dashboards shared by SC partners.

Second, cloud-based technology allows collecting, aggregating and storing data collected through multiple sources (e.g. IoT devices). Based on data lakes that are pools of structured (relational databases), semi-structured (CSV, XML, logs) and unstructured data (images, audio, video files), the main issue attached to this technology is the lack of data governance as there is no associated metadata, increasing the risk of security failures (data integrity).

At least, social media technologies offer several tools, technologies and methods of data mining to improve the commercial performance of the company allowing a greater customization of products and services [43]. The integration of the customer in the product design is a recent market trend that will lead to major transformations of SC in the future. In particular, it offers the opportunity to redefine supply chain processes from the outside-in and define more customer-centric supply chains, integrating product design at early stages.

As extensive interconnection is now technologically possible, the next step will be to combine data capturing tools and storing solutions with predictive analysis methods.

V. Conclusion

Visibility concerns top the project lists of supply chain managers across the globe as limited visibility means limited control of the SC. As manufacturing becomes increasingly digitized and data-driven, there is a growing hunger across the ecosystem for real-time information that increase the need for a visualization system of record for real-time communication, collaboration, and continuous improvement. The path to digitally enabled supply chains is still long to gain the full benefits, but supply chain visibility brings the opportunity to bring complete optimization to many aspects of business operations.

This paper presents basic requirements that will help build a framework for end-to-end supply chain visibility. We answer to several strategic questions covering **what** information is needed, **how and when** it has to be extracted from SC processes, and **which** technologies allow a digital transformation able to create value from better visibility in the entire supply chain network. These new ways to approach information and processes open the path to new opportunities for empirical research on this topic.

This paper suffers from several limits. First, our work needs to be deepened by observing practices and develop several business cases. Second, the current level of adoption of digital technologies by companies can be considered as a restraint to SC visibility. Several business cases reveal that companies are not in phase with the recent evolution of digital technologies. In particular, the capacity to analyze "soft data" such as the output of social media or other ways to capture information related to customers is still a major issue to improve the demand side. From the supplier side, the ability to catch weak signals and integrate them into decision-making systems also needs to be improved and offers many paths for further research.

References

[1] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big Data: The Next Frontier for Innovation*. McKinsey Global Institute: Competition and Productivity.

[2] Seufert, A., Schiefer, J. (2005). Enhanced business intelligence-supporting business processes with real-time business analytics. In *Database and Expert Systems Applications. Proceedings. Sixteenth International Workshop on* IEEE, 919-925.

[3] Waller, M.A., Fawcett, S.E. (2013). Data science, predictive analytics, and Big Data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77–84.

[4] Basole R.C., Bellamy, M.A., (2014). Supply Network Structure, Visibility, and Risk Diffusion: A Computational Approach. *Decision Sciences*, 45(4), 753-789.

[5] Jüttner, U., Peck, H., Christopher, M. (2003). Supply Chain Risk Management: Outlining an Agenda for Future Research, *International Journal of Logistics: Research & Applications*, 6(4), 197-210.

[6] Caridi, M., Moretto, A., Perego, A., Tumino, A. (2014). The benefits of supply chain visibility: A value assessment model. *International Journal of Production Economics*, 151, 1-19.

[7] Lee, H.L., Padmanabhan, V., Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science,* 43(4), 546-558.

[8] Teece, D.J., Pisano, G. (1994). The dynamic capabilities of firms : an introduction. *Industrial and Corporate Change,* 3(3), 537–556.

[9] Wu, L.Y. (2010). Applicability of the resource-based and dynamic capability views under environmental volatility. *Journal of Business Research* 63(1), 27–31.

[10] Brusset, X. (2016). Does supply chain visibility enhance agility? *International Journal of Production Economics*, 171, 46-59.

[11] Ngai, E., Chau, D., Chan T.L.A. (2011). Information technology, operational, and management competencies for supply chain agility: findings from case studies. *Journal of Strategic Information Systems* 20, 232–249.

[12] Yu, Z., Yan, H., Cheng, T.C.E. (2001). Benefits of information sharing with supply chain partners. *Industrial Management and Data Systems*, 101(3), 114–119.

[13] Choi, T.M., Sethi, S. (2010). Innovative quick response programs: a review. *International Journal of Production Economics*, 127(1), 1–12.

[14] Jeyaraj, A., Sethi, V. (2013). Implementation of Information Systems Infrastructures for Supply Chain Visibility in Organizations, *Journal of Information Science Technology*, 9(1), 3-23.

[15] Goh, M., De Souza, R., Zhang, A., He, W., Tan, P. (2009). Supply chain visibility: A decision making perspective, In: *Proceedings of the 4th IEEE Conference on Industrial Electronics and Applications*, Xian, China.

[16] Swaminathan, JM. Tayur, SM. (2003). Tactical Planning Models for Supply Chain Management. *Handbooks in Operations Research and Management Science*, 11, 423–454.

[17] Saint McIntire, J. (2014). *Supply Chain Visibility: From theory to practice*, Gower.

[18] Caridi, M., Crippa, L., Perego, A., Sianesi, A., Tumino, A. (2010). Do virtuality and complexity affect supply chain visibility? *International Journal of Production Economics*, 127, 372-383.

[19] Fawcett, S.E., Osterhaus, P., Magnan, G., Brau, Carter, M. (2007). Information sharing and supply chain performance: the role of connectivity and willingness. *Supply Chain Management: An International Journal*, 12(5), 358-368.

[20] Barratt, M., Oke, A. (2007). Antecedents of supply chain visibility in retail supply chains: a resource-based theory perspective. *Journal of Operation Management*, 25(6), 1217–1233.

[21] Evrard Samuel, K. (2010). Collaborative Demand Planning: Lessons from three Case Studies. In Wang L. and Koh L. (eds.), *Enterprise Networks and Logistics for Agile Manufacturing*, Springer-Verlag, London, 119-136.

[22] Zhou, H., Benton Jr., W. (2007). Supply chain practice and information sharing. *Journal of Operation Management*, 25(6), 1348-1365.

[23] Evans, P.C. Annunziata, M. (2012), *Industrial internet: pushing the boundaries of minds and machines*, available at: www.ge.com/docs/chapters/Industrial_Internet.pdf

[24] Montreuil, B. (2011). Toward a physical Internet: meeting the global logistics sustainability grand challenge, *Logistics Research*, 3, 71-87.

[25] Hessman, T. (2013). The dawn of the smart factory, *Industry Week*, February, 15-19.

[26] Akyuz, G.A. Rehan, M. (2009). Requirements for forming an 'e-supply chain', *International Journal of Production Research*, 47(12), 3265-3287.

[27] Shah, R., Goldstein, S.M., Ward, P.T. (2002). Aligning supply chain management characteristics and interorganizational information system types: an exploratory study. *Engineering Management, IEEE Transactions on*, *49*(3), 282-292.

[28] Fawcett, S. E., Magnan, G.M., McCarter, M.W. (2008). Benefits, barriers, and bridges to effective supply chain management. *Supply Chain Management: An International Journal*, *13*(1), 35-48.

[29] Gunasekaran, A., Ngai, E. W. (2004). Information systems in supply chain integration and management. *European Journal of Operational Research*, 159(2), 269-295.

[30] Pittman, A. (2015). What's new in Supply Chain Control Tower? *Supply Chain 24/7*, 29 octobre. http://www.supplychain247.com/article/whats_new_in_supply_chain_control_towers

[31] Otto, A. (2003). Supply Chain Event Management: Three Perspectives, *The International Journal of Logistics Management*, 14(2), 1-13.

[32] Ma, H.D. (2011). Internet of Things: objectives and scientific challenges, *Journal of Computer Science and Technology*, 26(6), 919-924.

[33] Marshall, T. (2012). Here's the thing about the internet of things, *Backbone Magazine*, February- March, 12-13.

[34] Khan, K. (2013). The transformative power of advanced analytics, *Supply Chain Management Review*, May/June, 48-49.

[35] Fu, Q., Zhu, K. (2010). Endogenous information acquisition in supply chain management, *European Journal of Operational Research*, 201, 454-462.

[36] Butner, K. (2010). The smarter supply chain of the future, *Strategy and Leadership*, 38(1), 22-31.

[37] Stadtler, H. (2005). Supply chain management and advanced planning-basics, overview and challenges. *European journal of operational research*, 163(3), 575-588.

[38] Zelbst, P.J., Green, K.W., Sower, V.E. Reyes, P.M. (2012). Impact of RFID on manufacturing effectiveness and efficiency, *International Journal of Operations & Production Management*, 32(3), 329-350.

[39] Visich, J.K., Li, S., Khumawala, B.M. and Reyes, P.M. (2009). Empirical evidence of RFID impacts on supply chain performance, *International Journal of Operations & Production Management*, 29(12), 1290-1315.

[40] Wang, L., Alexander, C.A. (2015). Big Data Driven Supply Chain Management and Business Administration. *American Journal of Economics and Business Administration*, *7*(2), 60-67.

[41] Kalfoglou, Y., Schorlemmer, M. (2004). Formal support for representing and automating semantic interoperability. In *The semantic web: Research and applications* (pp. 45-60). Springer Berlin Heidelberg.

[42] Li, L., Zhang, H. (2008). Confidentiality and information sharing in supply chain coordination. *Management science* 54 (8): 1467-1481.

[43] Wang, G., Gunasekaran, A., Ngai E.W.T., Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: certain investigations for research and application, *International Journal of Production Economics*, 176, 98-110.

Inventory Management

Analysis of End-of-Quarter Inventory in Hybrid Production Environments Using Simulation

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Lawrence Al-Fandi
Industrial Engineering Department
American University of the Middle East
Egaila, Kuwait
lawrence.alfandi@aum.edu.kw

Faisal Aqlan
Industrial Engineering Department
Penn State University, The Behrend College
Erie, PA, USA
fua11@psu.edu

Abstract— In this research, simulation and optimization techniques are used to study end-of-quarter (EoQ) inventory in hybrid manufacturing environments in order to minimize total inventory cost and improve service level. A fabrication/fulfillment process from a high-end server manufacturing environment is considered. The system is characterized by fast lead time and lengthy build processes. Main components of servers are tested ahead of time and stored in inventory as tested parts to be ready for prompt fulfillment of customer order. A critical aspect to complete an order is availability of parts in inventory. On the other hand, excess fabbed inventory will result in high holding cost. We propose a simulation-based framework to analyze EoQ inventory and effectively meet customer demand without carrying excessive inventory. The risk of revenue loss resulting from fabbing less than 100% of planned volume is investigated considering demand skew, fulfillment cycle time, availability of supply, gross physical inventory, and build plans versus real orders. The results of the case study show that fabbing less 100% of fabbed plan could result in missing valuable orders. As a result, since missing shipment is not an option for the company, other actions should be taken to minimize inventory and improve service levels.

Keywords— end-of-quarter inventory; fabrication test; order fulfillment; simulation modeling; high-end server manufacturing

I. Introduction

A production system is defined as a set of resources and procedures that interact to convert raw materials into finished products and deliver them to customers [1]. Production systems are modeled analytically and/or using computer simulations. Analytical models describe a system using mathematical or symbolic relationships. "Simulation models represent the events that could occur as a system operates by a sequence of steps in a computer program" [2]. Despite the fact that analytical models provide more accurate results, they are difficult to develop especially when the complexity of the system increases. On the other hand, simulation models can be used effectively to analyze complex production systems and study the impact of different factors (or scenarios) considering randomness and variability.

When the complexity of the system and inherent uncertainty increase, it is recommended that simulation models are used [3]. Many of today's manufacturing systems are highly complex. The manufacturers are facing intensifying competition and steady pressure for higher levels of customer service. Furthermore, the dynamic changes in market demands and companies' strategies require flexible introduction of new products and implementation of continuous improvements in their production and inventory management. Hybrid production control philosophy combines build-to-plan with make-to-order operations, which is referred to as configure-to-order (CTO) operations [4]. The main two benefits provided by CTO manufacturing environments are mass customization and quick response time to customer orders. CTO systems are used when the time it takes to build the customer order is negligible as compared to a long leadtime for producing the components. Examples of typical CTO environments include personal computers manufacturing and high-end server industry.

In high-end server manufacturing environments, the two components of the CTO systems are known as fabrication and fulfillment processes. The fabrication stage is a build-to-plan process, where components are procured, tested and assembled. The tested inventory is then kept in the final assembly until it is pulled by the fulfillment stage which is a make-to-order process.

In this research, we will use simulation modeling to study the End-of-Quarter (EoQ) inventory in a high-end server manufacturing environment. Simulation is used for studying different scenarios to minimize the inventory holding cost while achieving the required service levels. The company wants to efficiently meet customer demand without carrying excessive end of quarter inventory by assessing the risk on revenue if not building 100% of the planned volume. The factors that should be considered include: demand skewness, variability in fulfillment cycle times, supply availability, gross physical inventory, and build plan versus actual orders.

The rest of this paper is organized as follows: Section 2 provides a summary of the literature related to simulation modeling for inventory and production management. Section 3 presents a description of the system under study. In Section 4, the proposed framework is discussed which consists of the conceptual model and the simulation model. Section 4 also presents the results and analysis. And finally, Section 5 presents the conclusions and recommendations of this study.

II. Related Literature

Simulation modeling has been widely used to study manufacturing systems. Simulation techniques, especially discrete-event simulation (DES), can effectively be used to capture and analyze the complex behavior and interactions of complex systems with less effort when compared to analytical models. It can be effectively used to make accurate and sound decisions to identify the best alternative among several candidates [4].

Several researchers have used DES for different applications in manufacturing systems such as capacity analysis [5], production line consolidation [4], study of design changes [6], and evaluation of design alternatives [7].

According to [8], computer manufacturing industry is facing many challenges in its production and distribution processes due to the rapid advances in information technology. The use on simulation models for inventory management in server manufacturing was discussed in [9]. According to the authors, the highly complicated server manufacturing environment poses simulation modeling challenges such as two-stage fabrication/fulfillment process, complex server configurations, part tests with random yields, stochastic lead times and so on. The study developed a simulation-based tool that provides decision support at an operational level. The tool enables inventory planner to visualize projected inventory level over time. The use of simulation to study the consolidation of production lines in a high-end server manufacturing environment was discussed in [4]. Simulation modeling was combined with Lean manufacturing techniques for green transformation in high-end server manufacturing [10]. Due to high level of uncertainty that is associated with server manufacturing processes, process modeling and simulation are used to study and analyze variability of water and energy consumptions. A simulation study was conducted to determine the batch size for a fabrication test process in a high-end server manufacturing environment [11]. In another study [12], a simulation study was conducted to determine the release times for inventory in a pull system for high-end server industry. The objective of the study was to minimize the inventory quantities and waiting time in the production side stock areas while fulfilling customers' orders on time. The use of simulation to determine an optimal order allocation policy that involves multiple countries in a server manufacturing environment was discussed in [13]. The optimal order allocation policy takes various inventory replenishment strategies into account while minimizing the total costs. Even though many researches focused on the use of simulation to study high-end server manufacturing envirnment, non of theses studyies focused on the EoQ inventory and its impact on service level. In this study, we preset a simulation based framework to study the EoQ inventory in high-edn server manufacturing. The goal is to help decicion makers develop effective fabbing plans while minimizing inventory holding costs.

III. System Description

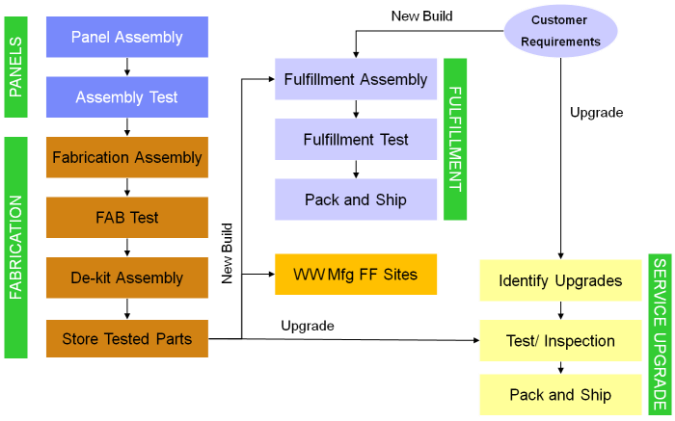
The system discussed in this study represents a high-end server manufacturing hybrid environment. The high-end server manufacturing facility produces high-end servers that are built with leading edge technology. The high-end server manufacturing environment is characterized by aggressive introduction cycles of new products (i.e., every two years), extreme demand skews, significant engineering changes, and

high inventory holding cost. An overview of the high-end server manufacturing system architecture is illustrated in fig. 1. The manufacturing environment is based on configure-to-order processes which is a combination of build-to-plan and make-to-order processes. This is also known as a fabrication/fulfillment strategy to respond to customer orders rapidly and minimize inventory holding costs.

- Fabrication: Components or subassemblies are produced, tested, and assembled based on a projected production plan and are kept in stock until an actual order is received from a customer.
- Fulfillment: Final products are assembled according to actual customer orders, such that no finished good inventory is kept.

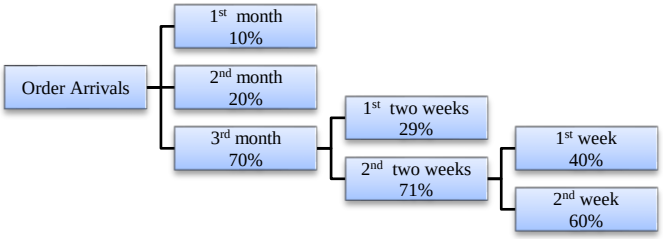
According to [4], the fabrication/fulfillment model provides the company with the flexibility of mass customization and the speed and efficiency of mass production. However, the randomness (i.e., random yields, system configuration, stochastic lead times, etc.) inherent to this model makes the inventory management and production planning a challenging problem, considering high inventory holding and missing opportunity costs. To deal with the uncertainty in customer demand, some of the key strategies that are implemented are as follows: 1) inventory sharing between different plants when there is a shortage of parts and components; 2) localized warehouse for suppliers at manufacturing sites; 3) flexible production planning for internal orders; 4) order fulfillment dashboard; and 5) information technology.

Fig. 1. An overview of server manufacturing environment structure [14]



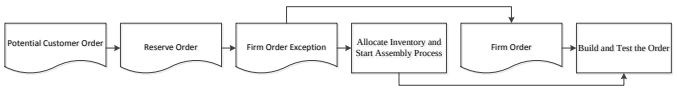
The high-end server manufacturing environment is characterized by skewed demand pattern where most of the customer orders arrive during the last month of the quarter (see fig. 2). This creates unbalanced use of resources and high inventory carrying costs.

Fig. 2. Order arrival pattern for high-end server manufacturing



Customer order starts a potential order and then a reserve order. The order is then moved to the next step, firm order exception, and then firm order. The order then gets processed and shipped to customer. If the order is cancelled by the customer, it is considered a cancelled order. If it cannot be shipped on time and customer is willing to receive it late, it is considered a backorder. Otherwise, it is considered a missed order. Fig. 3 shows the different stages of customer order in a high-end server manufacturing environment. Most of the orders get confirmed mainly in the last 2 weeks of the quarter. Furthermore, 85% of the total received orders get confirmed.

Fig. 3. Different stages of customer order in a high-end server manufacturing



The capacity/inventory tradeoff analysis in high-end server environments is important due the following reasons: 1) limited and expensive capacity which makes any savings in capacity of significant values, 2) high inventory carrying costs and expensive components, 3) fast lead time and lengthy build processes, 4) missing an order can cause the company to lose high expected income. For planning purposes, the quarter is divided into five cycles. The company needs to inform their suppliers by the middle of the quarter (cycle 3) the final plan to fulfill their orders. The production plan is based on forecasted sales for entire quarter and the capacity and staffing are determined based on forecasted quantity. To avoid the risk of missing customer shipment, the company runs 100% fab plan.

In the current system, a critical aspect to complete an order is the availability of fabricated (fabbed) commodities. Furthermore, excess Fab inventory will result in high holding cost. On the other hand, inventory shortage can lead to missing customer shipment will which cause the company to lose high expected income.

IV. Proposed Framework

A. Conceptual Model

Simulation modeling of production systems is a useful technique to study and improve systems performance. In this study, Discrete Event Simulation (DES) is used to study the tradeoff between capacity and inventory in a high-end server manufacturing environment. The framework of the simulation study (shown in fig. 4) consists of four main steps: data collection and input modeling, model building, model verification and validation, and what-if analysis. In the first step, historical data were collected and fitted into proper statistical distributions. An extensive data mining process was conducted to obtain the relevant historical data. The data was collected from multiple databases in addition to manual data collection though time studies. Once the simulation model was built, it was validated by comparing the simulation results to results from the actual systems. The last step was to perform what-if analysis to investigate the tradeoff between capacity and inventory and optimize the EoQ inventory and service levels. The What-If analysis algorithm is shown in fig. 5.

Fig. 4. Proposed framework for the simulation study

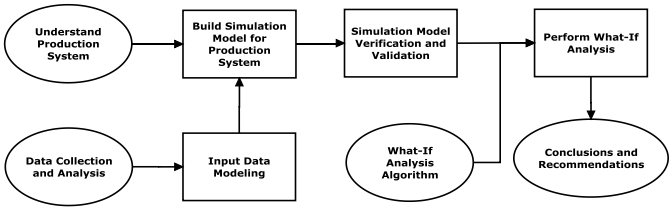
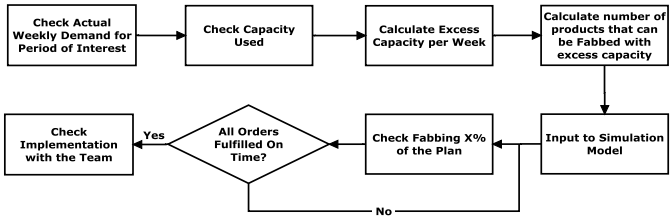


Fig. 5. What-If analysis algorithm



B. Simulation Model

Input data for the simulation model was collected for one quarter (or three months). The arrival patterns of customer orders and main components were modeled using the appropriate statistical distributions. The processing times and failure rate data were collected from the company's data base and through time studies. The assumptions are summarized in Table I.

TABLE I. Simulation model assumptions

Main Process	Assumptions
Visual Inspection I	○ 20% additional process time, 5% scrap
Fulfillment Test	○ 1.2 additional hours for regression testing ○ 85% of failures can be fixed in the process ○ 15% of failures are returned to First Line (fabrication and rework times are added)
Visual Inspection II	○ 1.2 additional hours for regression testing ○ 95% of failures can be fixed in the process ○ 5% of failures are returned to First Line (fabrication and rework times are added)
Shipping	○ Lead time to ship order to customer is 5 days

To study the End-of-Quarter inventory, the fabrication and fulfillment processes in the high-end server manufacturing environment shown in fig. 1 were considered. A detailed flow chart for the two processes is shown in fig. 6. The figure represents the conceptual model used to develop the simulation.

The simulation model was developed using Rockwell Arena® Software (v14.0). The number of replications used is 40

and the replication length was set to 90 days or one quarter. The base simulation model was verified and validated to ensure that it is working as intended and accurately represents the real system. For model verification, the flow of entities was monitored to ensure that the model is free from logical errors. The model was tuned several times in order to get the expected behavior. The system throughput, expressed as the number of products, from both simulation model and real system was statistically compared (see Table II). Throughputs of Product 1 and Product 2 orders reached shipping dock were used to validate the baseline model (see Table II). Simulation model was statistically indifferent from historical data, with a level of significance of 0.05.

Fig. 6. Detailed fabrication-fulfillment process in high-end server manufacturing

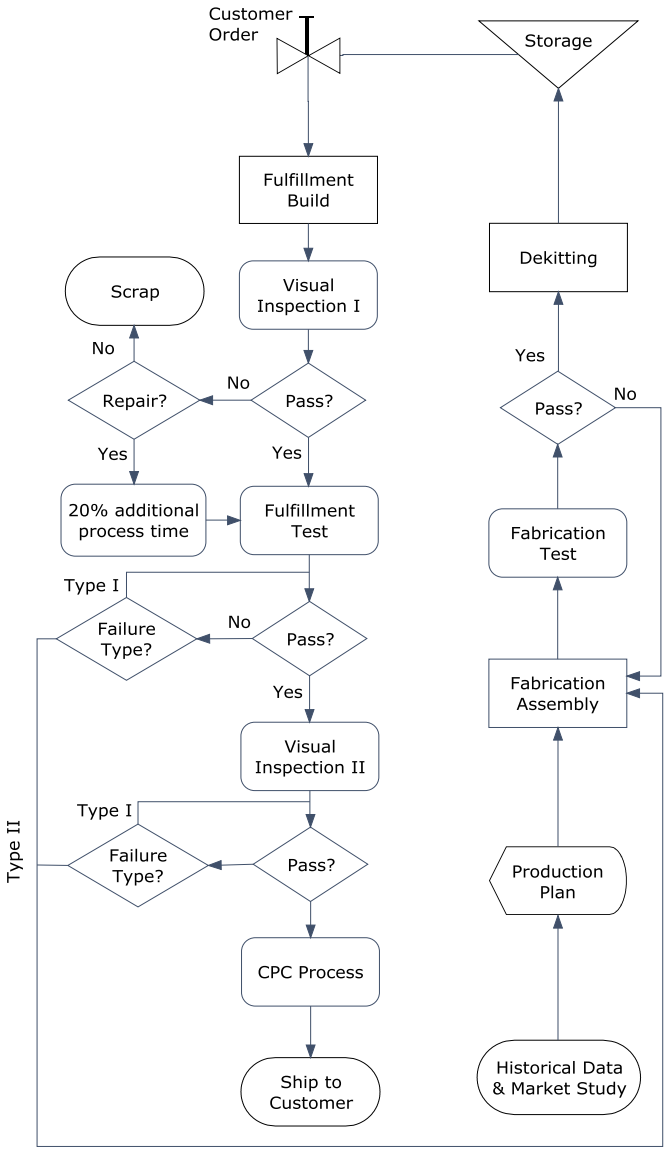


TABLE II. Validating simulation model

Shipping Dock (orders)	Simulation	Actual	% Difference	One sample t-test P-value
Throughput Product 1	127.05	125	1.64	0.19
Throughput Product 2	145.75	142	2.64	0.36

Simulation results for the product cycle times are shown in TABLE III. The scenario analysis considers potential outcomes if the company fabbed less than 100% of the plan. This will also answer the following questions:

1. Do we have buffer capacity to catch up at EOQ?
2. Not having enough capacity presents a high risk of missing shipment, how much is the risk?

The capacity is calculated in units of testing hours. The simulated average fill rate is =1- (1.4/272.8) = 99.487% (versus 99.5% actual fill rate). Statistical analysis was conducted to investigate the difference in fill rate between fabbing 100%, 99%, and 98% . In TABLE IV, if we fab less than 99% of fab plan, a significant risk of reduction in fill rate with 1.06% which means missing 2.83 shipments. Similarly, fabing 98% of the plan will reduce fill rate to 97.37 which means missing revenues of 5.55 extra orders on average compared to baseline scenario. It should be noted that the inventory holding cost was not considered in this study because it is much lowers than the selling price of the srver. In TABLE V, the two fabbing plans, 99% and 98%, are compared. The results show that the two fab plans are statustically different.

Based on the simulation results of this study, it was recommended that the company should fab 100% to avoid missing customer orders.

TABLE III. Simulation results for Product cycle times

Product	Average Days	Half Width
Product 1 Cycle time	10.03	1.91
Product 2 Cycle Time	9.05	2.21

TABLE IV. Comparison of base and alternative scenarios

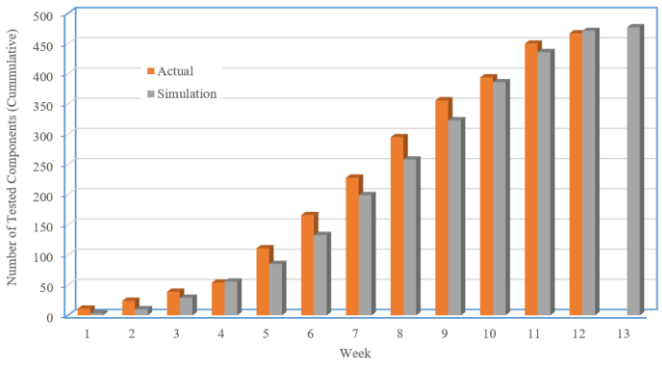
Base Scenario	Alternative Scenario	Sim Runs	Est. Mean Difference	Standard Deviation	95% C.I.
Fabbing 100%	Fabbing 99%	40	1.06	0.842	(0.788,1.33)
	Fabbing 98%	40	2.1	1.23	(1.71,2.49)

TABLE V. Comparison of 98% and 99% fab plans

Scenario 1	Scenario 2	Sim. Runs	Est. Mean Difference	Standard Deviation	95% C.I.
Fabbing 99%	Fabbing 98%	40	1.04	1.15	(0.677,1.41)

The developed simulation model can also be used to predict the number of components that should be tested in a gived period of the quarter. Fig. 7 shows an example of the number of fabbed components for product 1 (both simulation and ctual data). The figure shows the commulative fabbed components for an entire quarter.

Fig. 7. Number of fabbed components for product 1



I. Conclusions

In this paper, simulation was used to study the end-of-qurter inventory in a highprid push-pull production environment for high-end servers. The trade-off between fab plan versus fill rate was investigated. The effect of fabbing less than 100% was studied while assuming arrival of orders that require 100% of the forecasted fab plan. The fill rate was used as a measure of orders fulfilled within leadtime. The results showed that not fabbing 100% of the fabbed plan could results in missing valuable orders. Moreover, since missing shipment is not an option for the company, it is not recommended to fab less 100% of the fab plan. Future work will focus on studying the impact of order cancellations and alterations on inventory. In addition, new product introduction cycles and integrating manufacturing and remanufacturing inventories will be investigated.

References

[1] R. Askin, and J. Goldberg, Design and analysis of lean production systems. Wiley, London, 2002.

[2] G. Salvendy, Handbook of industrial engineering: Technology and operations management. 3rd edition. John Wiley & Sons, NY, 2001.

[3] T. Al-Hawari, F. Aqlan, and O. Al-Araidah, "Performance analysis of an automatic production system with queue length dependent service rate", International Journal of Simulation Modeling, vol. 9, no., pp. 184-194, 2010.

[4] F. Aqlan, S.S. Lam, and S. Ramakrishnan, "An integrated simulation-optimization study for consolidating production lines in a configure-to-order environment", International Journal of Production Economics, vol. 148, pp. 51-61, 2014.

[5] N. Gujarathi, R. Ogale, and T. Gupta, "Production capacity analysis of a shock absorber assembly line using simulation", Proceedings of Winter Simulation Conference, Washington, DC, 1213-1217, 2004.

[6] X. Zhiwei, and L. Yongxian, "Mechanical production line simulation and optimization analysis", Proceedings of IEEE International Conference on Automation and Logistics, Qingado, China, 2925-2930, 2008.

[7] S. Owens, and R. Levary, R., "Evaluating design alternatives of an extruded food production line using simulation", Simulation 78(10), 626-632, 2002.

[8] G. Lin, R. Breitwieser, F. Cheng, J.T., Eagen, and M., Ettl, "Product hardware complexity and its impact on inventory and customer on-time delivery", The International Journal of Flexible Manufacturing Systems, vol. 12, pp. 145-163, 2000.

[9] H. Cao, F. Cheng, H. Xi, M. Ettl, S. Buckley, and C. Rodriguez, "A simulation-based tool for inventory analysis in a server computer manufacturing environment", Proceedings of the 2003 Winter Simulation Conference.

[10] L. Al-Fandi, "A novel approach using lean and simulation modeling for effective green transformation for high-end server manufacturing", PhD Dissertation, Binghamton University, 2011.

[11] L. Al-Fandi and F. Aqlan, "Using simulation to determine the batch size for I/O drawer test process in a high-end server manufacturing environment", Proceedings of the 24th International Conference on Flexible Automation & Intelligent Manufacturing, 2014.

[12] L. Al-Fandi, S.S. Lam, and S. Ramakrishnan, "A simulation approach to determine inventory release times for a pull system", Proceedings of the Industrial Engineering Research Conference, 2010.

[13] L. Al-Fandi, B. Eck, S.S. Lam, P.F. Tsai, and K. Srihari, "Evaluating the impact on overall service level throug order allocation and inventory replenishment policies", Proceedings of the Annual National Conference of the American Society for Engineering Management, 2009.

[14] F. Aqlan, and S.S. Lam, "Supply chain optimization under risk and uncertainty: A case study for high-end server manufacturing", Computers and Industrial Engineering, vol. 93, pp. 78-87, 2016.

EOQ inventory model for disassembly systems

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Matthieu Godichaud, Lionel Amodeo,
Institut Charles Delaunay, LOSI, Université de Technologie de Troyes, UMR 6281, CNRS,
Troyes, France
{matthieu.godichaud, lionel.amodeo}@utt.fr

Abstract—Inventory policies for disassembly systems consist in determining reorder time interval and/or order quantity when deterministic, continuous and constant demand are considered with fixed ordering cost, variable operation cost and inventory cost. Disassembly systems brings new issue in inventory management due to unnecessary inventory that can be generated at each order. We developed in this context two models and their solutions with disposal option for inventory surplus. The first one without lost sales leads to simple form reorder decision. The second one takes into account the possibility to allow lost sales and we propose a solution procedure to find reorder time interval and order quantity.

Keywords—disassembly systems; inventory; lot sizing; economic order quantity.

I. Introduction

Disassembly planning problems in reverse logistic consist in determining the quantities and the timings of the disassembly of end-of-life products to generate components for which there are recycling demands or environmental requirements. The disassembly is a central activity in a reverse supply chain connecting the collection of end-of-life products at customer locations to the recycling of components with residual value (for reusing, remanufacturing, material recycling or energy recovering) or the processing and the conditioning of valueless components and material fractions. Before the disassembly of a product, each of its component or material fraction has to be allocated to a recovery channels, which generate demands and revenue, or allocated to others channels which reduce environmental impacts. The economic gain can be low in disassembly operations and a right planning method can save logistic costs and improve product recovery rate. One of the characteristic of the problems is that the disassembly of one unit of product generate one or more units of each component. It implies that the quantity of product disassembled is not necessarily exactly the number of demand for each component and unnecessary stock can be generated at each disassembly operation. The planning method have hence to be adapted to handle this characteristic based on the context (type of costs, demands, product recovery process and quality of the product).

We consider economic order quantity (EOQ) model which applies to deterministic, continuous and constant demand over a planning horizon with fixed cost each time an order is launched, variable cost which is function of the quantity of product ordered and inventory cost which is function of the number of stocked products. It can apply to disassembly systems when there are demands and inventory costs for the component and the number of product that can be collected is unlimited. However, as mentioned previously, disassembly operations generate unnecessary inventories when the demand are not the same for all components. Additional decisions have to be integrated in the model to handle the inventory surplus.

Two EOQ models for disassembly are proposed in this paper considering two options to handle inventory surplus. The first one consists in disposing the inventory surplus as soon as the order is received with a variable cost proportional to the quantity disposed. The second one consists in allowing lost sales and disposing the remaining inventory surplus. As far as the authors knows, disassembly EOQ problem is not treated in literature. A literature review on other disassembly planning problems is presented in the second section. The model with disposal is developed in the third section and the model with lost sales is developed in the fourth section. A numerical illustration is presented in the fifth section to validate the solution procedure.

II. Related works

Disassembly systems have been studied in various research areas and we focus on those related to disassembly planning. The basic approach to schedule disassembly operations on recovered products to served demands for their component is proposed in [1]. It applies to deterministic demands which vary over several periods of a finite horizon planning. The problem can be extended by considering various types of costs in particular fixed disassembly operation costs and inventory costs as in lot sizing problem [2]. Random parameters for the demands [3][4] or the disassembly yields [5] are inherent to disassembly systems due to variable state of the recovered end-of-life product and have an impact on planning decisions. The lot sizing problem can be more difficult to solve when we considered various disassembly structure with multiple level [6][7] or components commonalities [8]. The issue of unnecessary inventory of component has been treated in this context by accepting its accumulation throughout the planning horizon. It can be reduced by allowing lost sales or external procurements as in [9][8] but no decision is made to handle the surplus. As far we know, the EOQ problem in disassembly system has not been tackled in disassembly related literature and all the model encounter do not take into account the disposal decision of inventory surplus.

The EOQ basic problem and its extension have been widely analyzed in literature and extended review of their features can be found in [10][11]. A policy indicates the frequency of the order and the quantity ordered for each inventory. In assembly systems there is one demand for the final product and one order for each element of the bill of material to replenish its inventory.

By defining appropriate echelon inventory for each stage it can be shown that stationary and nested policies with zero inventory before an order is placed are optimal [12]. This class of policies is easy to implement and uses the result of the basic single item EOQ models [13]. Disassembly system are different from assembly system since there are several demands and all inventory are replenished simultaneously with one order. In distribution system there are as in disassembly systems several inventory supply by one upstream stage. In this case stationary and nested policy are not necessarily optimal especially when the downstream demands are not balanced [12]. Other policies, named integer-ration policy, perform better while remaining tractable in practice [14]. However, unlike disassembly system, each downstream stage can order exactly the quantity demanded until the next order from its upstream stage. As in disassembly system, there is one common ordering cost incurred to replenish several inventory in the joint replenishment problem. The problem is more complex to solve and heuristic approaches are more appropriate to determine the reorder intervals [15]. However in disassembly systems that we consider all the inventories related to one ordering cost are replenished simultaneously at each order and there is no separate ordering cost for each inventory.

Allowing lost sales in disassembly systems can be interesting in practice when the capacity of a disassembly center is limited and the revenues obtained from components can be low compared to inventory costs and disassembly costs [16]. In single inventory EOQ model with lost sales, the solution is either to have no lost sales or to lose all demands [10][17]. By combining lost sales with backorders it is no longer the case, backorders indicate how many times a customer will wait before he refuses the sale. Different models are then possible to represent the customer impatience [18]. As far we know, no model is proposed to analyze the disassembly EOQ problem with lost sales.

III. A disassembly economic order quantity (DEOQ) model and solution

The model applies to two-level product structures with the first level representing to product and the second level representing the components as illustrated on Fig. 1. The product contains a given number of each components (the label of each arc in Fig. 1), named yield and noted a_i thereafter, and the disassembly of one product generates all the components. All the demands have to be satisfied and they are considered as continuous and constant. There is a fixed cost generated each time a disassembly order is launched regardless of the quantity of product ordered. In this context, a stationary policy consists in disassembling a constant quantity of products at regular cycle time.

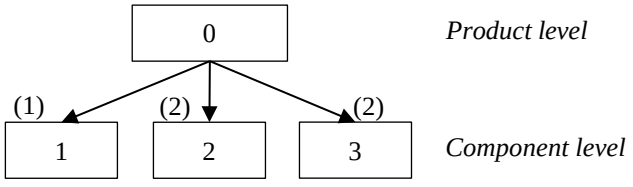
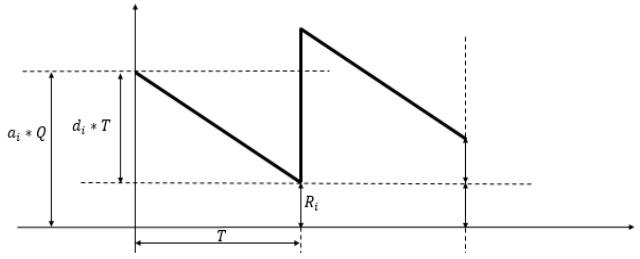
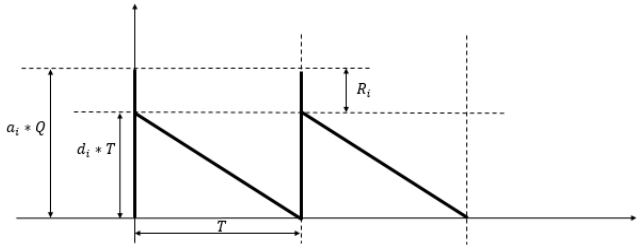


Fig. 1. An example of two-level product structure.

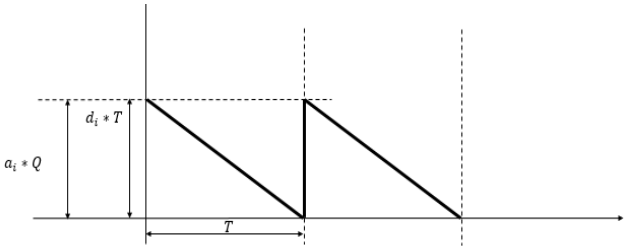
Fig. 2 presents the shape of the inventory evolution under this type of policy. For one component noted i, the slope of the inventory curve is equal to the demand per unit time (noted d_i). The quantity of each component received at the beginning of a cycle is equal to the quantity of product disassembled (noted Q) times the yield of the component. At the end of a cycle with a duration T, the component with the maximum ratio (demand/yield) is zero (Fig. 2c) however the inventory of others components is not zero and accumulates cycle after cycle (Fig. 2a). Disposal decisions have hence to be integrated in the model to have a stationary policy (Fig. 2b).



(a) Component inventory with inventory surplus R_i and without disposal



(b) Component inventory with disposal of R_i units



(c) Component inventory with maximal d_i/a_i ratio

Fig. 2. Component inventory evolution in disassembly (with and without disposal).

We consider the following other assumptions:

- components demands are independents, constant and continuous,
- each demand is characterized by a constant rate in units per unit time (per year for example),
- planning horizon is considered as infinite,
- the replenishments of part inventory (by disassembly operation) are instantaneous,
- there is a fixed cost for each disassembly operation incurred whenever an order is placed,

- there is an inventory holding cost for each part proportional to the inventory on-hand.

One of the characteristic of the disassembly systems is that the disassembly of one unit of an item generates one or more units of each of its component. It implies that, for each components, the quantity of product disassembled is not necessarily exactly the quantity demanded between two orders and unnecessary stock can be generated at each disassembly operation. By considering stationary demands, these inventory surplus accumulate with time. Disposal decisions have to be taken into account to ensure the stability of the disassembly system.

A disassembly order for a root item is defined in this paper as the launch of three types of operation:

- disassembly of a quantity of the root item,
- replenishment of the inventories of each items that succeed the root item in the bill of materials,
- disposing of inventory surplus of each items if the quantity received is superior to the demand until the next order (at least one these items has no disposal operation).

There are three disassembly operations in the disassembly systems of Fig. 1. For instance, an order of item 0 consists in:

- disassembling product 0,
- replenishing inventory of items 1, 2 and 3 ,
- disposing the inventory surplus of item 1 or 2 or 3 according to the ratio between the demand and the yield of these three items.

The problem is to determine policies for disassembly systems by considering disassembly operations costs and inventory costs. A policy sets the timing and the quantity associated to each order in the system. We restrict attention to policies that are stationary and have the zero-inventory property. In a stationary policy, the orders are repeated according to a cycle of time with a constant length. The zero-inventory properties means that an inventory is replenished only when it reach zero. This type of policy is commonly used in practice and studied in inventory systems [12].

The purpose of the model is to determine the cycle time (or the order quantity of product) to minimize the sum of setup cost, disassembly cost, inventory cost and disposal cost. The following notations are used:

- $i = 1, \dots, N$ are the index for the components,
- a_i is the yield of component i (number of component i in the product),
- d_i is the constant demand per unit time for component i,
- h_i is the inventory cost per unit and per unit time of component i,
- r_i disposal cost per unit of component i,
- k is the disassembly setup cost or fixed ordering cost for the product,
- c is the disassembly cost of the product,
- Q is the order quantity (the received quantity of each component after disassembly is $a_i\ Q$,
- T is the cycle time (time between two orders).

If we note G the component such as $d_G/a_G = \max_i\{d_i/a_i\}$ (the component which has zero inventory at the end of the cycle) then $T = a_G Q/d_G$. For other components, the inventory surplus is disposed and not kept in stock (since there is no fixed cost for disposal). The disposal quantity in each cycle is noted $\bar{R}_i$ and defined with:

$$\bar{R}_i = T(a_i\,[d_G/a_G] - d_i).$$

The proportion of disposal on the received quantity is then $1 - [d_j/a_j][d_G/a_G]$. The EOQ cost function represent a mean cost per unit time. The mean disposal quantity per unit time R_i and the remaining mean inventory I_i to serve the demand for each component i are:

$$R_i = a_i\,(d_G/a_G) - d_i \text{ and } I_i = d_i T/2.$$

The mean cost $C(T)$ is then defined in (1).

$$C(T) = \frac{k}{T} + c\frac{d_G}{a_G} + \sum\nolimits_i r_i\left(a_i\frac{d_G}{a_G} - d_i\right) + \sum\nolimits_i \frac{h_i d_i T}{2} \quad (1)$$

$C(T)$ has the same form as the basic EOQ model by replacing inventory cost by $\sum_i h_i d_i T/2$ and added a constant cost $\sum_i r_i(a_i\,[d_G/a_G] - d_i)$. After derivation of (1) according to T, the optimal cycle time T^* is for $dC(T)/dT = 0$:

$$T^* = \left(\frac{2k}{\sum_i h_i d_i}\right)^{1/2} \quad (2)$$

IV. DEOQ model with lost sales

A. Model statement

Inventory surplus and disposal of unnecessary components generated by disassembly can also be decreased by allowing lost sales. It also means that the sum of disassembly costs and inventory costs may be higher than the lost sale costs which can happen frequently in disassembly systems. The lost sales for a component correspond to a negative part of its inventory level at the end of the cycle time. The lost sales costs is proportional to the negative part of the inventory whereas the disposal cost is proportional to the positive part of the inventory. Each component has either lost sales costs or disposal costs during a cycle time and an economic balance have to find between the components with lost sales and the components with inventory surplus.

Contrary to the previous model without lost sales the cycle time T and the disassembly quantity Q are two distinct decision variables. In the proposed policy, if $a_i Q > d_i T$ then the component i has inventory surplus and the quantity $\bar{R}_i = a_i Q - d_i T$ is disposed at each cycle, if $d_i T > a_i Q$ then the component i has lost sales and the quantity $\bar{L}_i = d_i T - a_i Q$ is lost during a cycle. Under these conditions, using the Fig. 3 it is quite easy to show that the mean inventory per unit time, the mean disposal quantity per unit time and the mean lost sales per unit time are:

$$I_i = (\min\{a_i Q; d_i T\})^2/(2d_i T),$$

$$R_i = [a_i Q - d_i T]^+/T,$$

$$L_i = [d_i T - a_i Q]^+ / T,$$

with $[X]^+ = \max\{0, X\}$.

The other cost component are the fixed ordering cost which is k in each cycle and the disassembly cost which is cQ. The total mean cost $C(T, Q)$ is defined in (3) where p_i is the lost sales cost per unit.

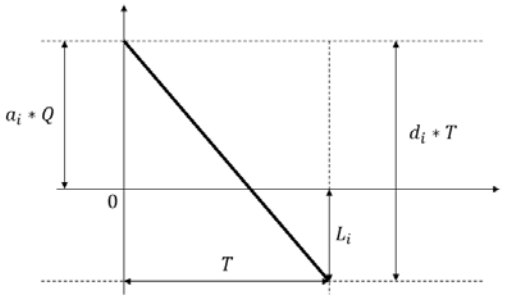


Fig. 3. Inventory evolution for a component with lost sales.

$$C(T,Q) = \frac{k}{T} + \frac{cQ}{T} + \sum_i h_i \frac{(\min\{a_i Q; d_i T\})^2}{2 d_i T} + \sum_i p_i \frac{[a_i Q - d_i T]^+}{T} + \sum_i r_i \frac{[d_i T - a_i Q]^+}{T} \quad (3)$$

Without loss of generality, the component can be numbered in decreasing order of the ratio $\frac{d_i}{a_i}$ (i.e. $\frac{d_N}{a_N} \le \cdots \le \frac{d_{i+1}}{a_{i+1}} \le \frac{d_i}{a_i} \le \cdots \le \frac{d_1}{a_1}$). For given values of T and Q, if $\frac{d_K}{a_K} < \frac{Q}{T} < \frac{d_{K-1}}{a_{K-1}}$ the $K-1$ first components are the component with lost sales, $i = 1 \dots K-1$, and the others are the components with inventory surplus, $i = K \dots N$. The total mean cost for a given ratio Q/T with $\frac{d_K}{a_K} < \frac{Q}{T} < \frac{d_{K-1}}{a_{K-1}}$ is noted $C_K(T, Q)$ and is defined in (4) which allow determining optimal values of T and Q.

$$C_K(T,Q) = \frac{k}{T} + c\frac{Q}{T} + \sum_{i=1}^{K-1} h_i \frac{(a_i Q)^2}{2 d_i T} + \sum_{i=K}^{N} h_i \frac{d_i T}{2} + \sum_{i=1}^{K-1} p_i \frac{d_i T - a_i Q}{T} + \sum_{i=K}^{N} r_i \frac{a_i Q - d_i T}{T} \quad (4)$$

To facilitate the analysis of (4), the following notations are used: $H_K = \sum_{i=1}^{K-1} h_i [a_i^2 / d_i]$, $U_K = \sum_{i=K}^{N} h_i d_i$, $P_K = \sum_{i=1}^{K-1} p_i a_i$, $R_K = \sum_{i=K}^{N} r_i a_i$, $V_K = \sum_{i=1}^{K-1} p_i d_i - \sum_{i=K}^{N} r_i d_i$ and the variables T and Q are replaced with the variables T and X with $X = Q/T$.

The cost function (4) is then equivalent, for $\frac{d_K}{a_K} < X < \frac{d_{K-1}}{a_{K-1}}$, to:

$$C_K(T,X) = \frac{k}{T} + H_K \frac{X^2 T}{2} + U_K \frac{T}{2} + (R_K - P_K + c)X + V_K \quad (5)$$

When $X = \frac{d_K}{a_K}$, the component K has no lost sales and no inventory surplus, the cost function, denoted by $C_K(T)$, depends only on T:

$$C_K(T) = \frac{k}{T} + H_K \left(\frac{d_K}{a_K}\right)^2 \frac{T}{2} + U_K \frac{T}{2} + (R_K - P_K + c)\frac{d_K}{a_K} + V_K \quad (6)$$

One can note that:

$$C_K\left(T, \frac{d_K}{a_K}\right) = C_K(T) \text{ and } C_K\left(T, \frac{d_{K-1}}{a_{K-1}}\right) = C_{K-1}\left(T, \frac{d_{K-1}}{a_{K-1}}\right).$$

The disassembly EOQ model with lost sales can then be written:

$$C(T,X) = \frac{k}{T} + H_K \frac{X^2 T}{2} + U_K \frac{T}{2} + (R_K - P_K + c)X + V_K \quad (7)$$

$$\text{for} \quad \frac{d_K}{a_K} \le X \le \frac{d_{K-1}}{a_{K-1}}, \forall\, K = 1 \dots N$$

B. Solution method

The solving approach to find optimal values of T and X consists in analyzing each piece of $C(T,X)$ noted $C_K(T,X)$ defined for $\frac{d_K}{a_K} \le X \le \frac{d_{K-1}}{a_{K-1}}$. We denoted by T_K and X_K the optimal values of $C_K(T,X)$ which have to satisfy $\partial C_K(T,X)/\partial X = 0$ and $\partial C_K(T,X)/\partial T = 0$:

- $\partial C_K(T,X)/\partial X = H_K X T + \left(R_K - P_K + c\right)$ then
$$X_K = \frac{P_K - R_K - c}{H_K T} \quad (8)$$
- $\partial C_K(T,X)/\partial T = -\frac{k}{T^2} + \frac{H_K}{2}X^2 + \frac{U_K}{2}$. By replacing X by X_K, we obtain:
$$\frac{\partial C_K(T,X)}{\partial T} = \frac{-2kH_K + (P_K - R_K - c)^2}{2H_K T^2} + \frac{U_K}{2}.$$

The existence of the solution depends on the sign of $-2kH_K + (P_K - R_K - c)^2$ and the constraint $\frac{d_K}{a_K} \le X \le \frac{d_{K-1}}{a_{K-1}}$:

- if $-2kH_K + (P_K - R_K - c)^2 > 0$, the cost function is increasing in T and to minimize $C_K(T,X)$, T has to be reduced as much as possible until the component $K-1$ has no lost sales and the cost function turns to $C_K\left(T, \frac{d_{K-1}}{a_{K-1}}\right)$,
- if $-2kH_K + (P_K - R_K - c)^2 < 0$, the optimal value T_K is given by (9) if $\frac{d_K}{a_K} \le X \le \frac{d_{K-1}}{a_{K-1}}$, otherwise T is on the limits.

$$T_K = \sqrt{\frac{2kH_K - (P_K - R_K - c)^2}{U_K H_K}} \quad (9)$$

The optimal value of T for X given is $T_K^*(X) = \sqrt{2k/[U_K + X^2 H_K]}$ and for $X = d_K/a_K$:

$$T_K^* = \sqrt{\frac{2k}{U_K + \left(\frac{d_K}{a_K}\right)^2 H_K}} \qquad (10)$$

One can note that for $K = 1$, $H_1 = 0$ and $U_1 = \sum_{i=1}^{N} h_i d_i$ then T_K^* correspond to optimal solution of (3) which is the disassembly EOQ without lost sales.

Based on equations (1), (2) and (5) to (10), the solution of (7) can be found by applying the following steps:

1. Set $K = 1$, $H_1 = P_1 = 0$, $U_1 = \sum_{i=1}^{N} h_i d_i$, $R_1 = \sum_{i=1}^{N} r_1 a_1$ and $V_1 = -\sum_{i=1}^{N} r_1 d_1$. Determine T^* with (2) and C^* with (1).

2. Set $K = K + 1$, $H_K = H_{K-1} + h_{K-1}(a_{K-1}^2/d_{K-1})$, $U_K = U_{K-1} - h_{K-1} d_{K-1}$, $P_K = P_{K-1} + p_{K-1} a_{K-1}$, $R_K = R_{K-1} - r_{K-1} a_{K-1}$ and $V_K = V_{K-1} + (p_{K-1} + r_{K-1}) d_{K-1}$.

3. If $-2kH_K + (P_K - R_K - c)^2 < 0$: determine T_K with (9) and X_K with (8), then if $\frac{d_K}{a_K} \le X_K \le \frac{d_{K-1}}{a_{K-1}}$: determine C_K with (10), then if $(C_K < C^*) : C^* = C_K$, $T^* = T_K$ and $X^* = X_K$.

4. Determine T_K with (15) and C_K with (11). If $(C_K < C^*)$ then $C^* = C_K$, $T^* = T_K$ and $X^* = d_K/a_K$.

5. If $K = N$ then the optimal cost is C^* otherwise go to step 2.

C. Numerical illustration

To illustrate the application of the solution procedure, we will use a numerical example for a product with five components and $k = 100$ and $c = 1$. The other data for each component are given in Table 1. The table 2 presents the expression of the cost function for each pieces defined on given interval on X. The first row corresponds to the cost if the component has no lost sales (and hence the others components have not) which mean that $X \ge 5$. The second line corresponds to the cost if the components 1 has lost sales and the others have not and so on.

TABLE I. Data for each component.

i	a_i	d_i	h_i	p_i	r_i	d_i/a_i
1	2	10	5	6	0.5	5
2	2	8	3	5	0.1	4
3	1	3	4	6	0,5	3
4	1	2	2	10	1	2
5	4	4	4	15	0.7	1

TABLE II. Expression of the cost function.

K	Expression of $C(T, X)$	Conditions
1	$(100/T)$ +53T+6.5X-12.1	$5 \le X$
2	$(100/T)$+X^2T+28T-6.5X+52.9	$4 \le X \le 5$
3	$(100/T)$+1.75X^2T+16T-16.7X+93.7	$3 \le X \le 4$
4	$(100/T)$+2.42X^2T+10T-23.2X+113.2	$2 \le X \le 3$
5	$(100/T)$+2.92X^2T+8T-34.2X+135.2	$1 \le X \le 2$

Table 3 presents the results of the procedure for the five iteration. The minimum cost is for $K = 4$ which means that the components 1, 2 and 3 have lost sales and the components 4 and 5 have inventory surplus. One also notes that none of the component has exactly zero inventory at the end of a cycle in this solution even component 4 since we have $X_4 > d_4/a_4$.

TABLE III. Optimal values

K	X_K	T_K	$C(T_K, X_K)$
1	5	1,37	166,00
2	4	1,51	159,56
3	3	1,77	156,29
4	**2,28**	**2,11**	**155,30**
5	1	3,03	167,08

V. Conclusion

Disassembly systems brings new issue in inventory management. The main fact is that one unit of product disassembled generate simultaneously one or more units of each component and hence unnecessary inventories. We have analyzed an inventory model with deterministic, continuous and constant demand with fixed cost each time an order is launched, variable cost function of the quantity of product ordered and inventory cost function of the number of stocked products. The unnecessary inventory surplus is disposed as soon as it is received and, in this context, the optimal reorder interval has a simple form as basic in single item EOQ models. We have extended the model to take into account lost sales and we have proposed a method to find the reorder interval.

Further researches can extend the results in several ways. The reorder quantity, obtained from the reorder interval, can be used as part of heuristics for stochastic demand models. The model can be modified to take into account separate decisions for each component at each order and setup costs for each components. Multi-stage disassembly systems are also under investigation.

References

[1] S.M. Gupta and K.N. Taleb, “Scheduling disassembly”, International Journal of Production Research, vol. 32, no. 8, pp.1857-1866, 1994.

[2] H.-J. Kim, D.-H. Lee and P. Xirouchakis, “Disassembly scheduling: literature review and future research directions”, International Journal of Production Research, vol. 45, no. 18-19, pp.4465-4484, 2007.

[3] H.-J. Kim and P. Xirouchakis, “Capacitated disassembly scheduling with random demand”, International Journal of Production Research, vol. 48, no. 23, pp. 7177-7194, 2010.

[4] Y. Barba-Gutiérrez and B. Adenso-Díaz, “Reverse MRP under uncertain and imprecise demand”, International Journal of Advanced Manufacturing Technology, 40(3-4), 413-424, 2009.

[5] K. Inderfurth and I.M. Langella, “Heuristics for solving disassemble-to-order problems with stochastic yields”, OR Spectrum 28, 73-99, 2006.

[6] H.-J. Kim, D.-H. Lee, P. Xirouchakis and O.K. Kwon, “A branch and bound algorithm for disassembly scheduling with assembly product structure”, Journal of the Operational Research Society, Volume 60, Issue 3, pp 419-430, 2009.

[7] Y. Barba-Gutiérrez, B. Adenso-Díaz, and S. M. Gupta, "Lot sizing in reverse MRP for scheduling disassembly", International Journal of Production Economics, vol. 111, no. 2, pp. 741-751, 2008.

[8] X. Ji, Z. Zhang, S. Huang, and L. Li, "Capacitated disassembly scheduling with parts commonality and start-up cost and its industrial application", International Journal of Production Research, vol. 54, no. 4, pp. 1-19, 2015.

[9] M. Godichaud, L. Amodeo and M. Hrouga, "Metaheuristic based optimization for capacitated disassembly lot sizing problem with lost sales", 2015 International Conference on Industrial Engineering and Systems Management (IESM), Seville, 2015, pp. 1329-1335.

[10] P. Zipkin, Foundations of inventory management. New York: McGraw-Hill; 2000.

[11] S. Axsater, Inventory control. Dordrecht: Kluwer Academic Publishers; 2000.

[12] J.A. Muckstadt and R.O. Roundy, Analysis of multistage production systems. Handbooks in OR and MS, vol. 4; 1993 [chapter 2].

[13] B. Abdul-Jalbar, A. Segerstedt, J. Sicilia and A. Nilsson "A new heuristic to solve the one-warehouse N-retailer problem", Computers & Operations Research, 37, pp.265-272, 2010.

[14] B. Abdul-Jalbar, J. Gutiérrez and J. Sicilia, "Integer-ratio policies for distribution /inventory systems", International Journal of Production Economics, vol. 93–94, pp. 407–15, 2005.

[15] K. Nilsson and A. Segerstedt, "Corrections of costs to feasible solutions of economic lot scheduling problems", Computers and Industrial Engineering, vol 54(1), pp. 155-68, 2008.

[16] M. Hrouga, M. Godichaud and L. Amodeo, "Heuristics for multi-product capacitated disassembly lot sizing with lost sales", IFAC-PapersOnLine, Volume 49, Issue 12, 2016, pp. 628-633.

[17] D.W. Pentico, M.J. Drake and C. Toews, "The deterministic EPQ with partial backordering: A new approach", Omega, vol. 37(3), pp. 624-636, 2009.

[18] D.W. Pentico and M.J. Drake, "A survey of deterministic models for the EOQ and EPQ with partial backordering", European Journal of Operational Research, vol. 214, pp. 179-198, 2011.

Blood products inventory routing problem with transshipment and substitution

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Imane HSSINI, Nadine MESKENS
LSM, Catholic University Of Louvain
Mons, Belgium
Imane.hssini@student.uclouvain.be
Nadine.meskens@uclouvain.be

Fouad RIANE
Laboratoire de Génie Industriel, Centrale Supélec Paris
Ecole Centrale de Casablanca
Casablanca, Morocco
Fouad.riane@centrale-casablanca.ma

Abstract— In this paper, we address the problem of blood products supply chain. We model this problem as an inventory routing problem while considering the particularities of blood products namely perishability, different shelf lives, substitution between blood groups. In addition, we allow transshipments (the transfer of products between hospitals) under certain conditions. We formulate this problem as an integer linear program (ILP) whose objective function is to minimize the costs of storage, distribution, obsolescence, transfer and substitution. Several decision variables were taken into account, namely the total quantity of the blood products delivered to a hospital, the quantities of products transferred between hospitals, the quantities of substitute products used besides the level of the stock at the end of every period and the number of containers transported by a vehicle. Computational tests were carried out in eight small sizes instances. The results show that the exact method provides optimal solutions but the complexity of the problem cannot be coped with unless we use approximative approaches.

Keywords—inventory routing problem; blood products; transshipment; linear programming; vendor managed inventory

I. Introduction

The whole blood and the labile blood products (platelets, red blood cells and plasma) are perishable products. If they are not used within their shelf life, they become obsolete. In addition, they are vital products and cannot be substituted. It is therefore necessary to maintain sufficient reserves to satisfy demands. However, the specifications of these products pose some constraints for blood stock managers, mainly relating to their perishable nature, their heterogeneity (different shelf life), storage and transport conditions, unpredictable demands, etc…. To that, one must add the complexity associated with the compatibility between the different blood groups (A, B, AB and O and Rhesus factor + or -). Indeed, a patient must be transfused with a blood product from a donor of the same blood group. However, when the same blood group of the patient is not available, another compatible blood group can be used - that is substitution between blood groups.

Mastering the blood products logistic, from the collection of blood and labile blood products until their administration for therapeutic use in healthcare facilities, is a health and economic necessity. Indeed, from an economic point of view, in Belgium for example, the selling price of the labile blood products is fixed by the competent authorities to ensure that the incomes can mainly cover the costs of production activities in transfusion establishment (organization of the collections, biological qualification, preparation of blood, storage and distribution to healthcare facilities). Any additional cost associated with these activities will be supported by the transfusion establishment. According to [1] the inventory management and the distribution are considered as the main elements of the production costs. This means that the optimization of storage and distribution can have positive financial outcome for the transfusion establishments. Therefore, it is interesting to focus on the cost improvement of these two processes. From a health perspective, a blood bag can save human lives - so far no substitutes exist – this is why we need to adopt effective stock management methods to mitigate risks of shortage.

Given this sanito-economic dimension, we have been interested to optimize simultaneously the processes of storage and distribution of blood products. This optimization must ensure the availability of the right product in good quantity at the right time, while reducing inventory levels and costs. This is the definition of optimization problems generally called Inventory Routing Problem (IRP). In fact, we treat the IRP problem in a Vendor Managed Inventory (VMI) context - the supplier decides the quantity to be delivered each day to each customer - with the possibility of transshipement or stock transfer - transfer of quantities of labile blood products from the surplus hospital to the deficit one-.

The aim of this research is to apply IRP methods to address one of the important problems encountered by managers of health facilities (hospitals or blood centers), which is the distribution of blood products. Indeed, we are studying a problem where a single supplier[1] is responsible for the supply of forty hospitals; it ensures the whole processes of the supply chain of blood products from the collection to the

[1] For the French-speaking part of Belgium, The Red Cross is considered as the unique supplier of the blood products. This ensures the smooth running of the blood collection process from blood donation until the delivery of various blood products to hospitals, passing through storage.

delivery. In this paper, we study the inventory routing problem as part of a two stage blood product supply chain (a single supplier with multiple hospitals). In reality, the demand of blood products is stochastic; however, to simplify our model we assume that the demand is deterministic. Referring to the classification of the IRP presented by [2], the problem of concern is a deterministic, multi-period, multi-product, multi-vehicle, multi-routing problem over finite planning horizon. The objective function is the minimization of inventory, transportation, substitution, transfer and obsolescence costs. This paper is organized as follows; we start reviewing related studies encountered in the literature. We then suggest a mathematical formulation for the problem that we solved to optimality for small size instances. We finally analyze the relevant computational results.

II. State of art

During the last two decades, the literature on the inventory routing problem has increased exponentially. Despite the abundant literature on the inventory routing, we did not find a lot of papers that address the inventory routing problem of blood products especially considering the possibilities of using substitutes and transfer of products between hospitals. To this end, we have introduced in this state of the art the articles dealing the IRP applied to perishable products since blood products share this relevant characteristic.

Indeed, the IRP researches related to perishable products are very rare. Of these, we include those performed by [3]; [4]; [5]; [6]; [7]; [8]; [9]; [10]; [11]; [12]; [13]; [14]; [15]; [16]; [17]. Among the researchers interested in the IRP with a single perishable product we quote:[4]; [5]; [9]; [10]; [15]; [16]; [6]. Each of them analyzed the problem in a different way according to the used variants (number and nature of product, nature of demand, size of fleet...) or the proposed solution. For examples: [10]; [15]; [6]; [14] have studied the IRP in the case of a deterministic demand, others such as [16]; [5] are interested in the case of a stochastic demand. In these articles, the authors have used several homogeneous vehicles to transport products; they targeted the minimization of costs (transport, storage and shortages) as the main objective function. In the case of multiple perishable products, [17]; [11]; [12] studied the IRP with deterministic demand. Besides the multiplicity of products, the stochasticity of demand was considered by researchers as [3]; [7]; [8] and [13].

Blood products have been the subject of some works such as those of [3]; [8]; [9]; [18]. Indeed, [9] studied the impact of the adoption of the VMI policy by the Austrian Red Cross blood service. The objective of this study was to minimize the distribution cost. They have developed a flexible vehicle routing system to deliver one blood product with deterministic demand. This study has been extended by [8] to cover the case of several blood products with a stochastic demand.

According to this state of art we noticed that the articles on the IRP applied to perishable products are few in number especially in the case of blood products. However, our research is distinguished from other research by taking into account the possibility of substitution between the different blood groups and also the possibility of transfer of stock between hospitals in case of need. In the following, we present the mathematical model.

III. Description and formulation of problem

In this problem, we assume that a single depot supplies, in each period of a given horizon, a set of hospitals with two heterogeneous blood products r (red blood cells and platelets) of different blood groups. These products are packed in identical bags and delivered in isothermal containers k with identical capacity Cap_k which means $Cap_k = Cap_{kr1} = Cap_{kr2}$. Given that the blood products are perishable, each product has a shelf life denoted sl_r which depends only on the type of product (does not depend on the group). Each product r has an age a. The product can be used only if $a \epsilon SL_r = \{1, \dots slr\}$. The depot is facing an assumed known demand, initiated by a set of geographically dispersed hospitals denoted H (indexed by i or j Є H). To deliver these products, the depot uses a fleet of heterogeneous vehicles V (indexed by v Є V), with a capacity Cap_v and a travel cost per kilometer CT_v. In the case of a stock-out situation or insufficient stock quantities, the transfer of the products between hospitals having a surplus inventory in a one of them and a deficit inventory in other can be achieved to satisfy the requests. If the whole request cannot be satisfied by considering transfers, one can use the substitution possibilities between groups of products to fulfil the demand according to the compatibility of the blood group. Substitutions and transfers generate costs. We consider these costs as penalties to avoid as much as possible the use of those alternatives. In what follows, we present the assumptions on which this model is based:

- Each tour must begin and finish at the depot and each hospital must be visited by one vehicle during each period t.
- The transfer of a product r of group g between hospital i and hospital j during a period t, is done only if the stock in the depot does not satisfy the entire demand of hospital j and If the stock of the hospital i is sufficient to satisfy both its own demand for product r of group g and that of hospital j. In this case, the entire quantity of the product r of the group g delivered to hospital j will be taken from the stock of the hospital i.
- In the case where a hospital's demand for a product r of group g cannot be fully satisfied either by the initial stocks of the depot or by those of the hospitals (the transfer is not possible), we deliver to this hospital, from the depot, the available quantity of this product and we complete the missing part with a product r of a group compatible with the ordered group. We call this product: Substitute Product.
- No preference intervenes in the choice of the substitute product, i.e. we can use any group g ' compatible with group g.
- The storage capacity of each hospital cannot be exceeded.

- In order to prioritize the transfer, we consider that the transfer cost is lower than the cost of substitution.
- Each expired product is physically removed from the stock and cannot be used to meet demand.
- A container can contain only one type of product (red blood cells or platelets). We do not transport both products in the same container.
- The shortage of the stock is not allowed.

To model this problem we use the parameters and the sets described in the following tables.

TABLE I : The used sets.

Set	Signification
H	The set of hospitals. H= {1,2,…, h}.
H+	The set of hospitals plus the depot (0 for depot, 1 and higher for hospitals).
V	The set of vehicles. V = {1,2,…, f}.
T	Planning horizon. T= {1,2,…,t'}(in days).
P	The set of blood products (r_1=red cells and r_2=platelets)
G	The set of blood groups(O^-,A^-,B^-,AB^-, O^+,A^+, B^+, AB^+)

TABLE II: The used parameters.

Parameter	Signification
D_{irgt}	The demand of hospital i of product r of group g during period t (in number of bags).
C_{ir}	The storage capacity of each hospital i for each product r (in number of bags).
IN_{rgajt}	The quantity in stock of the product r of group g and of age a, in each location j Є H^+ at the beginning of the period t (in number of bags).
$Dist_{ij}$	The distance between two locations i, j Є H^+ (in kilometers).
Cap_k	The capacity of container k (in number of bags)
Cap_v	The capacity of vehicle v (in number of containers).
φ_{ij}	The duration of the journey between two locations (in hours).
CS_{rj}	The holding cost of product r at location j Є H^+ (in euro).
CO_r	The obsolescence cost per product r (in euro).
Ct_{ij}	The transfer cost between location i Є H and j Є H (in euro).
Cs_r	The substitution cost of a product r (in euro)
$A_{gg'}$	Compatibility matrix. It is equal to 1 if the group g' is compatible with the group g and equal to 0 otherwise
τ_t	Total working hours per driver per each period t.

TABLE III : The decision variables.

Variable	Signification
Q_{irgavt}	The total quantity of the product r of group g and of age a, delivered to the hospital i during the period t, using the vehicle v.(in number of bags)
Q'_{irgavt}	The quantity of the product r of group g and of age a, taken from the stock of depot and delivered to the hospital i during the period t, using the vehicle v. This quantity does not contain substitutes; it only contains products whose group is the same ordered group (in number of bags)
$Su_{irg'avt}$	The quantity of the substitution product r of the group g'(substitute of the product r of group g) and age a, taken from the stock of depot and delivered to the hospital i during the period t, using the vehicle v. (in number of bags)
$T_{ijrgavt}$	The quantity of product r of group g and of age a, by the vehicle v, taken from the stock of the hospital i (never from the depot) and transferred to the hospital j during the period t. (in number of bags)
$q_{ijrgavt}$	The quantity of the product r of group g and of age a, remaining in vehicle v when moving from location i ∈ H^+ to location j ∈ H during period t. (in number of bags)
I_{rgajt}	The quantity in stock of the product r of group g and age a, at location j ∈ H^+ at the end of the period t. (in number of bags)
Nc_{ijrvt}	The number of containers containing the product r, and transported in the vehicle v during the travel between the location i ∈ H^+ and location j ∈ H during period t. (in number of containers)
x_{ijvt}	A binary variable set to 1 if location j is visited immediately after location i by vehicle v during each period t, 0 otherwise

The model is:

$$\text{Min} \sum_{i=0}^{h}\sum_{\substack{j=0\\ i\neq j}}^{h}\sum_{t=1}^{t'}\sum_{v=1}^{f} CT_v * Dist_{ij} * x_{ijvt}$$
$$+\sum_{j=0}^{h}\sum_{r\in P}\sum_{g\in G}\sum_{a=1}^{sl_r}\sum_{t=1}^{t'} CS_{rj} * I_{rgajt} + \sum_{j=0}^{h}\sum_{r\in P}\sum_{g\in G}\sum_{t=1}^{t'}\sum_{a>sl_r} CO_r * I_{rgajt}$$
$$+\sum_{i=1}^{h}\sum_{\substack{j=1\\ i\neq j}}^{h}\sum_{v=1}^{f}\sum_{r\in P}\sum_{g\in G}\sum_{t=1}^{t'}\sum_{a=1}^{sl_r} CT_{ij} * T_{ijrgavt}$$
$$+\sum_{j=1}^{h}\sum_{r\in P}\sum_{v=1}^{f}\sum_{g\in G}\sum_{a=1}^{sl_r}\sum_{t=1}^{t'} Cs_r * Su_{jrgavt} \qquad (1)$$

Subject to:

$$\sum_{v=1}^{f}\sum_{i=0}^{h} x_{ijvt} \leq 1 \qquad \forall\, j \in H, t \in T \qquad (2)$$

$$\sum_{\substack{i=0\\ i\neq j}}^{h} x_{ijvt} - \sum_{\substack{m=0\\ m\neq j}}^{h} x_{jmvt} = 0 \qquad \forall\, j \in H, v \in V, t \in T \qquad (3)$$

$$\sum_{i=0}^{h}\sum_{\substack{j=0\\ j\neq i}}^{h} x_{ijvt} * \varphi_{ij} \leq \tau_t \qquad \forall\, v \in V, t \in T \qquad (4)$$

$$\sum_{v=1}^{f}\sum_{\substack{i=0\\ i\neq j}}^{h} q_{ijrgavt} - \sum_{v=1}^{f}\sum_{\substack{m=0\\ m\neq j}}^{h} q_{jmrgavt} = \sum_{v=1}^{f} Q_{jrgavt}$$

$$\forall\ j \in H, r \in P, g \in G, a \in SL_r, t \in T \qquad (5)$$

$$\sum_{r\in P}\sum_{g\in G}\sum_{a=1}^{sl_r} q_{i0rgavt} = 0 \qquad \forall\ i \in H, v \in V, t \in T \qquad (6)$$

$$\sum_{g\in G}\sum_{a=1}^{sl_r} Q_{jrgavt} = \sum_{g\in G}\sum_{a=1}^{sl_r} Q'_{jrgavt} + \sum_{g\in G}\sum_{a=1}^{sl_r}\sum_{\substack{g'\in G\\ g'\neq g}} Su_{jrg'avt} * A_{gg'}$$

$$+ \sum_{\substack{i=1\\ j\neq i}}^{h}\sum_{g\in G}\sum_{a=1}^{sl_r} T_{ijrgavt} \qquad \forall\ j \in H, r \in P, v \in V, t \in T \qquad (7)$$

$$\sum_{v=1}^{f}\sum_{a=1}^{sl_r} Q'_{jrgavt} * \sum_{v=1}^{f}\sum_{i=1}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} = 0$$

$$\forall\ j \in H, j\neq i, r \in P, g \in G, t \in T \qquad (8)$$

$$\sum_{v=1}^{f}\sum_{a=1}^{sl_r}\sum_{\substack{g'\in G\\ g'\neq g}} Su_{jrg'avt} * \sum_{v=1}^{f}\sum_{i=1}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} = 0$$

$$\forall\ j \in H, j\neq i, r \in P, g \in G, t \in T \qquad (9)$$

$$\left(\left\lceil \frac{1}{Cap_k} * \left(\sum_{g\in G}\sum_{a=1}^{sl_r} q_{ijr_1gavt}\right)\right\rceil + \left\lceil \frac{1}{Cap_k} * \left(\sum_{g\in G}\sum_{a=1}^{sl_r} q_{ijr_2gavt}\right)\right\rceil\right)$$
$$\leq Cap_v * x_{ijvt}$$

$$\forall i \in H^+, j \in H, j\neq i, v \in V, t \in T \qquad (10)$$

$$\sum_{g\in G}\sum_{a=1}^{sl_r} IN_{rgajt} + \sum_{g\in G}\sum_{v=1}^{f}\sum_{a=1}^{sl_r} Q_{jrgavt} \leq C_{jr} \qquad \forall j \in H, r \in P, t \in T \qquad (11)$$

$$\sum_{v=1}^{f}\sum_{\substack{j=1\\ i\neq j}}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} \leq \sum_{a=1}^{sl_r} IN_{rgait} - D_{rgit} \qquad \forall i \in H, r \in P, g \in G, t \in T \qquad (12)$$

$$\sum_{v=1}^{f}\sum_{i\in H}\sum_{\substack{g'\in G\\ g'\neq g}}\sum_{a=1}^{sl_r} Su_{irg'avt} * A_{gg'} \leq \sum_{i\in H} D_{rgit} - \sum_{a=1}^{sl_r} IN_{rga0t}$$

$$\forall\ r \in P, g \in G, t \in T \qquad (13)$$

$$IN_{rga+1jt} = I_{rgajt-1} - \sum_{k\geq 1} I_{rgsl_r+k\,jt-1}$$

$$\forall\ j \in H^+, a \in SL_r, r \in P, g \in G, t \in T \qquad (14)$$

$$\sum_{a=1}^{sl_r} I_{rgait} = \sum_{a=1}^{sl_r} IN_{rgait} + \sum_{v=1}^{f}\sum_{a=1}^{sl_r} Q_{irgavt} - D_{rg\,it} - \sum_{a=1}^{sl_r}\sum_{\substack{j=1\\ j\neq j}}^{h}\sum_{v=1}^{f} T_{ijrgavt}$$

$$\forall\ i \in H, r \in P, g \in G, t \in T \qquad (15)$$

$$\sum_{a=1}^{sl_r} I_{rga0t} = \sum_{a=1}^{sl_r} IN_{rga0t} - \sum_{j=1}^{h}\sum_{a=1}^{sl_r} Q'_{jrgavt} - \sum_{j=1}^{h}\sum_{a=1}^{sl_r} Su_{jrgavt}$$

$$\forall\ r \in P, g \in G, v \in V, t \in T \qquad (16)$$

$$x_{ijvt} \in \{0,1\} \qquad \forall\ i, j \in H^+, i\neq j, v \in V, t \in T \qquad (17)$$

$$T_{ijrgavt}; Q_{jrgavt}; Q'_{jrgavt}; Su_{jrgavt}; IN_{rgajt} \geq 0; \quad \text{Integers}$$

$$\forall\ i, j \in H^+, i\neq j, v \in V, a \in SL_r, r \in P, g \in G, t \in T \qquad (18)$$

The objective function of the proposed model is defined by (1). This function minimizes the costs of transport, storage, obsolescence, substitution, and transfer at each period. Constraints (2) ensure that each hospital can be visited at most once by a maximum one vehicle during period t. The constraints (3) ensure the continuity of a tour. In this case, if a vehicle arrives at a hospital, it must leave, after serving it, to the next hospital or to the depot. Constraints (4) ensure that the total travel time of a vehicle should not exceed the planned total working hours in period. Constraints (5) indicate that the quantity of a product r of group g with age a delivered to the hospital j is the difference between the quantities of this product arriving at hospital j (from anywhere) and the quantities that leave hospital j (to anywhere). Constraints (6) indicate that the vehicle must be empty when returning to the depot. The constraints (7) show that the TOTAL quantity of a product r delivered to a hospital j can be composed of the products picked from the depot (products picked from the initial stock of the depot and / or the substitution products) OR the transferred products. The constraints (8) mean that the total quantity of a product r of group g delivered to a hospital must either be picked from the stock of the depot or transferred from a hospital; we cannot have both situations at the same time. The constraints (8) are nonlinear. In order to linearize it, we will introduce a binary variable B_{irgvt} and a parameter M. The constraints (8) will be replaced by the following constraints:

$$B_{jrgvt} = \begin{cases} 1, & \sum_{v=1}^{f}\sum_{i=1}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} = 0 \\ 0, & \sum_{v=1}^{f}\sum_{a=1}^{sl_r} Q'_{jrgavt} = 0 \end{cases}$$

$$\forall\ j \in H, i\neq j, r \in P, g \in G, t \in T \qquad (8a)$$

$$\sum_{v=1}^{f}\sum_{a=1}^{sl_r} Q'_{jrgavt} \leq M * B_{jrgvt} \qquad \forall\ j \in H, r \in P, g \in G, t \in T \qquad (8b)$$

$$\sum_{v=1}^{f}\sum_{i=1}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} \leq M * (1 - B_{jrgvt})$$

$$\forall\ j \in H, r \in P, g \in G, t \in T \qquad (8c)$$

The constraints (9) mean that the transfer and the substitution cannot be achieved at the same time (to satisfy a demand, we use either the substitution products or the

transferred products). Like constraints (8), the constraints (9) are nonlinear. To linearize it, we will use a binary variable B'_{irgvt} and a parameter M. The constraints (9) will be replaced by the following constraints:

$$B'_{jrgvt}=\begin{cases}1, & \sum_{v=1}^{f}\sum_{\substack{i=1\\ i\neq j}}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt}=0\\ 0, & \sum_{v=1}^{f}\sum_{a=1}^{sl_r}\sum_{\substack{g'\in G\\ g'\neq g}} Su_{jrg'avt}=0\end{cases}$$

$$\forall\ j \in H, r \in P, g\in G, t \in T \qquad (9a)$$

$$\sum_{v=1}^{f}\sum_{a=1}^{sl_r}\sum_{\substack{g'\in G\\ g'\neq g}} Su_{jrg'avt} \leq M*B'_{jrgvt}$$

$$\forall\ j \in H, r \in P, g\in G, t \in T \qquad (9b)$$

$$\sum_{v=1}^{f}\sum_{\substack{i=1\\ i\neq j}}^{h}\sum_{a=1}^{sl_r} T_{ijrgavt} \leq M*\left(1-B'_{jrgvt}\right)$$

$$\forall j \in H, r \in P, g\in G, t \in T \qquad (9c)$$

Constraints (10) ensure that the number of transported containers is less than or equal to the capacity of the vehicles. The constraints (10) are nonlinear. We replace it by the following three constraints integrating the decision variable Nc_{ijrvt}:

$$\sum_{g\in G}\sum_{a=1}^{sl_r} q_{ijr_1gavt} \leq Nc_{ijr_1vt}*Cap_k \qquad \forall\ i \in H, v \in V, t \in T \ (10a)$$

$$\sum_{g\in G}\sum_{a=1}^{sl_r} q_{ijr_2gavt} \leq Nc_{ijr_2vt}*Cap_k \qquad \forall\ i \in H, v \in V, t \in T \ (10b)$$

$$Nc_{ijr_1vt}+Nc_{ijr_2vt} \leq Cap_v*x_{ijvt}$$

$$\forall\ i \in H^+, j \in H, j\neq i, v \in V, t \in T \qquad (10c)$$

$$Nc_{ijrvt} \geq 0; \quad \text{Integer}$$

$$\forall\ i \in H^+, j \in H, j\neq i, r \in P, v \in V, t \in T \qquad (10d)$$

The constraints (11) impose to not exceed the total storage capacity of each hospital whatever the time period t. The constraints (12) limit the quantity that could be transferred from one hospital to another. The constraints (13) guarantee that the quantity of substitution products delivered to a hospital does not exceed the missing quantity of the ordered product picked from the stock of the depot. The constraints (14) modify the age of the products at each time period t and eliminate the quantities of the obsolete products from the stock. The constraints (15) and (16) define the inventory conservation at the hospitals and the depot respectively. The constraints (17) define the binary nature of decision variables. Constraints (18) require that these values are non-negative integers: stock levels in hospital and depot, quantities of products picked from the depot and delivered to a hospital, quantities of the substitution products and the quantities transferred from one hospital to another.

IV. Computational results

To validate the proposed mathematical model, we tested it on small size fictitious instances. We have randomly generated eight test instances with the following parameters: one depot delivers one or two products to a set of hospitals, (the number of hospitals varies between 3 and 5) by using two heterogeneous vehicles. The capacity of each vehicle is [Cap_{v1}=60, Cap_{v2}=140]; Hospital consumption by product is integer value randomly generated within the interval [1, 40]. We consider a planning horizon of three and five days (T=3 and T=5). The travel distance between locations i and j is given in kilometer. The inventory level at location i $\in H^+$ at the beginning of planning horizon is randomly generated as an integer between [0, 50]; the maximum working hours is set to 8 hours per day; The big number in constraints is set to M=100, the travel cost is 2 euro per kilometer and the storage cost at each location i $\in H^+$ is [r_1=100, r_2=100] in euro per period. The substitution, transfer and obsolescence cost are respectively 100 euro, 50 euro and 100 euro. The IRP model was tested using Cplex 12.7 on the Intel(R) Core(TM) i5-2450M CPU 2.50 GHz with 4 GB RAM. The results for the eight test instances are presented in Table 5. The table indicates the objective function values, the transportation, the storage, the obsolescence, the transfer and the substitution cost and the CPU computing time.

TABLE IV: Parameter values.

Instance	Number of hospitals	Number of products	Number of groups	Number of period
1	3	1	2	3
2	3	1	4	3
3	3	1	8	3
4	3	2	2	3
5	3	2	4	3
6	3	1	2	5
7	5	1	2	3
8	5	2	8	5

TABLE **V**: Results for the eight test instances.

Instance	Obsolescence cost (in euro)	Substitution cost (in euro)	Transshipment cost (in euro)	Storage cost (in €)	Transport cost (in €)	Total cost (in €)	CPU time (in sec)
1	2500	320	1500	35400	1248	40968	57.52
2	4500	0	3450	58900	1248	68098	996.60
3	5300	600	3150	89500	1208	99758	3143.64
4	3600	0	6250	17160	1248	28258	1158.10
5	2900	0	7750	15440	1248	27338	3875.30
6	1400	0	1800	39900	1390	44490	1001.23
7	2100	0	1550	37250	1200	37250	3154.26
8	-	-	-	-	-	-	1.0E75

The obtained results confirm that the computation time increases with the increase of the number of hospitals or the number of groups. Also, it should be noted that for very small size instances, the model presented in Section III can be solved by using Cplex solver in very small CPU time. However, the instance 8 cannot be solved in maximum allowed CPU time of 1.0E75 sec (This equates to one and a half hours of elapsed physical time). This implies that the resolution of the presented model is affected by any change in instance input parameters.

The use of the integer linear programming approach to model the problem of concern helped us understanding its dynamic. But we are faced of the inherent complexity of such problems. Our model cannot be solved by exact methods for real size instances where we should deliver two heterogeneous products of eight different groups and different age to about forty hospitals during five periods. We need then to consider approximative approaches to cope with the combinatorial complexity of the problem and to come up with efficient solutions in a reasonable computing time.

V. Conclusion

In this paper, we modeled the optimization problem of the blood products supply chain as an IRP with the objective of minimizing the costs of storage, distribution, obsolescence, transfer and substitution. The studied products are heterogeneous and perishable, each with its own characteristics in terms of shelf life. Hence, the need to respect some constraints in the storage and distribution such as the separation of products and delivery in specific insulated containers. Also, no shortage will be permitted because of the criticality of these products. We conducted an experimental analysis of the model on eight instances of small size. Since the IRP is NP-difficult [2], we are now looking for a heuristic approach to solve this problem for life instances of realistic sizes. Also we plan include additional constraints such as the stochastic nature of the demand.

References

[1] Charleston, S.C.M.P.o.t., The Cost of Blood: Multidisciplinary Consensus Conference for a Standard Methodology", Transfusion Medicine Reviews, Vol. 19,pp. 66-78,2005.

[2] Coelho, L.C., J.-F. Cordeau, and G. Laporte, Thirty Years of Inventory Routing", Transportation Science, Vol. 48,pp. 1-19,2014.

[3] Federgruen, A., G. Prastacos, and P. Zipkin, An allocation and distribution model for perishable products.", Operations Research, Vol. 34,pp. 75-82,1986.

[4] Azadeh, A., S. Elahi, M.H. Farahani, and B. Nasirian, A genetic algorithm-Taguchi based approach to inventory routing problem of a single perishable product with transshipment", Computers & Industrial Engineering, Vol. 104,pp. 124-133,2017.

[5] Coelho, L.C. and G. Laporte, Optimal joint replenishment, delivery and inventory management policies for perishable products", Computers & Operations Research, Vol. 47,pp. 42-52,2014.

[6] Jia, T., X. Li, N. Wang, and R. Li, Integrated Inventory Routing Problem with Quality Time Windows and Loading Cost for Deteriorating Items under Discrete Time", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Vol. 2014,pp. 1-14,2014.

[7] Chen, Y.M. and C.-T. Lin, A coordinated approach to hedge the risks in stochastic inventory-routing problem", Computers & Industrial Engineering, Vol. 56,pp. 1095-1112,2009.

[8] Hemmelmayr, V., K.F. Doerner, R.F. Hartl, and M.W.P. Savelsbergh, Vendor managed inventory for environments with stochastic product usage", European Journal of Operational Research, Vol. 202,pp. 686-695,2010.

[9] Hemmelmayr, V., K. Doerner, R. Hartl, and M.P. Savelsbergh, Delivery strategies for blood products supplies", OR Spectrum, Vol. 31,pp. 707-725,2009.

[10] Rusdiansyah, A. and D.-b. Tsao, An integrated model of the periodic delivery problems for vending-machine supply chains", Journal of Food Engineering, Vol. 70,pp. 421-434,2005.

[11] Zanoni, S. and L. Zavanella, Single-vendor single-buyer with integrated transport-inventory system: Models and heuristics in the case of perishable goods", Computers & Industrial Engineering, Vol. 52,pp. 107-123,2007.

[12] Le, T., A. Diabat, J.-P. Richard, and Y. Yih, A column generation-based heuristic algorithm for an inventory routing problem with perishable goods", Optimization Letters, Vol. 7,pp. 1481-1502,2013.

[13] Niakan, F. and M. Rahimi, A multi-objective healthcare inventory routing problem; a fuzzy possibilistic approach", Transportation Research Part E: Logistics and Transportation Review, Vol. 80,pp. 74-94,2015.

[14] Mirzaei, S. and A. Seifi, Considering lost sale in inventory routing problems for perishable goods", Computers & Industrial Engineering, Vol. 87,pp. 213-227,2015.

[15] Diabat, A., T. Abdallah, and T. Le, A hybrid tabu search based heuristic for the periodic distribution inventory problem with perishable goods", Annals of Operations Research, Vol.,pp. 1-26,2014.

[16] Soysal, M., J.M. Bloemhof-Ruwaard, R. Haijema, and J.G.A.J. van der Vorst, Modeling an inventory routing problem for perishable products with environmental considerations and demand uncertainty", International Journal of Production Economics, Vol. 164,pp. 118-133,2015.

[17] Al Shamsi, A., A. Al Raisi, and M. Aftab, Pollution-Inventory Routing Problem with Perishable Goods, in Logistics Operations, Supply Chain Management and Sustainability, P. Golinska, Editor. 2014, Springer International Publishing. p. 585-596.

[18] Kochan, C.G., S.S. Kulkarni, and D.R. Nowicki, Efficient Inventorying and Distribution of Blood Product During Disasters", International Series in Operations Research & Management Science, Vol.,pp. 185-204,2016.

Inventory deployment with uncertainty on production and lead-times

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Jean Respen
Dootix Sàrl
1630 Bulle, Switzerland
Email: jean@dootix.com

Nicolas Zufferey
GSEM - University of Geneva
1211 Geneva 4, Switzerland
Email: n.zufferey@unige.ch

***Abstract*—An international Swiss company is facing a complex inventory-deployment problem where expensive items of different models must be dispatched to wholesalers to finally reach the shops. Perturbations are expected at two levels (namely, production and transportation), and efficient reactions must be implemented to face to these uncertainties. Fast and efficient solution methods are proposed to solve realistic instances.**

I. Introduction

In this paper, a three-level inventory-deployment problem (P) is considered by a luxury company (artificially denoted as *LC* because of a non-disclosure agreement). A push approach is imposed by *LC*: expensive items of different products/models are manufactured and then shipped to the final shops, while satisfying various constraints. The supply chain network is made of a single plant, various wholesalers and several shops. There is a dedicated wholesaler for each shop. As the shops are not the property of *LC*, they are considered as the final points of interest. Perturbations are encountered at the production-plan level (as some suppliers might be late in delivering the components or the raw materials to the plant) and at the delivery level (because of dynamic lead-times between the various links of the network). Such perturbations can lead to shortages at the shops level, where any shortage results in a lost sale. The planning horizon is a year and a time unit is a week. For each week, the main decision to make is the number of items of each product to send from the plant to each shop (through its dedicated wholesaler). The goal consists in minimizing the shortage penalties.

There is no literature dealing with such a three-level problem with unreliable suppliers and transportation uncertainties. Note by the way that most of the literature involves only two levels. In addition, the perturbation on the production plan can lead to an unfeasible solution, which raises the need for a specific repair procedure. In contrast with [27] but in line with the *LC* policy, an integrated approach combining the design of the production plan and the inventory deployment is not possible.

This study aims at proposing dedicated solution methods and a simulator to improve the inventory-deployment policy of *LC*, while keeping the imposed push approach from the plant to the shops (i.e., no order can be transmitted from a downstream level to an upstream level). The simulator is used to evaluate the quality of a solution according to the above-described random events. Relying on [19], this work adds perturbations on the lead-times in a context where the only function to minimize is the shortage penalty.

The paper is organized as follows. Section II presents a literature review. In Section III, the two different perturbations faced by *LC* are first exposed, and a formal description of (P) is then proposed. Section IV describes the different solution methods used to provide a solution S, and the simulator designed to evaluate the quality of a solution S. The experiments are presented in Section V. Section VI concludes the paper and identifies possible future works.

II. Literature review

With respect to the current study, the three next paragraphs focus on uncertain lead-times, simulation and multi-level supply chains, respectively.

A multi-objective approach based on genetic algorithms is proposed in [10] for a two-level assembly problem. Service level (to the final client) is maximized when uncertainty affects the lead-times. In [22], inventory-management approaches are presented when the lead-times are stochastic and vary in a seasonal fashion. The study is extended in [23] in the case the company has to manage a planned maintenance period. In [8], an exhaustive literature review on static stochastic lead-times is proposed (where the term ”static stochastic” refers to independent and identically distributed lead-times, with its consequence of order crossovers). The authors propose interesting results concerning the time between two consecutive receipts of orders. In [15], a mixed-integer non-linear programming model is depicted for a reverse logistics problem with stochastic lead-times. Whereas the standard formulation uses a mixed-integer linear programming model, the author states that the quality of this model can be improved with a queuing model. The latter adds non-linear relationships, and therefore a non-linear model is exposed. Small examples are solved with this model, and larger cases are tackled using a genetic algorithm. In [11], the lead-times (associated with the orders of a replenishment policy) are random, in addition to a random demand. The authors investigate the impact of such an assumption on the lost sales. In [4], the impact of stochastic lead-times is studied, as well as its effect on the bullwhip effect. In [17], a review of inventory models is proposed, where the impacts of stochastic lead-times are presented.

As exposed in the above-mentioned papers, simulation has to be employed to accurately evaluate the quality of the

solutions. With regard to simulation, the reader is referred to [3], [13], [25] for various surveys. An inventory-deployment problem is tackled with an optimization-simulation approach in [6], based on genetic algorithms. The robustness of the method is measured with various supply-chain settings. In [12], a reverse supply-chain problem is studied in the case defective products are shipped back to the plant (from the clients level). Again, the robustness of the proposed approach is measured with a simulator.

In a three-level context (involving suppliers, warehouses and retailers) accounting for transportation costs, an inventory-management model is presented in [16]. Both centralized and decentralized ordering models are compared. In [7], the authors consider a two-level supply chain made of one warehouse and many retailers facing different compound Poisson demand processes. The order-up-to-S replenishment policy is used. A similar problem is investigated in [1], where the shortage costs are evaluated with recursive procedures. Distribution systems with stochastic demands are considered in [2], where echelon stock ([5]) and installation stock ([14], [18]) are evaluated. In [21], a mathematical formulation is developed for a multi-stage inventory and production control problem with capacity constraints. In the context of two-level inventory deployment, three strategies and managerial insights are designed in [29]. In [24], the performance of different multi-level inventory-management strategies are analyzed.

III. Presentation of problem (P)

Two different perturbations can impact the production plan:

- some models are not produced (for instance, because some suppliers did not deliver some components on time);
- because of a lower production rate (for instance, because a worker was unavailable for half a day), the expected production is not met for some products.

Let $\hat{p}_t^i$ (resp. p_t^i) be the planned (resp. actual) number of items of product i to produce (resp. produced) in week t. In other words, p_t^i accounts for the perturbation. Each week, because of delays from the suppliers level, at most δ_1 percent of models are not produced (i.e., $p_t^i = 0$ for these models instead of $p_t^i = \hat{p}_t^i$). Based on the situation of *LC*, such a perturbation has an impact for two weeks (i.e., $p_t^i = p_{t+1}^i = 0$) and this hole in the production plan is filled the week after (i.e., $p_{t+2}^i = \hat{p}_{t+2}^i + \hat{p}_t^i + \hat{p}_{t+1}^i$). On the remaining $1-\delta_1$ percent of the products (which are manufactured), the risk to be slowed down is δ_2 percent for each model. Therefore, at most δ_3 percent of $\hat{p}_t^i$ is produced in week $t+1$ instead of t. Let $\hat{P}_t = \sum_i \hat{p}_t^i$ and $P_t = \sum_i p_t^i$. The total number of unproduced items (because of a reduced production performance) at the end of week k is $Q_t = \hat{P}_t - P_t$. A threshold (given data) of ε percent of units can be slowed down among the total number $\hat{P}_t$ of items expected to be produced at week t. It means that if $Q_t \leq \varepsilon \cdot \hat{P}_t$, *LC* does not react. In contrast, if $Q_t > \varepsilon \cdot \hat{P}_t$, then $C_t = Q_t - \varepsilon \cdot \hat{P}_t$ units have to be compensated (i.e., other products are manufactured instead of the expected ones). Two types of models are eligible to fill the production plan:

- the products that are not slowed down in week t and planned in week $t+1$;
- the products that were not planned for week t (i.e., $\hat{p}_t^i = 0$) but are expected to be manufactured in week $t+1$ (i.e., $\hat{p}_{t+1}^i > 0$).

For these eligible products, as long as the expected quantity C_t is not reached, some units (randomly selected) are produced in week t. *LC* proposes to have $\delta_1 \in [5, 15]$, $\delta_2 \in [45, 55]$, $\delta_3 \in [5, 15]$, $\varepsilon \in [5, 15]$.

As the transportation of the items to the wholesalers and then to the shops can suffer different perturbations, dynamic perturbations were added to the model. *LC* defined that standard lead-times (from the factory to the wholesalers and from the wholesalers to the shops) are in the set $L = \{1, 2\}$ (in weeks). Let $\hat{L}_t$ denote the expected lead-time (from the factory to a wholesaler or from a wholesaler to a shop) at week t, and let L_t be its corresponding actual lead-time (i.e., involving the potential perturbation). The two following probability distributions are used for the lead-times.

- if $\hat{L}_t = 1$, then $P(L_t = 1) = 0.6$, $P(L_t = 2) = 0.3$, $P(L_t = 3) = 0.1$;
- if $\hat{L}_t = 2$, then $P(L_t = 1) = 0.2$, $P(L_t = 2) = 0.6$ and $P(L_t = 3) = 0.2$.

Therefore, if the expected lead-times are one for both trips from the factory to the wholesaler and from the wholesaler to the final shop, one can compute that the probability $P(L_{tot} > 2)$ that the actual total lead-time exceeds two is equal to 84%, which is significant.

Problem (P) can be formally presented as follows. N different products are manufactured in the plant. M is the total number of wholesalers. There are W weeks in a year (set to 47 by *LC*). At the end of each week t, p_t^i items of model i are shipped from the plant. As imposed by *LC*, all the items are weekly pushed (no inventory is kept in the plant). A set J of shops must be supplied by the wholesalers level, and each shop j is associated with exactly one wholesaler m_j. In other words, we can have $m_j = m_{j'}$ even if $j \neq j'$. A priority $w_j \in \{1, 2, 3\}$ is assigned to each shop j, and w^{max} denotes the priority of the most important shops (which is 3). For week t, I_t^{i,m_j} is the on-hand inventory of wholesaler m_j associated with model i. Each wholesaler delivers to its associated shops and tries to avoid storage. Let $x_t^{i,j}$ be a decision variable representing the number of units of product i arriving at shop j at the beginning of week t from its dedicated wholesaler m_j. Let $\hat{L}_t^{m_j}$ (resp. $L_t^{m_j}$) be the expected (resp. actual) lead-time (in weeks) from the plant to the wholesaler m_j at week t. Let $\hat{L}_t^{j,m_j}$ (resp. L_t^{j,m_j}) be the expected (resp. actual) lead-time (in weeks) from the wholesaler m_j to the shop j at week t. In other words, both $L_t^{m_j}$ and L_t^{j,m_j} account for the perturbation. It is assumed in this project that for each week t, model i and shop j, the demand $d_t^{i,j}$ is known. When the $x_t^{i,j}$ units reach the shop j, its on-hand inventory is updated as in Equation (1).

$$I_t^{i,j} = \max(0, I_{t-1}^{i,j} - d_{t-1}^{i,j} + x_t^{i,j}) \quad (1)$$

In other words, the current inventory (period t) is computed as the previous inventory (period $t-1$) plus the current number of delivered items minus the previous demand. A solution $\hat{S}$ is formulated as $\hat{S} = (\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_t, \ldots, \hat{S}_W)$,

where $\hat{S}_t = (\hat{S}_t^1, \hat{S}_t^2, \ldots, \hat{S}_t^i, \ldots, \hat{S}_t^N)$ is the solution for week t for each model i. For each week t and each model i, $\hat{S}_t^i = (x_t^{i,1}, x_t^{i,2}, \ldots, x_t^{i,j}, \ldots, x_t^{i,J})$ is a vector of decision variables involving each shop j.

The objective function f to minimize is the expected shortage penalty (for each product i, week t and shop j). A shortage appears if the planned inventory is below the demand (i.e., $\hat{I}_t^{i,j} < d_t^{i,j}$). Formally, let $\hat{B}_t^{i,j} = \max(0, d_t^{i,j} - \hat{I}_t^{i,j})$ be the expected shortage amount of model i in shop j at week t, and $\hat{v}_t^{i,j}$ be a binary value which is 0 if $x_t^{i,j} > 0$, and 1 otherwise. Equation (2) accurately expresses f. It is the sum of the shortage penalties, and a shortage that has started recently costs less than a shortage involving a longer time period. If a unit of model i (assumed to be in shortage) arrives in shop j in week t, the corresponding shortage penalty is fixed to 0 from that time period, even if the resulting on-hand inventory is not able to cover the demand. In other words, as soon as a model is represented in a shop, its shortage penalty disappears. For this purpose, the notation t' is used to determine the shortage penalty from the time it begins to the current time, in order to penalize the involved shop until the missing product is again represented in the assortment. This sensitive point is part of the strategy employed by *LC*, which focusses on the presence or absence (in contrast with the number of units) of product i in shop j. This strategy is particularly relevant as it concerns high-value items.

$$f = \sum_{j=1}^{J} w_j \sum_{t=1}^{W} \sum_{t'=1}^{t-1} (t - t') \sum_{i=1}^{N} \hat{B}_{t'}^{i,j} \cdot \hat{v}_t^{i,j} \qquad (2)$$

IV. Proposed simulation-optimization approach

As mentioned above, because of the perturbations, simulation is the only way to measure the actual value of any expected solution S. A discrete-event simulator (as the perturbations on the production plan and on the lead-times have to be generated) is used for this purpose, for which the input data is S, the production plan, the initial inventory of each product (in each shop and each wholesaler), and the demand. The proposed managerial approach consists in three steps:

1) use a solver to generate a solution S;
2) evaluate S with simulation;
3) the concerned *LC* decision-maker can then manually edit and simulate again the solution, based on her/his expertise and the non-modeled elements.

The solver is made of a constructive heuristic, which can involve or not simulation. In the former case, the solver accounts for uncertainty. A simulation requires roughly four minutes on the used computer. Because of managerial reasons, *LC* imposes a time-limit of 30 minutes to the solver. Consequently, greedy constructive heuristics are proposed for (P), instead of refined constructive algorithms (e.g., various types of ant algorithms as proposed in [31]) or advanced metaheuristics (e.g., tabu search with a learning process as developed in [20]). Note however that the various learning mechanisms employed in the two previous papers might constitute an interesting avenue of research in the context of problem (P). In the remaining part of this section, various solution methods are first proposed, and the simulator is then presented.

We design four solution methods (namely, *M1*, *M2*, *LTM1*, *LTM2*). The first phase of each method consists in generating p (parameter tuned to 100) solutions, where each solution is built with the constructive heuristic *Construct* presented in [19], and provided again in Algorithm 1 in order to better address the below-proposed extension. Each step of *Construct* consists of assigning one item to a randomly selected shop in the set of the most promising shops, which is computed based on the shortage potential and on the priority of each shop. Ties are always broken randomly. As in [9], *Construct* relies on a technique to speed up the search process by restricting the decision space (which is particularly relevant when a small time limit is imposed). Formally, $A_t^{i,w}$ is the set of shops with priority w and a positive demand for model i in week t, and A contains all the shops with priority w^{max}. For each model i and each week t, the manufactured units plus the available items at the wholesalers level are considered as the available items. In each step of *Construct*, the algorithm assigns one available item to a randomly selected shop in $A_t^{i,w}$. But if the total demand of model i of all the shops are satisfied for the considered week t (i.e., if $\mathcal{C} = \emptyset$), the remaining items are sequentially assigned to the shops with priority w^{max}.

Algorithm 1 Greedy constructive method *Construct*

For each week t and each model i, **do**:
1) For each w, build the set $A_t^{i,w}$.
2) Set $\mathcal{C}$ as the set $A_t^{i,w}$ with the largest priority w.
3) **For** each available item u, **do**:
   a) Assign u to a randomly chosen shop of $\mathcal{C}$.
   b) If the demand of a shop of $\mathcal{C}$ is satisfied, remove this shop from $\mathcal{C}$.
   c) If $\mathcal{C} = \emptyset$, go to step (2). But if all the $A_t^{i,w}$'s are empty, set $\mathcal{C} = A$.

The solution method *M1* generates a set of p solutions with *Construct*, and provides the one with the best f-value. Extending it, *M2* generates also a set of p solutions with *Construct*, then simulates the k (parameter tuned to 5) ones with the best f-value, and finally provides the solution with the best simulated value. Based on the information provided in the previous section, if both lead-times (from the factory to the wholesaler, and from the wholesaler to the final shop) are expected to be one, then the probability that we face a worst case scenario (i.e., the sum of both lead-times is above four) is significant, as it amounts to 15.75%. Therefore, we propose *LTConstruct* (to denote *Construct* with lead-times considerations), which is an extension of *Construct* where worst-case scenarios regarding the lead-times are anticipated. The difference between *Construct* and *LTConstruct* is the following: each $A_t^{i,w}$ involves more shops in *LTConstruct*, as it also contains the shops which do not have any demand for the corresponding week t, but have a demand for weeks $t+1$ and $t+2$, and have both expected lead-times (i.e., from the plant to the wholesaler and from the latter to the shop) equal to one. This augmentation of the $A_t^{i,w}$'s allows efficiently anticipating some bad scenarios that could definitely occur. Heuristics *LTM1* and *LTM2* work similarly to *M1* and *M2*, respectively, but with *LTConstruct* as the underlying constructive procedure (instead of *Construct*). In contrast with [30], no solution method employs filtering techniques based

on incompatible elements.

The simulator, detailed in Algorithm 2, returns a value $f(S)$ of a solution after having added the different perturbations to the initially planned solution $\hat{S}$. If necessary, the solution is repaired. Indeed, after the production-plan perturbation, the solution $\hat{S}$ might be unfeasible because the $\hat{p}_t^i$'s could differ from the p_t^i's (i.e., some units are not manufactured, and other units are produced instead). Let f^j be the contribution of shop j to the objective function f. To repair a solution, f^j is determined for each shop j, and all the so-obtained f^j's are then sorted increasingly. If for a given week, some initially expected items do not exist, it means that a $x_t^{i,j}$ must be reduced. It is performed by removing items one by one from the shops with the lowest f^j-value. In contrast, if some items must be added (as the missed production has to be compensated), they are added one by one to the shops with the highest f^j-value. The perturbations to the lead-times cannot lead to an unfeasible solution, as the same set of items is managed. A single simulation-run is based on the average results over $sim = 40$ replications to be reliable.

Algorithm 2 Discrete-event simulator

1) **Generate the lead-times perturbations**. For each week/wholesaler/shop, generate the actual lead-times (plant $\rightarrow$ wholesaler, wholesaler $\rightarrow$ shop).
2) **Generate the production-plan perturbations**. For each week t, do:
   a) For each model i, do:
      i) Use δ_1 to know if the production of i is canceled.
      ii) If YES, produce i in week $t+2$ instead of weeks t and $t+1$.
      iii) If NO, use δ_2 to know if the production of i is slowed down. If YES, use δ_3 to select the percentage of slowed-down items, and postpone their production to week $t+1$.
   b) Use ε to know if the missed production must be compensated or not.
3) **Possibly repair the solution.** If the resulting solution is unfeasible, perform the following repair procedure based on the shops' priorities. Formally, for each week t and each product i, do:
   a) Compute the number $E_t^i = p_t^i - \hat{p}_t^i$ of items that does not match the production plan.
   b) In the case of non-existing items (i.e., $E_t^i < 0$), remove items (sequentially) from the shops with the smallest f^j-value (i.e., decrease some $\hat{x}_t^{i,j}$ values to generate the corresponding $x_t^{i,j}$).
   c) In the case of extra-produced items (i.e., $E_t^i > 0$), add items (sequentially) to the shops with the largest f^j-value (i.e., increase some $\hat{x}_t^{i,j}$ values to generate the corresponding $x_t^{i,j}$).
4) **Evaluate the resulting solution** S: compute $f(S)$.

V. Results

Tests were performed on an Intel Quad-core i7 @ 3.4 GHz with 8 GB DDR3 of RAM memory, with the imposed time-limit of 30 minutes. We first describe the way we have generated realistic instances (based on the input from *LC*), and then we present the conducted experiments for such instances. Note that informal tests were performed using CPLEX on small-deterministic instances, but it cannot deal (by far!) with the instance size and time limit fixed by *LC*. Thus, no further investigation is performed in this direction.

An instance is made of the following information: the market size (determined with the number N of products, the number M of wholesalers, and the number J of shops with their associated priorities), the lead-times, the perturbation parameters (i.e., δ_1, δ_2, δ_3, ε, and the lead-times distribution), the production plan, the demand (following a normal distribution), and the initial inventory level for each wholesaler and each shop. In the so-called *standard* instance, we have $N = 165$, $W = 47$, $J = 192$, $M = 12$, $w^{max} = 3$, $\delta_1 = 10$, $\delta_2 = 50$, $\delta_3 = 10$, $\varepsilon = 10$. 31 additional instances were generated from the *standard* instance by varying δ_1, δ_2, δ_3, and ε.

Table I shows the results for the proposed methods for different values of δ_1, δ_2, δ_3 and ε. Two different values are used for these parameters: the first is lower than the *standard* value, and the second is the *standard* value (i.e., the value proposed by *LC*). The goal is to compare the performances of the algorithms and to measure the impact of lower perturbations on the objective. The first four columns give the values of each parameter. The fifth column, called $f^\star$, is the best f-value returned by one of the algorithms for the considered instance. For each instance, the last four columns show the average (over 10 runs) percentage gaps of the solution value obtained by each method (namely, *M1*, *M2*, *LTM1*, and *LTM2*, respectively) with respect to $f^\star$. The last row shows average values.

One can remark that the first row is the instance with the lowest perturbations, and the last row is the instance with the largest perturbations (which is the *standard* instance). The results show that the methods can be ranked as follows: *M1* $<$ *M2* $<$ *LTM1* $<$ *LTM2*. Such results highlight on the one hand that the use of *LTConstruct* instead of *Construct* is beneficial (i.e., the anticipation of bad lead-times scenarios is helpful), and on the other hand that if a method accounts for uncertainty (i.e., if it uses the simulator as one of its components), then the results are improved.

VI. Conclusion and future works

In this paper, we have formulated and modeled a complex problem faced by a well-known Swiss luxury brand *LC*. Because of the significant size of the instances and of the time-limit constraints, we have proposed appropriate solution methods which efficiently combine constructive algorithms with the use of a simulator. Future works include: (1) the development of quick local search procedures or other constructive methods; (2) the consideration of additional perturbations to more accurately represent the supply chain of *LC*; (3) the integration of the job-scheduling decisions in the decision process, within an integrated logistics paradigm (e.g., [26], [28]).

TABLE I: Results (percentage gaps) on realistic instances

δ_1	δ_2	δ_3	ε	$f^\star$	***M1***	***M2***	***LTM1***	***LTM2***
5	40	5	5	118,455,640	0.39%	0.24%	0.27%	0.06%
5	40	5	5	116,009,071	0.32%	0.27%	0.17%	0.01%
5	40	5	10	116,319,227	0.30%	0.37%	0.22%	0.05%
5	40	5	10	114,378,271	0.48%	0.35%	0.21%	0.17%
5	40	10	5	126,394,561	0.33%	0.39%	0.22%	0.09%
5	40	10	5	124,527,362	0.41%	0.35%	0.21%	0.18%
5	40	10	10	123,427,873	0.42%	0.25%	0.18%	0.19%
5	40	10	10	116,243,452	0.48%	0.39%	0.22%	0.15%
5	50	5	5	109,385,858	0.40%	0.38%	0.11%	0.02%
5	50	5	5	118,228,656	0.30%	0.20%	0.15%	0.15%
5	50	5	10	122,884,718	0.33%	0.31%	0.15%	0.04%
5	50	5	10	122,655,917	0.46%	0.37%	0.27%	0.20%
5	50	10	5	127,615,936	0.42%	0.33%	0.20%	0.19%
5	50	10	5	116,941,971	0.50%	0.35%	0.27%	0.19%
5	50	10	10	113,940,549	0.43%	0.25%	0.13%	0.00%
5	50	10	10	114,251,147	0.34%	0.31%	0.24%	0.12%
10	40	5	5	120,605,238	0.39%	0.27%	0.11%	0.00%
10	40	5	5	124,630,881	0.49%	0.23%	0.24%	0.12%
10	40	5	10	125,276,585	0.42%	0.24%	0.12%	0.14%
10	40	5	10	119,964,402	0.49%	0.30%	0.22%	0.19%
10	40	10	5	117,811,209	0.44%	0.28%	0.11%	0.20%
10	40	10	5	114,319,331	0.33%	0.31%	0.29%	0.20%
10	40	10	10	116,297,081	0.34%	0.27%	0.17%	0.14%
10	40	10	10	118,306,872	0.36%	0.37%	0.16%	0.14%
10	50	5	5	123,513,148	0.48%	0.36%	0.25%	0.14%
10	50	5	5	120,230,897	0.38%	0.25%	0.24%	0.16%
10	50	5	10	114,486,963	0.34%	0.37%	0.24%	0.07%
10	50	5	10	121,170,295	0.48%	0.26%	0.11%	0.20%
10	50	10	5	122,206,444	0.30%	0.28%	0.24%	0.10%
10	50	10	5	112,475,111	0.41%	0.26%	0.16%	0.17%
10	50	10	10	126,955,027	0.33%	0.40%	0.25%	0.13%
10	50	10	10	108,497,299	0.40%	0.32%	0.20%	0.16%
Averages					**4.00‰**	**3.10‰**	**2.00‰**	**1.30‰**

REFERENCES

[1] S. Axsäter. Optimization of order-up-to-S policies in two-echelon inventory systems with periodic review. *Naval Research Logistics*, 40(2):245–253, 1993.

[2] S. Axsäter and L. Juntti. Comparison of echelon stock and installation stock policies for two-level inventory systems. *International Journal of Production Economics*, 45(1):303–310, 1996.

[3] J. Banks. *Handbook of simulation*. Wiley Online Library, 1998.

[4] D. C. Chatfield, J. G. Kim, T. P. Harrison, and J. C. Hayya. The bullwhip effectimpact of stochastic lead time, information quality, and information sharing: a simulation study. *Production and Operations Management*, 13(4):340–353, 2004.

[5] A. J. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490, 1960.

[6] J. Daniel and C. Rajendran. A simulation-based genetic algorithm for inventory optimization in a serial supply chain. *International Transactions in Operational Research*, 12(1):101–127, 2005.

[7] R. Forsberg. Optimization of order-up-to-S policies for two-level inventory systems with compound Poisson demand. *European Journal of Operational Research*, 81(1):143–153, 1995.

[8] J. C. Hayya, U. Bagchi, J. G. Kim, and D. Sun. On static stochastic order crossover. *International Journal of Production Economics*, 114(1):404–413, 2008.

[9] A. Hertz, D. Schindl, and N. Zufferey. Lower bounding and tabu search procedures for the frequency assignment problem with polarization constraints. *4OR*, 3 (2):139 – 161, 2005.

[10] F. Hnaien, X. Delorme, and A. Dolgui. Multi-objective optimization for inventory control in two-level assembly systems under uncertainty of lead times. *Computers & Operations Research*, 37(11):1835–1843, 2010.

[11] G. Janakiraman and R. D. Roundy. Lost-sales problems with stochastic lead times: convexity results for base-stock policies. *Operations Research*, 52:795 – 803, 2004.

[12] S. Kara, F. Rugrungruang, and H. Kaebernick. Simulation modelling of reverse logistics networks. *International Journal of Production Economics*, 106(1):61–69, 2007.

[13] A. M. Law and W. D. Kelton. *Simulation modeling and analysis*, volume 2. McGraw-Hill New York, 1991.

[14] H. L. Lee and K. Moinzadeh. Two-parameter approximations for multi-echelon repairable inventory models with batch ordering policy. *IIE Transactions*, 19(2):140–149, 1987.

[15] K. Lieckens and N. Vandaele. Reverse logistics network design with stochastic lead times. *Computers & Operations Research*, 34(2):395–416, 2007.

[16] A. Madadi, M. E. Kurz, and J. Ashayeri. Multi-level inventory management decisions with transportation cost consideration. *Transportation Research Part E: Logistics and Transportation Review*, 46(5):719–734, 2010.

[17] S. Minner. Multiple-supplier inventory models in supply chain management: A review. *International Journal of Production Economics*, 81:265–279, 2003.

[18] K. Moinzadeh and H. L Lee. Batch size and stocking levels in multi-echelon repairable systems. *Management Science*, 32(12):1567–1581, 1986.

[19] J. Respen, N. Zufferey, and Ph. Wieser. Three-level inventory deployment for a luxury watch company facing various perturbations. *Journal of the Operational Research Society*, http://dx.doi.org/10.1057/s41274-016-0136-9, 2017.

[20] D. Schindl and N. Zufferey. A Learning Tabu Search for a Truck Allocation Problem with Linear and Nonlinear Cost Components. *Naval Research Logistics*, 62 (1):32 – 45, 2015.

[21] A. Segerstedt. A capacity-constrained multi-level inventory and production control problem. *International Journal of Production Economics*, 45(1):449–461, 1996.

[22] E. A. Silver and N. Zufferey. Inventory control of raw materials under stochastic and seasonal lead times. *International Journal of Production Research*, 43:5161–5179, 2005.

[23] E. A. Silver and N. Zufferey. Inventory control of an item with a probabilistic replenishment lead time and a known supplier shutdown period. *International Journal of Production Research*, 49(4):923–947, 2011.

[24] A. Svoronos and P. Zipkin. Estimating the performance of multi-level inventory systems. *Operations Research*, 36(1):57–72, 1988.

[25] S. Terzi and S. Cavalieri. Simulation in the supply chain context: a survey. *Computers in Industry*, 53(1):3–16, 2004.

[26] S. Thevenin, N. Zufferey, and R. Glardon. Model and Metaheuristics for a Scheduling Problem Integrating Procurement, Sale and Distribution. *Annals of Operations Research*, http://dx.doi.org/10.1007/s10479-017-2498-z, 2017.

[27] S. Thevenin, N. Zufferey, and M. Widmer. Metaheuristics for a scheduling problem with rejection and tardiness penalties. *Journal of Scheduling*, 18 (1):89 – 105, 2015.

[28] S. Thevenin, N. Zufferey, and M. Widmer. Order acceptance and scheduling with earliness and tardiness penalties. *Journal of Heuristics*, 22 (6):849 – 890, 2016.

[29] D.-Q. Yao, X. Yue, S. K Mukhopadhyay, and Z. Wang. Strategic inventory deployment for retail and e-tail stores. *Omega*, 37(3):646–658, 2009.

[30] N. Zufferey. *Heuristiques pour les Problèmes de la Coloration des Sommets d'un Graphe et d'Affectation de Fréquences avec Polarités*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, 2002.

[31] N. Zufferey. Optimization by ant algorithms: Possible roles for an individual ant. *Optimization Letters*, 6 (5):963 – 973, 2012.

Lean Supply Chains

A new metric to evaluate the supply chain agility

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Selmen Boubaker
Laboratoire Génie Industriel, Centralesupélec, Université Paris Saclay, France
selmen.boubaker@centralesupelec.fr

Ziad Jemai
OASIS-ENIT University of Tunis El manar, Tunisia

Evren Sahin
Laboratoire Génie Industriel, Centralesupélec, Université Paris Saclay, France

Yves Dallery
Laboratoire Génie Industriel, Centralesupélec, Université Paris Saclay, France

Abstract— The business environment in which companies operate in becomes more and more dynamic, changing and unpredictable. Hence, supply chains are often asked to deal with sudden pressures and unplanned events, i.e. *situations* that need agility. With the increasing interest of the concept of agility, decision makers need practical metrics and indicators that enable them to assess supply chain agility. In this paper, we propose a new agility metric, the Supply Chain Agility Metric (SCAM), to measure the ability of a supply chain to respond to unplanned situations. Such a measurement can reflect the agile capabilities of a SC and boost to make the needed improvements.

Keywords— Supply chain agility, Agility Evaluation, agility metric, Situations needing agility,

I. Introduction

In today's extremely changing and volatile market landscape, supply chains are often exposed to growing challenges. With the growing competency, dealing rapidly and effectively with unpredictable events and situations is a survival obligation and a key to a high performance. Therefore, the agility, defined as the ability to rapidly respond to changes in market and customer demands [1-3], become a highly-recommended competency for supply chains. The literature related to supply chain agility has been largely increased in the two last decades. Starting by defining the concept and its relation with supply chain performance [5, 6], researchers defined the characteristics of agile supply chains [3, 7].More lately, many works have been published proposing approaches to measure the supply chain agility. Due to the high complexity of the external environment and the large number of related parameters, such a measurement remains very complex.

The aim of our study is to propose a generic and practical agility metric that would enable to measure quantitatively the ability of a supply chain to respond to unplanned situations. Hence, this work is intended to constitute the first step of a more comprehensive SC agility assessment model that can be used for benchmarking and agility enablers improvement purposes.

The paper is organized as follows: we present first the existing literature related to the evaluation of supply chain agility, the basic concepts and proposed approaches. Second, we define a new metric to measure the agility level of the supply chain for any situation that needs agility. Finally, we give a conclusion and perspectives for further researches.

II. Literature Review

In this section, we present first, the basic concepts linked to supply chain agility literature. Second, we explain the different approaches proposed to evaluate agility in supply chains.

A. Basic concepts

Many researchers tried to define the factors related to the supply chain agility.

[3] defines the key characteristics of an agile supply chain that are responsiveness, competency, flexibility and quickness. Researchers later called them "supply chain agile capabilities", a term introduced by [8] as the building blocks for supply chain strategy and a source of competitive advantage. [9] defines the agility enablers of a system as technological (e.g., information integration) and structural tools (e.g., Customer/marketing sensitivity) and techniques needed so that the system can be able to adapt to changing market conditions. In the supply chain agility literature, researchers used the terms enables-attributes and attributes to refer to the same concept.

[6] defines the agility drivers as pressures of the external environment on supply chain boosting to improve their competitiveness and overall business performance (e.g., Customer requirement, technological innovation). In our study, we use the word *"situation"* to refer to the pressure of the external environment of the supply chain needing agility.

Based on the works of [1, 3], [10] draw a conceptual model (Figure 1) using drivers, enablers and agile capabilities to reach agile supply chain goals that were cost, time, function and robustness.

B. Approaches for supply chain agility evaluation

The literature proposed many approaches to evaluate the agility level in the supply chain. Most of these works are based on questionnaires answered by managers and experts. To avoid vagueness and for more objectivity, many of works used the fuzzy logic for the information gathering[10-13]. In these approaches, linguistic qualifications are translated to numerical distributions allowing to obtain quantitate evaluations. Other decision making methods such as AHP were also used to gather the information needed for the supply chain agility measurement. In this section, we present first methods using the fuzzy logic. Second, we present those using AHP models.

The first approach to measure the agility dedicated to the supply chain was proposed by [10] which introduces the agility supply

chain index. The index was calculated selecting 24 agile sub-attributes grouped into 4 main attributes: collaborative relationships, process integration, information integration and customer/marketing sensitivity. The attributes were related, in

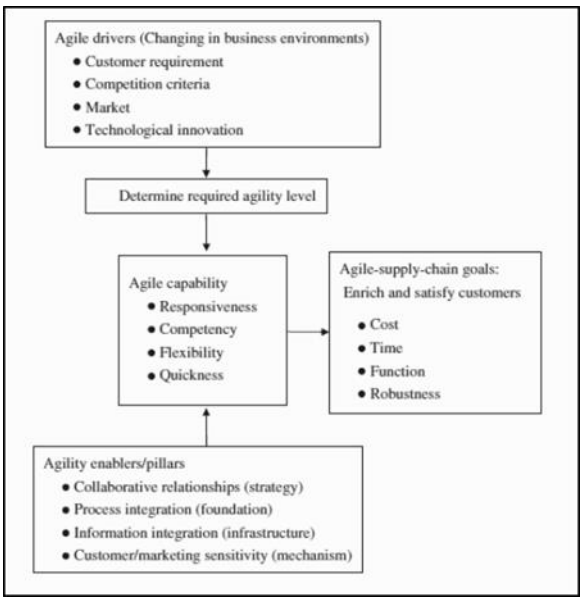


Figure 1: Conceptual model of SC agility [10]

one hand, to the four agility main capabilities mentioned by [7] : responsiveness, competency, flexibility and quickness, in the other hand, to the agile drivers. They used the fuzzy logic to approximate the managers' linguistic qualifications of attributes' ratings and weights. The FAI, fuzzy agility index is an aggregation of these fuzzy weights and ratings. Finally, The FAI is translated to an appropriate linguistic level which qualified the agility of the supply chain. [10] proposed also a fuzzy merit - importance index of each attribute to identify the main obstacles to improve the agility level.

Figure 2 present the framework used by [10]. The approach was applied to measure the supply chain agility in a Taiwan based international IT products company. The supply chain was labeled very agile. Analyzing the fuzzy merit index of each attribute, they identify the factors requiring most urgent improvements that were related to the decision-making facilitation and the demand capture. This approach permitted to have a good supply chain agility evaluation for benchmarking and improvements evaluation. It gives also the urgent points needing improvements to reach a better agility level but the list used by [10] was not exhaustive of all attributes related to the supply chain agility. This list was updated with [14].They used the same approach but with a more exhaustive list of attributes related to the supply chain agility: 86 SC attributes related to five SC main-attributes (Virtual enterprise organization, collaborative relation-ship, strategic management, knowledge and information technology management, customer and market sensitivity). The approach was applied on an Indian organization which was highly agile but the analyses of the attributes shows that many improvements can be done.[15] used the same approach but they choose the trapezoidal fuzzy number set in place of the triangular set used by [10] and [14] and the approach was applied to a large-scale auto-mobile manufacturing company in India. Choosing the right fuzzy set remain a research area for mathematicians. [16] compared the response of the system with various membership functions used for the input and the output of variables and conveyed that the Triangular membership function was superior.

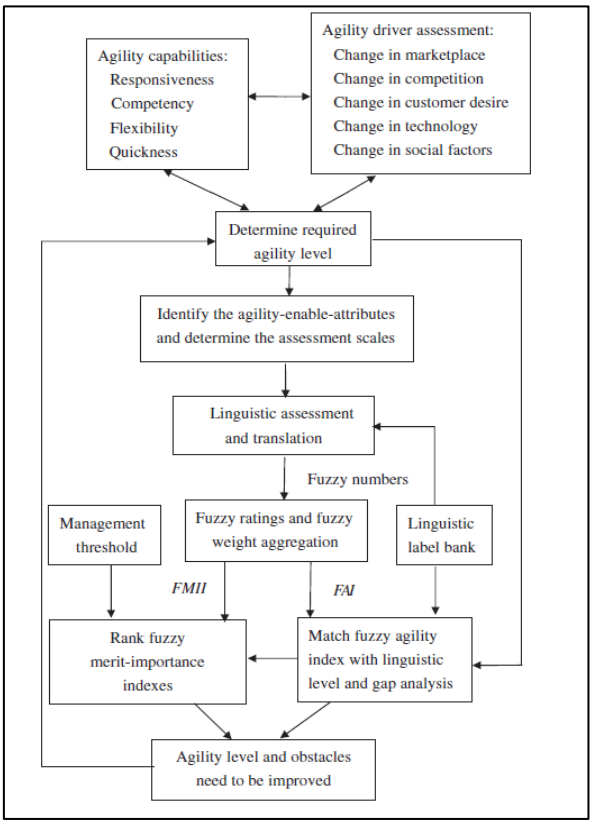


Figure 2:Framework of CS agility evaluation [10]

[4] proposed an agility matrix (Figure 3) to evaluate the level of agility in the supply chain using an adaptive neuro-fuzzy interference system, an approach that integrates both neural networks and fuzzy logic principles. The position of the supply chain in the matrix is determined by the level of readiness calculated with the fuzzy logic using the SC attributes and the and the level of capability (calculated with the fuzzy logic using the SC capabilities). This approach permit to have a qualitative qualification of the agility level in the supply chain using a very complex approach and it does not reveal the improvements ways to have a better agility in the supply chain.

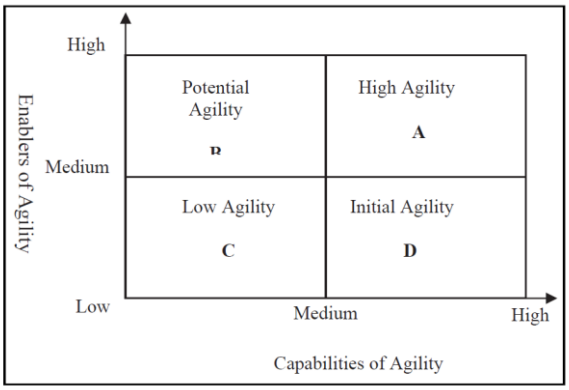


Figure 3: The agility matrix [4]

[17] proposed an approach to measure the performance of agility based on the SCOR model and using the fuzzy logic for the data gathering. For each of the basic processes in the supply chain in the SCOR model (Sourcing, Making and Delivery), They gather the linguistic qualifications of the Capabilities, the providers, performances and final goals of the supply chain related to the agility. Using the data envelopment analysis, they calculate an interval efficiency score for each process and for the global supply chain reflecting the level of agility.

[13] developed an approach based on association rule mining ,which is a rule-based machine learning method for discovering interesting relations between variables in large databases, to evaluate the supply chain agility. they defined 8 agility attributes: Flexibility, Profitability, quality, innovation, proactivity, speed of response, and robustness. Using the fuzzy logic, they converted the rates of each attribute to a convenient unidimensional framework, studying the correlation between the attributes using a rule mining algorithm they generate if-then rules between the eight used attributes allowing to determine analytical evaluation of agility in the supply chain.[12] applied the same approach in an Indian electronic manufacturing organization. Resulting with 174 rules helping the decision makers to evaluate agility in supply chains in an 'affective manner'. Contrary to the other methods presented in this paragraph, this approach doesn't result with a global evaluation of the supply chain agility but it gives rules linking the agility attributes. For example, one of the rules identified by [12] was : 'when a supply chain has medium flexibility AND medium speed of response then we can conclude that the supply chain high profitability, innovativeness AND medium robustness.

[18] proposed an approach based on complex network theory. Using AHP method, they calculated the weights of number of indicators related to the agility in the 3 type of nodes in the supply chain: Suppliers, Manufacturers and Distributors. They used Delphi method, Data mining to get initial agility evaluation matrix with the weight of indicators. They calculated the agility membership degree of every indicator to finish with the supply chain network evaluation, calculated considering the weights of each of the 3 nodes in the supply chain. Figure 4 present the different steps leading to the evaluation results.

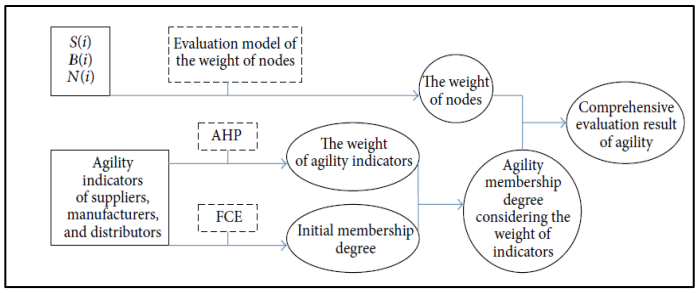


Figure 4:Steps of the proposed approach [18]

In most of papers existing in the literature, researchers have underlined the fact that the measurement of SC agility is a complex task [10, 19]. This is mainly due to the vagueness of the concept of agility, the identification of various situations that need agility and the quantitative measurement of parameters related to these situations. Furthermore, most of approaches developed so far are based on questionnaires answered by managers and specialists rating the level of application and weights of enablers related to SC agility. Due to the vagueness of the concept, such qualitative judgements can be criticized. In this study, we propose a more relevant and factual agility metric that would enable to measure the ability of a supply chain to respond to unplanned situations. As explained in Section 3, this work can be considered as a first step for a more comprehensive SC agility assessment model.

III. Proposed SC Agility Metric (SCAM)

As reported in various studies [10, 12, 13], practitioners are interested in evaluating the agility of their SC. For this purpose, we propose a 4-steps SC agility assessment model. It would mainly consist in:

1. Identifying the various situations that require SC agility and the associated enablers
2. Defining a SC agility metric (SCAM) to assess the agility related to each situation
3. Aggregating the (SCAM)s related to situations identified in step 1.
4. Assessing the (overall) SC agility level

Indeed, a supply chain faces various sudden situations that can affect its competitive capabilities and market position. Hence, agility is often defined as the ability to quickly respond to changes caused by these situations [3]. The response to these situations is made through the activation of various agility enablers such as the reduction of fixed planning horizons, the reduction of supply, production and distribution lead times, the acceleration of decision-action loops, etc. Thus, a KPI measuring the Supply Chain agility must be able to consider the heterogeneity of situations that need agility and measure quantitatively the "ability" of the Supply Chain to responds to each situation by activating the appropriate agility enablers.

Such a metric is expected to give a factual assessment of the readiness, flexibility, speed and responsiveness of the SC regarding the occurrence of a specific unplanned situation. Having identified the various situations and calculated the agility metric related to each situation, an aggregation model can then be used to calculate the overall SC agility.

The present paper focuses on the first two steps of the SC agility assessment model introduced above. We first propose a classification of situations faced by supply chains and then, we define a new SCAM measuring the ability of a supply chain to respond to these situations.

A. Situations requiring SC agility

The literature on situations requiring SC agility is broad and proposes various situations [20-22]. [23] identifies up 109 situations classified within 13 groups based on the type of the change caused by the situation. The groups are:

- Uncertainty in demand
- Product life cycle
- Marketing
- Gaining new market/tenders/customers
- Introduction of new distribution channels
- Uncertainty in the internal back-ground of the company
- Fluctuation of exogenous factors
- Competition
- Uncertainty / fluctuation of suppliers
- Uncertainty / fluctuation in production
- Uncertainty / fluctuation in logistics
- Uncertainty / fluctuation in transport
- Uncertainty in IT

When we examine the different situations in terms of impacts generated on the supply chain, we can distinguish two types of impacts. A first group of situations generates a change in quantity (e.g. Consumer demand fluctuation, exchange rate fluctuation) while a second group is related to a change in time (e.g. Shortened product lifecycle, increased environmental protection). For these two dimensions (quantity and time) situations can affect either a global variation or a change in the repartition between two or more products, markets, distribution channels... etc. (Figure 5)

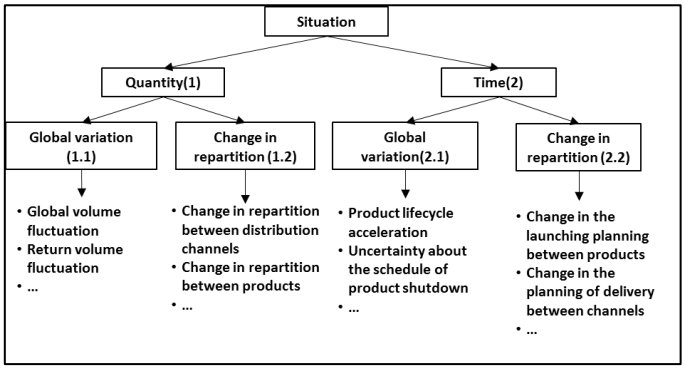


Figure 5: Types of situations

B. Proposed SCAM related to a situation

As situations affect the supply chain in terms of (a variation in) quantity and / or time, the SCAM must be able to quantify the time needed by the supply chain to respond to a variation in a parameter (measured in quantity or time) that is related to a situation. Therefore, we propose the following definition: the SCAM is the number of days required to the SC, to be prepared or/and to respond to a variation of a parameter (expressed in quantity or time) related to a situation by activating the appropriate enablers. The variation can be of a certain amplitude, expressed in percentage or in absolute. The amplitude of the variation is to be fixed to correspond to the specifications of the evaluated supply chain and the agility level targeted. (e.g., the time needed to prepare a variation of 20% of global demand, the time needed to prepare a delay of one month of supplier delivery).

One has to notice that the 11th reversion of the supply chain operations reference model (SCOR) reserved a part for the supply chain agility. Hence, the "upside supply chain flexibility" is defined here as the number of days needed to achieve a sustainable 20 % increase in quantities delivered. We argue in this paper, that the proposed SCOR agility indicators are limited to measure agility in situations related to demand fluctuations only. Hence, the SCAM proposed in this paper extends the measure of agility to various other situations (including demand fluctuations).

The time we measure is the time needed to anticipate the coming event and to be prepared to respond to the situation by achieving the adequate organizational and technical improvements.

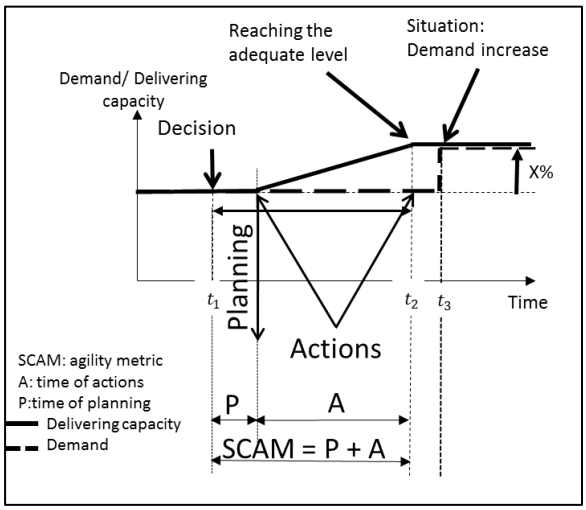


Figure 6: calculation of the agility indicator in the case of detected situation

As an example, in Figure 6, we present a situation characterized by a sustained increase of x% of the demand volume. At t_1 , we expect an increase of demand starting at t_3. The time of planning (P) correspond to the time needed to study and plan the actions allowing to maintain the service level of the SC, while the time of actions (A) correspond to the time of execution of the planned improvements (i.e. recruit/hire/ additional labor if needed, acquire the needed manufacturing

equipment, negotiate contracts with suppliers and distributors, …etc.). The (SCAM) corresponding to the situation is so the time between the detection of the eventual event (t_1) event and (t_2) the time the SC reaches the adequate delivering capacity allowing to satisfy the new demanded quantity. As a first attempt, the (SCAM) focus just on the time needed to improve the capacity allowing to respond to the situation and we don't examine the way of evolution of concerned parameters (Uniform, asymmetric …).

IV. Discussion and conclusions

With the high complexity of the market, the huge competency, and the rapid change of the economic and geopolitics environment, supply chains are often asked to deal with sudden external pressures. Thus, the agility, defined as the ability of a system to respond effectively to unplanned situations, become a competitive advantage and a strategic axis. Enterprises, nowadays, are investing in technological and organizational tools permitting to have agile supply chains. Therefore, supply chain managers and decision makers need to dispose of practical metrics permitting to evaluate this ability within their supply chains.

Many approaches have been proposed in the literature trying to measure the agility level of supply chains. Most of methods are based on questionnaires done with the supply chain managers and specialists where they were asked to evaluate agility enablers: their weights and their application level within the supply chain. Those estimations were used with aggregation models to give a general evaluation of the supply chain agility. Such a measurement is dependent on managers' judgement. With the vagueness that still characterize the concept of agility, these approaches can be criticized.

In this study, we propose a factual agility metric that would enable to measure the ability of a supply chain to respond to unplanned possible situation by evaluating its impact, either as a variation in quantity or time. This metrics helps to:

- Measure the SC agility facing a critical situation with the existing configuration, parameters and strategies,
- Reveal the impact of agility enablers' activation on the supply chain performance.

Moreover, we propose an approach allowing to evaluate of the agility level of the supply chain based on the aggregation of the (SCAM)s related to the most critical situations that can occur. That approach can lead to a good KPI for benchmarking and strategy monitoring. Therefore, we aim, as a perspective, to develop this approach with a real case illustration, identify the most critical situations for the evaluated SC and use the (SCAM)s related to each situation with the suitable aggregation method to obtain an overall assessment for the supply chain agility.

References

1. Sharp, J.M., Z. Irani, and S. Desai, *Working towards agile manufacturing in the UK industry.* International Journal of Production Economics, 1999. **62**(1-2): p. 155-169.
2. Christopher, M. and D. Towill, *An integrated model for the design of agile supply chains.* International Journal of Physical Distribution & Logistics Management, 2001. **31**(4): p. 235-246.
3. Christopher, M., *The Agile Supply Chain: Competing in Volatile Markets.* Industrial Marketing Management, 2000. **29**(1): p. 37-44.
4. Hosseini, S.S.M., et al., *A novel framework for agility assessment in supply chain considering enablers and capabilities.* Kuwait Journal of Science & Engineering, 2010. **37**(2b): p. 161-182.
5. van Hoek, R.I.H., A.; Christopher, M., *Measuring agile capabilities in the supply chain.* International Journal of Operations & Production Management, 2001. **21**(1-2): p. 126-147.
6. Yusuf, Y.Y., et al., *Agile supply chain capabilities: Determinants of competitive objectives.* European Journal of Operational Research, 2004. **159**(2): p. 379-392.
7. Giachetti, R.E., et al., *Analysis of the structural measures of flexibility and agility using a measurement theoretical framework.* International Journal of Production Economics, 2003. **86**(1): p. 47-62.
8. Morash, E.A., *Supply Chain Strategies, Capabilities, and Performance.* Transportation Journal, 2001. **41**(1): p. 37-54.
9. Gunasekaran, A., *Agile manufacturing: Enablers and an implementation framework.* International Journal of Production Research, 1998. **36**(5): p. 1223-1247.
10. Lin, C.T., H. Chiu, and P.Y. Chu, *Agility index in the supply chain.* International Journal of Production Economics, 2006. **100**(2): p. 285-299.
11. Vinodh, S. and S. Aravindraj, *Evaluation of leagility in supply chains using fuzzy logic approach.* International Journal of Production Research, 2013. **51**(4): p. 1186-1195.
12. Vinodh, S., N.H. Prakash, and K.E. Selvan, *Evaluation of agility in supply chains using fuzzy association rules mining.* International Journal of Production Research, 2011. **49**(22): p. 6651-6661.
13. Jain, V., L. Benyoucef, and S.G. Deshmukh, *A new approach for evaluating agility in supply chains using Fuzzy Association Rules Mining.* Engineering Applications of Artificial Intelligence, 2008. **21**(3): p. 367-385.
14. Vinodh, S. and M. Prasanna, *Evaluation of agility in supply chains using multi-grade fuzzy approach.* International Journal of Production Research, 2011. **49**(17): p. 5263-5276.
15. Samantra, C., et al., *Agility appraisal for integrated supply chain using generalized trapezoidal fuzzy numbers set.* International Journal of Advanced Manufacturing Technology, 2013. **68**(5-8): p. 1491-1503.

16. Zhao, J. and B.K. Bose, *Evaluation of membership functions for fuzzy logic controlled induction motor drive.* 2002. **1**: p. 229-234.
17. Khalili-Damghani, K. and M. Tavana, *A new fuzzy network data envelopment analysis model for measuring the performance of agility in supply chains.* International Journal of Advanced Manufacturing Technology, 2013. **69**(1-4): p. 291-318.
18. Xu, N.R. and J.B. Liu, *Research on Evaluation on Agility of Agile Supply Chain Network Based on Complex Network Theory*. Mathematical Problems in Engineering, 2015. **2015**: p. 10.
19. Fayezi, S., A. Zutshi, and A. O'Loughlin, *Understanding and Development of Supply Chain Agility and Flexibility: A Structured Literature Review.* International Journal of Management Reviews, 2016. **00**: p. 1-30.
20. Iskanius, P., *An agile supply chain for a project-oriented steel product network*, in *the Faculty of Technology*. 2006, University of Oulu.
21. Jain, V., L. Benyoucef, and S.G. Deshmukh, *What's the buzz about moving from 'lean' to 'agile' integrated supply chains? A fuzzy intelligent agent-based approach.* International Journal of Production Research, 2008. **46**(23): p. 6649-6677.
22. Tseng, Y.H. and C.T. Lin, *Enhancing enterprise agility by deploying agile drivers, capabilities and providers.* Information Sciences, 2011. **181**(17): p. 3693-3708.
23. Hüsgen, A.L., *Technical report ,Developing Context-Specific Supply Chain Agility : Concept and case studies of situations and enablers of agility*. 2015, Centrale Paris.

Sustaining lean through lean experts' roles

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Florian Magnani,
IOM Department &
LCFC Laboratory
ESCP Europe, Paris,
France & Arts et Métiers
ParisTech, Metz, France
florian.magnani@edu.escpeurope.eu

Ali Siadat
LCFC Laboratory
Arts et Métiers
ParisTech
Metz, France
ali.siadat@ensam.eu

Emmanuel Caillaud
ICUBE Laboratory
Université de Strasbourg
Strasbourg, France
Emmanuel.caillaud@unistra.fr

Olivier Gaudichau
Lean expert and manager
PSA Group
Paris, France
Olivier.gaudichau@mpsa.com

Abstract— Lean practices, firstly described through the Toyota Production System, are constantly evolving and spreading in various organizations. Though lean practices have received great attention, the difficulty lies in establishing a common definition of thoses practices and in its transference. Lean representatives, or lean experts, who function as the safeguards of lean practices, usually drive this process. Lean experts act in order to transfer lean expertise but also to anchor adoption of lean into the organizational context. However, in the existing research, the characteristics and the roles of lean experts are not clearly stated: are they teachers, facilitators or technical experts? The present paper aims to define lean expertise with particular attention to its human dimension and to depict lean experts' roles in relation to adoption levels of the environments in which they intervene. In order to do so, this exploratory study is built on a qualitative method and references three company departments at various stages of lean adoption.

Keywords— Lean systems, Management of Competences, Expert's role, Qualitative study

I. Introduction

This Lean research has become a global movement [1], but researchers and practitioners are still working towards getting along with the adoption phase. Successful lean adoption is often limited, due in part to an enterprise's failure to fully perform lean practices on a daily basis. For example, the learning practices that lead to continuous improvement - versus radical improvement – seem to be as significant as the practices that drive an organization's profit motives. Educating all members of an organization is a common feature of lean adoption practices [2]. Through experience-based learning activities, lean experts intend to create environments that facilitate the furthering of their employees' education, meanwhile, developping their employees' capabilities in order to deepen their knowledge of their work and of strategies geared towards improving it. Lean experts drive transference of these lean behaviors by engaging employees in the collective improvement practices featured in lean adoption [3]. However, discussions about the nature of these relationships between experts and employees are often absent in the lean literature, on both individual and group levels. Observations call for exploring the roles of individual actors and groups in decision-making models, which is, in our case, lean adoption. Few studies explore the behavioral dynamics of actors within operation systems [4].

Reading through the lean literature, we found that there is a surprisingly large variety of conceptual models which have been taken when referring to issues related to lean adoption. Lean has evolved from a set of methods to a complex sociotechnical system [5]. This evolution may explain the difficulties behind maintaining momentum during a lean adoption. More recent research discusses the direct influence that internal lean experts have on the adoption process [6], [7], during which they are responsible for modifying and maintaining individual and group behaviors within the company. Some studies focus on the influence these representatives' formal roles have on the initiation of the adoption process; meanwhile, they largely overlook the influence of their informal roles within organizational networks during adoption. In addition, lean experts' roles are not clearly defined: are they teachers [8], facilitators [7] or technical experts [9]? When roles are abstract, it makes it difficult to outline the expected competencies related to lean expertise. Questions concerning the criterion needed to properly gauge the levels of lean expertise remain unanswered. This study aims to identify lean experts' roles and focus on how these roles affect the company's success in lean adoption.

After reviewing the existing literature on lean expertise, its related competencies, and lean experts' roles, we outlined our research objectives and the methodology used to conduct our research. The exploratory aim of our research allowed us to present preliminary results so as to build theory on the human (i.e soft) dimension of lean expertise.

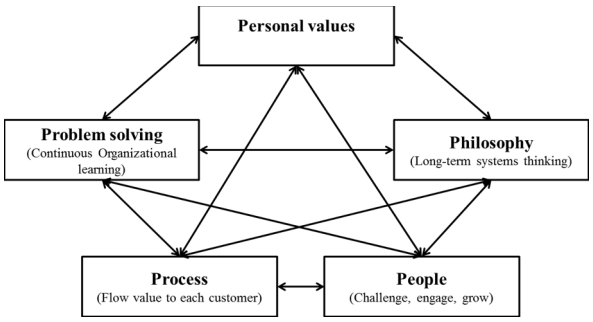
II. Literature review

A. Lean expertise

Researchers usually define lean as a complex organizational innovation, one that had the same strategic significance as that of Taylorism but has since undergone a

change and is now studied as its own separate type of innovation [10], [11]. First, lean was viewed exclusively as a technical system differing from mass production by its work design, flow improvement and quality assessment. Then, relationships between lean and socio-technical theory were found [5], [10]. Resulting research asserts that lean's social and technical subsystems interact simultaneously to improve both the operational and financial benefits of an organization [12]. Liker's model [13] is often referred to by researchers and practitioners to depict lean as a socio-technical system (see Fig.1). The technical sub-system has been extensively studied, which explains why research is able to provide a sound definition of the required technical expertise of lean (such as Pull system, Quality Management, Productive maintenance, Elimination of Waste). However, research on the social subsystem is oftentimes reduced to human resources practices at an organizational level [14]. In fact, research on Operation Management is often conducted solely on an organizational level [4], leaving the soft features of the social subsystem at an individual level underestimated.

Fig. 1. Components of a lean system (adapted from Liker & Ross, 2016)



Looking at expertise research, significant advances have been made and have discredited the myth that expertise is innate for gifted people. Numerous studies have documented that expertise is learned over time through experience and/or practice and emphasize that deliberate practice is a key element in developing sound expertise [15]. In lean literature, deliberate practice systematically pushes a person beyond their current level of performance [13]. Rother [16] conceptualized lean katas, or deliberate improvement practices, which were designed to develop new routines aligned with organizational purposes. These routines aim to develop new competencies, but such competencies have yet to be fully described.

B. Lean competencies

In many professional fields, certification helps researchers and practitioners to assess a particular set of competencies. Oftentimes, certification is obtained through a standard set of education, testing and experience requirements, typically administered by a central authority. What makes lean expertise assessment difficult lies in the diverse definition of lean practices and related competencies. For example, the specific requirements for certification vary from one company to another based on the company's idea and understanding of lean expertise. It engenders difficulties to compare levels of professionals and the effectiveness of training programs. Laureani and Antony [9] proposed a standard for Lean Six Sigma certification, mainly rooted in hard competencies. Lean and Six Sigma are often merged together which make it difficult to assess lean competencies as a single entity. Focusing solely on lean, we reviewed articles dealing with lean competencies and compared them to Liker and Convis' [17] lean competencies framework: Commit to self-development, Coach and Develop others, Support Daily Kaizen, Create Vision and Align Goals. These were the first of lean's soft competencies to be delineated. These competencies demonstrate that lean encompasses not only problem-solving methods, but also largely human interaction methods.

The "commit to self-development" competency does not appear in research: only when we studied articles dealing with Toyota did we see employees pushing themselves to continuously develop their knowledge, with specific regard to standardization and systematizing problem-solving skills [18]. Concerning Lean experts, they aim to transfer lean practices, which have two fundamental purposes: equip employees with the means to improve their tasks and to self-develop [8]. The difficulty lies in the way in which Toyota practices have intertwined hard with soft competency development. Problem solving, for example, is a means to developing awareness, gathering information, communicating and continuously improving. Ergo, human (i.e soft) competencies call for a more precise definition.

C. Lean experts' role

Current research is expanding the role of experts in organizations, whether it is to define the content of their knowledge, currently more interactional than technical [19], or to define the scope of their influence and action [15]. From a technical perspective, lean experts begin as process configurators and eventually become process controllers. In contrast, from a more interactional perspective, lean experts' roles are first and foremost about organizational change and then about organizational learning [20]. In this case, they share "sense makers" characteristics as they perform actions that propel the Lean adoption [21]. During a lean adoption process, lean experts often act as change agents while employees and managers act as the change recipients.

Current research contends that lean experts embody a "teacher" role, using every opportunity to develop people, engaging in knowledge transfer without giving the answers, and motivating people without being punitive [13]. They are usually called "sensei" [22] which is defined as "*a master of lean knowledge as a result of years of experience in transforming the gemba (i.e the environment in which they intervene)*". Compared to technical experts, a sensei's expertise is needed to solve open problems, meaning problems that are hard to define and lack a known solution. Faced with an open problem, these sensei have managers explore issues by tackling a number of smaller, closed problems.

Lean experts' working experiences are rooted in strong hands-on knowledge and their ability to identify teachable

moments when they arise. These experiences highlight the lean expert's role in developing "Knowledge work". Knowledge workers are those employees who are responsible for exploring and generating ideas and concepts rather than concentrating solely on implementing or managing existing processes. Generally, knowledge workers have high degrees of expertise, education and/or experience and the primary purpose of their job involves the creation, distribution or application of knowledge. Knowledge workers conduct non-routine problem solving practices that require a combination of convergent, divergent and creative thinking [23]. Oscillating between knowledge worker and change agent qualities, the lean expert's role in relation to adoption levels remains largely unstudied.

The current literature allowed us to describe lean technical competencies, though not always as clearly as espected, but still lacked a precise description of its soft competencies. By studying lean experts who embody these hard and soft competencies, we hope to gain a greater understanding of the influence that soft competencies have on the lean adoption process. In doing so, we aim to propose a framework of internal lean experts' roles in a company context related to lean adoption levels.

III. Methodology

Considering the lean adoption process within an organization as the result of a collective and cumulative learning process driven by employees [6], it seems appropriate to explore the behavioral role of individual actors within an organization. Taking lean experts as the starting point and having them transfer the necessary knowledge to the company [20], our research question is: what are the lean experts roles during the lean adoption? Through an exploratory approach, this article intends to fulfill the following two goals: (1) to contribute to the definition of lean expertise by clarifying lean soft competencies and (2) to depict lean experts' roles at different stages of the adoption process.

In order to study the specific soft competencies of lean experts and their roles during a lean adoption, we chose to study the interactions lean experts have with the employees of the organization. To do so, we decided to conduct an exploratory longitudinal study. According to Yin [24], the distinctive need for case studies arises out of the desire to understand complex social phenomena because case study methods allow investigators to retain the holistic and meaningful characteristics of real-life events. Case study research consists of a detailed investigation of these phenomena, within their context. Moreover, we conducted this case study in order to generate hypotheses and build preliminary theory. Since our research endeavor aims to develop new hypotheses about the transference of lean expertise, we opted for a case study research strategy involving triangulation among a variety of different sources of data [25]. These data include formal (21) and informal interviews with lean experts, analysis of focus groups (4 meetings), and on-site observations of lean experts' interventions. These data are part of a three-year immersion research in a single company starting in January 2014 and ending in January 2017.

The case considered in this study is PSA Group, an international automotive company, where lean adoption was part of a consistent strategy over a period of sixteen years. This process can be illustrated in five different phases: initiation, generation, diffusion, unification, and reconfiguration (cf. [6]). During these phases, the adoption's scope evolved as well as lean experts' roles. At first, they helped identify and align the best practices. Then, they offered training and support to the departments of interest. At that time, the network of lean experts was extensive: teams of lean experts started to form in the Manufacturing, Research & Development (R&D), Sales & Marketing (S&M), Program and Human Resources (HR) departments. In order to aggregate and create a vision around the next phase of the adoption process, the PSA Excellence System (PES) was constructed and a PES experts team was created at the corporate functions under the command of the CEO. Since this initial effort, the PES experts' team has been dissolved and the number of lean experts at PSA Group has diminished. In this study, we take a closer look at lean experts' roles and focus on the three remaining experts' teams attached to three departments: S&M, Manufacturing, and R&D. These three departments had various lean maturity levels, lean experts' profiles and intervention approaches. Through this research, we intend to delineate the lean expert's soft competencies and the role they play during the lean adoption process.

IV. Findings

A. Descriptive findings

Lean adoption at PSA Group has taken on various forms, through limited duration projects or through introducing a management system. The Manufacturing department could be considered proficient because its members arrive at an autonomous execution of the PSA Excellence System. The R&D department could be considered an advanced beginner as a result of its rigorous training, but only partial application and execution of lean methods. Finally, the S&M department could be seen as a novice due to the low number of employees who are not entirely convinced of the usefulness of the PSA Excellence System. These Departments' dedication in the lean adoption and their maturity levels were the foundations upon which we built our differentiation criteria. Change types and lean experts' interventions at PSA Group called for a need to focus on the role of lean experts.

In PSA's Manufacturing department, the decision to adopt lean was made in 2000 when an international survey showed that the company was far from achieving the same levels of success as its competitors. However, some traces of lean were found before 2000, such as Just-in-time and Quality Circles. Between 2002 and 2005, a partnership was formed with Toyota, giving PSA Group open-access to the features of its production system. Some employees from PSA Group were sent to their new partner's company to learn from its operating

Toyota Production System. They were then expected to report back to PSA Group and implement what they learned and observed. These employees became the lean experts and henceforth, took action in a more normative way than the others. This type of intervention was only possible when the employees became aware of the performance gap between PSA operating system and TPCA's (joint-venture between Toyota and PSA Group). They implemented a number of technical practices, as well as management practices such as management control and policy deployment. As more and more employees became convinced of the benefits of PES, these experts' roles transformed from mere implementers to the technical specialists and the safeguards of PES. This change could be explained by the emergence of a worldwide network of lean leaders, or regular employees that took the role of implementers, within PSA Group's factories.

In the R&D department, the Lean adoption began in 2002 when the partnership between the PSA Group and Toyota was confirmed. Some engineers went to work with Toyota to better understand the way they approached design and product conception and how the two were linked to production. They were trained by Toyota coordinators during the plant's conception and they brought back these competencies in order to transfer them to their company. This explains why most of their actions revolved around training and guidance in lean products and process development. Sometimes, they conducted surveys assessing adoption maturity and adherence of employees in order to decide which areas needed improvement. These engineers became the lean experts of the R&D department. They were mainly trainers and facilitators, meaning that the change mainly came from their employees, not from them.

In the S&M department, the lean adoption officially began in 2007 with the goal in mind to transfer lean practices to departments other than Manufacturing. A team of lean experts was created. Some experts worked alongside the sales representative network in order to improve the way they operate activities. Other lean experts assisted department directors in transforming the processes of which they were in charge. Today, these experts struggle to engage employees to embrace lean practices on a daily basis. The Quality Management System (QMS), already in place before 2007, is what employees adhere by today as opposed to PES. Even still, lean experts intervene through a push/pull change type, focusing mainly on visual management and standardized work as the entryway to lean practices in their environment. These experts essentially act as coaches or change agents in that they understand the current situation of their recipients and diagnose if lean practices can help them solve their daily problems and make changes according to this diagnosis and the willingness of the person concerned. They were the ones mainly using soft competencies to transfer lean practices.

B. Analysis

During the interviews, a clear dichotomy surfaced between institutional experts, sometimes self-proclaimed, and emergent experts, recognized as experts by peers. A correlation may exist between the institutional expert and their assertive interventions. Interestingly, emergent experts in the S&M department seem to have a greater impact on how employees perceive lean and its adoption. These experts' interventions oscillate between push and pull tactics depending on the environment in which they intervene. We were told that they act in a coaching manner when interacting with a manager: *"We do not want to go beyond what the manager wants to see accomplished within his department. Once we understand what the manager wants, with his help, we can carefully analyze his interactions with other employees and his work environment. When possible, we help the manager map out steps towards achieving his goals so that they have a positive impact on his interactions with his employees and his work environment."* (PES Manager at S&M, Personal interview in French, September 2016). They are often recognized as coaches because they only use their expertise to help managers find harmony between what they want to do in their environment and what they are allowed to do. They push the manager towards a more realistic observation of his own behaviors by asking key questions rather than giving advice or training. In turn, their soft competencies help transfer the appropriate technical practices which can then be adapted to the environment of the manager. In this case, these experts appear at a low maturity level in order to follow the actors' pace of appropriation and, accordingly, protect the adoption from deviation. They possess a high and extensive level of expertise including soft and technical competencies.

At an average maturity level, experts support the employees' development and transfer their knowledge in a Just-In-Time manner – the knowledge necessary at the right time and in the right amount for the receiver. Their actions attempt to bring about adoption initiatives while canalizing efforts of appropriation. These experts act as facilitators, encouraging alignment of individual practices around a collective dynamic of improvement. They possess a medium level of expertise, especially in regard to technical competencies. Finally, when employees reach high maturity levels, lean experts bring their own technical expertise in a way that develops their employees and gives some insight into how to run the system and keep it from deviating. Beyond reinforcing organizational alignment, their interventions provide individual expertise acquisition for employees. By transferring their knowledge and challenging employees, they themselves also improve. These mentors are in charge of ensuring that the rigorous performance of lean practices by employees is aligned with lean methodology, while also respecting the fact that employees are the experts of their own work. They often provide insight and new perspectives that expand employees' understanding of the practices and often give pointed feedback. In comparison with coaches that tend to focus more on generic issues, these mentors look at more targeted, job-specific issues. To do so, they possess a specific but high narrow level of expertise of hard competencies. When they are confronted with difficulties, they call on *"magicians"* who *"possess softer competencies"* (PES Manager at Manufacturing, Personal interview in French, July

2016) and who can more effectively transfer lean practices in more sensitive environments.

In addition, during the study, three focus groups were conducted with lean experts and permitted to synthesize the overall competencies that lean expertise regroups together. At PSA Group, the competencies are organized into three categories: behavioral competencies, job-related competencies and lean competencies. The following table gathers the general agreement about lean competencies:

TABLE I. Description of Lean competencies

Efficiency types	Related competencies
Operational	Just-in-time Jidoka Standardization Production Control Elimination of Muri, Mura, Muda
Organizational	Structuration of people organization Nominal management Improvement management
Relational	Problem resolution / PDCA[1] Human resource development Systemic interactions

The first set of lean competencies is related to operational efficiency: the technical competencies commonly found in the present literature [1], [11]. The second set of lean competencies is related to organizational efficiency: the soft competencies that leaders and managers possess and use to support routinized learning. They are closely associated with Liker and Convis' framework [17]. The last set of lean competencies is related to relational efficiency: the soft competencies that outline the steps towards building capabilities. This capability-building process was constructed as a combination of problem resolution from an individual perspective, human resource development from a group perspective, and systemic interactions from a system perspective. Using this set of competencies, the two aspects practitioners need to focus on are understanding (1) what the lean experts know and consequently do and (2) how they interact within the context of their interventions.

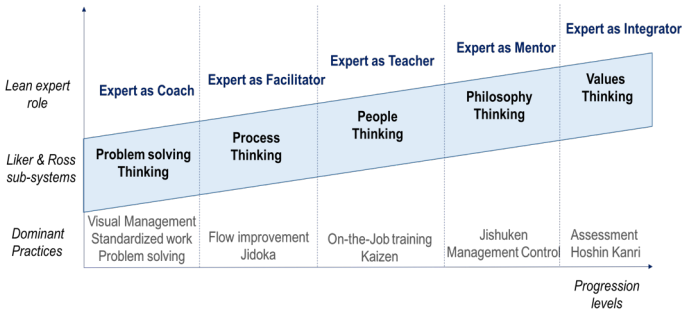
V. Discussion

The role of lean experts evolved according to the progression level of lean adoption and, indirectly, through the feedback given by managers (dominant and local perceptions). The role played by lean experts was crucial to understanding the Lean adoption process. We began by detailing an experts' typology based on their level of maturity. Lean experts' role can be summarized as follows: at a low level of maturity, they are coaches. At an average level of maturity, they are facilitators or teachers. At a high level of maturity, they are mentors. The level of lean soft competencies seems to be inversely related to level of progression. It can be explained by the dissemination of expertise throughout the organization by lean experts. Being the safeguards of lean expertise, lean experts are catalysts in successful lean adoption, passing on the responsibility to local managers and directors. We would like to add another role, one that is taken on by a sensei: the integrator role. The integrator role consists of two parts: the "critical observer" and the "reflective learner". It unlocks the potential of other roles by reinforcing their interrelatedness, while simultaneously making use of their distinctive features. However, the integrator role of the sensei, which is not a universal role, is supposed to give him the possibility of adaptation in any circumstance. His ability to adapt is dependent upon the environment in which he intervenes and whether the dominant lean practices are already established.

The numerous articles dealing with lean practices without coming to a common understanding of those practices proves the need for one specific definition. Failure to do so results in greater difficulties for practitioners to successfully carry out the adoption process. We then depict the evolution of dominant practices used by experts according to progression levels (see Fig. 2): all methods are not equal in terms of expected outcomes and they require soft or technical competencies, usually both. It seems important to consider the environment maturity (i.e. employees' and managers' willingness) before prescribing certain methods. Some issues concerning the positioning of experts arose. There remains a clear dissonance between actors and experts, between those who work and those who improve [3]. We notice that, at Toyota, there is no segmentation between work and improvement, such as actions and learning. There is also no explicit designation of experts and non-experts. For example, Taichi Ohno, the first lean expert, was given the title of expert because of his high involvement and actions which challenged the status quo of the organization while developing his associates. The job-related background of the experts, in addition to their own Lean expertise, seems to be underestimated in current literature.

Links between the expert and the manager is shown through the concept of applied knowledge development: Learn, Do, Self-learn, Teach. The lean expert, from his insight and understanding of his field, develops the learner, but also continue to develop himself through knowledge transfer mechanisms. The lean expert aspires to educate other employees who will eventually carry the knowledge, transfer it, enrich it, and, consequently, develop themselves. In order to do so, we propose in Fig.2 a depiction of the roles experts take on throughout the adoption process, building upon the subsystems proposed by Liker and Ross [13] :

Fig. 2. Linkages between experts' roles, experts' dominant thinkings and dominant practices



[1] Plan, Do, Check, Act

VI. Conclusion

Conducting this research, we proposed a framework of lean expertise so as to unveil the lean soft competencies, which will in turn help us to define the typologies of internal lean experts in a company context related to lean adoption levels. This research was based on a single case study, so there is still a need to generalize or test what this study brings and, thus clarify what lean expertise is and what we should expect from lean experts (competencies, attitudes) to foster lean adoption. Organization use lean experts as safeguards of lean expertise and catalysts to embrace the lean system in its entirety. This study shows that lean experts roles during the lean transformation is highly influential in the adoption's success.

Acknowledgment

We would like to thank the two anonymous reviewers for their insightful comments which were taken into account in the current version of this article. We would also like to thank all the lean experts from PSA Group who assisted us during our research project.

References

[1] N. V. K. Jasti and R. Kodali, "Lean production: literature review and trends," *Int. J. Prod. Res.*, vol. 53, no. 3, pp. 867–885, Feb. 2015.

[2] T. Fujimoto, *Evolution of Manufacturing Systems at Toyota*, 1st edition. New York: Productivity Press, 1999.

[3] A. C. Alves, J. Dinis-Carvalho, and R. M. Sousa, "Lean production as promoter of thinkers to achieve companies' agility," *Learn. Organ.*, vol. 19, no. 3, pp. 219–237, Jun. 2012.

[4] A. Wieland, R. B. Handfield, and C. F. Durach, "Mapping the Landscape of Future Research Themes in Supply Chain Management," *J. Bus. Logist.*, vol. 37, no. 3, pp. 205–212, Sep. 2016.

[5] G. A. Marodin and T. A. Saurin, "Implementing lean production systems: research areas and opportunities for future studies," *Int. J. Prod. Res.*, vol. 51, no. 22, pp. 6663–6680, Nov. 2013.

[6] F. Magnani, "Historical evolution of a Lean system: case study of the PSA Group," *Logistique Manag.*, vol. 24, no. 3–4, pp. 199–206, Oct. 2016.

[7] M. Maalouf and B. Gammelgaard, "Managing paradoxical tensions during the implementation of lean capabilities for improvement," *Int. J. Oper. Prod. Manag.*, vol. 36, no. 6, pp. 687–709, May 2016.

[8] J. Liker and M. Ballé, "Lean Managers Must be Teachers," *J. Enterp. Transform.*, vol. 3, no. 1, pp. 16–32, Jan. 2013.

[9] A. Laureani and J. Antony, "Standards for Lean Six Sigma certification," *Int. J. Product. Perform. Manag.*, vol. 61, no. 1, pp. 110–120, Dec. 2011.

[10] B. Dankbaar, "Lean Production: Denial, Confirmation or Extension of Sociotechnical Systems Design?," *Hum. Relat.*, vol. 50, no. 5, pp. 567–584, May 1997.

[11] J. P. Womack, D. T. Jones, and D. Roos, *The Machine That Changed the World: The Story of Lean Production*, Reprint edition (2000). New York: Free Press, 1990.

[12] N. P. Jayamaha, J. P. Wagner, N. P. Grigg, N. M. Campbell-Allen, and W. Harvie, "Testing a theoretical model underlying the 'Toyota Way' – an empirical study involving a large global sample of Toyota facilities," *Int. J. Prod. Res.*, vol. 52, no. 14, pp. 4332–4350, Jul. 2014.

[13] J. K. Liker and K. Ross, *The Toyota Way to Service Excellence: Lean Transformation in Service Organizations*. New York: McGraw Hill Higher Education, 2016.

[14] R. Shah and P. T. Ward, "Lean manufacturing: context, practice bundles, and performance," *J. Oper. Manag.*, vol. 21, no. 2, pp. 129–149, Mar. 2003.

[15] H. Collins and R. Evans, *Rethinking Expertise*. Chicago: University of Chicago Press, 2007.

[16] M. Rother, *Toyota Kata: Managing People For Improvement, Adaptiveness, and Superior Results*, 1st edition. New York: McGraw-Hill Professional, 2009.

[17] J. Liker and G. L. Convis, *The Toyota Way to Lean Leadership: Achieving and Sustaining Excellence through Leadership Development*, 1st edition. New York: McGraw-Hill, 2011.

[18] M. A. Maginnis, "The Impact of Standardization and Systematic Problem Solving on Team Member Learning and Its Implications for Developing Sustainable Continuous Improvement Capabilities," *J. Enterp. Transform.*, vol. 3, no. 3, pp. 187–210, Jul. 2013.

[19] A. Johri, "Impressions in action: the socially situated construction of expertise in the workplace," *J. Organ. Ethnogr.*, vol. 4, no. 1, pp. 44–63, Feb. 2015.

[20] C. Herron and C. Hicks, "The transfer of selected lean manufacturing techniques from Japanese automotive manufacturing into general manufacturing (UK) through change agents," *Robot. Comput.-Integr. Manuf.*, vol. 24, no. 4, pp. 524–531, Aug. 2008.

[21] K. E. Weick and R. E. Quinn, "Organizational Change and Development," *Annu. Rev. Psychol.*, vol. 50, no. 1, pp. 361–386, 1999.

[22] Lean Enterprise Institute, *Lean Lexicon: A Graphical Glossary for Lean Thinkers*, 5 Spi. Brookline, Mass: Lean Enterprise Institute,US, 2003.

[23] W. Reinhardt, B. Schmidt, P. Sloep, and H. Drachsler, "Knowledge Worker Roles and Actions—Results of Two Empirical Studies," *Knowl. Process Manag.*, vol. 18, no. 3, pp. 150–174, Jul. 2011.

[24] R. K. Yin, *Case Study Research: Design and Methods*, Fifth Edition edition. Los Angeles: SAGE Publications, Inc, 2013.

[25] J. W. Cox and J. Hassard, "Triangulation in Organizational Research: A Re-Presentation," *Organization*, vol. 12, no. 1, pp. 109–133, Jan. 2005.

A Leverage Potential of Chilled Distribution Center Operations

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Jirapat Wanitwattanakosol, Saran Jangnilwong and Sirikorn Santirojanakul
College of Arts, Media and Technology
Chiang Mai University
Chiang Mai, Thailand
jirapat.w@cmu.ac.th

Abstract— Modern trade in Thailand has leaded to a rapid and massive increase in stores, especially for a convenience store type. A chilled distribution center (CDC) is established to control quality, freshness, quantity and time in temperature controlled supply chain. An objective of this research is to propose enhancement operations in CDC by using management and information technology perspectives. The basic selected quality control tools were adopted to detect causes of wastage events. The Elimination, Combine, Rearrange, and Simplify (ECRS) concept was applied to determine efficient working steps in any process. A web database was built on agile software development for handling credit note documents. The results showed that the proposed approach could eliminate and simplify operations. Hence, operation times decreased from 710 minutes to 210 minutes. Moreover, the CDC could accurately monitor reverse logistics employee performances.

Keywords— chilled distribution center; credit note; lean; information technology

I. Introduction

Modern trade in Thailand has leaded to a rapid and massive increase in stores, as well as foreign direct investment (FDI) and a shift in term of the ratio of traditional and modern trade [1]. Moreover, the Thai retailing industry will continue to develop due to drastic competition among existing players, as well as the implementation of the ASEAN Economic Community set to begin at the end of 2015, which will expand the territory of retailing competition both in terms of opportunities and threats [2]. Reference [3] described four types of food retail stores as (1) hypermarkets or discount stores (selling space of 6,000 to 12,000 m^2), (2) cash and carry stores (membership-based wholesalers), (3) supermarkets (selling space of 800 to 1,500 m^2 with a focus on fresh food and grocery products) and (4) convenience stores (selling space of 30-300 m^2). The convenience store sells items and offers services. It is often located alongside busy roads in densely-populated areas.

The empirical study is motivated by a chill chain problem faced from a leading convenient store business in Thailand. Chill chain integrity disciplines are rigorously implemented from supplier to a chilled distribution center (CDC) and to retail shelf by considering three key issues as cost, partnerships and food safety and hazard analysis and critical control point (HACCP) [4].The CDC is a distribution center in temperature controlled supply chain. problem is to handle a credit note (CN) issue at a CDC. The CN document will be categorized in five types as switch, incomplete, surplus, lost and uncommon. The company's name has not been disclosed in order to protect the confidentiality. It is fictitiously referred to as a BBB company. This company has more than nine thousand branches covered all regions of Thailand. Convenience food stores offer a variety of fresh food, ready to eat meals, freshly baked bakery and brewed coffee and beverages. Nine CDCs are also established not only to control quality and freshness products, but also distribute specified quantity for each branch in the right time. The BBB's chill chain keeps two to six degree Celsius temperature controlled with ten day shelf life.

One of the CDCs is located in Chiang Mai province, where is in the Northern part of Thailand, has responded to handle eight hundred and eighty two branches based on Northern fourteen provinces such as Chiang Mai, Tak, Phare, Lampang, etc. If products are shipped to the branch, some products will be returned to the CDC with CN documents. Reverse logistics department plays an important role to consider the status of a CN document. These CN products are sold with discount price at the CDC. Traditional wastes and information management wastes are detected in the CDC operations.

This research aims to propose enhancement operations in CDC by using management and information technology (IT) perspectives. An academic motivation is to apply existing and well-known concepts for a real case. The rest of this paper is organized into different sections. Traditional wastes and information management wastes are reviewed based on the lean principle in Section 2. Section 3 outlines the research methodology. Results and discussion are described in Section 4. Finally, Section 5 concludes this research and scopes for future research.

II. Seven Wastes from Manufacturing Environments to IT

Innovations in IT related to logistics have emphasized using information as a supporting element of the value-added process to provide better visibility of physical goods [5]. IT can play a vital role in timely response to change market requirements for lean operating companies [6]. Reference [7] proposed the Australian case study by applying lean manufacturing principles and transformed the business practices for adopting latest IT systems to increase productivity. Reference [8] investigated the feasibility of adopting a web-based information system to the increased efficiency and productivity

of manufacturing operations. IT performance evaluation has not received much attention in the literature on lean. Reference [9] demonstrated new demands on the type of IT support needed to successfully implement lean production in practice. However, the landscape of IT, the complexity and dynamics of technology, the interdependency of its functional silos, and conflicting priorities when working with the business, all combine to create distinctive obstacles for applying lean philosophy [10].

Lean IT is the extension of lean manufacturing and lean services principles to the development and management of IT products and services. Although lean principles are generally well established and have broad applicability, their extension from manufacturing to IT is only just emerging [11]. Reference [12] investigated results of applying lean IT partnering by using six case studies (Achmea-Atos, HEINEKEN-HP, Achmea-Avanade, Achmea-SAP, Sogeti-Ministry of Economic Affair, KPN-Accenture). However, insufficient empirical research has been carried out to determine the adaptation of lean IT to be contingent upon application context.

Reference [13] developed an approach for supporting information management improvement by applying lean. Fig. 1 illustrates value-flow models as applied to a production system and information management system. This research also compared the traditional wastes with types of information management wastes as shown in Table 1. It was found that three type of traditional wastes are no analogous dimensions.

Reference [10] advocated that "Quality at the source" plays an important role for lean IT influenced of the lean principles pyramid. Without quality information, the results will always be inferior, confusion, frustration, and errors to meet end user expectations. Moreover, Reference [14] compared seven wastes from manufacturing environments to IT as shown in Table 2.

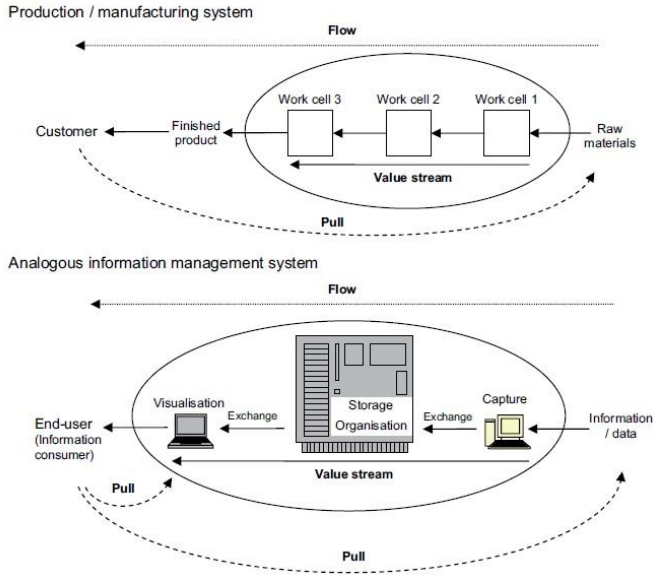


Fig. 1. Value flow models.
Source: [13]

TABLE I. Waste Comparisons of Manufacturing and Information Management

Manufacturing Systems	Information Management
Overproduction	Flow excess
Waiting	Flow demand
Extra processing	Failure demand
Defects	Flawed flow
Transport	
Inventory	
Motion	

Source: [13]

TABLE II. Comparing wastes in manufacturing and IT

Waste	Manufacturing Systems	IT
Inventory	Excess raw material, work in process, and finished goods	Excessive information causing searching and version control problems, excessive backlog and work in progress
Overproduction	Producing more or sooner than the customer needs	Excessive email, reports, system alerts, etc., that are not read or acted upon
Delays	Waiting for material from the upstream work center	System downtime, unnecessary work flow steps
Transportation	Moving materials from one location to another	Handoff of information across organization boundaries and multiple systems, security barriers to information flow
Over processing	Doing more work than the customer wants (e.g. unnecessary product features)	Redundant data, unnecessary transactions, unused reports, software features that users do not need
Motion	Bending, lifting, twisting, walking, packing/unpacking, searching for stuff	Searching for information, reentering data, excessive keystrokes, frequently shifting priorities
Defects	Defects that require correction or cause scrap	Incorrect, untimely, and confusing information that causes bad decisions

Source: [14]

III. Research Methodology

"What are appeared waste activities in CDC?" is a research question. This question was solved by combing management and IT steps. The Ishikawa diagram and the flowchart were selected as tools for this research. Quality control tools lies the idea to improve a knowledge management system by using knowledgeable employees suggestions [15]. The flowchart is a widely used technique for representing a workflow that represents an algorithm or process in the steps as boxes of various kinds and their order by connecting these with arrows. The Ishikawa diagram is designed to examine the causes of a specific issue, problem or process [16]. The knowledge of employees and researchers were listed potential causes under the categories of the unique problem. To overcome respondent bias [17], multiple respondents per case were asked to participate. To ensure that participants were suitable to fill in the evaluation sheet, they were selected based on their own understanding of the questions as well as their familiarity with the concepts surveyed.

Next, the integrated lean and IT focused on eliminating any waste in traditional and IT by implementing of Elimination, Combine, Rearrange, and Simplify (ECRS) concept. ECRS is an effective approach in process improvement represents the four core concepts as stated in Table 3 [18]. ECRS is one of classical simplified approach in a continuous improvement process [19]. One of the relationship between management and IT procedures is manual entry data from management activities could bring into IT part. IT could reduce the time associated with manual tasks. Agile software development approach aims to quickly release software in a rapidly changing requirements environment by using an iterative development for customers [20]. This approach has four steps as scope, define, develop and evaluate.

TABLE III. Concept of ECRS

Abbreviation	Meaning
E	Eliminate waste found in manufacturing such as waiting time, unnecessary movement and work step
C	Combine unnecessary work steps to reduce the number of working step and total processing time
R	Rearrange any process step for reducing distance of moving or the number of movement
S	Simplify or propose easier method for working or introduce new equipment

Source: [17]

IV. Results and Discussion

A. Quality Control Tools Application

A number of CN documents at reverse logistics department are triggered by third party logistics (3PL) activities. Once 3PLs distribute assignment products to convenient branches, CN issues are always reported from a convenient manager via a specified online software. CDC officers convert CN data in Excel spreadsheets. However, it is not easy to query when the data grows and evolves over time. Thus, printed CN documents are submitted to the CDC. A supervisor of reverse logistics department checks CN product data and considers CN status. Officers count each CN product category and record in Excel. On the one hand, some CN items are sold at cheap prices. On the other hand, remaining CN items are stored at CDC. An above explanation was summarized as shown in Fig. 2. Based on the exploration from CN history, it was found that CN issues on August, 2016 comprised 755 EA/day or 13,073.46 Baht/day (See Fig. 3).

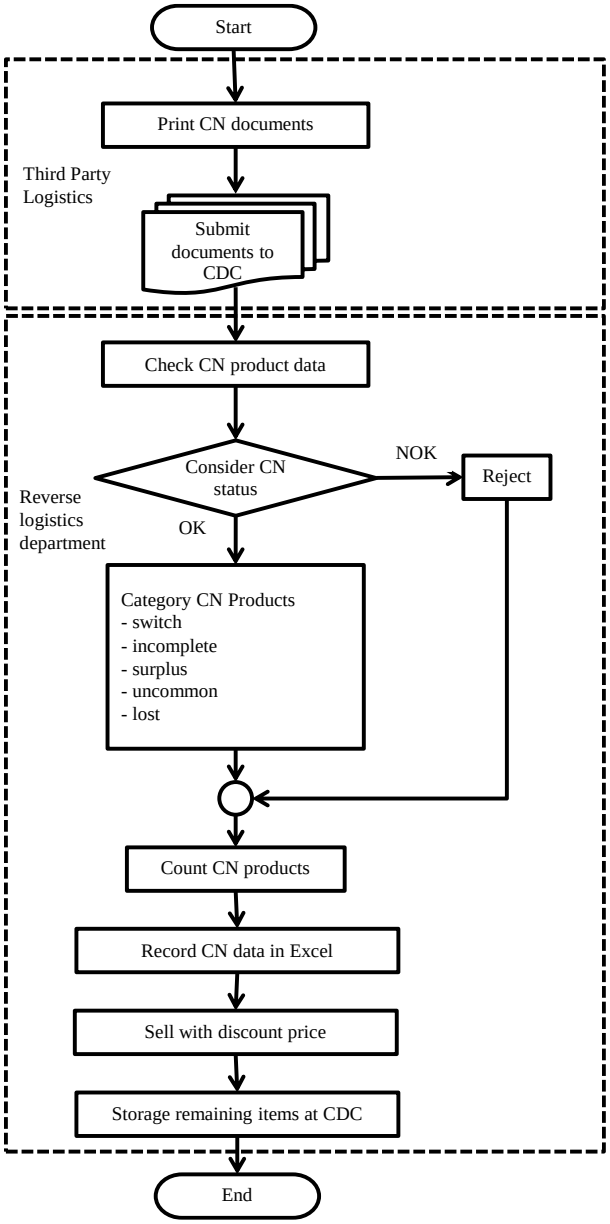


Fig. 2. A flow chart of CN event.

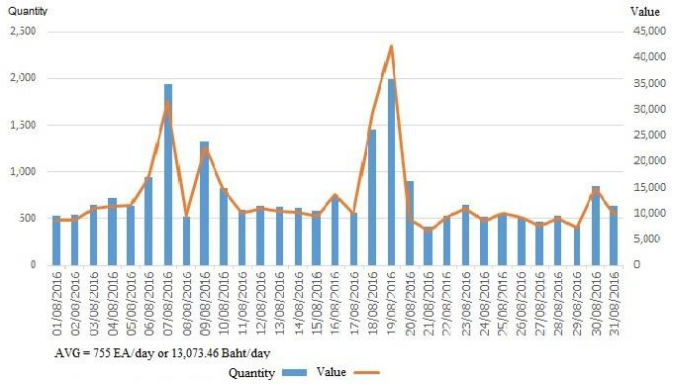


Fig. 3. CN data history on August, 2016.

In general, the Ishikawa diagram comprises at least 5M (Man, Machine, Method, Material and Measurement) used in manufacturing sector. However, this research proposed causes with a listing of potential causes under the four categories as man, material, method and branch as shown in Fig. 4. Problems were discussed with the senior manager and engineers to find root causes. Two questions were arisen to determine root causes as "How likely is this cause to be the major source of the issue or variation?" and "How easy would it be to fix or control?" A meeting pointed out that man (low skill) and method (manual operation and time consuming) should be prioritized in the research project. These causes had a greater effect than the others on the problem.

B. Agile Software Development

Some manual entry were found as wastage activities comprised data sharing, considering CN status, daily report and shipping accuracy and picking speed calculation. Results of agile software development was briefed regarding the prototyping approach. A use case diagram was drawn for a scope phase from either functional requirements or non-functional requirements by interviewing users as depicted in Fig. 5. Next, a relational database management system was used at define phase to create an entity–relationship (ER) diagram comprised CNCDC table, Product table, Pick_load table, Speed_detail table and Speed_picker table (See Fig. 6). This prototype was developed and released on PHP CSS Html and JavaScript in bootstrap framework. The prototype was implemented for testing in the evaluate phase. The developed product was released dealing with generally accepted by the users. Fig .7 shows a dash board of CN summary week reports.

The web application for managing information and eliminating waste on "E" or "Elimination" and "S" or "Simplify" operations from ECRS concept was contributed to remove wastage operations in CDC as inventory, delays and motion of traditional wastes and inventory, delays, over processing, motion and defect of IT wastes.

It was found that operation time was reduced by using this web application. CN data in CSV format from warehouse terminal system were shared to the web database. All CN data were recorded status without human error. Key performance indicators (shipping accuracy and picking speed) were compared with CN records of each operator in order to find the best practice of operator performance. This calculation by using Excel was eradicated from the CDC operations. Each report was summarized in daytime and it could export in Excel spreadsheet format.

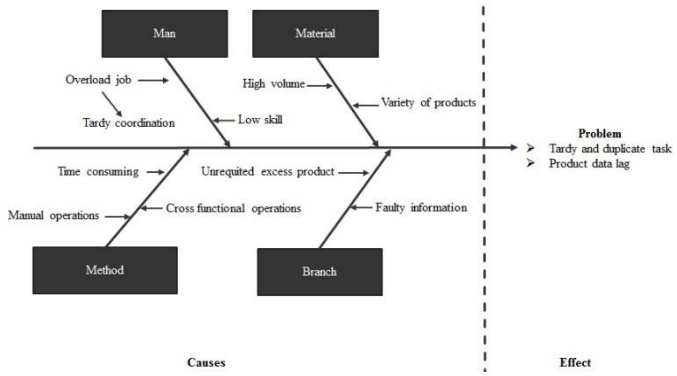


Fig. 4. Ishikawa diagram.

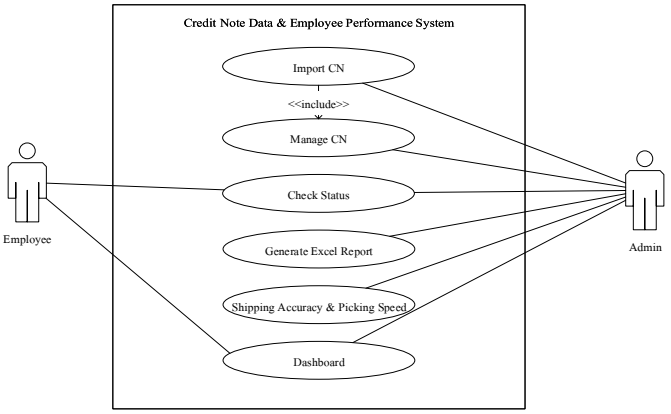


Fig. 5. Use case diagram.

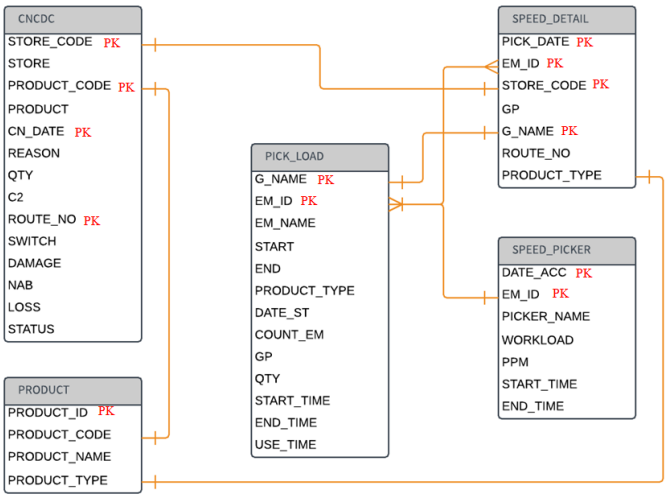


Fig. 6. ER diagram.

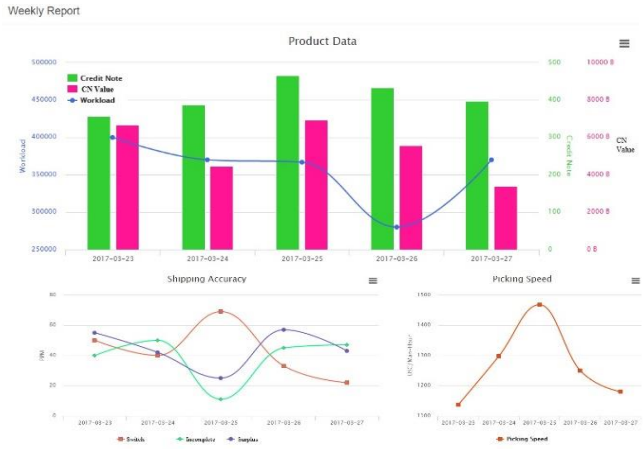


Fig. 7. A dash board.

V. Conclusions

This study aimed to propose enhancement operations in CDC by using management and IT perspectives. This research employed the concept of ECRS for dealing with CN issues. Selected QC tools were applied to explore steps of work and root causes of specified problems. The web-based database application was built on agile development to eliminate wastes in traditional and IT sides. On the basis of a case study, it was shown that this research can reduce operation times from 710 minutes to 210 minutes as shown in Table 4. Moreover, the CDC could accurately monitor reverse logistics department employee performances. It should be noted that a "Low skill" issue was solved by comparing workload and actual operating time of each operator. The best practice was used as work instruction for CDC operators. Moreover, CDC engineers could leverage sorting and arranging skills of operators for improving overall distribution efficiency.

A practical implication of this research is that the engineer could consider operator performances. Although there is no nothing new at an academic implication, this research has the merit to apply existing and well-known concepts for a real case. Low skill, manual operation and time consuming were solved by eliminating processes as considering CN status, summarized daily report and shipping accuracy and picking speed calculation. Data sharing in database was also simplified as operation time decrease average two minute per distributed trip.

TABLE IV. Summary of a testing system (unit: minute)

Operation	Old	New	Different
Data sharing on average 70 trips	350	210	140
considering CN status	30	0	30
daily report	90	0	90
shipping accuracy and picking speed calculation	240	0	240
Total	710	210	500

References

[1] R. Shannon, "The transformation of food retailing in Thailand 1997–2007". Asia Pacific Business Review, vol.15, pp. 79-92, 2009.

[2] R. Shannon, "The expansion of modern trade food retailing in Thailand", The International Review of Retail, Distribution and Consumer Research, vol.24, pp. 531-543, 2014.

[3] Y. Srimanee, and J.K. Routray, "The fruit and vegetable marketing chains in Thailand: policy impacts and implications", International Journal of Retail & Distribution Management, vol.40, pp. 656-675, 2012.

[4] D. Smith and L. Sparks, "Temperature controlled supply chains", in Food Supply Chain Management, M.A. Bourlakis and P.W.H. Weightman, Eds. New Jersey: John Wiley & Sons, 2008, pp. 179-198.

[5] I. Lewis, and A. Talalayevsky, "Logistics and information technology: a coordination perspective", Journal of Business Logistics, vol.18, pp. 141, 1997.

[6] G.P. Dimitropoulos, "Agility index of automatic production systems: Reconfigurable logic and open source as agility enablers", Computers in Industry, vol. 60, pp. 248-256, 2009.

[7] J.P. Mo, "The role of lean in the application of information technology to manufacturing", Computers in industry, vol.60, pp. 266-276, 2009.

[8] B. Tjahjono, "Supporting shop floor workers with a multimedia task-oriented information system", Computers in Industry, vol.60, pp. 257-265, 2009.

[9] J. Riezebos, W. Klingenberg, and C. Hicks, "Lean production and information technology: connection or contradiction?", Computers in Industry, vol.60, pp. 237-247, 2009.

[10] M.A. Orzen, and T.A. Paider, The Lean IT Field Guide: A Roadmap for Your Transformation, Florida: CRC Press, 2015.

[11] A. Asefeso, Lean It (Key to Improvement of Carbon Footprints), Swindon: AA Global Sourcing Ltd, 2013.

[12] V. Blijleven, Y. Gong, and H. van der Zee, Lean IT Partnering. Van Haren, 2015.

[13] B.J. Hicks, "Lean information management: Understanding and eliminating waste", International journal of information management, vol.27, pp. 233-249, 2007.

[14] S.C. Bell, and M.A. Orzen, Lean IT: Enabling and sustaining your lean transformation. Florida: CRC Press, 2011.

[15] G.C. Johannsen, "Total Quality Management in a Knowledge Management perspective", Journal of Documentation, vol.56, pp. 42-54, 2000.

[16] D. Donndelinger, and B. Van Dine, "Use the Cause-and-Effect diagram to manage conflict", Quality Progress, vol.29, pp. 136, 1996.

[17] C. Bowman, and A. Véronique, "Using single respondents in strategy research", British Journal of Management vol.8, pp. 119–131, 1997.

[18] K. Wajanawichakon, and C. Srimitee, "ECRS's principles for a drinking water production plant", IOSR Journal of Engineering, vol.2, pp. 956-960, 2012.

[19] J. Wanitwattanakosol, T. Nitiwitaya, and P. Khuwuthyakorn, "Process improvement for checking jigs and CNC programs", Proceedings of the 18th International Conference on Industrial Engineering, 2016.

[20] D. Greer, and Y. Hamon, "Agile software development", Software: Practice and Experience, vol.41, pp. 943-944, 2011.

Machine Scheduling

Scheduling job shops with operators in a free assignment-changing mode

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Imène Benkalai
Université du Québec à Chicoutimi,
G7H 2B1, Saguenay (QC), Canada
Email: imene.benkalai1@uqac.ca

Pierre Baptiste
École polytechnique de Montréal
H3T 1J4, Montréal (QC), Canada
Email: pbaptiste@polymtl.ca

Djamal Rebaine
Université du Québec à Chicoutimi,
G7H 2B1, Saguenay (QC), Canada
Email: djamal_rebaine@uqac.ca

***Abstract*—This paper deals with the problem of assigning operators, within a free changing mode, in a job-shop environment given an order for processing the set of tasks. We seek a schedule that minimizes the overall finish time, known as the makespan. Within this model, a task needs an operator during the entire duration of its processing. We show that the problem is NP-hard in the strong sense if the number of operators is greater than 3, provide a lower bound of the makespan value, and exhibit polynomial time algorithms for the cases involving one and two operators, respectively.**

I. Introduction

Integrating human resources and their different characteristics is quite a recent field of research in scheduling theory. In general, the problems of this field have proven to be difficult to solve, this difficulty being due to the addition of operator constraints to scheduling environments that were already difficult for the most part.

The analysis of real-life production systems induced that numerous links exist between human and material resources in the aforementioned environments, especially on the operational level [5], [28]. Moreover, the human actor has always played a crucial role in economy. The reason is that, on the one hand, he is consumer and producer at the same time, and, on the other hand, he is able to face unpredictable events despite the advances that have been achieved in production equipments [5]. Furthermore, it has become ever more obvious that the classical scheduling models are of little help in the real world applications [23]. Indeed, to model more accurately a real-life situation, additional features are required among which are human operators. All of this to say that it is important to take into account the human component when making decisions related to the management of material resources in order to reduce the gap between theoretical research and practical applications, but also to design effective optimization tools.

As usual with the birth of a new area, studies in this field of scheduling have started with simplified assumptions. The few developed models had the objective of modeling ever more accurately the constraints encountered in practice, especially in industrial systems.

Let us recall that in classical scheduling, it is implicitly assumed that the number of machines is equal to the number of operators, which would allow an operator to be dedicated to one machine during all the scheduling process. This is rarely true in real systems. This paper deals with scheduling job shops in which the number of operators is less than the number of machines. We assume that the assignment of operators to machines is done in a free changing mode. In other words, we assume that the assignment of an operator to a given machine may be changed at any time. We also assume that the processing sequence of jobs through the machines is known in advance.

Finally, in order to model a situation often encountered in practice, we assume that operators have different work rates, *i.e.* an operator may be able to process tasks faster/or slower than another operator. To the best of our knowledge, this problem has not been addressed yet.

This paper is organized as follows. Section 2 presents the adopted framework and notation while a description of the problem under study is presented in Section 3. A review of related works is the subject of Section 4. Section 5 is devoted to the NP-hardness of the problem, and the presentation of two polynomial-time algorithms for the case of two and one operators, respectively. Finally, we present in Section 6 a conclusion as well as perspectives for future work.

II. Framework and notation

As we pointed out earlier, scheduling problems under resource constraints are in general more complex than the classical ones, for every task needs for its processing, in addition to the machines available in the scheduling environment, a given number of additional resources[1].

The classical scheme of notation presented in [23] was extended by Blazewicz *et al.* to specify the characteristics of scheduling problems with resource constraints [12]; the latter notation is the most widely used.The parameter $\beta_2 \in \{\varnothing,\ res\ \lambda\delta\rho\}$ allows to describe these additional characteristics.

If $\beta_2 = \varnothing$, there are no resource constraints. Otherwise, if $\beta_2 = res\ \lambda\delta\rho$, the problem has specific constraints of some resources. $\lambda,\ \delta,\ \rho \in \{., k\}$, represent respectively the resources types, their limited quantities and the needs in resources. The latter are arbitrary if equal to ., else, they are equal to a given number k.

Once a first assignment is fixed, operators can change their assignment in time according to some modes, namely :

[1]That are often available in an amount that is not enough to handle all the tasks simultaneously.

- *Free assignment changing mode (f)* where an operator can interrupt the processing of a task at any time to handle another task. The interrupted task could later be resumed by this same operator of by another one.
- *Periodic assignment changing mode (p)* where the processing of tasks is performed over a given number of fixed periods. An operator remains assigned to only one machine during the length p of a period. He can interrupt the processing of a task on machine i at the end of a period to handle another task of another machine h. The interrupted task could later be resumed by this same operator of by another one.
- *End of task changing mode (e)* where an operator who processes a task j on a machine i can never interrupt the processing of a named task before its completion. The operator can then either start processing another task on the same machine or start processing a task on another machine or simply remain idle.

We extended the above-mentioned notation to be able to describe the characteristics of our problem by adding the following symbols :

- S, to indicate that the problem is subject to a fixed job sequence S.
- v_h, to indicate that the operators have different performance levels.

In our works, we derived a set of lower bounds to be able to assess the quality of our solution methods, they are of the form $\beta_{oui}(k)$ where :

- o precises the constraints of the environment and is equal to s, a or . if the problem is subject to a fixed job sequence S, to a fixed assignment A or considers both aspects simultaneously, respectively.
- u precises the assignment changing mode and can be equal to e, f, p as defined above or . if the assignment is fixed.
- $i = 1, 2, 3$ precises the studied objective where 1 is for C_{max}, 2 is for $\sum C_j$ and 3 is for L_{max}.
- k is the number of operators.

III. Problem description

The job shop scheduling problem can be described as follows. Let $J = \{1, 2, \ldots, n\}$ be a set of n jobs to be processed by a set $M = \{M_1, M_2, \ldots, M_m\}$ of m machines. Each job comprises m operations O_i, $i = 1, \ldots, m$, each of which has to be processed by a machine M_i, $i = 1, \ldots, m$. In addition, every job has its own routing through the system, *e.g.* in a three machine job shop, a job may need to be processed on machines M_1, M_3 and M_2, in that order while another job would need to visit machine M_2 first before being processed by machines M_1 and M_3.

In our model, we assume that a task requires the presence of one of the k operators, $k < m$, during its entire processing. In other words, a machine will remain inactive as long as there is no available operator to run it. Thus, the processing times of tasks are not directly affected by the intervention of operators. The impact of sharing operators is rather modeled by waiting times that are due to the fact that the number of operators is less than the number of machines. The given data p_{ij} and C_{ij} represent respectively the processing time and the finish time of operation (i, j) *i.e.* the processing of job j on machine M_i.

We assume that we have ℓ types of operators, k_h operators of each type, $h = 1, \ldots, \ell$, with $\sum_{h=1}^{\ell} k_h = k$. An operator of type ℓ has a work rate v_ℓ [2]. The assignment change of these operators on the different machines is operated according to the free mode. In other words, an operator can interrupt the processing of a task at any time to process another task. The interrupted task may be resumed by the same operator or by another one. Following the notation presented in Section 2, we denote our problem $J_m|res\ 1k1, S, f, v_h|C_{max}$.

The goal is to find a feasible schedule of tasks so as to minimize the makespan. In a job shop environment, we note the given processing sequence of tasks $\pi = (\pi_{ij})$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. The goal is to find an assignment of operators a^* such that

$$C_{\max}(\pi, a^*) = \min_{a \in A} C_{\max}(\pi, a),$$

where A denotes the set of feasible assignments, a^* is in A, and $C_{max}(\pi, a)$ is the overall completion time of the sequence π under the assignment a.

IV. Related works

Job shops problems have been extensively studied in the scheduling literature[22]. The reason for that is twofold : first they allow to model many real-life situations, which makes them very useful on the practical side, second, they are among the most difficult combinatorial optimization problems, so their study may provide useful theoretical results for the advance of research.

Indeed, the classical job shop problem, denoted $J_m||C_{max}$, is NP-hard in the strong sense for even for $m = 2$ [23].

In what follows, we present a brief review of the literature dedicated to the job shop problem with operators.

Scheduling problems with operators in general and job shop problems in particular being relatively recent and difficult to solve, we noted that only a few studies dealt with them. The difficulty of these problems concerns mainly the modeling of the human resources' characteristics, such as the experience of an operator, his qualifications, the variable workload that is assigned to him, etc.

In the classical scheduling models, it is implicitly assumed that the number of operators is equal to the number of machines. Therefore, an operator can operate exclusively the machine to which he is assigned. However, this assumption is rarely true in practice where the number of operators is generally smaller than the number of machines [7].

[2]The obtained results in terms of complexity are identical for cases where operators have equal performance levels.

In each and every study that was dedicated to scheduling problems with operators, the authors defined different models based on various assumptions. These assumptions allowed to model with more or less precision the interactions that exist between human and material resources. Since Vickson's works in 1980 [26], [27], the majority of studies that were undertaken in the field of scheduling with operators dealt with processing times that were controllable by the amount of assigned resources. Solving these problems induced to assign a surplus of operators to the tasks in order to reduce the processing time of the latter. Unlike previous works, the model proposed by Cheurfa studied a cyclical flow shop with a number of operators that is less than the number of machines [14]. In his work, the author defines the concept of "sharing" operators and studies a decision problem rather than an optimization problem, with operators with equivalent performance levels. The intervention of the latter is limited to assembling, control and disassembling operations [9].

Analyzing the literature allowed us to note that a fair part of the works that dealt with scheduling with operators was dedicated to identical parallel-machine environments. We may cite for example the works of Zouba [28]. In his PhD thesis, the author studied the simultaneous handling of scheduling the tasks and assigning the operators in an identical parallel-machine environment with a number of operators that is less than the number of machines and the possibility for an operator to share his time between different machines. In that model, it mainly comes to study the impact of sharing the operators on the processing times of the tasks. In [6], that model was refined by assuming that the ratio between real and theoretical processing times is a linear function of the occupancy rates of tasks by the operator. Even if it doesn't allow to predict the exact processing times, this new model allows to forecast the global increase in the overall duration of a schedule [9].

Around the end of the 1990s and the beginning of the 2000s, many works dealt with the OWMM concept (One-Worker-Multiple-Machine) given its popularity, especially in just-in-time production systems. We may cite for example Cheng*et al.* [13] in which the authors studied a two-machine flow shop in order to minimize the makespan when the intervention of the operator is limited to setup operations at the beginning and at the end of tasks. Let us also cite the PhD thesis of Baki where the author studies flow shop and open shop problems with one operator [4]. The latter is in charge of setup operations at the beginning of a sequence of tasks, hence the use of the batch concept consisting in grouping tasks and then scheduling them. Several objectives functions based on completion times and/or due dates were considered. For further details on scheduling with human resources, see e.g. [4], [5].

Let us observe most scheduling problems are NP-hard. The addition of resource constraints may only make them more difficult to solve. As a result, many studies dealt with special cases such as restricted number of machines to $m = 2$. For more details, see e.g. [4], [6].

Finally, let us mention the model of flow shop we proposed in [7] with a number of operators smaller than the number of machines, and where tasks need an operator for the total duration of their processing. In that paper, we designed heuristics, proposed a lower bound, and compared the relative importance of the human and material resources by comparing a simultaneous approach with a phase-by-phase approach. Initially allowing only a change of assignment at the end of a task, this model was extended to take into consideration a free assignment changing mode in [9] and even adapted to minimize the maximum lateness (L_{max})[8].

Regarding the job shop, considering its inherent complexity, relatively few papers were dedicated to its study in the context of the integration of human resources. However, these studies managed to provide high quality solution methods that prove to be quite competitive.

The problem of job shop with operators with a number of operator less than the number of machines was first introduced by Agnetis *et al.* in [1]. The authors provided complexity results for minimal cases and derived a set of effective solutions approaches that include a pseudo-polynomial dynamic programming approache, a fully polynomial time approximation scheme (FPTAS), an enumeration scheme based of a generalized disjunctive graph that was integrated in a branch-and-bound procedure and a dynamic programming based heuristic.

The dynamic programming algorithm outperformed by far the branch-and-bound and the designed heuristic showed robustness in providing results of fairly good quality.

Let us also mention the genetic algorithm of Mencia *et al.* [20]. The authors devised a schedule generation scheme inspired form that of Giffler and Thompson [15]. In that algorithm, they assign tasks to operators in order to form a feasible solution. Their scheme guaranties that it can produce an optimal solution. They used that scheme as a decoder for a rather classical genetic algorithm that used a job order crossover [11]. Since the research on this problem was in its early days, the only method they could compare their approach with was that of Agnetis *et al.* [1]. They claim that their genetic algorithm was produced quite competitive results. In addition to that, since it was rather fast, it had a lot of room for improvement. This genetic algorithm was later enhanced by integrating a weak Lamarckian evolution and a narrowing-search-space technique in [21].

Agnetis *et al.* also proposed two heuristics and a mixed integer programming formulation for a job shop with operators that was used to model a case of handicraft production[2]. Their study showed that they could obtain close-to-optimal solutions by exploiting the characteristics of the problem in their heuristics.

Sierra *et al.* developed other schedule generation schemes that can be used in different settings as heuristic search, branching strategies or decoders in evolutionary algorithms. They work on refining their previously developed scheme [20] by making it faster and more efficient. They used the resulting algorithm in an exact approach that included pruning by means of dominance rules[25].

Lastly, let us mention the memetic algorithms of Mencia *et al.*[22]. They devised them with two different schedule builders and integrated a local search step by means of a tabu search procedure. Their neighborhood structure was based on reversing arcs in a disjunctive graph. Their algorithms were tested over and extensive set of instances and produced high quality results in short time.

Since we are dealing with the makespan, here, we focused on that objective in the related works review.

Before closing this section, let us mention that few papers were published on minimizing the mean flow time; for more details, see [24], [19], [18]. Also, in a very recent study, we dealt with the maximum lateness objective [10].

V. Complexity results and polynomial-time algorithms

As with any combinatorial optimization problem, we first start the analysis with its complexity status. In what follows, we present the complexity results for cases with one, two and more than three operators.

A. The k-operator case, $k \geq 3$

Theorem V.1. *The problem $J_m|res\ \ell k_h 1, S, f, v_h|C_{\max}$ is NP-hard in the strong sense for $k \geq 3$ even for the case in which all the v_h are equal.*

Proof: In [10], we have proven the flow shop problem with the same environment features as our problem at hand to be strongly NP-hard for $k \geq 3$ for cases where v_h values are different but also when they are all equal $\forall h = 1, \ldots, \ell$.

Since the flow shop is a special case of the job shop where all job routings are the same, it follows that $J_m|res\ \ell k_h 1, S, f, v_h|C_{\max}$ is strongly NP-hard for $k \geq 3$ for all values of v_h, $h = 1, \ldots, \ell$. ■

Next, we develop a lower bound for this case, it is based on the fact that at the beginning and at the end of the scheduling, only one operator is busy. The rest of the time, we would ideally have all the operators working full-time. This gave us the following result:

$$\beta_{of1}(k) = p_{1\pi(1,1)} + p_{m\pi(m,n)} + \frac{1}{k}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} - \left(p_{1\pi(1,1)} + p_{m\pi(m,n)}\right)\right],\ k \geq 3$$

is a lower bound for $J_m|res\ \ell k_h 1, S, f, v_h|C_{\max}$ when $k \geq 2$.

B. The two-operator case

Theorem V.2. *The problem $J_m|res\ 2k_h 1, S, f, v_h|C_{\max}$ is solvable in polynomial time.*

Proof: The complexity of the problem is showed via the restriction method. Let us first note that the ordering constraints imposed by the processing sequence of tasks on the machines as well as by S form a precedence graph. Therefore, the problem can be seen as a two-uniform-parallel-machine problem with precedence constraints with preemption, denoted as $Q_2|prec, prmp|C_{\max}$, where the machines, the tasks and the precedence graph represent respectively the $k = 2$ operators, the $m * n$ operations, and the precedence relations between these operations.

We know that $Q_2|prec, prmp|C_{\max}$ problem is solvable in polynomial time.

Indeed,

$$Q_2|prec, prmp|C_{\max} \propto Q_2|prec, prmp, r_j|L_{\max}$$

and $Q_2|prec, prmp, r_j|L_{\max}$ is solvable in $O(m^2n^2)$ [17]. It follows that $J_m|res\ \ 2k_h 1, S, f, v_h|C_{\max}$ is solvable in polynomial time. ■

Let us now present a polynomial solution method for $J_m|res\ 2k_h 1, S, f, v_h|C_{\max}$.

Let $Succ(i,j)$ be the set of successors of the operation $\pi(i,j)$.

The *level* $L_s(T_{\pi(i,j)})$ of a task $T_{\pi(i,j)}$[3] at the time s is the minimum time that is necessary for processing the said task as well as all its successors from the time s forward. For our problem, $L_s(T_{\pi(i,j)}) = \left(\sum_{(x,y)\in Succ(i,j)} p_{x,\pi(x,y)}\right) + p_{i,\pi(i,j)}$.

The algorithm in $O(m^2n^2)$ presented on Figure 1 is an adaptation of the algorithm of Horvath *et al.* [16]. It provides a solution with a minimum makespan for $k = 2$ in the case where operators are identical but also when they have different performance levels. In practice, we would first have to assign values to the performance levels of operators based on their speed in processing jobs before applying the algorithm. These values could be assessed by the person in charge of the job shop.

$s \leftarrow 0$;
$h \leftarrow$ the number of free operators at time s;
$j \leftarrow$ the number of tasks at the highest level at time s;
Do
{
  If($j \leq h$)
  { The j fastest operators process the j tasks at the same rate (*) ; }
  Else
  { The j tasks are processed at the same rate by the h operators(*); }
  If($\exists$ free operators)
  { Assign the tasks of next highest level; }
  If(($\exists\ T$ that ends at t) or ($\exists\ T, T'$ / $L_s(T) > L_s(T')$ and $L_t(T) = L_t(T')$))
  { $s \leftarrow t$; }
} **While**(Not all tasks have been processed)

To build a solution with the free assignment-changing mode, the j tasks that share the h operators during the steps (*) receive the same processing amount from the said operators;

Each shared interval is divided into j sub-intervals;

Each of the j tasks is scheduled in h sub-intervals, each time with a different operator;

Fig. 1. Algorithm for solving $J_m|res\ 2k_h 1, S, v_h|C_{max}$.

[3]The task at the j^{th} position on the i^{th} machine in the given sequence S.

C. The one-operator case

Theorem V.3. *The problem $J_m|res\ 111, S|C_{\max}$ is solvable in polynomial time.*

Proof: We use the same proof technique as we previously did for $k = 2$. We use for the reduction the $1|prec, prmp|C_{\max}$ problem which is solvable in polynomial time. Indeed,

$$1|prec, prmp|C_{max} \propto 1|prec, prmp, r_j|L_{max}$$

and $1|prec, prmp, r_j|L_{\max}$ is solvable in $O(m^2n^2)$ [3]. ■

As we previously mentioned, the OWMM (One Worker Multiple Machine) concept is very popular, especially in just-in-time production systems and is often encountered in practice.

In the case of one operator, the sum of all the processing times $\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}$ is an obvious lower bound. We notice that every *non delay*[4] has a makespan equal to this lower bound. Its optimality is hence established. Such a solution can be computed in $O(mn)$.

An optimal scheduling policy could be to process without interruption all the operations of job J_1, then those of job J_2 and so on until job J_n.

VI. Conclusion and future works

In the current economical environment, which is strongly competitive, it becomes ever more important for the good functioning of companies to have an efficient management of their resources with adequate decision-making systems.

In this context, and knowing that classical scheduling models that do not consider operators are of little help in the real world applications [23], it is necessary to develop models that are more realistic, particularly by taking into consideration human resources, so as to reach a higher levels of efficiency in resource-management and at the same time reduce the gap between academical research and its real-world applications.

In this paper, we have studied the assignment of a number of operators that is less than the number of machines in a job shop environment with a given sequence of tasks and a free assignment-changing mode in order to minimize the makespan. The operators we considered have different performance levels. We studied the complexity of our problem in the case with one, two and more than three operators. When $k \geq 3$, our problem is NP-hard in the strong sense, even if the operators have equivalent performance levels. In that case, we derived a lower bound that can be used to asses the solution quality of future methods that could be designed for that case.

We provided polynomial-time solution method for both cases with one and two operators.

For future research, given the fact that the problem at hand is strongly NP-hard for $k \geq 3$, it would be interesting to develop heuristic methods for that case but also to test the provided algorithms on instances driven from real-world situations when $k > 2$.

It would also be of interest to consider other objectives that are important in practice such as the mean flow time and also include additional features to our model such as sequence-dependent setup times release dates.

[4] A solution in which no machine is left idle when there is an available job and an available operator to process it.

References

[1] A. Agnetis, M. Flamini, G. Nicosia, and A. Pacifici. A job shop problem with one additional resource type. *Journal of scheduling*, 14:225–237, 2010.

[2] A. Agnetis, G. Murgia, and S. Sbirilli. A job shop scheduling problem with human operators in handicraft production. *International Journal of Production Research*, 52:13:3820–3831, 2014.

[3] K. R. Baker, E. L. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan. Preemptive scheduling of a single machine to minimize maximum cost subject to release dates and precedence constraints. *Operations Research*, 31(2):381–386, 1983.

[4] M.F. Baki. *Some Problems in One-Operator Scheduling*. PhD thesis, University of Waterloo, 1999.

[5] P. Baptiste, V. Giard, A. Hait, and F. Soumis. *Gestion de production et ressources humaines*. Presses Internationales Polytechnique, 2005.

[6] P. Baptiste and A. Munier. Ordonnancement sur machines parallèles avec partage d'opérateurs. In *10ème Congrès International de Génie Industriel-CIGI 2013*, 2013.

[7] I. Benkalai, P. Baptiste, and D. Rebaine. Ordonnancement d'ateliers de type flow shop avec contrainte d'opérateurs. In *11ème Congrès International de Génie Industriel-CIGI 2015*, 2015.

[8] I. Benkalai, P. Baptiste, and D. Rebaine. Ordonnancement d'ateliers de type flow shop avec opérateurs en mode d'affectation libre. In *17ème Conférence ROADEF de la Société Française de Recherche Opérationnelle et Aide à la Décision*, 2016.

[9] I. Benkalai, D. Rebaine, and P. Baptiste. Assigning operators in a flow shop environment. In *Information Systems, Logistics and Supply Chain*, 2016.

[10] I. Benkalai, D. Rebaine, and P. Baptiste. Ordonnancement d'ateliers de type job shop avec opérateurs. In *18ème Conférence ROADEF de la Société Française de Recherche Opérationnelle et Aide à la Décision*, 2017.

[11] C. Bierwirth. A generalized permutation approach to jobshop scheduling with genetic algorithms. *Operations Research Spectrum*, 17:87–92, 1995.

[12] J. Blazewicz, K. Ecker, E. Pesch, G. Schmidt, and J. Weglarz. *Handbook on scheduling : From theory to applications*. Springer, 2007.

[13] T.C.E. Cheng, G. Wang, and C. Sriskandarajah. One-operator-two-machine flowshop scheduling with setup and dismounting times. *Computers & Operations Research*, 26:715–730, 1999.

[14] M. Cheurfa. *Gestion des ressources humaines en production cyclique*. PhD thesis, ENSM Saint-Étienne, 2005.

[15] B. Giffler and G. L. Thompson. Algorithms for solving production scheduling problems. *Operations Research*, 8:487–503, 1960.

[16] E. C. Horvath, S. Lam, and R. Sethi. A level algorithm for preemptive scheduling. *J. Assoc. Copmut. Mach.*, 24(1):32–43, 1977.

[17] E. L. Lawler. Preemptive scheduling of precedence-constrained jobs on parellel machines. In *Deterministic and stochastic scheduling, Proceedings of the NATO advanced study and reasearch institute on theoretical approaches to scheduling problems*, pages 101–123. D. Reidel publishing Co., 1982.

[18] C. Mencia, M. R. Sierra, M. A. Salido, J. Escamilla, and R. Varela. Solving the job shop scheduling problem with operators by depth-first heuristic search enhanced with global pruning rules. *Artificial Intelligence Communications*, 28:365–381, 2015.

[19] C. Mencia, M. R. Sierra, and R. Varela. An efficient hybrid search algorithm for job shop scheduling with operators. *International Journal of Production Research*, 51:17:5221–5237, 2013.

[20] R. Mencia, M. R. Sierra, C. Mencia, and R. Varela. Genetic algorithm for job shop with operatos. In *Ferrandez J. M., Alvarez Sanchez J. R., de la Paz F., Toledo F. J. (eds.) New challenges on bioinspired applications. IWINAC. Lecture Notes in Computer Science*, volume 6687. Springer, Berlin, Heidelberg, 2011.

[21] R. Mencia, M. R. Sierra, C. Mencia, and R. Varela. A genetic algorithm for job shop scheduling with operators enhanced by weal lamarckian evolution and search space narrowing. *Natural Computing*, 13:179–192, 2014.

[22] R. Mencia, M. R. Sierra, C. Mencia, and R. Varela. Memetic algorithms for the job shop scheduling problem with operators. *Applied soft computing*, 34:94–105, 2015.

[23] M. Pinedo. *Scheduling : Theory, algorithms and systems*. Prentice Hall, 2002.

[24] M. R. Sierra, C. Mencia, and R. Varela. Optimally scheduling a job shop with operators and total flow time. In *Lozana J. A., Ganaz J. A., Moreno J. A. (eds.) Advances in Artificial intelligence. CAEPIA. Lecture Notes in Computer Science*, volume 7027. Springer, Berlin, Heidelberg, 2011.

[25] M. R. Sierra, C. Mencia, and R. Varela. New schedule generation schemes for the job shop problem with operators. *Journal of intelligent manufacturing*, 26:511–525, 2015.

[26] R.G. Vickson. Choosing the job sequence and processing times to minimize total processing plus flow cost on a single machine. *Operations Research*, 28(5):1155–1167, 1980.

[27] R.G. Vickson. Two single machine sequencing problems involving controllable job processing times. *AIIE transactions*, 12(3):258–262, 1980.

[28] M. Zouba. *Ordonnancement de machines parallèles identiques avec des contraintes de ressources humaines*. PhD thesis, École Polytechnique Montréal, 2009.

Existing results and new developments for the single processor scheduling problem with time restrictions

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Rachid Benmansour
University of Valenciennes and Hainaut Cambrésis, LAMIH – UMR CNRS 8201
Le Mont Houy
F-59313 Valenciennes CEDEX 9, France
Institut National de Statistique et d'Économie Appliquée, Rabat, Morocco
Email: r.benmansour@insea.ac.ma

Oliver Braun
Trier University of Applied Science, Environment Campus Birkenfeld
Campusallee
D-55768 Birkenfeld, Germany
Email: o.braun@umwelt-campus.de

Said Hanafi
University of Valenciennes and Hainaut Cambrésis, LAMIH – UMR CNRS 8201
Le Mont Houy
F-59313 Valenciennes CEDEX 9, France
Email: said.hanafi@univ-valenciennes.fr

***Abstract*—Single processor scheduling with time restrictions (STR) is a new scheduling problem with wide practical applications and has been formulated at first by Braun, Chung and Graham in 2014. In the tradition of Ron Graham's worst-case analysis from the 1960's (that became popular as analysis of approximation factors of approximation algorithms and in the world of online-algorithms under the notion competitive analysis), the authors provide a worst-case analysis of List Scheduling for this STR problem. Since the publication of the STR problem in 2014, several problems around the basic STR problem have been investigated. Recently, an interesting analogy of the STR problem with the Parallel machine scheduling problem with a single server (PSS) has been published. We review the existing results and give an outlook on new research possibilities.**

I. Introduction

In the single processor scheduling problem with time restrictions (STR), n independent jobs have to be processed on a single processor, such that during any time period of length $\alpha > 0$ the number of jobs being executed is less than or equal to a given integer value B. Practically, this constraint reflects the condition that each job needs exactly one of B additional resources for being processed and that a resource has to be renewed after the processing of a job has been finished.

As an illustration of the single processor scheduling problem with time restrictions, let imagine that each job being processed requires the use of one of B identical subprocessors, that can be other external resources such as machines, containers, tanks, or personnel. Furthermore, each subprocessor that has been used needs a certain amount of time α to reset (e.g. cool down, cleaned, re-filled, re-loaded, updated, etc.) before it can be used again. Hence, it is never possible to process more than B jobs during any interval $[x, x+\alpha), \forall x \in R^+$.

Another view on the same problem is that there are B identical processors and one server that has to be used to load or unload the processors. This problem is well known in the literature under the name parallel machine scheduling problem with a single server (PSS).

In both problems, it is assumed that the jobs are simultaneously available for processing at the beginning of the planning horizon, and their processing times are fixed and known in advance. Formally, the problems can be stated as follows.

We are initially given a set $S = \{S_1, S_2, \ldots, S_n\}$ of jobs. Each job S_i has associated with it a length s_i. In the STR problem, the jobs are all processed sequentially on a single processor. Only one job can be worked on at any point in time. A job S_i will be processed during a (semi-open) time interval $[\alpha, \alpha + s_i)$ for some $\alpha \geq 0$.

Given some permutation π of S, say $\pi(S) = T = (T_1, T_2, \ldots, T_n)$ with corresponding job lengths $(t_1, t_2, \ldots, t_n)$, the jobs are placed sequentially on the real line as follows. The initial job T_1 in the list T begins at time 0 and finishes at time t_1. In general, T_{i+1} begins as soon as T_i is completed, provided the following constraint is always observed:

$$\text{For every real } x \geq 0, \text{ the unit interval } [x, x+1) \text{ can intersect at most } B \text{ jobs.} \quad (1)$$

In the PSS problem, there are m parallel and identical processors. Before its processing, each job S_i must be loaded on a common server which needs a certain amount of setup time s_i. We assume that the processing times of all jobs are identical and need α time units. During the loading operation, both the processor and the server are occupied. Another view on the same problem is that each job has a processing time s_i and needs to be unloaded after its processing which takes α time units for each job.

In Section 2, we give a review on existing results of the single processor scheduling problem with time restrictions (STR). Section 3 reviews the related results of the parallel processor scheduling problem with single server (PSS). Finally, in Section 4, we give an outlook on further research possibilities.

II. Single processor scheduling with time restrictions

Constraint (1) reflects the condition that each job needs one of B additional resources for being processed and that a resource has to be renewed after the processing of a job has been finished. The preceding procedure results in a unique placement (or *schedule*) of the jobs on the real line. We define the *finishing time* $\tau(T)$ to be the time at which the last job T_n is finished. For a given job set S, we seek to find those permutations $T = \pi(S)$ which minimize the finishing time $\tau(T)$.

In [7] it has been shown that the problem is NP-hard in general (the authors polynomially reduce the NP-hard problem PARTITION to a special case of the STR problem).

Recently, it has been shown that the STR problem is NP-hard even for $B = 2$ by showing the equivalence between the STR problem and the parallel machines scheduling problem with a single server [5]. In another recently published paper [17], the authors claim to prove the same by reducing Cardinality Constrained Partition to the STR problem.

Let us denote by $\tau_w(S)$ the largest possible finishing time for any permutation of S, and let $\tau_o(S)$ denote the optimal (i.e., the shortest possible) finishing time for any permutation of S. The following bounds from [7]–[9] are similar in spirit to some of the bounds in the very early *worst-case analysis* literature of scheduling algorithms [11], [12]:

Theorem 1: For $B = 2$ and any set S,

$$\tau_w(S) - \frac{4}{3}\tau_o(S) \leq 1.$$

This bound is best possible.

Theorem 2: For $B \geq 3$ and any set S,

$$\tau_w(S) - \left(2 - \frac{1}{B-1}\right)\tau_o(S) \leq \frac{B}{B-1}.$$

This bound is best possible.

In [8], [9], the authors provide an analysis of the LPT-heuristic, where the jobs are ordered non-increasingly according to the Longest-Processing-Time-first (LPT) algorithm and show the following upper bounds:

Theorem 3: For $B \geq 2$ and any set S,

$$\tau_{LPT}(S) - \left(2 - \frac{2}{B}\right)\tau_o(S) \leq 1.$$

This bound is best possible.

Theorem 4: For $B = 2$ and any set S,

$$\tau_{LPT}(S) - \tau_o(S) \leq \begin{cases} \frac{1}{2}(1 + l_1 - l_{n-1} - l_n) & \text{if } n \text{ is odd,} \\ \frac{1}{2}(l_1 + l_2 - l_{n-1} - l_n) & \text{if } n \text{ is even.} \end{cases}$$

This bound is best possible.

The authors of [9] state that an easy improvement for the case $B = 2$ is a heuristic H where the schedule is started with the smallest job and then *LPT* is performed. In that way the bound $\tau_H(S) - \tau_o(S) \leq 1/2$ can be achieved. The authors of [16] prove the same result and show that this bound is asymptotically optimal.

In [16], the authors show that for $B \geq 5$ there exists a permutation of the jobs which can be processed within a factor of $5/4$ of the optimum (plus an additional small constant). For $B = 3$ the corresponding factor equals $\frac{3}{2}$ and for $B = 4$ it is $\frac{4}{3}$.

In [17], the authors give a polynomial time approximation scheme (PTAS) together with a Dynamic Programming (DP) algorithm for $B \geq 2$.

Two different mixed integer linear programming formulations for the solution of the STR problem have been published so far [2]–[6], [15]. The first one is based on a time-indexed formulation, and the second one is based on a formulation with assignment and positional date variables.

The time-indexed formulation (TIF) is based on time-discretization where time is divided into periods. Period t starts at time t and ends at time $t+1$. The planing horizon H must be part of the input data. The particularity of the STR problem is that in an optimal schedule there is no idle time between the first B jobs since during the interval $[0, C_B)$ no time interval of duration α is able to intersect more than B jobs. For the jobs at the following positions (i.e. $j \geq B+1$), the starting time of the job at position j cannot be smaller than $C_{j-B} + \alpha$ (otherwise, the interval $[C_{j-B}, C_{j-B} + \alpha)$ would intersect at least $B+1$ jobs). In the case $C_{j-1} > C_{j-B} + \alpha$, the completion time of the job S_j is set to $C_j = C_{j-1} + p_{[j]}$.

$\min C_{max}$

s.t. $C_{max} \geq \sum_{t=0}^{H-s_i}(t+s_i)x_{it}, \forall i \in \{1,\ldots,n\}$,

$\sum_{t=0}^{H-s_i} x_{it} = 1, \forall i \in \{1,\ldots,n\}$,

$\sum_{i=0}^{n}\sum_{s=t-s_i+1}^{t} x_{is} = 1,\ t \in H$,

$\sum_{i=1}^{n}\sum_{s=t}^{t+\alpha-1} x_{is} \leq B - 1 + \sum_{i=0}^{n} x_{it}, t \in H$,

$\sum_{i=1}^{n}\sum_{s=t}^{t+\alpha} x_{is} \leq B - 1 + Lx_{0,t}$

$+L(1 - \sum_{i=0}^{n} x_{i,t+1}) + L(1 - \sum_{i=1}^{n} x_{i,t+\alpha}), \forall t \in H$,

$x_{it} \in \{0,1\}, \forall i \in N \cup \{0\}, t \in [0, H - s_i]$.

As for the formulation with assignment and positional date variables, the binary variable x_{ij} corresponds to the assignment of job i to the position j (i.e. x_{ij} equal to 1 if and only if the job i is assigned to position j).

$\min C_n$

s.t. $\sum_{i=1}^{n} x_{ij} = 1, \forall i \in \{1,\ldots,n\}$,

$\sum_{i=1}^{n} x_{ij} = 1, \forall j \in \{1,\ldots,n\}$,

$C_j = C_{j-1} + \sum_{i=1}^{n} s_i x_{ij}, \forall j \in \{1,2,\ldots,B\}$,

$C_j \geq C_{j-1} + \sum_{i=1}^{n} s_i x_{ij}, \forall j \in \{B+1,\ldots,n\}$,

$C_j \geq C_{j-B} + \alpha + \sum_{i=1}^{n} s_i x_{ij}, \forall j \in \{B+1,\ldots,n\}$,

$C_0 = 0$,

$C_j \geq 0, \forall j \in \{1,\ldots,n\}$,

$x_{ij} \in \{0,1\}, \forall i, j \in \{1,\ldots,n\}$.

Computational experiments were conducted on an Intel i7 2.8 GHz computer with 16GB of RAM under Windows 7 operating system and we used CPLEX 12.6.3 as the integer programming solver. The APV formulation outperforms the TIF formulation by far. and all of the problem instances could be solved optimally with the APV formulation. Therefore, only the results of the APV formulation are reported in Table I of this survey.

TABLE I. CPU TIMES OF 10 PROBLEM INSTANCES FOR THE APV FORMULATION WITH $\alpha = 10$, $B = 2,5,10$ AND $n = 100,500$

$B=$	2		5		10	
$n=$	100	500	100	500	100	500
1	0.70	49.05	0.25	86.64	0.30	12.36
2	0.77	42.92	1.08	113.32	0.22	18.35
3	0.62	45.62	0.23	95.77	0.23	12.74
4	0.70	32.21	0.83	112.10	0.27	23.16
5	0.80	39.56	0.80	102.25	0.23	8.84
6	0.78	47.89	1.01	167.20	0.22	36.60
7	0.64	50.74	1.11	86.83	0.27	16.38
8	0.75	41.83	1.08	107.08	0.27	21.75
9	0.83	43.39	0.30	123.59	0.22	18.67
10	0.59	40.97	0.27	72.09	0.25	17.14

III. PARALLEL PROCESSOR SCHEDULING PROBLEM WITH SINGLE SERVER

The PSS problem is denoted $P,S1 \mid s_i, p'_i \mid C_{max}$ in the notation of Kravchenko and Werner [14], where s_i is the loading duration of the job i and p'_i is its corresponding processing time. The PSS problem (with related results to the STR problem) is well studied in the literature and can be stated as follows: There are m identical parallel machines which must process n jobs. Each job has a deterministic processing time p'_i. Before its processing, the job must be loaded on a common server which is called a setup s_i. During the loading operation, both the processor and the server are occupied. The objective (we are mainly interested in) is minimizing the completion time of the last job that finished its processing, i.e. the makespan.

The parallelism of the PSS problem to the STR problem is as follows: The single server corresponds to the single processor. And the m parallel machines of the PSS problem correspond to the B additional resources of the STR problem. The setup times s_i of the PSS problem correspond to the processing times s_i of the STR problem. And, finally, the processing times p'_i of the PSS problem correspond to α of the STR problem. So, the STR problem corresponds to the PSS problem with equal processing times $p'_i = \alpha$.

Hall et al. [13] give an overview of complexity results to the PSS problems. Transferred to the STR problem, they show that even for $B = 2$ and $\alpha = 1$ or α is a constant, the problem is NP-complete. In more detail, they show that there exists a solution to a specific instance of the scheduling problem with $C_{max} \leq$ a threshold value C if and only if there exists a solution to the Numerical Matching with Target Sums (which is known to be unary NP-complete).

Brucker et al. [10] prove a similar result (binary NP-complete) by showing that there exists a solution to a specific instance of the scheduling problem with $C_{max} \leq$ a threshold value C if and only if there exists a solution to the Equal Size Partition (which is known to be binary NP-complete).

Abdekhodaee et al. [1] prove the same result (binary NP-complete) by showing that there exists a solution to a specific instance of the scheduling problem with $C_{max} \leq$ a threshold value C if and only if there exists a solution to the restricted partitioning problem (partitioning into sets with equal cardinality, which is known to be binary NP-complete).

Kravchenko and Werner [14] provide a worst-case analysis of List Scheduling and LPT and a pseudopolynomial algorithm for the situation when the setup times are all equal to $s_i = 1$. Transferred to the STR problem: Single processor, B resources, $s_i = 0$ for all Jobs $S_i, i = 1,\ldots,n$, arbitrary α_i. They show that List Scheduling has a worst-case performance ratio of $3-2/m$ and LPT has a worst-case performance ratio of $2-1/m$.

IV. OUTLOOK ON FURTHER RESEARCH POSSIBILITIES

The STR problem is optimally solvable for up to 500 jobs with the APV formulation of Benmansour, Braun, and Hanafi [5]. There exists also the PTAS of Zhang, Ye, Chen and Chen [17], but the authors do not report any computational results. Benmansour, Braun, and Hanafi developed also a Variable Neighbourhood Search (VNS) algorithm to solve STR [6]. It can be expected that more sophisticated heuristics will sole the STR problem optimally for larger problem instances than currently reported (i.e. for more than $n = 500$ jobs).

In the field of worst-case analysis, the existing results are as depicted in Tables II and III.

TABLE II. $B = 2$ RESOURCES

Heuristic	Reference	Bound
List Scheduling	[7]	$\tau_w(S) - \frac{4}{3}\tau_o(S) \leq 1$
LPT	[8], [9]	$\tau_{LPT}(S) - \tau_o(S) \leq 1$
H	[9]	$\tau_H(S) - \tau_o(S) \leq \frac{1}{2}$
T	[16]	$\tau_H(S) - \tau_o(S) \leq \frac{1}{2}$

TABLE III. $B \geq 3$ RESOURCES

Heuristic	Reference	Bound
LPT ($B = 3$)	[8], [9]	$\tau_{LPT}(S) - \frac{4}{3}\tau_o(S) \leq 1$
H3 ($B = 3$)	[16]	$\tau_{H3}(S) - \frac{3}{2}\tau_o(S) \leq c$
LPT ($B = 4$)	[8], [9]	$\tau_{LPT}(S) - \frac{3}{2}\tau_o(S) \leq 1$
H4 ($B = 4$)	[16]	$\tau_{H4}(S) - \frac{4}{3}\tau_o(S) \leq c$
LPT ($B = 5$)	[8], [9]	$\tau_{LPT}(S) - \frac{8}{5}\tau_o(S) \leq 1$
H5 ($B = 5$)	[16]	$\tau_{H5}(S) - \frac{5}{4}\tau_o(S) \leq c$
List Scheduling	[7]–[9]	$\tau_w(S) - \left(2 - \frac{1}{B-1}\right)\tau_o(S) \leq \frac{B}{B-1}$
LPT	[8], [9]	$\tau_{LPT}(S) - \left(2 - \frac{2}{B}\right)\tau_o(S) \leq 1$
H ($B \geq 5$)	[16]	$\tau_H(S) - \frac{5}{4}\tau_o(S) \leq c$

It will be interesting to develop and model other realistic problem settings. We want to give a specific example to make our ideas clear. As a problem instance, we take $n = 5$ jobs $S_1, S_2, \ldots, S_5$ jobs with processing times $s_1 = 7, s_2 = 7, s_3 = 4, s_4 = 2, s_5 = 1$ and $\alpha = 10$ as recovery time for each one of the jobs.

1) In the STR formulation, the time α for a resource to recover is the same for all jobs and for all processing times. An interesting modification is to set α_i differently dependent of the job S_i, e.g. α_1 of job S_1 is $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$, $\alpha_4 = 4$, $\alpha_5 = 5$. This reflects the situation that - independently of the actual processing time - a job needs always the same amount to recover, but each job can have a different recovery time. The B-constraint 1 would have to be adjusted to that new situation. Transferred to the PSS formulation (where the server fulfills a load task) this situation reflects that each job S_i might have a different setup time s_i before its processing and then a different processing time α_i. And transferred to the PSS formulation (where the server fulfills an unload task) this situation reflects that each job S_i might have a different processing time s_i before it is unloaded which takes α_i time units.
2) Another modification is to change in the STR formulation the setting where all jobs have the same recovery time α to a setting where α depends on the processing time of the jobs. As an example, it might be reasonable that the recovery period is as longer as the processing time is. For the given problem instance, we could as an example think of $\alpha_i = 2 \cdot s_i$ so that we would get $\alpha_1 = 14$, $\alpha_2 = 14$, $\alpha_3 = 8$, $\alpha_4 = 4$, $\alpha_5 = 2$. Again, transferred to the PSS formulation (where the server fulfills a load task): The setup time s_i is always a constant factor larger (or smaller) than the actual processing time α_i of the job. Or, transferred to the PSS formulation (where the server fulfills an unload task): The unloading time α_i is always a constant factor larger (or smaller) than the actual processing time s_i.

It will be also interesting to investigate other objectives (for both, the preemptive and the non-preemptive case), such as L_{max}, $\sum C_j$, $\sum w_j C_j$, $\sum U_j$, $\sum w_j U_j$, $\sum T_j$, and $\sum w_j T_j$.

V. Conclusion

We gave a summary of existing results for the single processor scheduling problem with time restrictions and the related parallel processor scheduling problem with a single server. Moreover, we gave a sketch of future research possibilities in that area. In an extended version of this paper, it will be interesting to check which of the existing results for the parallel processor scheduling with single server can be transferred to the problem setting with single processor scheduling and time restrictions. In that way, additional insights in both problems will be gained.

References

[1] Abdekhodaee A.H., Wirth A., Scheduling parallel machines with a single server: some solvable cases and heuristics, Computers & Operations Research, volume 29, pp 295–315 (2002).

[2] Benmansour R., Braun O., Allaoui H., Modeling the single-processor scheduling problem with time restrictions as a parallel machine scheduling problem in: Z. Hanzalek, G. Kendall, B. McCollum, P. Sucha (Eds.), Proceedings of the 7th Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2015), Prague (Czech Republic), ISSN 2305-249X, pp 325-330 (2015).

[3] Benmansour R., Braun O., Artiba A., On the single-processor scheduling problem with time restrictions, in: I. Kacem, P. Laroche, Z. Roka (Eds.), Proceedings of the 2014 International Conference on Control, Decision and Information Technologies (CoDIT), Metz (France), ISBN 978-1-4799-6773-5, pp 242-245 (2014).

[4] Benmansour R., Braun O., Artiba A., Mixed integer programming formulations for the single processor scheduling problem with time restrictions, in: I. Kacem, S. Martin, Z. Roka, A. Sava (Eds.), Proceedings of the 45th International Conference on Computers & Industrial Engineering (CIE45), Metz (France), ISBN 978-1-5108-1745-6, pp 288-296 (2016).

[5] Benmansour R., Braun O., Hanafi S., The single processor scheduling problem with time restrictions: complexity and related problems, under review (2017).

[6] Benmansour R., Braun O., Hanafi S., A variable neighborhood search for the single processor scheduling problem with time restrictions, working paper (2017).

[7] Braun O., Chung F., Graham R., Single-processor scheduling with time restrictions, Journal of Scheduling, volume 17, pp 399–403, (2014).

[8] Braun O., Chung F., Graham R., Bounds on single processor scheduling with time restrictions, in: T. Fliedner, R. Kolisch, A. Naber (Eds.), Proceedings of the 14th International Conference on Project Management and Scheduling, Munich (Germany), ISBN 978-3-00-045630-5, pp 48-51 (2014).

[9] Braun O., Chung F., Graham R., Worst-case analysis of the LPT algorithm for single processor scheduling with time restrictions. OR Spectrum, volume 38, pp 531–540 (2016).

[10] Brucker P., Dhaenens-Flipo C., Knust S., Kravchenko S.A., Werner F., Complexity results for parallel machine problems with a single server, Journal of Scheduling, volume 5, pp 429–457 (2002).

[11] Graham R.L., Bounds on certain multiprocessing anomalies, Bell System Technical Journal, volume 45, pp 1563–1581 (1966).

[12] Graham R.L., Bounds on multiprocessing timing anomalies, SIAM Journal on Applied Mathematics, volume 17, pp 416–429 (1969).

[13] Hall N.G., Potts C.N., Sriskandarajah C., Parallel machine scheduling with a common server, Discrete Applied Mathematics, volume 102, pp 223–243 (2000).

[14] Kravchenko S.A., Werner F., Parallel machine scheduling problems with a single server, Mathematical and Computer Modelling, volume 26, pp 1–11 (1997).

[15] Wilbaut C., Benmansour R., Braun O., Hanafi S., Iterative relaxation-based heuristic for the single-processor scheduling problem with time restrictions, in: Proceedings of the International Conference on Industrial Engineering and Systems Management (IESM2015), Sevilla (Spain), ISBN 978-2-9600532-6-5, pp 496-501 (2016).

[16] Zhang A., Ye F., Chen Y., Chen G., Better permutations for the single-processor scheduling with time restrictions, Optimization Letters, volume 11, pp 715–724 (2016).

[17] Zhang A., Ye F., Chen Y., Chen G., On the NP-hardness of scheduling with time restrictions, arXiv:1703.00575vl [cs.DS] 2 Mar 2017 (2017).

Unrelated parallel machines with sequence dependent setup times and machine eligibility

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Fernando Carrasco*, Paz Perez-Gonzalez*, Victor Fernandez-Viagas* and Jose Manuel Framinan*
* Industrial Management, University of Seville
Avda de los descubrimientos 41092, Sevilla (Spain)
Email:reffmcd@gmail.com, pazperez@us.es,
vfernandezviagas@us.es, framinan@us.es

***Abstract*—This paper addresses an unrelated parallel machines with sequence dependent setup times and machine eligibility, minimising the total tardiness (i.e. $Rm|M_j|s_{ijk}|\sum T_j$). This problem is a real manufacturing problem inspired in a metal-mechanic company of south Spain. To the best of our knowledge, this problem has not been tackled in the literature so far. In this paper, we model the problem by an mixed integer linear programming model. In addition we adapt existing constructive methods applied to similar problems in order to give approximate solutions in short times.**

I. Introduction

In this paper we consider a problem inspired in the situation presented in a metal-mechanic company, where sheets are bent in folder machines. In this company folder machines are all different, some of them are supplied manually and the rest by a robot. Features of the machines imply different processing times for the same piece on each machine (unrelated parallel machines). In addition, not all machines are capable to process all pieces, implying machine eligibility for the jobs. Due to design aspects of the pieces, different setup times are required to process each piece, implying machine and sequence dependent setup times. As each piece has a due date related to the customer orders, the production decision maker is focused in the minimization of the total tardiness to increase the satisfaction of the customers. Therefore, according to [1], the problem under consideration can be denoted by $Rm|M_j, s_{ijk}|\sum T_j$.

Daily scheduling implies more than 200 pieces to be processed by the company. Due to needs of dynamic production scheduling in the facility, the idea is to implement some constructive heuristic which provides in short times approximate solutions for instances with more than 200 jobs. As this problem has not been tackled in the literature by constructive heuristics, we adapt those applied to similar problems. More specifically we analysed constructive methods for unrelated parallel machine problems with machine eligibility, and unrelated parallel machine problems with setup times.

The structure of the paper is as follows: we review the related literature in Section II. The problem is modelled by a mixed integer linear programming model in Section III. Constructive heuristics identified in the literature are described in Section IV. In order to determine the quality of the solutions provided by the heuristics, we compare the solutions to the optimal solutions provided by the mathematical model for small instances in Section V. Conclusions and future research lines are presented in Section VI.

II. Related literature

Unrelated paralell machine problems considering machine eligibility has been scarcely studied in the literature. To the best of our knowledge, only $Rm|M_j|\sum C_j$ is polynomial, since $Rm||\sum C_j$ is [2]. The first paper considering a scheduling problem on unrelated parallel machines and machine eligibility is [3]. They analyse the feasibility of the problem $Rm|M_j, s_{ijk}, r_j, d_j|-$ by graph representation.

Regarding the makespan objective, reference [4] proposes a heuristic for small instances of the problem $Rm|M_j|C_{max}$, providing a computational study comparing the results to the optimal solutions. In [5] authors consider a more complicated problem than the $Rm|M_j|C_{max}$ including resource constraints, i.e. additional resources must be available to process jobs. They propose a mathematical programming model and two genetic algorithms. More recently, the same authors tackle a more constrained problem in [6], by adding to the previous problem sequence dependent setup times, release dates and precedence constraint. They apply two metaheuristics, a genetic algorithm and an artificial immune system algorithm. $Rm|M_j, s_{ijk}|C_{max}$ is considered by [7], including the possibility of producing imperfect items and rework processes to achieve an acceptable quality level for these imperfect items. Authors apply some meta-heuristics (genetic and bees algorithms) to provide approximate solutions. $Rm|M_j|C_{max}$ including controllable processing times is addresses by [8]. They solve optimally small size instances of the problem by a integer linear programming model, and propose two ant colony optimization metaheuristics to provide approximate solutions for big size instances.

Regarding due dates related objectives, $Rm|M_j, \text{family-setup}|\max T_j$ is considered by [9]. In this paper, authors consider setup times only when a job is scheduled after a job of different type. They propose a heuristic based on tabu search. $Rm|M_j, r_j, s_{jk}|\sum w_j T_j$ is solved optimally for small instances by [10], by an integer programming model based on the vehicle route problem. Finally, Reference [11] describes mathematically the problem $Rm|M_j, r_j, s_{ijk}, prec|\sum L_j$ by an integer programming model to solve the problem optimally for small instances. Additionally, they provide two metaheuristics,

more specifically a genetic algorithm and an ant colony optimization algorithm.

Regarding unrelated parallel machine problems considering sequence dependent setup times, $Rm|s_{ijk}|C_{max}$ has been considered by [12], where a metaheuristic, Meta-RaPS is proposed to provide approximate solutions to large size instances. Additionally, reference [13] considers an Ant Colony Optimization algorithm for the same problem. Reference [14] propose two Genetic Algorithms, and reference [15] a Simulated Annealing method. In both studies their proposals are compared and outperform to previous methods from the literature. Reference [16] presents an immune-inspired algorithm, using the Al-Salem instances [17] to compare their proposal to previous methods from the literature. Finally, reference [18] provides an integer linear programming model to solve optimally up to 30 jobs instances for the problem $Rm|s_{ijk}|C_{max} + \sum w_j T_j$.

III. Integer Linear Programming Model

The problem under consideration can be modelled as an adapted version of the mixed integer linear programming (MILP) model proposed by [14] for $Rm|s_{ijk}|C_{max}$, based on [19].

The data of the model is presented in the following:

i	index for machines
j, k	index for jobs
N	Set of jobs
N^*	Set of jobs union a dummy job 0
M	Set of machines
p_{ij}	processing time of job j
s_{ijk}	sequence-dependent setup time on machine i when job j is scheduled before job k
d_j	due date of job j
V	a big number

Decision variables are:

X_{ijk}	$\begin{cases}1 & \text{if job } j \text{ precedes job } k \text{ on machine } i\\ 0 & \text{otherwise}\end{cases}$
C_{ij}	Completion time of job j on machine i
T_j	Tardiness of job j

$$Minimize \sum_{j \in N} T_j \tag{1}$$

s.t.

$$T_j \geq C_{ij} - dj \tag{2}$$
$$\forall i \in M,\ \forall j \in N$$

$$\sum_{i \in M} \sum_{\substack{j \in N^* \\ j \neq k}} X_{ijk} = 1 \qquad \forall k \in N \tag{3}$$

$$\sum_{i \in M} \sum_{\substack{k \in N \\ j \neq k}} X_{ijk} \leq 1 \qquad \forall j \in N \tag{4}$$

$$\sum_{k \in N} X_{i0k} \leq 1 \qquad \forall i \in M \tag{5}$$

$$\sum_{\substack{h \in N^* \\ h \neq j}} X_{ihj} \geq X_{ijk} \tag{6}$$
$$\forall j,\ k \in N,\ j \neq k,\ \forall i \in M$$

$$C_{ik} + V(1 - X_{ijk}) \geq C_{ij} + s_{ijk} + p_{ik} \tag{7}$$
$$\forall j \in N^*,\ \forall k \in N,\ j \neq k,\ \forall i \in M$$

$$C_{i0} = 0 \qquad \forall i \in M \tag{8}$$

$$X_{ijk} \in \{0, 1\} \tag{9}$$
$$\forall j \neq k \in N^*,\ \forall i \in M$$

$$T_j \geq 0 \qquad \forall j \in N \tag{10}$$

$$C_{ij} \geq 0 \qquad \forall j \in N,\ \forall i \in M \tag{11}$$

Equation (1) is the objective to minimize, i.e., the total tardiness. Constraint set (2) computes the tardiness of each job j. Constraint set (3) ensures that each job has only one predecessor and it is assigned to only one machine. Constraint set (4) ensures that the maximum number of successors of job j is one. Constraint set (5) ensures that at least one job is assigned for each machine i. Constraint set (6) checks, for each machine i, if job j is scheduled before job k on that machine, j has at least one predecessor (i.e. at least the dummy job). Constraint set (7) imposes that, that a job must not start to be processed until its predecessor has been processed. Constraint set (8) guarantees that the dummy job is the first to be assigned on each machine i. Constraint set (9) define the binary variable about precedences between jobs. Remaining sets (10, 11) define the continuous variables of the model explained previously.

IV. Constructive heuristics

In order to provide approximate solutions for big (more realistic) size instances of the problem, we consider different constructive heuristics from the literature, in order to determine their performance for our problem. We have considered the following constructive heuristics:

- Al-Salem and Armacost [17]. This method has been developed for the problem $Rm|M_j|C_{max}$, so it has been adapted to $\sum T_j$. Additionally, as they do not consider setup times, we add a scheduling phase, trying to improve the result according to the sequence dependent setup times. This method is described in Section IV-A.
- Rabadi et al [12]. Authors develop a metaheuristic called MetaRaPS for the problem $Rm|s_{ijk}|C_{max}$. The first step of this method is the constructive heuristic described here. This method does not consider machine eligibility, so the constructive heuristic can provide unfeasible solutions for our proposed problem. Then, we modify some steps in order to obtain feasible solutions for our problem. The adapted method is described in Section IV-B.

A. Constructive heuristic: Al-Salem

Constructive heuristic proposed by [17], depends on a given parameter $\alpha > 0$. The algorithm is described in Algorithm 1, being C_i the maximum completion time of jobs assigned to a machine i; $i^1 = arg\min_{i \in M} p_{ij}$ is the index of the machine where a given job j has the shortest processing time; and $i^2 =$

$arg\min_{i \in M-\{i^1\}} p_{ij}$ be the machine where this jobs has the second shortest processing time. Additionally,

$$\rho_j = \frac{p_{i^1,j}}{p_{i^2,j}}$$

and

$$p_j = \frac{1}{|M_j|} \sum_{i \in M_j} p_{ij}$$

Algorithm 1 Constructive heuristic: Al-Salem

```
procedure ALSALEM(α)
    Step 1
    Set PJ = ∅
    Set C_i = 0, ∀i
    for j = 1,...,n do
        if |M_j| = 1 then
            Assign j to its eligible machine i(j)
            C_{i(j)} = C_{i(j)} + p_{i(j)j}
        else
            if ρ_j ≤ α then
                Assign j to i^1
            else
                PJ = PJ ∪ {j}
            end if
        end if
    end for
    Step 2
    Set PJ = J_2 ∪ ··· ∪ J_m such as
    J_h = {j ∈ PJ/|M_j| = h}
    for l=2,...,m do
        Sort jobs in J_l in non-increasing order of p_j
        i(j) = min_{i∈Mj}{C_i + p_ij}
        Assign j to i(j)
        C_{i(j)} = C_{i(j)} + p_{i(j)j}
    end for
end procedure
```

This constructive heuristic provides solutions which do not consider sequence dependent setup times. Actually, sequence is not relevant in this method since it is developed to minimize makespan. Therefore, we have scheduled jobs assigned on each machine following the constructive heuristic proposed by [20] for the problem $1|s_{jk}|\sum w_j T_j$. This method schedules jobs according the following priority index on each machine i, being scheduled first the job with the highest value of the index.

$$I_j(C_{il}, l) = \frac{1}{p_{ij}} exp - \left\{ \frac{\max\{d_j - p_{ij} - C_{il}, 0\}}{k_1 \overline{p}_i} \right\} exp \left\{ \frac{-s_{lj}}{k_2 \overline{s}_i} \right\}$$

where C_{il} is the completion time of the last job l scheduled in machine i, $\overline{s}_i$ is the average setup time of jobs assigned to machine i, $\overline{p}_i$ is the average processing time of jobs assigned to machine i, and k_1 and k_2 depend on the due date tightness factor τ; the due date range factor R; and the setup time severy factor η in the following way:

$$k_1 = \begin{cases} 4.5 + R & R \leq 0.5 \\ 6.0 - 2R & R > 0.5 \end{cases}$$

$$k_2 = \tau / 2\sqrt{\eta}$$

TABLE I. DUE DATE GENERATION FOR UNRELATED PARALLEL MACHINES

Due date generation	Reference
$U[d - w/2, d + w/2]$	[22]
$U[\frac{n\overline{p}}{m}(1 - \tau - R/2), \frac{n\overline{p}}{m}(1 - \tau + R/2)]$	[23]
$U[n\overline{p}(1 - \tau - R/2), n\overline{p}(1 - \tau + R/2)]$	[24]
$U[C^1_{max}(1 - \tau - R/2), C^1_{max}(1 - \tau + R/2)]$	[9]
$U[C^2_{max}(1 - \tau - R/2), C^2_{max}(1 - \tau + R/2)]$	[25]
$U[C^3_{max}(1 - \tau - R/2), C^3_{max}(1 - \tau + R/2)]$	[26]
$n^2\overline{p}(1 - U[0, 1])$	[27]

with τ, R and η defined as follows:

$$\tau = 1 - \frac{\overline{d}}{n(\overline{p}_i + \beta\overline{s}_i)};\ R = \frac{d_{max} - d_{min}}{n(\overline{p}_i + \beta\overline{s}_i)};\ \eta = \frac{\overline{s}_i}{\overline{p}_i}$$

with $\overline{d}$ the average due date of jobs. Finally, β depends on the variability coefficient $c_v = \frac{V[s]}{\overline{s}_i^2}$ (with $V[s]$ is the variance of the setup times), and n_i (the number of jobs assigned in the machine i) according to a graphic given by [20]. The algorithm obtained applying the method to AlSalem is denoted AlSalem_impr.

B. Constructive heuristic: Rabadi

Constructive heuristic developed by [12] does not consider machine eligibility, so it cannot be implemented as it is since it provides unfeasible solutions. Therefore, we have adapted it, indicating with $Step^*$ those modified steps respect to the original method in Algorithm 2.

Let $[AP]_i$ be the adjusted processing time matrix for each machine i, such as $AP_{ijk} = s_{ijk} + p_{ij}$, $\forall j \in N^*$, $\forall k \in N$. Let S be the schedule in the format {jobs in machine 1; jobs in machine 2;...; jobs in machine m}, $J_h = \{j \in N/|M_j| = h\}$, i.e the set of jobs with h machines eligible, $l(i)$ the last scheduled job in machine i, L_i the load of machine i, and finally $pos(i)$ the position of $l(i)$ in machine i.

V. COMPUTATIONAL EXPERIMENTS

The mathematical model as well as the constructive heuristics have been tested experimentally. For each instance of the problem, we have the following data: number of jobs n and machines m; processing times, p_{ij}; setup times s_{ijk}; due dates d_j; and the set of eligible machines by a job j, M_j. We follow Vallada [14], which provides a test bed for $Rm|s_{ijk}|C_{max}$, providing different values for n, m, and the distributions to generate p_{ij} and s_{ijk}. We do not use directly this test bed since we add new parameters, and consequently, more instances. To generate the due dates and machine eligibility we have analysed the related literature in Section V-A and Section V-B in order to find a suitable method of generation. The proposed test bed is explained in Section V-C.

A. Due dates generation

We have reviewed unrelated parallel machine literature in order to determine a suitable way to generate the due dates. Table I contains the proposals (most of them based on [21]). We have unified the notation in the following way: Tardiness factor τ, and range factor R are controlled parameters; $\overline{p} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}}{n \cdot m}$; $p_j = \min_{1 \leq i \leq m} p_{ij}$.

Algorithm 2 Rabadi algorithm

procedure RABADI
  Step 0^*
  Set $S = \emptyset$
  $J_h = \{j \in N / |M_j| = h\}$
  for i=1,...,m **do**
    $l(i) = 0$, $pos(i) = 0$
  end for
  Step 1^*
  $i^* = arg\min_{i \in M} L_i$
  Step 2^*
  for $l = 1, \ldots, n$ **do**
    if $J_l \neq \emptyset$ **then**
      $j^* = arg\min_{j \in J_l}\{AP_{il(i)j}\}$ (ties are broken arbitrarily)
      if $i^* \in M_{j^*}$ **then**
        Go to Step 3
      else
        Go to Step 1^* with $M = M - \{i^*\}$
      end if
    end if
  end for
  Step 3
  $pos(i^*) = pos(i^*) + 1$
  Assign j^* to machine i in S in position $pos(i^*)$
  Step 4
  $L_{i^*} = L_{i^*} + AP_{i^* l(i^*) j^*}$
  for $i = 1, \ldots, m$ **do**
    Eliminate column and row of $[AP]_i$ corresponding to job j^*
  end for
  Delete j^* from its set J_h
  Step 5
  if $J_h = \emptyset\ \forall h$ **then**
    STOP
  else
    Go to Step 1^*
  end if
end procedure

In the expression given by [22], the parameter d is $d = \overline{p}(1-\tau)n/m$. Additionally, $w = \frac{R}{m}\sum_{j=1}^{n} p_j$. Reference [9] introduces machine eligibility and setup times to estimate the makespan. Authors consider two types of setup times dependent on machines, s_{i1j} and s_{i2j}.

$$C^1_{max} = \frac{\sum_{j=1}^{n}(\sum_{i \in M_j} p_{ij} + s_{i1j} + s_{i2j})/|M_j|}{m}$$

In the expression provided by [25], the estimation of C_{max} is given considering sequence dependent setup times:

$$C^2_{max} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(p_{ij} + \sum_{k=1}^{n} s_{ijk}/n)/m}{m}$$

Reference [26] considers setup times dependent on machines, s_{ij}, with $s_j = \min_{1 \leq i \leq m} s_{ij}$, and their estimation of makespan is given by:

$$C^3_{max} = \frac{\sum_{j=1}^{n} p_j + s_j}{m}$$

We have selected the due date generation procedure from the proposal by [9], since it has been developed considering both, setup times and machine eligibility. We adapt the expression, given to estimate the makespan, to estimate the completion time for each job j, in order to obtain a more realistic estimation:

$$C_j = \frac{\sum_{i \in M_j} p_{ij} + \frac{\sum_{k=0^n} s_{ikj}}{n}}{|M_j|} \cdot \frac{n}{m}$$

So we compute an estimation of the makespan from the maximum value of C_j. Note that we have modified the setup times by sequence and machine depedent setup times s_{ijk}.

B. Machine eligibility generation

We have reviewed unrelated parallel machine literature with machine eligibility constraints, in order to identify how to generate this data. Reference [11] generates a binary matrix $n \times m$. If element on row j and column i is 1, it means that machine i is capable of processing job j. Authors impose at least, one element 1 in each row. References [6], [5] use a similar generation, imposing that, at least, there are $m/2$ elements 1 in each row. Reference [28] generates machine eligibility based on a constant me, which is defined relative to 10. To specify the eligibility of a particular job on a particular machine, authors generate random numbers in the range [1,10]. If the generated number is smaller than me, then the machine is set as ineligible to process that job; else it is set as eligible. They set $me = 3$ in their experiments. In reference [4], jobs are equally allocated to each group of machines. For instance, for the two machines problem, there are three possible groups: only machine 1; only machine 2; and both machines 1 and 2. Then if there are 30 jobs, ten jobs would be allocated to each group creating the machine eligibility. This allocation is randomly performed. In reference [29] the number of eligible machines of each job $|M_j|$ depends on the number of machines. They consider two options, $|M_j| \sim U[1, m/2]$ and $|M_j| \sim U[1, m]$. They do not specify what happens if a machine is not eligible.

We have generated machine eligibility combining the method provided by [28], controlling the probabilities by a constant denoted p in this paper, and [29], controlling the number of eligible machines for each job. Therefore, for each instance, for each job, we have generate machine eligibility in the following way:

Algorithm 3 Machine eligibility procedure

procedure MACHINEELIGIBILITYPROCEDURE(p)
  for $j = 1, \ldots, n$ **do**
    Generate a random number $k \in [1, 100]$
    if $k \leq p$ **then**
      $|M_j| \in U[1, m/2]$
    else
      $|M_j| \in U[1, m]$
    end if
    Select randomly $|M_j|$ machines eligible for j
  end for
  if $\exists i \in M$ such as $i \notin M_j\ \forall j \in N$ **then**
    Assign randomly a job to i
  end if
end procedure

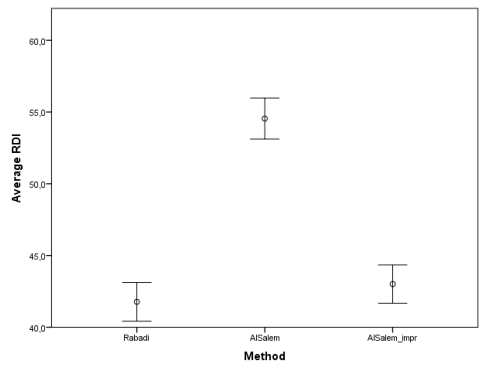


Fig. 1. Average RDI for the methods

C. Test bed generation

Based on Vallada test bed for $Rm|s_{ijk}|C_{max}$ [14], we have generated small instances (due to the computational effort needed to solve exactly the problem by the MILP model), similar to those proposed by Vallada, but including due dates and eligibility. Therefore, $n = \{6, 8, 10, 12\}$; $m = \{2, 3, 4, 5\}$; processing times are generated as usual in scheduling literature $p_{ij} \sim U[1, 99]$; and we have considered two of the four levels proposed by Vallada for setup times: $s_{ijk} \sim U[1, 49]$ and $U[1, 124]$. Due dates have been generated as explained in Section V-A, considering $\tau \in \{0.2, 0.4\}$ and $R \in \{0.2, 0.4\}$. Finally, machine eligibility for each job has been generated as explained in Section V-B with $p \in \{80, 50, 100\}$. 10 instances per combination have been constructed, in total 3840 instances, which have been solved by the MILP presented in Section III using Gurobi; and the two constructive heuristics (one of them with two versions) adapted to the proposed problem described in Section IV.

D. Results

Gurobi has solved optimally all 3840 small instances. In order to determine the best adaptation between the constructive heuristics, we compare the objective value provided for each method M for each instance I, FO_{IM}, by the relative deviation index [30] as follows:

$$RDI_{IM} = \frac{FO_{IM} - Best_I}{Worst_I - Best_I}$$

Where $Best_I$ is the total tardiness value provided by the MILP (optimal value) and $Worst_I$ is the worst total tardiness value provided by the methods. Methods applied are Rabadi, AlSalem and AlSalem_impr. Rabadi algorithm provides the best results for small instances, as it can be observed in Figure 1, although there are not statistical differences among Rabadi and Al-Salem. Table II shows average data by number of jobs n and number of machines m, including the total average for each method and computational times given in seconds, being almost instantaneous all of them. It can be observed that methods provide better results as n increases and m decreases. It can be expected that, for bigger instances where n increases significantly, the results provided can be closer to the optimal value.

TABLE II. COMPUTATIONAL RESULTS FOR SMALL INSTANCES

		RDI		
n	m	Rabadi	AlSalem	AlSalem_impr
6	2	44.4508	47.0517	47.4661
	3	43.0624	55.9578	47.7698
	4	49.8563	55.9369	57.8781
	5	49.8654	54.4721	54.7481
8	2	37.3454	45.9549	38.5092
	3	44.6044	54.3796	44.3473
	4	50.7664	59.0433	52.0538
	5	51.2225	59.6964	53.8863
10	2	33.3194	41.6681	31.0362
	3	34.9185	58.3701	34.9130
	4	45.4435	61.1494	46.0878
	5	46.8338	57.2866	49.1193
12	2	28.0467	43.3521	23.1971
	3	27.7139	60.7610	22.9281
	4	40.3675	55.1409	37.7834
	5	40.4987	62.4576	46.3964
	Total	41.7697	54.5424	43.0075
Computational Time		$2 \cdot 10^{-6}$	$85 \cdot 10^{-6}$	$443 \cdot 10^{-6}$

VI. CONCLUSION

In this paper we tackle a unrelated paralell machine problem with machine eligibility, to minimize the total tardiness, where setup times are sequence and machine dependent. This problem has been inspired in the real situation of a manufacturing process in a Spanish company. As the problem has not been considered in the literature previously, in this preliminary work we propose to solve it exactly by a mixer linear programming model, and approximately by two constructive heuristics adapted from the literature about similar problems on unrelated parallel machines. On one hand, we have adapted a method from the eligibility literature (called in this paper AlSalem algorithm), and, on the other hand, we have adapted a method from the sequence and machines dependent setup times (called in this paper Rabadi algorithm). Results revel that both adaptations provide similar results regarding to effectiveness and efficiency. Future research lines are focused on developing a method designed particularly for the problem, since results may be improved due to the differences among the solutions provided by the adapted methods and the optimal solutions.

REFERENCES

[1] R. Graham, E. L. Lawler, J. K. Lenstra, and A. Rinnooy Kan, "Optimization and heuristic in deterministic sequencing and scheduling: a survey," *Annals of Discrete Mathematics*, vol. 5, pp. 287–326, 1979.

[2] J. Y. Leung and C.-l. Li, "Scheduling with processing set restrictions : A survey," *Intern. Journal of Production Economics*, vol. 116, pp. 251–262, 2008.

[3] S. Kedad-Sidhoum and M. Mekhilef, "Unrelated parallel machine scheduling under constraints: Stability number of an associated graph," in *IEEE Symposium on Emerging Technologies & Factory Automation, ETFA*. IEEE, 1997, pp. 436–441. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0030651462&partnerID=tZOtx3y1

[4] A. Al-Salem and R. L. Armacost, "Unrelated machines scheduling with machine eligibility restrictions," *Engineering Journal of University of Qatar*, vol. 15, pp. 193–210, 2002. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0347653632&partnerID=tZOtx3y1

[5] M. Afzalirad and M. Shafipour, "Design of an efficient genetic algorithm for resource-constrained unrelated parallel machine scheduling problem with machine eligibility restrictions," *Journal of Intelligent Manufacturing*, jun 2015. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84932157064&partnerID=tZOtx3y1

[6] M. Afzalirad and J. Rezaeian, "Resource-constrained unrelated parallel machine scheduling problem with sequence dependent setup times, precedence constraints and machine eligibility restrictions," *Computers & Industrial Engineering*, vol. 98, pp. 40–52, 2016.

[7] M. Rambod and J. Rezaeian, "Robust meta-heuristics implementation for unrelated parallel machines scheduling problem with rework processes and machine eligibility restrictions," *Computers & Industrial Engineering*, vol. 77, pp. 15–28, nov 2014. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84907680845&partnerID=tZOtx3y1

[8] C. Low and G.-H. Wu, "Unrelated parallel-machine scheduling with controllable processing times and eligibility constraints to minimize the makespan," *Journal of Industrial and Production Engineering*, vol. 1015, no. February, pp. 1–8, 2016. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/21681015.2016.1139005

[9] J.-F. Chen, "Minimization of maximum tardiness on unrelated parallel machines with process restrictions and setups," *The International Journal of Advanced Manufacturing Technology*, vol. 29, no. 5, pp. 557–563, jan 2006. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-33745698694&partnerID=tZOtx3y1

[10] K. Sudtachat, "Unrelated parallel machine scheduling with sequence dependent setup time job precedence, release time and machine eligibility," in *CIE 2014 - 44th International Conference on Computers and Industrial Engineering and IMSS 2014 - 9th International Symposium on Intelligent Manufacturing and Service Systems, Joint International Symposium on "The Social Impacts of Developments in Informat*. Computers and Industrial Engineering, 2014, pp. 2324–2332. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84923914305&partnerID=tZOtx3y1

[11] M. Afzalirad and J. Rezaeian, "Design of high-performing hybrid meta-heuristics for unrelated parallel machine scheduling with machine eligibility and precedence constraints," *Engineering Optimization*, pp. 1–21, jun 2015. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84930395455&partnerID=tZOtx3y1

[12] G. Rabadi, R. Moraga, and A. Al-Salem, "Heuristics for the unrelated parallel machine scheduling problem with setup times," *Journal of Intelligent Manufacturing*, vol. 17, pp. 85–97, 2006.

[13] J.-P. Arnaout, G. Rabadi, and R. Musa, "A two-stage Ant Colony Optimization algorithm to minimize the makespan on unrelated parallel machines with sequence-dependent setup times," *Journal of Intelligent Manufacturing*, vol. 21, no. 6, pp. 693–701, feb 2010. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-78650175630&partnerID=tZOtx3y1

[14] E. Vallada and R. Ruiz, "A genetic algorithm for the unrelated parallel machine scheduling problem with sequence dependent setup times," *European Journal of Operational Research*, vol. 211, no. 3, pp. 612–622, jun 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221711000142

[15] K.-C. Ying, Z.-J. Lee, and S.-W. Lin, "Makespan minimization for scheduling unrelated parallel machines with setup times," *Journal of Intelligent Manufacturing*, vol. 23, no. 5, pp. 1795–1803, nov 2012. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84870880734&partnerID=tZOtx3y1

[16] R. O. M. Diana, M. F. de França Filho, S. R. de Souza, and J. F. de Almeida Vitor, "An immune-inspired algorithm for an unrelated parallel machines' scheduling problem with sequence and machine dependent setup-times for makespan minimisation," *Neurocomputing*, vol. 163, pp. 94–105, sep 2015. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84930274201&partnerID=tZOtx3y1

[17] A. Al-Salem and R. L. Armacost, "Unrelated machines scheduling with machine eligibility restrictions," *Engineering Journal of University of Qatar*, vol. 15, pp. 193–210, 2002. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0347653632&partnerID=tZOtx3y1

[18] P. L. Rocha, M. G. Ravetti, G. R. Mateus, and P. M. Pardalos, "Exact algorithms for a scheduling problem with unrelated parallel machines and sequence and machine-dependent setup times," *Computers & Operations Research*, vol. 35, no. 4, pp. 1250–1264, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054806001614

[19] A. Guinet, "Scheduling sequence-dependent jobs on identical parallel machines to minimize completion time criteria," *International Journal of Production Research*, vol. 31, no. 7, pp. 1579–1594, 1993. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0027625565&doi=10.1080%2F00207549308956810&partnerID=40&md5=03bc7

[20] Y. H. Lee, "A heuristic to minimize the total weighted tardiness with sequence-dependent setups," *IIE Transactions (Institute of Industrial Engineers)*, vol. 29, no. 1, pp. 45–52, 1997. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030735290&partnerID=40&md5=526d4383f282e21456aaa282694e7c09

[21] C. N. Potts and L. N. Van Wassenhove, "A decomposition algorithm for the single machine total tardiness problem," *Operations Research Letters*, vol. 1, no. 5, pp. 177–181, 1982.

[22] V. Suresh and D. Chaudhuri, "Minimizing maximum tardiness for unrelated parallel machines," *International Journal of Production Economics*, vol. 34, no. 2, pp. 223–229, mar 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0925527394900388

[23] C.-F. Liaw, Y.-K. Lin, C.-Y. Cheng, and M. Chen, "Scheduling unrelated parallel machines to minimize total weighted tardiness," *Computers & Operations Research*, vol. 30, no. 12, pp. 1777–1789, oct 2003. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0042311378&partnerID=tZOtx3y1

[24] D. G. Na, D. W. Kim, W. Jang, and F. F. Chen, "Scheduling unrelated parallel machines to minimize total weighted tardiness," in *2006 IEEE International Conference on Service Operations and Logistics, and Informatics*. IEEE, nov 2006, pp. 758–763. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-36049048634&partnerID=tZOtx3y1

[25] S.-W. Lin, C.-C. Lu, and K.-C. Ying, "Minimization of total tardiness on unrelated parallel machines with sequence- and machine-dependent setup times under due date constraints," *The International Journal of Advanced Manufacturing Technology*, vol. 53, no. 1-4, pp. 353–361, jul 2010. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-79955573105&partnerID=tZOtx3y1

[26] J.-H. Lee, J.-M. Yu, and D.-H. Lee, "A tabu search algorithm for unrelated parallel machine scheduling with sequence- and machine-dependent setups: minimizing total tardiness," *The International Journal of Advanced Manufacturing Technology*, vol. 69, no. 9-12, pp. 2081–2089, jul 2013. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84892373938&partnerID=tZOtx3y1

[27] M. Naderi-Beni, E. Ghobadian, S. Ebrahimnejad, and R. Tavakkoli-Moghaddam, "Fuzzy bi-objective formulation for a parallel machine scheduling problem with machine eligibility restrictions and sequence-dependent setup times," *International Journal of Production Research*, vol. 52, no. 19, pp. 5799–5822, may 2014. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84906785434&partnerID=tZOtx3y1

[28] O. Alagöz and M. Azizolu, "Rescheduling of identical parallel machines under machine eligibility constraints," *European Journal of Operational Research*, vol. 149, no. 3, pp. 523–532, sep 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S037722170200499X

[29] R. Gokhale and M. Mathirajan, "Scheduling identical parallel machines with machine eligibility restrictions to minimize total weighted flowtime in automobile gear manufacturing," *The International Journal of Advanced Manufacturing Technology*, vol. 60, no. 9-12, pp. 1099–1110, oct 2012. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84861703325&partnerID=tZOtx3y1

[30] V. Fernandez-Viagas and J. M. Framinan, "NEH-based heuristics for the permutation flowshop scheduling problem to minimise total tardiness," vol. 60, pp. 27–36, 2015.

Resolution of a Job-Shop problem with Precedence Constraints between jobs (JSPC)

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Matthieu Gondran, Philippe Lacomme,
Enzo Mura, Nikolay Tchernev,
Université Clermont Auvergne, (LIMOS, UMR CNRS 6158),
63177 Aubière, France
gondran@isima.fr, placomme@isima.fr, enzo.mura@etudiant.univ-bpclermont.fr, tchernev@isima.fr

Abstract— This paper concerns the resolution of job-shop scheduling problem with precedence constraints between jobs. The resolution scheme is based on a GRASP×ELS framework which takes advantages of an indirect modelization of the solutions. The experiments prove the method competes with the best published methods of the Job-Shop and provides solutions less than 1% of the lower bound for small and medium scale instances.

Keywords—Job-Shop, Scheduling, Precedence Constraints, GRASP×ELS

I. Introduction

A. Definition of the Job-Shop

The Job-Shop Scheduling Problem (JSSP) is a well-known optimization problem that can be formulated as follows: a set of n jobs ($i = 1..n$) is define by an ordered sequence of machine operation. Each operation O_{ik} is related with a specific job i and a rank k (in the sequence of operations), has a duration p_{ik}, has to be processed on a machine belonging to a set of m machines ($j = 1..m$) and can begin only if all the previous machine operations of the sequence are finished. Each machine can process only one operation at a time. The objective is to schedule each operation on the machines, with respect to the ordered sequence such that the makespan (Cmax) is minimized. According to [5] the problem is well-known to be NP-hard.

A helpful problem representation is the disjunctive graph model due to Roy and Sussmann [10]. Using that model, any instance of the job shop scheduling problem can be modeled by a directed graph $G = (V; A; E)$, where V is the set of nodes, A the set of conjunctive arcs between two consecutive machine operations belonging to the same job, and E the set of disjunctive arcs between two operations belonging to the same machine. V contains nodes modeling machine operation, a source node 0 and a sink node $*$. The source node 0 is connected to the first operation of each job, and the sink node $*$ to the last.

A solution is modeled by an acyclic graph that contains conjunctive arcs and disjunctive arcs between operations using the same machine. An optimal solution is defined by an acyclic graph with the minimal makespan Cmax providing the earliest starting time of each operation thanks to a longest path algorithm.

The classical job-shop scheduling problem makes the following assumptions:

- Each job consists of a finite number of operations.
- The processing time for each operation using a particular machine is known.
- There is a sequence of operations that has to be achieved to complete each job.
- Each operation is performed on one machine only.
- A machine can process only one operation at a time.
- Preemption is not allowed.
- Due date constraints are not defined.
- There is no setup.

B. Definition of the Job-Shop with precedence constraints

Each job i is related to S_i a set of jobs which must be finished before starting the job i. S_i defines a new set of constraints between the last operation of a job $j \in S_i$ and the first operation of the job i. Let us consider an example of JSSP composed of 3 jobs with 3 operations defined in Table 1. For each job, this table introduces the set of operations and for each operation the machines (m_1, m_2 or m_3) and the processing time on the machines.

Table 1 Example of JSSP instance data

Op→ Jobs↓	O_{i1}	O_{i2}	O_{i3}	S
Job 1	(m_1, 10)	(m_2, 35)	(m_3, 25)	{2,3}
Job 2	(m_1, 15)	(m_3, 16)	(m_2, 12)	∅
Job 3	(m_3, 11)	(m_1, 12)	(m_2, 21)	∅

If $S_1 = \{2,3\}$ it implies that O_{11} must be scheduled after both O_{33} and O_{23} which defines two extra disjunctive arcs in the graph as stressed in red on the Figure 1.

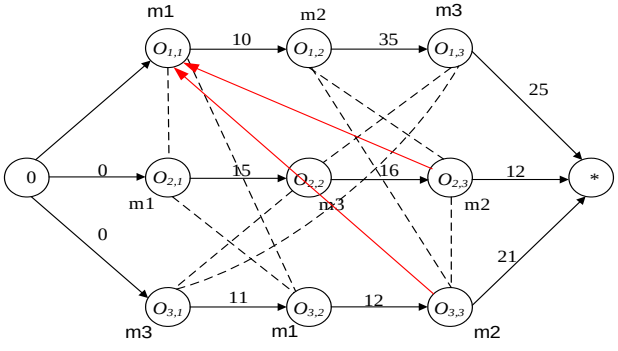


Fig. 1. Disjunctive graph of the job-shop with precedence constraints (problem modelling).

A solution of the problem is fully defined by a fully oriented acyclic graph in which the earliest starting time of operation have been obtained using any longest path algorithm.

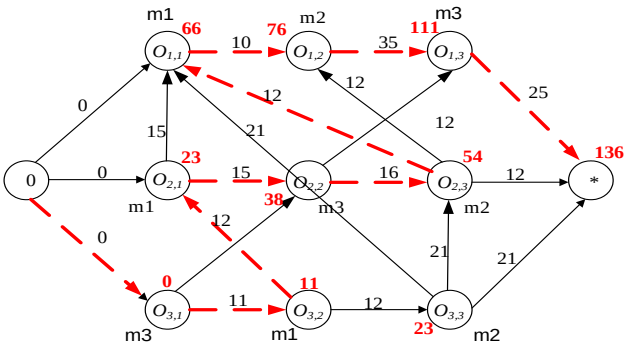


Fig. 2. Example of one solution.

The makespan value is 136 (Figure 2) and the earliest starting times (in bold red in the graph) define a semi-active solution. The dotted bold red arcs defines the longest path in the graph. The gantt chart modeling is introduced on the Figure 3.

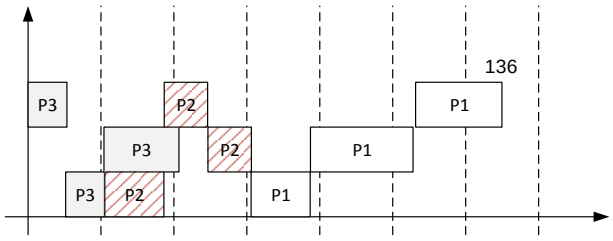


Fig. 3. Gantt chart of one solution.

II. Proposition

A. Key-points for an efficient resolution scheme

The resolution scheme we promote is based on:

- a Quasi-Direct Representation of Solution;
- an efficient local search taking advantages of the longest path analysis.

The solution search space must be investigated efficiently, avoiding premature convergence and trying to favor non explored search space area (detection of clones, memory management avoiding similar curves in the search space…). Let us note the publication of Cheng [2] which introduces a classification of the mapping functions between one object and a solution. According to the authors three types of mapping function (Figure 4) can be defined:

- 1-to-1 mapping function where one and only one object of the coding space is linked to one and only one solution;
- n-to-1 mapping function where several objects in the coding space can be linked to the same solution;
- 1-to-n mapping function where one object in the coding space can be linked to several solutions.

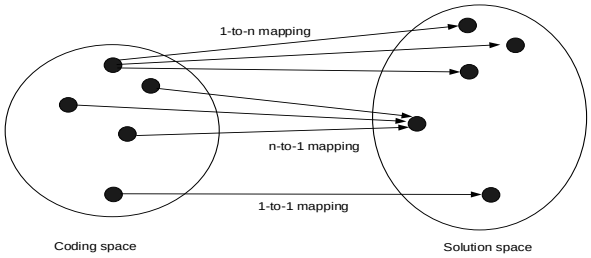


Fig. 4. The mapping function classification (Cheng *et al.*, 1996).

In the past decades, numerous proposals dedicated to the JSSP have been introduced including but not limited to the Bierwirth's proposition which takes advantage of a vector with repetitions where each job number appears iteratively, and where the number of occurrences is equal to the number of machine operations of the job. This vector defines a n-to-1 mapping function [1], and this representation is usually referred to as 1) Bierwirth's vector or 2) vector with repetition or 3) permutation job list. The main advantages of Bierwirth's proposal concerns the local search which could be achieved in the coding space by definition of a basic but strongly efficient modification of the vector avoid time consuming modification of the graph.

B. Extension of the Bierwirth's vector

For the JSSP with precedence constraints, a Bierwirth's vector can lead to cyclic graphs due to the precedence constraints between jobs. The figure 5 provides a fully oriented disjunctive graph obtained thanks to the vector λ =[3 3 2 2 3 1 1 2 1] but which encompasses a cycle due to the sequence O_{23}, O_{11} and O_{12}.

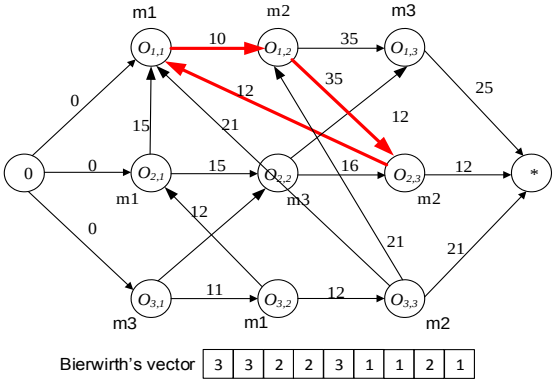


Fig. 5. A Bierwirth's vector leading to an unprofitable cyclic graph.

In order to avoid the generation of unprofitable Bierwirth vector, there is a great interest in generating Bierwirth vector leading to acyclic graph only. To build such extension of the Bierwirth vector, the heuristic building the Bierwirth vectors is based on 4 steps:

- definition of a graph G';
- jobs partitioning into d-partition computed in G';
- for each d-partition d_i a vector V_i is generated by randomly inserting all jobs belonging to the d-partition d_i;
- all the vectors V_i are combined together to form the extended Bierwirth's vector $\lambda = (V_1, V_2, \dots, V_i, \dots, V_{max})$.

The graph $G' = (V', E')$ where V' is the set of nodes representing the jobs (and not the operations of each job as the graph G), and E' is the set of disjunctions between the jobs according to the set of constraints S_i for each job i. Each arc is valued by 1. If a job i has $S_i = \{j, k, l\}$, there are arcs from j, k and l to i. If a job i has $S_i = \{\emptyset\}$, there is one arc from O to i. The Figure 6 illustrates the graph G' of the Table 1.

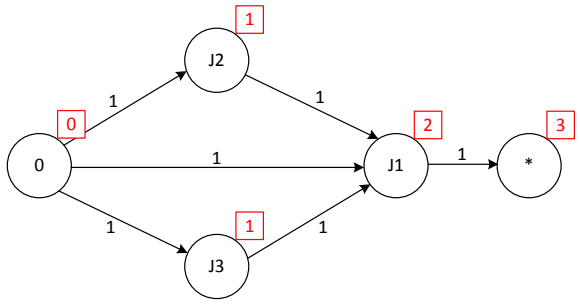


Fig. 6. Definition of a graph G'.

In order to partition the graph G', the longest path between each job i and the node O is computed. Every job with the longest path valuing 1 belongs to the same partition d_1, every job with a longest path valuing 2 belongs to d_2, and so on. The value of the longest path between each job and the node 0 is stressed on the Figure 6. Because, the longest path is computed in a graph G' which encompasses only the job precedence constraints, the d-partition is a pre-processing.

```
Algorithm 1: compute_partition()
 1.Input
 2.   G: graph of the pickup nodes
 3.Output
 4.   V': d-partition of V
 5.Begin
 6.   k := 0; Vk := ∅
 7.   while (N ≠ ∅) do
 8.      for all i ∈ N do
 9.         if (Γi⁻ = ∅) then
10.            Vk = Vk ∪ {i}
11.            N = N\{i}
12.         end if
13.      end for
14.      k := k+1
15.   end while
16.   return V'
17.End
```

The procedure compute_Partition(), Algorithm 1, first starts with an empty d-partition at line 6 and the subsets V_k are iteratively filled, starting from $k = 0$. Γ_i^- defines the set of predecessor of nodes which includes the previous machine operation in the job sequence and the disjunctive arcs defining job precedence constraints. For example in the graph of figure xxx, $\Gamma_{O_{11}}^- = \{O_{23}, O_{33}\}$. Considering the current iteration k, all the nodes $i \in N$ (where N is the set of machine operation) such that $\Gamma_i^- = \emptyset$ are inserted into the subset V_k and these nodes are then removed from N before starting the next iteration.

Considering a d-partition of the graph G', each partition d_i is associated to a vector V_i, which is a classical Bierwirth's vector randomly generated with all jobs belonging to d_i: each job is repeated as many time it has operations in its sequence. For example, $V_1 = (2,2,3,2,3,3)$.

All vectors V_i are combined end-to-end in the lexicographic order to generate the extended Bierwirth's vector.

C. Local search

The algorithm loops along the vertices of the critical path from the last operation to the first one. At each iteration, a neighbor can be created if the vertices α and β are two operations of different jobs and if there is no precedence constraint between these jobs. The neighbor is generated by trying to permute the operations in the Bierwirth's vector. The algorithm computes the maximal right position of α, *i.e.* $\lambda^+(\alpha)$ in the vector, and the minimal left position of β *i.e.* $\lambda^-(\beta)$. If $\lambda^+(\alpha) > \lambda(\beta)$ then the permutation of α and β is achieved by a right-shifted of α to $\lambda^+(\alpha)$ and by a left-shifted of β to $\lambda^-(\beta)$ as illustrated on Figure 7.

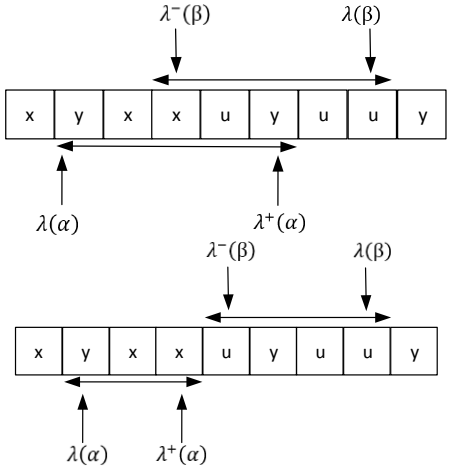


Fig. 7. Two possible situations.

If $\lambda^+(\alpha) < \lambda^-(\beta)$, a swap of α and β leads to a vector which must be repaired. The reparation scans all the successors of α from $\lambda(\alpha)$ to $\lambda(\beta)$, switches them after the $\lambda(\beta)$, searches all the predecessor of β and next switches them before $\lambda(\alpha)$.

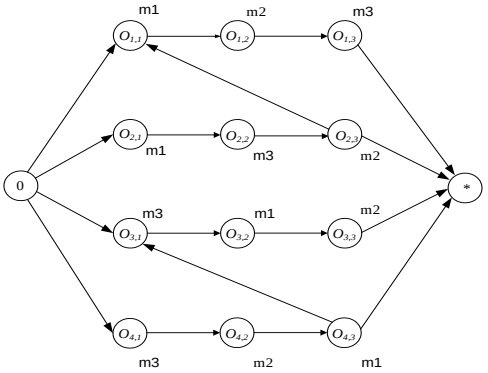


Fig. 8. Graph with 4 jobs and 2 precedence constraints.

The example in Figure 9 stresses the reparation steps to obtain the Bierwirth's vector of the graph in Figure 8. The different reparation steps consider both a precedence constraint from the job 2 and the job 1 and a precedence constraint from the job 4 to the job 3. The first step consists in swapping the second machine-operation of job 2 and the third machine-operation of

job 3. Then the second step switches all the jobs 4 before the job 3. Finally the procedure switches all the jobs 1 after the job 2.

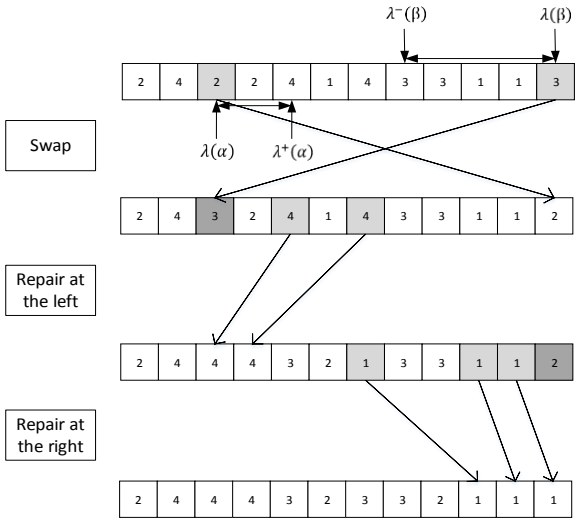


Fig. 9. Example of reparation on Bierwirth's vector

D. Principle of GRASP×ELS

The metaheuristic used in this article comes from an Evolutionary Local Search – ELS from Wolf and Merz [11] where a neighborhood is generated from a solution S. The algorithm generate each neighbor by permuting two random operations in the vector submitted next to a repair procedure if required. A GRASP (Greedy Randomized Adaptive Search Procedure) iteration build an initial solution using a greedy randomized heuristic and improve the solution with local search. The hybridization of GRASP with ELS improve the solution with the ELS neighborhood instead of using directly a local search.

III. Numerical experiments

A. Job-Shop instances

The method is compared with the state-of-the-art algorithms proposed by (Gonçalves *et al.*; Huang; Kamrul *et al.*; Rego and Duarte; Gao *et al.*)[4][6][7][8][9]. To favor fair performance analysis the running times have been scaled by the speed factor presented in Table 2. This factor takes into account the MFlops (Million Floating-point operation per second) performance of each processor and are computed using [3]. Table 2 gives the information available about these algorithms including the operating system, the language and the computer. Our proposal's algorithm runs on one Intel Core i7-3770 with 3.40 Ghz computer under Windows 7 with 16 Go of memory, this computer has about 2529 MFlops and thus the scale factor of Gonçalves *et al.*'s computer [6] is evaluated at $0.55 \approx (1402/2529)$. Table 2 gives the speedup factor to apply to computational time of a method to obtain a scale computational time, which could be fairly compared with the method we introduced. Please note that information is not available for Kamrul [8].

The experiments are achieved considering the GRASP×ELS introduced in section II.D and applies to solve 40 Laurence's instances and it is compared with the following state-of-the-art methods:

- Hybrid Genetic Algorithm (HGA) using Parameterized Active schedules (HGA-Param. Active), HGA using Non-delay schedules (HGA-Non-delay) and HGA using Active schedules (HGA-Active) introduced by Gonçalves [6].
- 4 methods proposed by Kamrul [8]: Genetic Algorithm (GA) and three versions of Memetic Algorithm (MA) noted as follows: Partial re-ordering rule with GA: MA(PR); Gap reduction rule with GA: MA(GR); and Gap reduction and restricted swapping rule with GA: MA(GR-RS);
- A filter-and-fan approach (F&F) proposed by Rego and Duarte[9];
- MA method proposed by Gao *et al.* [4].

Table 2 Relative performances of computers (N/A: information Not Available).

	(Gonçalves et al, 2005)	Huang(2005)	(Kamrul *et al.*, 2009)	(Rego and Duarte, 2009)	(Gao *et al.*, 2011)	Our proposal
Computer	1.333 GHz AMD Thunderbird CPU	PIV 2.8GHz	N/A	PIV 1.7 Ghz	PIV 1.8Ghz	Intel® Core™ i7-3770CPU @ 3.40GHz
OS	Windows Me	N/A	N/A	Windows 2000 Pro	N/A	Windows 7
Language	Visual Basic 6.0	C	N/A	Visual C++ 6.0	C++	C++
MFlops	1402	2440	N/A	1393	2178	2529
Speed factor	0.55	0.96		0.55	0.86	1

Table 3 Parameter settings.

	Parameter definition	GRASP×ELS parameters Sequential
nd	Number of neighborhood	20
Nc_max	Maximal number of unproductive iteration during mutation	40
Ne	Number of ELS iterations	50
Np	Number of GRASP iterations	100
	Total computational effort :	100000

For the incumbent problem, a design of experiments has been tuned with different combinations of parameters to identify one acceptable set of parameters, which are introduced in Table 3.

The GRASP×ELS is benchmarked over the 40 instances of Laurence and it is compared with 11 states of the art methods:

- Gonçalves *et al.* have proposed 3 methods;
- Huang have proposed 2 methods;
- Kamrul *et al.* have proposed 4 methods;
- Rego and Duarte have proposed 1 method;
- Gao *et al*. have proposed 1 method

Table 4 Relative performances of methods.

				Small *0<x<76*	*Medium* *76<x<101*	*Large* *101<x<200*	*Strongly Large* *x>200*
(Gonçalves et al., 2002)	*1*	*HGA_Param Active*	Avg. Dev.	0.00	0.06	1.10	0.62
	2	*HGA_Non_delay*	Avg. Dev.	0.27	0.85	2.48	1.41
	3	*HDA_Active*	Avg. Dev.	0.24	0.25	2.72	1.41
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	36.74	115.06	522.54	1538.52
(Huang, 2005)	*4*	*SB*	Avg. Dev.	1.55	2.10	7.51	3.55
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	N/A	N/A	N/A	N/A
	5	*MSB*	Avg. Dev.	0.96	1.48	3.68	2.44
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	<1s	<1s	<1s	<1s
(Kamrul et al.2009)	*6*	*GA*	Avg. Dev.	0.62	2.09	6.28	5.72
	7	*MA (PR)*	Avg. Dev.	0.62	2.05	6.20	5.54
	8	*MA (GR)*	Avg. Dev.	0.00	0.31	3.40	2.27
	9	*MA (GR.RS)*	Avg. Dev.	0.00	0.06	2.70	1.61
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	N/A	N/A	N/A	N/A
(Rego and Duarte,2009)	*10*	*F&F*	Avg. Dev.	0.00	0.12	0.55	0.58
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	0.60	0.71	6.20	5.04
(Gao et al.,2011)	*11*	*MA*	Avg. Dev.	0.00	0.00	0.79	0.67
			Scale time T*	N/A	N/A	N/A	N/A
			Scale time TT	N/A	N/A	N/A	N/A
Our proposal	*12*	*GRASPxELS*	Avg. Dev.	0.10	0.33	6.48	5.64
			Scale time T*	<1s	2.304	10.932	17.85
			Scale time TT	<1s	2.538	10.932	17.94

For the small scale instances the GRASP×ELS provides an average deviation of 0.10% outperforming 6 methods (HGA_Non_Delay and HAD_Active of Gonçalves, SB and MSB of Huang and GA and MA (PR) of Kamrul). For the medium scale instances, the GRASP×ELS provides an average deviation of 0.33 % and outperforms 5 methods (HGA_Non_Delay of Gonçalves, SB and MSB of Huang and GA and MA (PR) of Kamrul). For the large scale instances the GRASPxELS provides an average deviation of 6.48% and outperforms 1 method (SB of Huang). For the strongly large scale instances the GRASP×ELS provides an average deviation of 5.64% and outperforms 1 method (GA of Kamrul).

B. Experiments on JSSP with Precedence Constraints

The GRASP×ELS is then compared with 25 instances created for the job-shop with precedence constraints between jobs. The instances 1 to 20 encompass a lot of constraints compared to the number of jobs (more than one constraints per job) and the instances 21-25 encompass less precedence constraints. Table 4 gives a comparative study between our proposal and exact solution of the instances (Column $h^*(x)$ for the optimal solution and $t^*(x)$ the total time) obtained using CPLEX and a linear formulation.

For each instance x, the first replication permits to obtain $h(x,1)$ i.e. the best solution found, $t^*(x,1)$ the time at which the solution $h(x,1)$ has been found, $t(x,1)$ the total computational time, the gap $g(x,1)$ from $h(x,1)$ to the lower bound $lb(x)$.

To provide a fair comparative study and because we promote a meta-heuristic, each instance x is solved using r replications where:

N_{job} :	the number of jobs.
$N_{Precedences}$	the number of precedence constraints between jobs.
$N_{variables}$:	the number of variables in the linear program.
$N_{constraints}$:	the number of constraints in the linear program.
$h^*(x)$:	the optimal solution.
$t^*(x)$:	the computational time required to find the optimal solution.
$h^*(x,r)$:	the best solution found during the r replications.
$t^*(x,r)$:	the best computational time required to obtain the best solution.
$\overline{t(x,r)}$:	the average computational time to execute the method.
N_{optim} :	the number of optimal solution found over 20 replications.
$g^*(x,r)$:	the best gap from $h^*(x,r)$ to the lower bound $lb(x)$.

Let us note $\overline{g^*(.,r)}$ the estimation of the average gap considering the 25 instances, $\overline{t^*(.,r)}$ and $\overline{t(.,\mathrm{r})}$ the estimation of the average computational time.

The average gap $\overline{g^*(.,r)}$ between the best solution found over the 20 replications and the optimal solution is 0.0% for all instances and the average computation time to find the best solutions is about 0.02 seconds.

Table 5 Experiment on Job-Shop with precedence constraints.

	Exact resolution						***(GRASP×ELS)***				
	N_{job}	$N_{Precedences}$	$N_{variables}$	$N_{constraints}$	$h^*(x)$	$t^*(x)$	$h^*(x,\mathrm{r})$	$t^*(x,\mathrm{r})$	$\overline{t(x,\mathrm{r})}$	N_{optim}	$g^*(x,r)$
LMT_01	8	9	96	144	545	0.02	545	0.00	0.00	20/20	0.00
LMT_02	7	7	83	122	426	0.10	426	0.00	0.00	20/20	0.00
LMT_03	8	6	107	158	408	0.01	408	0.00	0.00	20/20	0.00
LMT_04	7	6	102	149	634	0.01	634	0.00	0.00	20/20	0.00
LMT_05	10	8	125	184	645	0.01	645	0.00	0.00	20/20	0.00
LMT_06	10	8	130	193	404	0.14	404	0.00	0.00	20/20	0.00
LMT_07	12	12	333	493	784	0.25	784	0.00	0.01	20/20	0.00
LMT_08	10	10	158	234	731	0.01	731	0.00	0.00	20/20	0.00
LMT_09	8	6	204	293	721	0.02	721	0.00	0.00	20/20	0.00
LMT_10	12	11	277	407	737	0.02	737	0.00	0.00	20/20	0.00
LMT_11	12	12	492	721	1145	0.04	1145	0.00	0.01	20/20	0.00
LMT_12	10	10	386	561	1162	0.07	1162	0.00	0.01	20/20	0.00
LMT_13	12	12	360	523	916	0.04	916	0.00	0.01	20/20	0.00
LMT_14	16	17	576	848	1069	0.43	1069	0.02	0.31	20/20	0.00
LMT_15	15	24	548	811	1193	0.03	1193	0.00	0.06	20/20	0.00
LMT_16	14	16	818	1191	1751	0.14	1751	0.00	0.03	20/20	0.00
LMT_17	20	24	1075	1587	1681	0.28	1681	0.00	0.02	20/20	0.00
LMT_18	16	18	1091	1591	1712	0.38	1712	0.14	2.10	4/20	0.00
LMT_19	18	20	1555	2280	3302	0.10	3302	0.00	0.08	20/20	0.00
LMT_20	20	22	1373	2011	2187	0.58	2187	0.16	1.16	16/20	0.00
LMT_LC_01	8	3	114	166	450	0.02	450	0.00	0.00	20/20	0.00
LMT_LC_02	10	6	199	290	505	0.03	505	0.00	0.00	20/20	0.00
LMT_LC_03	8	4	119	170	382	0.04	382	0.00	0.00	20/20	0.00
LMT_LC_04	12	7	418	603	762	0.33	764	0.16	1.51	5/20	0.00
LMT_LC_05	10	6	434	628	988	0.40	988	0.00	0.05	20/20	0.00
$\overline{g^*(.,r)}$											0.00
$\overline{t(.,\mathrm{r})}$									0.19		
$\overline{t^*(.,r)}$								0.02			

IV. Concluding remarks

In this paper, we have introduced a new problem, the job-shop scheduling problem with precedence constraints between jobs. A GRASP×ELS metaheuristic with a new neighborhood system is proposed and benchmarked on the 40 Laurence's job-shop instances. The algorithm competes with state of the art methods in terms of both results and computational time which is lower than 18 seconds in average for the biggest instances. Next the algorithm is applied on two sets of instances of the job shop with precedence constraints between jobs, one set that encompasses a lot of precedence constraints and one with only few precedence constraints. Our proposal found the optimal solutions for both set of instances in a short computation time.

References

[1] Bierwirth, C. A. (1995). Generalized permutation approach to jobshop scheduling with genetic algorithms. OR Spektrum, 17, 87–92.

[2] Cheng, R., Gen, M., & Tsujimura, Y. (1996). A tutorial survey of job-shop scheduling problems using genetic algorithms – I representation. Computers and industrial engineering, 30, 983–997.

[3] Dongarra, J. (2011). Performance of various computers using standard linear equations software. Report CS-89-85, University of Tennessee, 2011.

[4] Gao, L., Zhang, G., Zhang, L., & Li, X. (2011). An efficient memetic algorithm for solving the job shop scheduling problem. Computers & Industrial Engineering, 60, 699–705.

[5] Garey, M. R., Johnson, D. S., & Seth, R. (1976). The complexity of flowshop and jobshop scheduling. Mathematics of Operations Research, 1, 117-129.

[6] Goncalves, J. F., de Magalahaes Mendes, J. J., & Resende, M. G. C. (2005). A hybrid genetic algorithm for the job shop scheduling problem. European Journal of Operational Research, 1 (67), 77–95

[7] Huang, Z. (2005). A Modified Shifting Bottleneck Procedure for job-shop Scheduling. College of Computer Science and Technology

[8] Kamrul Hasan S. M., Ruhul Sarker, Daryl Essam (2009). Memetic algorithms for solving job-shop scheduling problems. Memetic Computing (pp 69–83), vol. 1.

[9] Rego, C., & Duarte, R. (2009). A filter-and-fan approach to the job shop scheduling problem. European Journal of Operational Research, 194, 650–662.

[10] Roy, B., & Sussmann, B. (1964). Les problèmes d'ordonnancement avec contraintes disjonctive, In: Note DS No. 9 bis, SEMA, Paris, 1964.

[11] Wolf, S., & Merz, P. (2007). Evolutionary local search for the super-peer selection problem and the p-hub median problem. In M. Blesa, C. Blum, B. Naujoks, A. Roli, G. Rudolph, & M. Sampels (Eds.), Lecture Notes in Computer Science. Springer. Vol. 4771, pp. 1–15.

Resolution of a job-shop with routing considering a quality of service for customer: JSSP with Routing

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Matthieu Gondran, Philippe Lacomme, Nikolay Tchernev,
Université Clermont Auvergne, (LIMOS, UMR CNRS 6158),
63177 Aubière, France
gondran@isima.fr, placomme@isima.fr, tchernev@isima.fr

Abstract— The Job-Shop scheduling Problem (JSP) is a well-known NP-hard scheduling problem and the Job-Shop Scheduling Problem with Transport (JSPT) is an extension including a vehicle scheduling problem to transport job between machines. This problem is commonly solved considering the makespan minimization only, and the routing problem is not part of the objective function but defined only as a constraint set which must hold. In this paper, the objective consists in both makespan minimization and quality of service maximization defining the JSP with Routing (JSPR). The quality of service considers the Total Duration (or Flow Time), the Total Riding Time, and the Total Waiting Time of jobs into buffers. A GRASP×ELS is defined to solve both the classical JSPT and the JSPR. The method competes with the best published methods for the JSPT and provides high quality solutions for the JSPR.

Keywords— JSPT, Scheduling, Routing, Quality of Service

I. Introduction to the Job-Shop with transport

A. Scheduling with transport

The Job-Shop scheduling Problem (JSP) has received a considerably amount of attention in the recent years. The Job-Shop Problem with Transport (JSPT) is an extension where a heterogeneous fleet of vehicles is assumed to transport the jobs from one machine to the next one.

The JSPT is first introduced by Bilge and Ulusoy [1] and formalized by Knust [11]. Brucker and Strotmann [2] tackle a job-shop problem with several homogeneous transport robots. Hurink and Knust [10] introduce a highly efficient disjunctive graph for the job-shop with one single robot and derived some specific properties to a proper definition of neighborhoods in local search procedure. Zhang et al. [19] investigate the JSPT with constraints on the transportation time, the capacity of the buffers, and add time-lags between successive operations belonging to the same job. Umar et al [17] propose a Hybrid Multi-Objective Genetic Algorithm for the JSPT considering the makespan, the delays and the transport time of the Automated Guided Vehicle (AGV). More recently, Nouri et al. [14] draw up an inventory of the state-of-art of the JSPT by classifying all the papers published between 1995 and 2014. But no one article tackles the quality of service or use approach from the routing problem community. In this paper, the vehicles are assumed to be either a robot or AGV.

The remainder of the paper is organized as follows. The next subsection will be the definition of the problem. The Section 2 defines the framework promoted, and at last Section 3 deals with computational results.

B. Definition

A JSP consists of scheduling a set of n jobs $J = \{J_1 \dots J_n\}$ which require to be processed on a set of m machines $M = \{M_1 \dots M_m\}$. Each job is composed by an ordered set of n_i machine-operations $[O_{i,1}, O_{i,2} \dots O_{i,n_i}]$ and a machine-operation $O_{i,j}$ refers to one machine μ_{ij} for a duration $pt_{i,j}$. Only one machine-operation can be processed on each machine at the same time. In the JSPT, jobs have transfer requests from one machine to another one and these requests are handled by a fleet of r vehicles $R = \{R_1, \dots, R_r\}$. The vehicle r has a capacity C_r. A vehicle r performing a loaded transport between machines M_i and M_j has a trip duration $t^r_{i,j}$, a vehicle r performing an unloaded transport between machines M_i and M_j has a trip duration $v^r_{i,j}$. A classical objective of the JSPT is to minimize the total duration, i.e. the makespan, of the schedule.

In this paper, the proposed evaluation takes into account the integration of the routing criteria in the objective function referring to the Quality of Service (QoS). The QoS is defined by the Total Duration, the Total Riding and the Total Waiting Time.

II. PROBLEM OF INTEREST: JOB-SHOP WITH ROUTING

A. Disjunctive graph for the Job-Shop with transport

The modeling of the JSPT is a generalization of the disjunctive graph introduced by Roy and Sussmann [16] for the JSP. Each job is an ordered sequence of machine-operations, where each machine-operation is modeled by a node, and precedence constraints are modeled by conjunctive arcs between machine-operations. The graph includes also two dummy nodes referred to as 0 and *. Each machine-operation is modeled by a vertex of the graph. An edge leaving an operation $O_{i,j}$ of duration $pt_{i,j}$ is weighted by the value of $pt_{i,j}$. Every couple of machine-operations $O_{i,j}$ and $O_{y,x}$ which must be processed on the same machine $\mu_{i,j} = \mu_{y,x}$, is gathered by a disjunctive edge which must be transformed into an arc to define a fully oriented disjunctive graph.

In the JSPT, before each machine-operation $O_{i,j}$, a node is added to the graph to represent the delivery operation $D_{i,j}$ (excepted for the first machine-operation of each job). After each machine-operation $O_{i,j}$, a node is added to model the pickup operation $P_{i,j}$. An arc from a machine-operation $O_{i,j}$ to a pickup operation $P_{i,j}$ is weighted by $pt_{i,j}$ plus the loaded time which is assumed to be 0. Arcs from a delivery operation $D_{i,j}$ to machine-operation $O_{i,j}$ is weighted by 0 since the unloaded time is assumed to be 0. Each job i requests to be transferred from the machine $\mu_{i,j}$ processing the machine-operation $O_{i,j}$ to the machine $\mu_{i,j+1}$ processing the machine-operation $O_{i,j+1}$. A transfer request is fully defined by an ordered sequence of transport operations achieving a path from the pickup operation $P_{i,j}$ to the delivery operation $D_{\mathrm{i,j+1}}$. A transport operation $T_{(\mathrm{O}_{i,j},\mathrm{O}_{y,x})}$ is modeled on the graph by an arc between two pickup/delivery operations and falls into four configurations (P,D), (P,P), (D,D), (D,P) where P is a pickup operation and D a delivery operation. Moreover the transport operation $T_{(\mathrm{O}_{i,j},\mathrm{O}_{y,x})}$ is defined by a departure time $B_{i,j}$ on the machine μ_{ij}, an arrival time $A_{y,x}$ of the vehicle r on the machine μ_{yx}, and a vehicle r assigned to the transport. Let us define $S_{i,j}$ the starting time of the machine-operation $O_{i,j}$, and $C_{i,j}$ its finishing time. As the loaded and unloaded times are null, the departure time $B_{i,j}$ is assumed to be equal to the starting time of the pickup operation $P_{i,j}$ and the arrival time $A_{y,x}$ is equal to the starting time of the delivery operation $d_{i,j}$.

In a JSPT with several vehicles of non-unitary capacity, the resolution of the problem requires:

- the assignment of vehicles to both pickup and delivery operations;
- the creation of disjunctions between all machine operations requiring the same machine;
- the creation of disjunction between all pickup and delivery operations assigned to the same vehicle.

A solution of JSPT is given by assigning one vehicle to each transport operation and by oriented all undirected arcs into directed ones (between machine-operations, and between pickup/delivery operations), giving the trip of each vehicle and the ordered sequence of the jobs on each machine.

Figure 1 stresses an example, for a JSPT, of a non-oriented disjunctive graph where the machine disjunctions are modeled by edges but that no encompasses pickup/delivery operation disjunctions since theses disjunctions are vehicle assignment dependent.

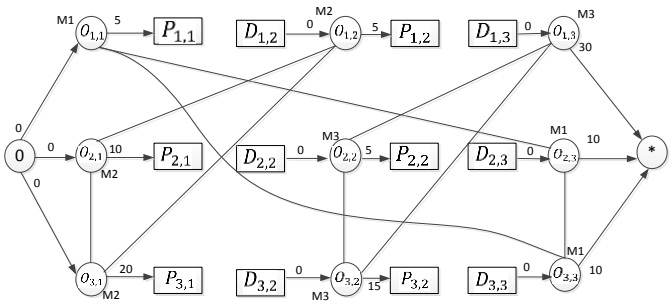


Fig. 1. A non-oriented disjunctive graph

B. Quality of Service in the JSPR

For a job i the Total Duration TD_i is defined by the difference between the finishing time of its last machine-operation and the starting time of its first machine-operation (that is the flow time of a job). The Riding Time corresponds to the elapsed time between the departure time of the vehicle r from the machine μ_{ij}, where the job i is loaded, and the beginning of the machine-operation $O_{i,j+1}$ of job i at machine μ_{ij+1}. The waiting time of a job i in an output buffer is defined by the difference between the finishing time of his machine-operation $O_{i,j}$ and the beginning of the transport operation, the waiting time in the output buffer is the difference between its delivery in the output buffer and the beginning of the machine-operation.

Here, the objective is to determine a feasible schedule which minimizes the following function:

$$F = \alpha \times C_{max} + \beta \times (TD + TRT + TWT).$$

- $C_{max} = \max_{j=1,n}\{C_j\}$ where C_j is referred to as the completion time of the last operation O_{j,n_j} of the job J_j.
- $TD = \sum_{i=1}^{n}(C_{i,n_i} - S_{i,1})$ where $S_{i,1}$ is starting time of the first operation $O_{i,1}$ and C_{i,n_i} denotes the completion time of the last operation O_{i,n_i} of the job J_i.
- $TRT = \sum_{j\in L_i}(S_{i,j+1} - B_{i,j})$ where $S_{i,j+1}$ is the starting date of the machine operation $O_{i,j+1}$ of the job J_j, and $B_{i,j}$ is the starting time pickup operation of $P_{i,j}$, and L_i the set of pickup operations.
- $TWT = \sum_{i=1}^{i=n}\sum_{j\in L_i}(C_{i,j} - B_{i,j}) + \sum_{i=1}^{i=n}\sum_{j\in u_i}(st_{i,j} - A_{i,j})$ Where U_i is the set of delivery operations.

C. Routing evaluation for the JSPT with Quality of Service:

A positive arc from one machine operation to the following pickup operation gives the minimal delay between the earliest starting time of the machine operation and the earliest starting time y of the pickup operation. Hence, every pickup operation has a $[y;\infty[$ time-window constraint that defines the earliest starting time of the pickup operation, which is correlated to the machine operation finishing time. Similarly, a delivery operation must be achieved before the maximal starting time x of the machine operation where the delivery is processed. So every delivery operation is bounded by x and has a $]-\infty;x]$ time-window constraint. This type of time-windows arises in routing problems including but not limited to the Dial-a-Ride problem (DARP) where both maximal route duration and maximal customer riding time must be addressed.

The DARP is a problem in which a set of transportation requests has to be handled by a fleet of vehicles. Each request corresponds to a customer that requires transport from a pickup node to a delivery node with respect of both time windows on pickup and delivery nodes. Moreover, each trip starts and ends at the depot, and the number of customers transported by a vehicle at the same time must fit the vehicle capacity. If a vehicle arrives earlier on a pickup or delivery point, it must wait, if a vehicle arrives late, the solution is not feasible. A solution of the DARP is a set of trips where all requests are achieved respecting every time-windows.

Cordeau and Laporte [6] define the Quality of Service (QoS) for the Dial-And-Ride Problem (DARP) taking into account the Total Duration, the Total Riding Time and the Total Waiting time. They propose a greedy approach to find a solution which is not left-shifted but a compromise between makespan minimization and QOS maximization. This approach includes 4 steps, minimizing sequentially the four criteria (Makespan, TD, TRT and TWT) while the deterioration of the objective function is prevented using maximal time-lags which are iteratively added into the graph.

In the same trend that Cordeau and Laporte [6], the proposed approach TLH (Time-Lag based Heuristic) for the JSPR takes into account the QoS and contains four steps to iteratively minimize the Makespan, TD, TRT and TWT. At each step, to avoid the deterioration of the previously minimized criteria, some maximal time-lags are added on the disjunctive graph. A maximal time-lag constraint can be formalized as following: $S_j \leq S_i + TL_{j,i}$ with $S_i \leq S_j$. S_i and S_j are the starting time of, respectively, the operations O_i and O_j. It means that the operation O_j cannot start after the starting date of the operation O_i plus the value $TL_{j,i}$ of the maximal time-lag. A maximal time-lag is represented on a graph by an arc from O_j to O_i with a negative value. Adding maximal time-lag constraints in a JSP can lead to unprofitable cyclic graph (Caumond et al. [3]) depending on the two operations which are concerned by the maximal time-lag and/or on the maximal time-lag value. Building non-trivial acyclic graph containing maximal time-lag with a lower value requires cycle detection.

The proposed routing evaluation is composed of 4 steps:

Step 1. This step leads to the minimization of the makespan by computing the earliest starting time of each operation thanks to a tuned Bellman-Ford's algorithm (see Lacomme et al. [12] for more details). The makespan is constrained by adding a maximal-time lag between the dummy nodes * to 0 which values $-C_{max}$.

Step 2. The Total Duration is minimized by adding maximal time-lags between the last operation of each job and the first job operation. These time-lags refer as to D on the Figure 2. The time-lags are randomly and iteratively added.

Step 3. The Total Riding Time is minimized by adding maximal time-lags between a machine-operation $O_{i,j}$ and the pickup operation $P_{i,j-1}$. On the Figure 2, these time-lags refer as to the A. The time-lags are iteratively added following the reverse trips of the vehicles.

Step 4. The Total Waiting Time is optimized by adding maximal time-lags between machine-operations $O_{i,j}$ and delivery operations $D_{i,j}$ and between pickup operations $P_{i,j}$ and machine-operations $O_{i,j}$. These time-lags refer as to the B and C on the Figure 2. The time-lags are iteratively added following the reverse trips of the vehicles.

For the four steps, the maximal time-lags are iteratively added and after each insertion the graph is evaluated to avoid cyclic graph. A time-lag between $O_{i,j}$ and $O_{x,y}$, must be the lowest and its value is initialized to a minimal value, which is the shortest path between the $O_{i,j}$ and $O_{x,y}$. The graph is evaluated and if a cycle is detected, the value of the time-lag is increased of the value of the cycle length until no cycle is detected, in this case, the following time-lag is inserted into the graph and so on. For example, if after adding a time-lag TL, the graph contains a cycle and the sum of the values of the arcs constituting the cycle is 10, to the value of the time-lag must be increased of 10.

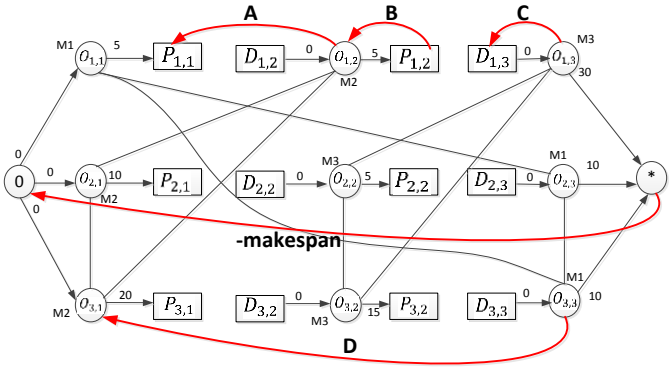


Fig. 2. The different maximal time-lags

D. Framework description

Cheng et al. [5] highlight the importance of an ad hoc quasi direct representation of solutions with two separated space: the coding space and the solution space linked by a mapping function. Efficient mapping functions fall into three categories of mapping function: a 1-to-1 mapping function where one and only one object of the coding space is linked to one and only one solution; a n-to-1 mapping function where several objects in the coding space can be linked to the same solution; and a 1-to-n mapping function where one object in the coding space can be linked to several solutions.

According to Chassaing et al. [4], the key-points for an efficient resolution scheme are the following:

- a quasi-direct representation of solution based on the Bierwirth vector;
- a mapping function to evaluate a solution which considers the makespan and the QoS (see I.C)
- a local search alternatively working on coding space and solutions space;
- a metaheuristic favoring search space exploration.

As already pointed out, solution of the JSPT is obtained by evaluating a fully acyclic oriented disjunctive graph which defines the disjunctions between machine-operations and between the pickup/delivery operations assigned to the same vehicles. To obtain a solution of the JSPR, a fully oriented acyclic graph must be evaluated, and the graph for the JSPR can be represented in the coding space similarly to the JSPT methods, and the same local searches can be used for both problems.

The Bierwirth's vector is a commonly used representation for the JSP, referred to as vector by repetition based on a topological order, providing only acyclic graph and defining a n-to-1 mapping function. This vector has been generalized by Lacomme et al. [13] for the JSPT and is suitable for the JSPR. In the extended Bierwirth's vector (MTS) each job appears once for each machine-operation, pickup and delivery operations. The first occurrence of j in the vector refers to the first machine-

operation of the job j, the second occurrence refers to the pickup operation, the third occurrence is the delivery operation, the fourth one refers to the second machine-operation and so on. Lacomme et al. [13] introduce a second vector OA concerning the assignment of the vehicles to the pickup operations. The first occurrence of a job j in MTS referring to a pickup operation is assigned to the vehicle in the first position in OA, the second operation MTS referring to a pickup operation is assigned to the vehicle in the second position in OA, and so on. A delivery operation $D_{i,j}$ is assumed to be assigned to the same vehicle than the pickup operation $P_{i,j}$.

This representation of the coding space is suitable with the mapping function presented section II.C: the TLH evaluation to find new solutions which are not left-shifted but which takes into account the QOS.

One of the other advantages of vector by repetition concerns the solution neighborhoods used as local searches and based on the modification of the longest path in the graph which permits to define well-tuned operator to the vectors MTS and OA. One of the most powerful local search for the JSP is the Dell'amico and Trubian [7]'s neighborhood which exchanges machine operation at the end or at the start of a Grabowski's block belonging to the critical path of the Graph. A Grabowski et al. [9]'s block is defined as a maximal successive machine-operations. Lacomme et al. [13] has been extended the neighborhood system of Dell'amico and Trubian [7] for the JSPT, taking into account machine-operations, pickup and delivery operations. The proposal consists in switching two machine-operations belonging to the same machine block, or either in switching two successive pickup/delivery operations or in updating the vehicle assignment.

The metaheuristic is based on a GRASP×ELS which has been successfully applied for decades to various optimization problems including vehicle routing problems or scheduling problems. The GRASP×ELS proposed by Prins [15] combines the positive features of the GRASP (Greedy Randomized Adaptive Search Procedure) initially introduced by (Feo and Resende [8]) and of the ELS (Evolutionary Local Search) which was investigated by Wolf and Merz [14].

III. Numerical experiments

A. Benchmarks

To the best of our knowledge, no instance is openly available and dedicated to the JSPR with several vehicles of non-unitary capacity. This is why we reuse here the Bilge and Ulusoy [1] instances which have been first introduced for the JSPT with two vehicles of unitary capacity. These instances are grouped into two sets: dataset 1 with 40 instances with $t_{ik}/p_k > 0.25$ and dataset 2 with 42 instances with $t_{ik}/p_k \leq 0.25$ where t_{ik} is the average travelling time and p_k is the average processing time.

The GRASP×ELS has been implemented in C++ and the experiments are performed on a 3.60 GHz Intel Core i7-4790 computer with 16,0 GB RAM under windows 7, which leads to 2671 MFlops. The parameters are the following: number of neighborhoods in the ELS is 30, the number of ELS iteration is 60 and the number of GRASP iteration is 150. The function to minimize is defined by:

$$h(y) = \alpha \times h_{cmax}(y) + \beta \times h_{cost}(y).$$

with

$h_{cmax}(y) = C_{max}$ (or makespan)
$h_{QOS}(y) = 1/h_{cost}(y)$
$h_{cost}(y) = TD(y) + TRT(y) + TWT(y)$

Because the objective is to minimize first the makespan, we consider $\alpha = 10000$ and $\beta = 1$. Thanks to the objective function $h(y)$ and GRASP×ELS, the three approaches are encompassed, the first one solves the JSPT with the objective function $h(y) = h_{cmax}(y)$ ($\alpha = 1$, $\beta = 0$); the two others approaches concern the JSPR: the first one is sequential, i.e. $h(y) = h_{cmax}(y)$ and the TLH evaluation is applied on the best solution found; whereas the second approach is integrated, $h(y) = \alpha \times h_{cmax}(y) + \beta \times h_{cost}(y)$ with $\alpha = 10\,000$ and $\beta = 1$.

To provide a fair comparative study and because we promote a metaheuristic, each instance is solved using 5 replications. Due to constraints in the publication length, only the average values are given. All the detail results are available on: http://fc.isima.fr/~gondran/JSPT/tlh.html .

Let us note

- $\overline{h^*_{cmax}(x,n)}$: the best C_{max} found for the n replications of the instance x
- $\overline{h^*_{cmax}(.,n)}$: the average $\overline{h^*_{cmax}(x,n)}$ for all instances
- $\overline{h^*_{cost}(x,n)}$: the best h_{cost} found for the n replications for the instance x
- $\overline{h^*_{cost}(.,n)}$: the average $\overline{h^*_{cost}(x,n)}$ for all instances
- $gap^*(.,n)$: is the average gap between $\overline{h^*_{cmax}(x,n)}$ and the best known lowerbound
- $\overline{st^*(.,n)}$: is the average scalar time to find the best solution
- $\overline{stt(.,n)}$: is the average scalar total time to the best replication
- $\overline{gap^*(.,.)}$ is the average value of $\overline{gap^*(.,n)}$ for all the replications
- $st^*(.,.)$ is the average value of $\overline{st^*(.,n)}$ for all the replications
- $\overline{stt(.,.)}$ is the average value of $\overline{stt(.,n)}$ for all the replications

B. Several unitary capacity vehicles

According to Nouri et al. [14], 16 approaches have been published for the JSPT using the Bilge and Ulusoy [1] instances. Table 1 provides a comparative study of 16 methods from the state-of-art and GRASP×ELS solving the JSPT ($\alpha = 1$, $\beta = 0$) for the first set of benchmark instances with $t_{ik}/p_k > 0.25$. 11 methods from the state-of-art is compared to the proposed algorithm for the second set of instances with $t_{ik}/p_k \leq 0.25$ in the Table 2.

For the 40 instances with $t_{ik}/p_k > 0.25$, the GRASP×ELS provides an average gap of 31%. This gap is comparable to many other methods (STW, UGA, MAS, MSB, FMAS, HRA), other publications are a better (AGA, PGA, SALS, MA, ASL, TS_SPMA) of 2%, and with others methods, the GRASP×ELS has a gap more important: DE, GATS+HM but for some instances of these methods, the best solutions is lower than the lower bound given by Bilge and Ulusoy [1]. Let us note the running time of the approach is very short. A usual question is to wonder whether the solution gap between the GRASP×ELS and the lower bound could be reduced by increasing the number of

iterations. Several experiments have been achieved to assign same running time than the best published methods, but no significant improvement was obtained. The modeling of the problem used by the approach is dedicated to problems with several non-unitary capacity vehicles and encompasses problems with unitary capacity several vehicles. A better modeling for problem with unitary capacity vehicles is proposed by Lacomme [12] where each pickup and delivery nodes between two successive machine-operations are gathered to constitute only one node, reducing the search space of the solution and improving the efficiency of the metaheuristic. Nevertheless, this representation cannot include problems with several non-unitary capacity vehicles.

Table 1 Comparison of the literature methods on the 40 instances with $t_{ik}/p_k > 0.25$ for two unitary capacity vehicle.

Method	Ref	$\overline{gap^*(.,n)}$	$\overline{st^*(.,n)}$	$\overline{stt(.,n)}$	Nb_replications	$\overline{gap^*(.,.)}$	$\overline{st^*(.,.)}$	$\overline{tt(.,.)}$
STW	Bilge and Ulusoy 1995	0.30	/	/	/	/	/	/
UGA	Ulusoy et al. 1997	0.27	/	/	20	0.32	/	/
AGA	Abdelmaguid et al. 2004	0.26	/	/	20	/	/	/
PGA	Reddy and Rao 2006	0.26	/	/	20	/	/	/
SALS	Deroussi et al. 2008	0.26	2.22	/	/	/	/	/
MA	Larabi 2010	0.26	/	/	5	/	/	/
DE	Babu et al. 2010	0.14	/	/	/	/	/	/
FDE	Kumar et al. 2011	0.26	/	/	20	/	/	/
MAS	Erol et al. 2012	0.75	/	/	/	/	/	/
MSB	Q. Zhang et al. 2013	0.35	395.76	/	/	/	/	/
TS_SPMA	Y. Zheng et al. 2014	0.26	/	/	/	/	/	/
ALS	Baruwa et al. 2015	0.26	503.10	3240.00	/	/	/	/
FMAS	Sahin et al. 2015	0.45	/	/	/	/	/	/
GATS	Zhang et al. 2012	0.28	/	/	20	/	/	/
GATS+HM	Nouri et al. 2016	0.21	12.08	/	10	/	/	/
HRA	K. Zheng et al. 2016	0.70	/	/	/	/	/	/
GRASPxELS		0.31	3.21	11.39	5	0.31	2.29	11.44

Table 2 Comparison of the literature methods on the 42 instances with $t_{ik}/p_k \leq 0.25$ for two unitary capacity vehicle.

Method	Ref	$\overline{gap^*(.,n)}$	$\overline{st^*(.,n)}$	$\overline{stt(.,n)}$	Nb_replications	$\overline{gap^*(.,.)}$	$\overline{st^*(.,.)}$	$\overline{tt(.,.)}$
STW	Bilge and Ulusoy 1995	0.07	/	/	/	/	/	/
UGA	Ulusoy et al. 1997	0.05	/	/	20	0.06	/	/
AGA	Abdelmaguid et al. 2004	0.06	/	/	20	/	/	/
PGA	Reddy and Rao 2006	0.06	/	/	20	/	/	/
MEMA	Lacomme et al. 2013	0.06	/	/	5	/	/	/
DE	Babu et al. 2010	-0.05	/	/	/	/	/	/
FDE	Kumar et al. 2011	0.06	/	/	20	/	/	/
MAS	Erol et al. 2012	0.21	/	/	/	/	/	/
TS_SPMA	Y. Zheng et al. 2014	0.06	/	/	/	/	/	/
FMAS	Sahin et al. 2015	0.07	/	/	/	/	/	/
HRA	K. Zheng et al. 2016	0.11	/	/	/	/	/	/
GRASPxELS		0.07	0.24	8.40	5	0.07	1.68	8.41

Table 3 Comparison of the three approaches for two unitary capacity vehicles.

	JSPT		JSPR sequential				JSPR integrated			
	$\overline{h^*_{cmax}(.,n)}$	$\overline{h^*_{cost}(.,n)}$	$\overline{h^*_{cmax}(.,n)}$	$\overline{Gap_{cmax}}$	$\overline{h^*_{cost}(.,n)}$	$\overline{Gap_{cost}}$	$\overline{h^*_{cmax}(.,n)}$	$\overline{Gap_{cmax}}$	$\overline{h^*_{cost}(.,n)}$	$\overline{Gap_{cost}}$
$t_{ik}/p_k > 0.25$	112.6	958.9	112.6	0.00%	742.8	29.8%	112.3	-0.23%	622.7	55.5%
$t_{ik}/p_k \leq 0.25$	173.7	1412.6	173.7	0.00%	1081.7	29.3%	173.7	0.00%	659.7	108.3%

Table 4 Comparison of the three approaches for two non-unitary capacity vehicles.

	JSPT		JSPR sequential				JSPR integrated			
	$\overline{h^*_{cmax}(.,n)}$	$\overline{h^*_{cost}(.,n)}$	$\overline{h^*_{cmax}(.,n)}$	$\overline{Gap_{cmax}}$	$\overline{h^*_{cost}(.,n)}$	$\overline{Gap_{cost}}$	$\overline{h^*_{cmax}(.,n)}$	$\overline{Gap_{cmax}}$	$\overline{h^*_{cost}(.,n)}$	$\overline{Gap_{cost}}$
$t_{ik}/p_k > 0.25$	101.4	849.6	101.4	0.00%	721.7	17.04%	101.2	-0.10%	594.3	41.07%
$t_{ik}/p_k \leq 0.25$	173.3	1380.4	173.3	0.00%	1119.5	23.16%	173.3	0.00%	672.8	99.15%

For the 42 instances with $t_{ik}/p_k \leq 0.25$ (Table 2) no fair comparative study is possible for the GRASP×ELS with the literature since the computational times are not provided for previous researches. Nevertheless in few seconds, the proposed approach finds results close to the state-of-art results.

Tables 3 presents the average value of the makespan (columns $\overline{h^*_{cmax}(.,n)}$) and the cost (columns $\overline{h^*_{cost}(.,n)}$) for the best replications of each instance for the three approaches: JSPT, JSPR sequential and JSPR integrated considering both datasets. The two rows correspond to the two datasets. The $\overline{Gap_{cmax}}$ is defined as the gap of the makespan between the JSPR (sequential or integrated) and the JSPT. $\overline{Gap_{cmax}} = \left(JSPR\left(\overline{h^*_{cmax}(.,n)}\right) - JSPT\left(\overline{h^*_{cmax}(.,n)}\right)\right) / JSPT\left(\overline{h^*_{cmax}(.,n)}\right)$. Similarly the $\overline{Gap_{cost}}$ is defined as the cost gap between the JSPR (sequential or integrated) and the JSPT. $\overline{Gap_{cost}} = \left(JSPR\left(\overline{h^*_{cost}(.,n)}\right) - JSPT\left(\overline{h^*_{cost}(.,n)}\right)\right) / JSPT\left(\overline{h^*_{cost}(.,n)}\right)$.

A $\overline{Gap_{cmax}} = 0\%$ means the two approaches are able to find similar solutions considering the makespan, which is the case between the JSPT and the JSPR sequential for both datasets and between the JSPT and the JSPR integrated for the data set 2. Let us note that the JSPR integrated as a $\overline{Gap_{cmax}} = -0.23\%$, meaning that it is able to find better solutions than the JSPT considering the makespan. The $\overline{Gap_{cost}}$ of the JSPR integrated are respectively equal to 29.9% and 29.3% for the first and the second dataset: the JSPR sequential improves the solution compared to the JSPT considering the QOS criteria. The $\overline{Gap_{cost}}$ between the JSPR integrated and the JSPT are respectively equal to 55.5% and 108.3%, showing the efficiency to the JSPR integrated to find high quality solutions. In the second dataset, the process times have been doubled and transportation time have been halved, this can explain why the $\overline{Gap_{cost}}$ of the JSPR integrated has been double for the dataset 2.

C. Several non-unitary capacity vehicles

To the best of our knowledge no result are available for JSPT with several non-unitary capacity vehicles. It is why Bilge and Ulusoy [1] instances have been extended to several non-unitary capacity vehicles. The processing times and travelling times are not changed but we consider two vehicles with a capacity of two jobs each, i.e. a vehicle can transport two jobs during its trip.

Similarly to the Table 3, the Table 4 provides the average values of the makespan and of the cost for the three different approaches (JSPT, JSPR sequential, JSPR integrated), and the average gaps between the JSPR approaches and the JSPT. Let us notice that the average gap $\overline{Gap_{cmax}}$ between the JSPT and the JSPR is null for both datasets for the JSPR sequential and for the JSPR integrated. Nevertheless for the dataset 1, the JSPR integrated is able to find solutions with a better makespan than the JSPT: $\overline{Gap_{cmax}} = -0.10\%$. Concerning the $\overline{h^*_{cost}(.,n)}$, the JSPT integrated finds the best solutions for both datasets, the gaps $\overline{Gap_{cost}}$ with the JSPT are respectively 41.07% and 99.15% whereas the gaps between the JSPR sequential and the JSPT are respectively 17.04% and 23.16% These results highlight the efficiency of the integrated approach to find high quality solution considering the makespan minimization and the QOS maximization.

IV. Concluding remarks

This paper introduces a routing evaluation for a Job-Shop Scheduling Problem with Routing with a fleet of vehicles. The promoted approach takes advantages, of a GRASP×ELS metaheuristic with an indirect representation of the solutions and an evaluation of the solution by a tuned algorithm from both scheduling and routing problems. This approach permits to maximize the quality of service considering the Total Duration, the Total Riding Time and the Total Waiting Time without deteriorating the makespan. The advantages to use a routing algorithm in the JSPT are clearly highlighted through the different tables. These results push us into considering that it is possible to minimize simultaneously the makespan and maximizing the Quality of Service and they highlight the importance of an integrated approach.

References

[1] Bilge, Ü., & Ulusoy, G. (1995). A time window approach to simultaneous scheduling of machines and material handling system in an FMS. Operations Research, 43(6), 1058-1070.

[2] Brucker, P., & Strotmann, C. (2002, April). Local search procedures for job-shop problems with identical transport robots. In Eight International Workshop on Project Management and Scheduling, Valencia (Spain).

[3] Caumond, A. (2006). Le problème de job-shop avec contraintes: modélisation et optimisation (Doctoral dissertation, Université Blaise Pascal-Clermont-Ferrand II).

[4] Chassaing, M., Fontanel, J., Lacomme, P., Ren, L., Tchernev, N., & Villechenon, P. (2014). A GRASP× ELS approach for the job-shop with a web service paradigm packaging. Expert Systems with Applications, 41(2), 544-562.

[5] Cheng, R., Gen, M., & Tsujimura, Y. (1996). A tutorial survey of job-shop scheduling problems using genetic algorithms—I. Representation. Computers & industrial engineering, 30(4), 983-997.

[6] Cordeau, J. F., & Laporte, G. (2003). A tabu search heuristic for the static multi-vehicle dial-a-ride problem. Transportation Research Part B: Methodological, 37(6), 579-594.

[7] Dell'Amico, M., & Trubian, M. (1993). Applying tabu search to the job-shop scheduling problem. Annals of Operations research, 41(3), 231-252.

[8] Feo, T. A., & Resende, M. G. (1995). Greedy randomized adaptive search procedures. Journal of global optimization, 6(2), 109-133.

[9] Grabowski, J., Nowicki, E., & Zdrzałka, S. (1986). A block approach for single-machine scheduling with release dates and due dates. European Journal of Operational Research, 26(2), 278-285.

[10] Hurink J. and Knust S. (2005), Tabu search algorithms for job-shop problems with a single transport robot. European Journal of Operational Research; 162 (1): 99-111

[11] Knust, S. (1999), Shop-Scheduling Problems with Transportation. Ph. D. Thesis, Fachbereich Mathematik /Informatik Universität Osnabrück

[12] Lacomme, P., Larabi, M., & Tchernev, N. (2007). A disjunctive graph for the job-shop with several robots. In MISTA conference (Vol. 20, pp. 285-292).

[13] Lacomme, P., Larabi, M., & Tchernev, N. (2013). Job-shop based framework for simultaneous scheduling of machines and automated guided vehicles. International Journal of Production Economics, 143(1), 24-34.

[14] Nouri, H. E., Driss, O. B., & Ghédira, K. (2016). A Classification Schema for the Job Shop Scheduling Problem with Transportation Resources: State-of-the-Art Review. In Artificial Intelligence Perspectives in Intelligent Systems (pp. 1-11). Springer International Publishing.

[15] Prins, C. (2009). A GRASP× evolutionary local search hybrid for the vehicle routing problem. In Bio-inspired algorithms for the vehicle routing problem (pp. 35-53). Springer Berlin Heidelberg.

[16] Roy, B., & Sussmann, B. (1964). Les problemes d'ordonnancement avec contraintes disjonctives. Note ds, 9.

[17] Umar, U. A., Ariffin, M. K. A., Ismail, N., & Tang, S. H. (2015). Hybrid multiobjective genetic algorithms for integrated dynamic scheduling and routing of jobs and automated-guided vehicle (AGV) in flexible manufacturing systems (FMS) environment. The International Journal of Advanced Manufacturing Technology, 81(9-12), 2123-2141.

[18] Wolf, S., & Merz, P. (2007, October). Evolutionary local search for the super-peer selection problem and the p-hub median problem. In International Workshop on Hybrid Metaheuristics (pp. 1-15). Springer Berlin Heidelberg.

[19] Zhang, Q., Manier, H., & Manier, M. A. (2014). A modified shifting bottleneck heuristic and disjunctive graph for job shop scheduling problems with transportation constraints. International Journal of Production Research, 52(4), 985-10

Multitasking scheduling with energy cost

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Junkai He†, Feifeng Zheng‡, Chengbin Chu§

*School of Economics & Management, Tongji University, China
Email: mingliu@tongji.edu.cn

†Glorious Sun School of Business & Management, Donghua University, China
Email: hejunkai@163.com

‡Glorious Sun School of Business & Management, Donghua University, China
Email: ffzheng@dhu.edu.cn

§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
School of Economics & Management, Tongji University, China
Email: chengbin.chu@centralesupelec.fr

***Abstract*—In this work, we investigate a multitasking scheduling problem with energy cost. Various practical domains exist multitasking operations, such as manufacturing and service industries. Besides, energy-oriented production is an urgent requirement under finite-resource environment. This paper is motivated by multitasking scheduling on a shared machine, which has continuous available time in planning horizon. Our main works include: (i) embedding energy into multitasking scheduling, (ii) formulating a bi-objective integer programming model, (iii) devising an epsilon constraint algorithm and a non-dominated sorting genetic algorithm II (NSGA-II). We minimize the total weighted completion time and total energy cost simultaneously. The epsilon constraint method can produce exact solutions for small-size instances, while NSGA-II can solve for large-size experiments. Numerical experiments have been conducted to validate our proposed model and algorithms.**

I. Introduction

Multitasking scheduling is a growing interest in machine scheduling domain. This concept originates in computer systems in which multiple tasks share a single processing resource such as CPU (e.g., Chan et al., 2008; Rogers et al., 2014; Zhang et al., 2014; Mostafa and Kusakabe, 2016). The same phenomenon can be found in reality. One of the examples is garment design industries, in which the designers can use multiple browsers to serve various personalized customers. Another example is that one house can be rented for different merchants with daytime and night services.

The rapid development of global economy demands more energy, while the shortage of finite resources limits the development of many countries. As the industrial production consumes the most energy, an energy-oriented production attracts huge attention in recent years.

In our concerned problem, we study the shared resource which can provide continuous service. See Figure 1, we divide the time horizon into alternate working periods of length a and b, and require each job to be processed either completely within in a periods or completely in b periods. For example, a job which is partly processed in period a cannot be processed in period b, however, it can be processed in any following period a. This design allows machine to hold an unfinished job, at the same time, shift to another job, which can be considered as multitasking (see Hall et al., 2016).

We equivalently consider this shared machine as two identical parallel machines with complementary available time. Figure 1 illustrates the basic idealogy of proposed problem, in which each shaped rectangle represents the non-available time of one machine. Besides, the energy cost in different time periods vary within a certain range.

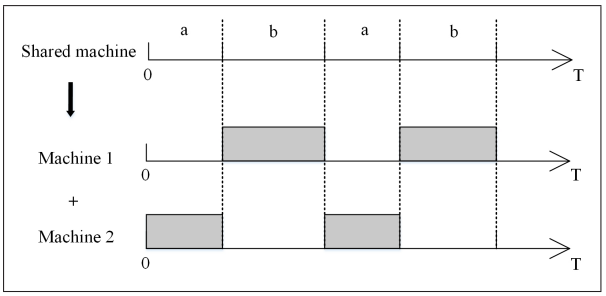


Fig. 1. Framework of proposed problem.

To the best of our knowledge, few papers study on multitasking scheduling, and fewer papers consider energy cost at the same time. We are the first to study multitasking scheduling with energy cost. Our contribution mainly includes: a mixed bi-objective integer programming model is formulated for this newly proposed problem. We design an exact epsilon constraint algorithm to solve this model. Numerical experiments have been conducted and computational results demonstrate the validity of proposed model. We also devise a non-dominated sorting genetic algorithm II (NSGA-II) to produce nearly optimal solutions within a short duration.

The reminder of this paper is organized as follows. Section 2 reviews the related literature. In Section 3, we describe the proposed problem in detail and formulate an mixed integer programming model. Section 4 presents the exact solutions by epsilon constraint method. Basic design of NSGA-II is shown in Section 5, and finally Section 6 concludes this work and future research directions.

II. Literature review

In the applications of operation management, many works investigate the characteristics and influences of widely existed multitasking environments. Seshadri and Shapira (2001) and

Monsell (2003) propose that task switching and corresponding cost need to be considered when dealing with multitasking. Shakeri and Logendran (2007) develop a mixed integer programming model for scheduling tasks in multitasking environments, for which the number of completed tasks is not a good measure. Gaglioppa et al. (2008) study multitasking, multistage production plan and scheduling methods in manufacturing industries, in which a manufacturing system outputs several products in one process. Constraint-based functions are designed to solve the proposed mixed integer programming. Watson and Strayer (2010) and Sanbonmatsu et al. (2013) show that multitasking might degrade workers' performance. Coviello et al. (2014) study the impact of multitasking for a single worker, and show that he can complete more jobs with the increase of the pick-up rate until this rate reaches some certain value, after which the productivity decreases as the pick-up rate increases. Hall et al. (2015) introduce the switching cost function for each job to capture multitasking behavior. They investigate the multitasking scheduling problems with the objectives of total weighted completion time, maximum lateness, and number of late jobs. They show that the first two problems are solvable while the third problem is NP-complete. Hall et al. (2016) address two multitasking models featured with alternate period processing and shared processing, examine the complexity of the multitasking scheduling problems and propose fast algorithms. Zhu et al. (2017a) study multitasking scheduling problems with a rate-modifying activity considering human operators, and conduct four classical objectives to illustrate the theorems and algorithms. Zhu et al. (2017b) generalize their work to the setting with multiple rate-modifying activities. Liu et al. (2017) investigate a joint multitasking scheduling and common due date assignment problem on a single machine, and devise an integer programming model with the objective to minimize the total penalties.

Recently, energy saving attracts large attention due to the fact that some of energy resources are non-renewable. Park et al. (2009) show that energy consumption has grown by as much as 300% over last 50 years, which burdens the limited resources. Gahm et al. (2015) review the previous works and develop a research framework for energy-efficient scheduling. The three dimensions: energetic coverage, energy supply and energy demand are introduced and used to classify the literature. Rager et al. (2015) study the energy-oriented parallel machine scheduling problem aiming at minimizing the overall energy cost. Genetic algorithm and heuristics are conducted to solve this problem. Ding et al. (2015) formulate a bi-objective model with the objectives of production efficiency and energy saving. Mansouri et al. (2016) investigate green scheduling of two-machine flow shop to find the trade-off between makespan and energy cost. Thus we conduct time-of-use (TOU) as the measurable indicator to record diverse energy cost.

Our equivalent problem is featured with parallel machines and periodic available time. Liu et al. (2011) study online scheduling on parallel machines with one machine periodically unavailable. They consider fixed available time and speed-dependent available time problems, design algorithms and show some lower bounds on competitive ratio. Wang and Liu (2013) deal with an integrated optimization model for preventive maintenance in a single machine. The objective is to minimize the total expected weighted completion time of jobs. To solve this problem, they develop a branch and bound algorithm. Wang and Liu (2014a) investigate an integrated bi-objective optimization problem with preventive maintenance in two-stage hybrid flow shop. As machine maintenance on first stage results in unavailability, the scheduling objectives are minimizing the makespan and the unavailability of first stage machine. A multi-objective tabu search method is designed to compare with a well-known multi-objective genetic algorithm. Wang and Liu (2015) study the integration optimization of production scheduling and preventive maintenance with multiple dependent resources. The objective is to simultaneously minimize the makespan for the production aspect, the unavailability of the machine system. Ayoobian and Mohsendokht (2016) aim towards improvement of safety systems reliability in a nuclear power plant. Periodical maintenance programs can be formulated in terms of a multi-objective optimization where unavailability and cost are considered, genetic algorithm is used to find the best series of answers in the form of Pareto frontier curve. Other related studies refer to Wang and Liu (2014b), Bohlin et al. (2015), Cheng et al. (2016) and Ruiz-Torres et al. (2016).

Although multitasking and energy-saving have been attracting attentions these years, we can hardly find papers linking them together, which is an increasing demand in machine scheduling domain.

III. Model formulation

A. Problem description

We assume the job set is J, then there are $|J|$ jobs to be processed on one shared machine, which has continuous alternate working periods. Besides, interruption is allowed in this multitasking scheduling.

This shared machine can be decomposed into two identical parallel machines with periodic non-available time, thus each job can be assigned on any one of two parallel machines. See Figure 2, machine 1 is available within time intervals a, that is, this machine is unavailable for processing jobs within time intervals b. Similarly, machine 2 is available within time intervals b but unavailable within intervals a.

We consider the planning horizon of this shared resource equals to T. In any period a or b, it consists several time steps (10 minutes for one time period in practice), for example, $a = 3t, b = 4t$, then $a+b = 7t$ represents a working cycle. Besides, the energy cost in different time steps are various, thus each job can choose low-cost time steps within its corresponding machine available time.

See Figure 2, we observe that some jobs may be interrupted during processing because (i) one machine can only process within its available time. For example, if job 1 is unfinished in period a on machine 1, it must resume in any one of the following periods a; (ii) energy-oriented production requires each job to be processed within low-cost time steps. For example, machine 2 stops processing job 2 and finds low-cost time steps to resume processing with the purpose to reduce total energy cost.

We assume job j is of integer processing time $p_j \in Z^+$, which is known in advance. The most important thing we need to decide is the total occupied time steps of each job, i.e.,

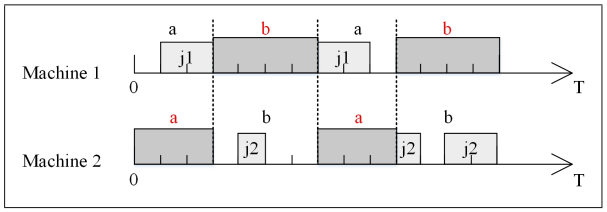


Fig. 2. Illustration of job interruption.

TOU, so as to minimize the total weighted completion time and total energy cost simultaneously. Thus we can find the trade-off between production efficiency and energy cost.

As mentioned above, we can describe this problem as $P2|multitasking|(\sum_{i\in J}\omega_i c_i, \sum_{i\in J}\sum_{t\in T}\theta_t f_{i,t})$, in which $P2$ represents 2 parallel machines, and $multitasking$ is the characteristic of this machine scheduling problem. The objective $\sum_{i\in J}\omega_i c_i$ represents total weighted completion time, while $\sum_{i\in J}\sum_{t\in T}\theta_t f_{i,t}$ represents total energy cost, in which w_i is the weight of job i, c_i is the completion time of job i, θ_t is the energy cost of time step t, and $f_{i,t}$ represents the occupied time steps of job i. The model formulation is shown as follows:

B. Bi-objective integer model

Indices:

- i, j: indices of jobs;
- m, n: indices of parallel machines;
- t: index of periodic time steps.

Parameters:

- J: set of jobs;
- M: set of parallel machines;
- T: set of time steps;
- ω_i: weight of job i where $i \in J$;
- θ_t: energy cost of time step t where $t \in T$;
- p_i: processing time of job i where $i \in J$;
- $\beta_{m,t}$: a binary variable equals to 1 if machine m is available in time step t, 0 otherwise where $m \in M, t \in T$;
- $\overline{M}$: a sufficiently large integer.

Variables:

- $X_{i,m}$: a binary variable equals to 1 if job i is assigned on machine m, 0 otherwise where $i \in J, m \in M$;
- s_i: start time of job i where $i \in J$;
- c_i: completion time of job i where $i \in J$;
- $\alpha_{i,j,m}$: a binary variable equals to 1 if jobs i and j are assigned to the same machine m, and job i is processed before j, 0 otherwise where $i, j \in J, i \neq j, m \in M$;
- $f_{i,t}$: a binary variable equals to 1 if job i occupies time step t, 0 otherwise where $i \in J, t \in T$.

Objective function:

$$Obj_1 = min \sum_{i\in J} \omega_i c_i \tag{1}$$

$$Obj_2 = min \sum_{i\in J}\sum_{t\in T} \theta_t f_{i,t} \tag{2}$$

Subject to:

Assignment constraints

$$\sum_{m\in M} X_{i,m} = 1, i \in J \tag{3}$$

$$X_{i,m} \geq \alpha_{i,j,m}, i, j \in J, i \neq j, m \in M \tag{4}$$

$$X_{j,m} \geq \alpha_{i,j,m}, i, j \in J, i \neq j, m \in M \tag{5}$$

$$\alpha_{i,j,m} + \alpha_{j,i,m} \leq 1, i, j \in J, i \neq j, m \in M \tag{6}$$

$$\alpha_{i,j,m} + \alpha_{j,i,m} \geq 1 - \overline{M}(2 - X_{i,m} - X_{j,m}), \quad i, j \in J, i \neq j, m \in M \tag{7}$$

Machine available time

$$\sum_{i\in J} f_{i,t} \leq 1, t \in T \tag{8}$$

$$\sum_{t\in T} f_{i,t} = p_i, i \in J \tag{9}$$

$$f_{i,t} \leq \overline{M}(1 - X_{i,m}) + \overline{M}\beta_{m,t}, i \in J, m \in M, t \in T \tag{10}$$

$$f_{i,t} + f_{j,t} \leq 1 + \overline{M}(2 - X_{i,m} - X_{j,n}), \quad i, j \in J, i \neq j, m, n \in M, m \neq n, t \in T \tag{11}$$

Processing constraints

$$s_i \geq 0, i \in J \tag{12}$$

$$c_i \geq t f_{i,t}, i \in J, t \in T \tag{13}$$

$$c_i \leq s_j + \overline{M}(1 - \alpha_{i,j,m}), i, j \in J, i \neq j, m \in M \tag{14}$$

$$c_i \geq s_j + 1 - \overline{M}\alpha_{i,j,m}, i, j \in J, i \neq j, m \in M \tag{15}$$

Ranges

$$X_{i,m}, f_{i,t}, \alpha_{i,j,m} \in \{0, 1\}, i, j \in J, i \neq j, m \in M, t \in T \tag{16}$$

$$s_i, c_i \in Z^+, i \in J \tag{17}$$

Equation (1) represents the objective of minimizing total weighted completion time, while equation (2) is of minimizing total energy cost. Constraint (3) ensures each job i can be assigned on one exact machine. Constraints (4)-(7) describe the linking relationships of binary variables $X_{i,m}$ and $\alpha_{i,j,m}$, while

constraints (8) and (9) show the relationship between jobs i and j if they are processed on same machine m in sequence. Combining these above 6 constraints together can express the job-processing sequence in detail. Constraints (10)-(13) are machine available time constraints. Specifically, constraint (10) ensures for any time step t, at most one job can be processed. Constraint (11) is the definition of processing time of job t, which equals to the sum of its occupied time steps. Constraint (12) shows that only if job i has been assigned on machine m, at the same time, machine m is available in time step t, then job i can occupy this time step or not. Constraint (13) guarantees only one machine can processed in each time step t. Constraint (14) ensures the start time of each job i. Constraint (15) searches for the maximal occupied time step t of job i to record its completion time. Finally, constraints (16) and (17) are the ranges of binary variables and integer variables.

IV. Exact solution approach

In this section, we conduct a series of experiments for different instances to analyze and evaluate the performance of our proposed model. Commercial solver CPLEX 12.6 is used and called in Matlab2014b and executed on a personal computer (Intel Core i5 with 4 GB RAM) with Microsoft Windows 7 operating system.

A. The epsilon constraint method

We first utilize epsilon constraint method to solve this model, whose core ideology is to transform a bi-objective problem into a series of single preferred objective problems and limit another by epsilon constraints. Then this model can be formulated as follows, assuming Obj_1 is selected as the preferred objective.

min Obj_1, s.t., Constraint (3)-(17), and $Obj_2 \leq \epsilon_k$

in which ϵ is a parameter with the bound of Ideal point (Obj_1^I, Obj_2^I) and Nadir point (Obj_1^N, Obj_2^N). The following for single objective problems can be optimally solved considering the epsilon constraints.

min Obj_1, s.t., Constraint (3)-(17)

min Obj_2, s.t., Constraint (3)-(17)

min Obj_1, s.t., Constraint (3)-(17), and $Obj_2 \leq Obj_2^I$

min Obj_2, s.t., Constraint (3)-(17), and $Obj_1 \leq Obj_1^I$

Several iterations can show the trade-off between two objectives, and ϵ is initialized by Obj_2^N for the first iteration. The value of ϵ in kth (k >1) iteration ϵ_k is updated as $Obj_{2,k-1} - \delta$, where $Obj_{2,k-1}$ is obtained for $(k-1)$th iteration, and δ is a constant which set as the minimum step value of Obj_2, refers to Wu et al. (2015).

B. Optimal solutions by CPLEX

We then adopt an epsilon constraint method to optimally solve our proposed bi-objective model, which can calculate the trade-off between two proposed objectives. All the parameters are generated randomly based on the practical information, and the total weight of every job equals to 1, i.e., $\omega_1 + \omega_2 + ... + \omega_J = 1$. One 7-job experiment is set as follows to analyze the effect of this method:

- Number of jobs, $J = 7$;
- Number of parallel machines, $N = 2$;
- Planning time horizon, $T = 35t$. Assuming there are five cycles for total planning, in which period $a = 3t$ for machine 1 and period $b = 4t$ for machine 2, thus the length of one cycle equals to $a + b = 7t$.
- Processing time, $p_i = [5, 6, 2, 4, 4, 3, 2]$;
- Job weight, $w_i = [0.15, 0.15, 0.05, 0.2, 0.1, 0.15, 0.2]$;
- Energy cost matrix, in which one line represents one cycle (including 7 time steps).

$$\begin{pmatrix} 3 & 2 & 3 & 4 & 4 & 2 & 2 \\ 5 & 1 & 5 & 5 & 4 & 1 & 2 \\ 2 & 4 & 1 & 4 & 1 & 4 & 3 \\ 4 & 4 & 5 & 5 & 2 & 4 & 1 \\ 1 & 4 & 3 & 3 & 5 & 4 & 4 \end{pmatrix}$$

We first demonstrate the optimal solutions for each single objective. Figure 3 represents the scheduling strategies for total weighted completion time and total energy cost based on the above parameter set, the number in each shaped rectangle represents job index. The completion time of objective 1 is c_i=[14, 24, 26, 10, 21, 6, 2], thus the minimum total weighted completion time is 12.4, while the minimum total energy cost is 69. We can easily observe the scheduling solutions are totally different, thus the two objectives reveal diverse optimization abilities.

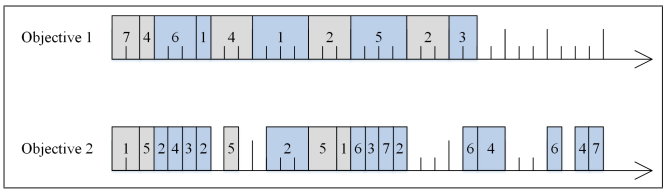


Fig. 3. Scheduling strategies for single objective under 7-job instance.

Then we use epsilon constraint method to conduct the experiment based on the same input. The Pareto frontier is shown in Figure 4 in which the point $7-1$ (represents the first obtained point in 7-job instance, see Table 1 and Figure 4) represents the Ideal point of objective 1, while the Nadir point objective 2. If we use a smooth curve the draw this point together, then every point on this curve has the same decision paly outs. Table 1 records the solution combination of produced points. Besides, the calculating time is 354 seconds.

TABLE I. Results of epsilon constraint method under 7-job instance

Pareto frontier	Obj_1	Obj_2
7-1	12.4	82
7-2	12.45	81
7-3	12.5	78
7-4	12.9	77
7-5	13.2	74
7-6	13.3	73
7-7	13.75	72
7-8	13.95	71
7-9	14.35	70
7-10	15.2	69

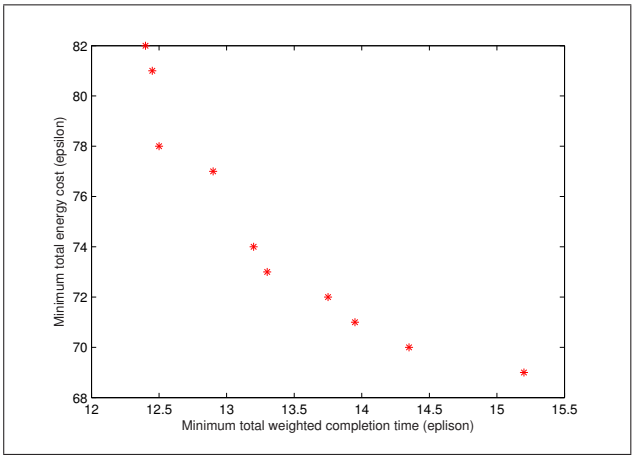


Fig. 4. Pareto frontier of epsilon constraint method under 7-job instance.

We try to use large-size instances to examine our proposed model, however, CPLEX becomes very time consuming. Considering two parallel machine scheduling problem with total completion time is already NP-hard (Pinedo, 2005), our problem is more complicated, thus we need to devise fast heuristics for large-size experiments.

V. Heuristic Algorithms

In this section, we will devise several heuristic algorithms to resolve our problems under large size instances, including a non-dominated sorting genetic algorithm II (NSGA-II) and a problem-based greedy algorithm for comparison.

NSGA-II, proposed by Kalyanmoy et al. (2002), is a widely used genetic algorithm for solving multi-objective problems. Basic framework of NSGA-II is shown as pseudo code in Algorithm 1 below.

Algorithm 1 Framework of NSGA-II

Require: Parameters for proposed problem
Ensure: Solution set
population=pop;
generation=gen;
P=Initialize( );
P=Non-domination sort(P);
for *k=1:gen* **do**
  pool=gen/2;
  tour=2;
  parent_ P=Selection(P, pool, tour);
  offspring_ P=Crossover_ Mutation(parent_ P);
  intermediate_ P=parent_ P $\bigcup$ *offspring_ P*;
  F = Non-domination sort(intermediate_ P);
  P=$\emptyset$;
  i=i+1;
  while $|P| + |F_i| \leq pop$ **do**
    P=P $\bigcup$ *F_i*;
    i=1;
  end while
  k=pop-$|P|$;
  P=P $\bigcup$ *Selection(P, F_i, k)*;
end for

In Algorithm 1, the function *Initialize()* randomly generates several feasible solutions. The initial population is sorted by *Non-domination sort()* function. For each generation, the offspring solutions are generated by genetic algorithm operators, such as crossover and mutation, defined in *Crossover_ Mutation()* function, after which *Test()* function is used to check feasibility. Then the combination process is needed and non-domination level $\{F_1, F_2, ...F_i, ...\}$ can be sorted by *Non domination sort()* function. In order to meet the requirement of *pop*, a binary selection policy, i.e., *Selection(P, F_i, k)* is used to realize this demand. The main points we consider are as follows.

Population initialization. We generate the individual solutions which must be feasible, thus the following constraints need to be considered: (i) each machine can only process within its own available time periods, (ii) each job can be only assigned to one machine, (iii) if one time step is occupied by one job, no other jobs can be processed in this time step. All the conditions above are embedded into function $Initialize()$ to produce feasible solutions.

Objective description. Based on the produced original solutions, the two objective values can be calculated.

Genetic operators. (i) Two parent solutions are randomly selected to generate two offsprings after mating selection; (ii) we use a specified crossover probability, and conduct *binary crossover method* in which two random genes in the same part are selected between two individuals; (iii) a mutation procedure is conducted to help this algorithm escape from local optimal, with mutation probability, the gene can be chosen in any part and change into a feasible value.

We will do further experiments based on the preliminary design of NSGA-II, so as to solve our proposed problem within an acceptable time. More fast algorithms will be designed to compare with NSGA-II.

VI. Conclusion

We study multitasking phenomenon with energy cost. A mathematical model is formulated to describe this comprehensive problem. We optimally get the scheduling solutions using epsilon constraint method. For large-size instances, we plan to device NSGA-II to resolve this problem in multiple segments, and design some fast heuristics to make comparison with NSGA-II. Numerical experiments demonstrate the validity of proposed model, which shows incremental contribution in daily operations under multitasking environments, thus can be applied into practice.

Future research directions may contain: (i) consideration of due date related objectives, (ii) improvement on the proposed algorithm, (iii) consideration of stochastic situation or maintenance of this share machine.

Acknowledgment

This work is supported by the National Science Foundation of China (71571134, 71428002, 71432007, and 71571135). The work is also supported by the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program.

References

[1] N. Ayoobian, M. Mohsendokht, Multi-objective optimization of maintenance programs in nuclear power plants using Genetic Algorithm and Sensitivity Index decision making. Annals of Nuclear Energy, 88: 95-99, 2016.

[2] M. Bohlin, M. Wärja, Maintenance optimization with duration-dependent costs. Annals of Operations Research, 224(1): 1-23, 2015.

[3] E. Chan, F. G. V. Zee, P. Bientinesi, G. Quintana, SuperMatrix: A multithreaded runtime scheduling system for algorithms-by-blocks. ACM Sigplan Symposium on Principles and Practice of Parallel Programming, PPOPP 2008, Salt Lake City, Ut, Usa, February. DBLP, 123-132, 2008.

[4] G. Q. Cheng, B. H. Zhou, L. Li, Joint optimisation of production rate and preventive maintenance in machining systems. International Journal of Production Research, 54(21): 1-17, 2016.

[5] D. Coviello, N. Persico, P. K. Goldberg, Time allocation and task juggling. American Economic Review, 104(2): 609-623, 2014.

[6] J. Y. Ding, S. Song, C. Wu, Carbon-efficient scheduling of flow shops by multi-objective optimization. European Journal of Operational Research, 248(3): 758-771, 2015.

[7] F. Gaglioppa, L. A. Miller, S. Benjaafar, Multitask and multistage production planning and scheduling for process industries. Operations Research, 56(4): 1010-1025, 2008.

[8] C. Gahm, F. Denz, M. Dirr, Energy-efficient scheduling in manufacturing companies: A review and research framework. European Journal of Operational Research, 248(3): 744-757, 2015.

[9] N. G. Hall, J. Y. T. Leung, C. L. Li, The effects of multitasking on operations scheduling. Production and Operations Management, 24(8): 1248-1265, 2015.

[10] N. G. Hall, J. Y. T. Leung, C. L. Li, Multitasking via alternate and shared processing: Algorithms and complexity. Discrete Applied Mathematics, 208(C): 41-58, 2016.

[11] D. Kalyanmoy, P. Amrit, A. Sameer, T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2): 182-197, 2002.

[12] M. Liu, F. Zheng, C. Chu, Y. Xu, Optimal algorithms for online scheduling on parallel machines to minimize the makespan with a periodic availability constraint. Theoretical Computer Science, 412(39): 5225-5231, 2011.

[13] M. Liu, S. Wang, F. Zheng, C. Chu, Algorithms for the joint multitasking scheduling and common due date assignment problem. International Journal of Production Research, Accepted.

[14] S. A. Mansouri, E. Aktas, U. Besikci. Green scheduling of a two-machine flowshop: Trade-off between makespan and energy consumption. European Journal of Operational Research, 248(3): 772-788, 2016.

[15] S. Monsell, Task switching. Trends in Cognitive Sciences, 7(3): 134-140, 2003.

[16] S. M. Mostafa, S. Kusakabe, Towards reducing energy consumption using inter-process scheduling in preemptive multitasking OS. International Conference on Platform Technology and Service, DOI: 10.1109/PlatCon.2016.7456832, 2016.

[17] C. W. Park, K. S. Kwon, W. B. Kim, B. K. Min, S. J. Park, I. H. Sung, Y. S. Yoon, K. S. Lee, J. H. Lee, J. Seok, Energy consumption reduction technology in manufacturing-A selective review of policies, standards, and research. International Journal of Precision Engineering & Manufacturing 10(2): 151-173, 2009.

[18] M. L. Pinedo. Scheduling: Theory, algorithms, and systems. Springer, 2008.

[19] M. Rager, C. Gahm C, F. Denz, Energy-oriented scheduling based on Evolutionary Algorithms. Computers & Operations Research, 54: 218-231, 2015.

[20] T. G. Rogers, M. OConnor, T. M. Aamodt, Learning your limit: managing massively multithreaded caches through scheduling. Communications of the ACM, 57(12): 91-98, 2014.

[21] A. J. Ruiz-Torres, G. Paletta, R. M. Hallah, Makespan minimisation with sequence-dependent machine deterioration and maintenance events. International Journal of Production Research, 1-18, 2016.

[22] D. M. Sanbonmatsu, D. L. Strayer, N. Medeirosward, Who Multi-Tasks and Why? Multi-Tasking Ability, Perceived Multi-Tasking Ability, Impulsivity, and Sensation Seeking. PLOS ONE, 8(1): e54402, 2013.

[23] S. Seshadri, Z. Shapira, Managerial allocation of time and effort: the effects of interruptions. Management Science, 47(5): 647-662, 2001.

[24] S. Shakeri, R. Logendran, A mathematical programming-based scheduling framework for multitasking environments. European Journal of Operational Research, 176(1): 193-209, 2007.

[25] S. Wang, M. Liu, A branch and bound algorithm for single-machine production scheduling integrated with preventive maintenance planning. International Journal of Production Research, 51(3): 847-868, 2013.

[26] S. Wang, M. Liu, Two-stage hybrid flow shop scheduling with preventive maintenance using multi-objective tabu search method. International Journal of Production Research, 52(5): 1495-1508, 2014a.

[27] S. Wang, M. Liu, Two-machine flow shop scheduling integrated with preventive maintenance planning. International Journal of Systems Science, 47(3): 1-19, 2014b.

[28] S. Wang, M. Liu, Multi-objective optimization of parallel machine scheduling integrated with multi-resources preventive maintenance planning. Journal of Manufacturing Systems, 37: 182-192, 2015.

[29] J. M. Watson, D. L. Strayer, Supertaskers: Profiles in extraordinary multitasking ability. Psychonomic Bulletin & Review, 17(4): 479-485, 2010.

[30] F. Zhang, J. Cao, W. Tan, S. U. Khan, Evolutionary scheduling of dynamic multitasking workloads for big-data analytics in elastic cloud. IEEE Transactions on Emerging Topics in Computing, 2(3): 338-351, 2014.

[31] Z. Zhu, F. Zheng, C. Chu, Multitasking scheduling problems with a rate-modifying activity. International Journal of Production Research, 55(1): 296-312, 2017a.

[32] Z. Zhu, M. Liu, C. Chu, J. Li, Multitasking scheduling with multiple rate-modifying activities. International Transactions in Operational Research, DOI: 10.1111/itor.12393, 2017b.

Bi-objective optimization of a reentrant flow shop scheduling with exact time lag considering energy cost

Ming Liu[1], Xin Liu[2], FeiFeng Zheng[3] and Feng Chu[4]

***Abstract*— In this paper, we study a bi-objective optimization of the two-machine reentrant flow shop scheduling problem with an exact time lag considering energy consumption cost. This problem is proposed by Amrouche et al. (2016), in which each job must be operated from the first machine M_1 to the second machine M_2 and then back to the first machine M_1 with an exact time lag l_j between the two operations on the first machine. The two objectives are: (1) minimization of the total energy consumption cost and (2) makespan. To describe the problem precisely, we propose a bi-objective mixed integer programming formulation. An ϵ-constraint method and NSGA-II approach are proposed to obtain the optimal Pareto front and approximate Pareto solutions for the problem. The frameworks of the proposed methods are presented in this paper.**

I. INTRODUCTION

With the development of technology and industry, new research issues keep emerging in the field of flow shop scheduling. The usual flow shop scheduling problem assumes that jobs are operated on all machines following a preset sequence, and a job visit each machine only once (c.f. Yang and Chern (2000), Yang et al. (2008), Wu and Lee (2009), Wang et al. (2015)). However, this assumption may be inconsistent with some production conditions in reality. Reentrant flow shop scheduling problem has been attracting more and more attention with the assumption that a job can visit a certain machine more than once since it was first proposed by Graves (1983).

Generally, the objective of reentrant flow shop scheduling problem is to minimize customer dissatisfaction and related cost, such as the number of tardy jobs, makespan and total tardiness etc. Environmental effects and energy consumption cost are seldomly considered in the previous relevant studies. As people become more and more concerned about the environmental problems and energy wast nowadays, it is necessary to take energy consumption cost into consideration.

In this paper, we study a bi-objective optimization of the two-machine reentrant flow shop scheduling problem with an exact time lag. The problem is proposed by Amrouche et al. (2016), where each job must be operated from the first machine M_1 to the second machine M_2 and then back to the first machine M_1 with an exact time lag l_j between the two operations on the first machine. Amrouche et al. (2016) prove the NP-hardness of the problem and propose heuristics for the problem. We extended their work by establishing a bi-objective mathematical formulation and considering energy consumption cost. The main contributions of this paper are presented as follows.

(1) We consider two objectives optimization of minimizing the total energy consumption cost and the makspan simultaneously;

(2) A mixed integer programming formulation is established;

(3) An ϵ-constraint method and NSGA-II approach are proposed to obtain the optimal and approximate Pareto solutions for the problem.

The rest of this paper is structured as follows. Section 2 reviews the literature on reentrant flow shop scheduling problem and shop scheduling problem with an exact time lag and energy consumption. Section 3 describes the problem, and a mixed integer programming formulation is proposed. In Section 4, we propose solution methods to obtain the Pareto solutions for the problem. Section 6 concludes the work and indicate further research directions.

II. LITERATURE REVIEW

There has been a limited number of researches on reentrant flow shop scheduling problem (c.f. Hwang and Sun (1998), Jing et al. (2009), Chu et al. (2010), Huang et al. (2014), Sangsawang et al. (2015), Zhou et al. (2016), Amrouche et al. (2016)). Moreover, the number of previous relevent researches focusing on reentrant flow shop with exact time lags is relatively few.

A. reentrant flow shop

The reentrant flow shop scheduling problem is first proposed by Graves (1983), and the author proposes a heuristic algorithm for the problem. Lev and Adiri (1984) investigate the V-shop problem, in which jobs are operated on m machines in a preset sequence $M_1, M_2, ..., M_{m-1}, M_m, M_{m-1}, ..., M_2, M_1$. They prove that this problem is NP-hard, and also propose polynomial algorithms for some special cases. Wang et al. (1997) study a chain-reentrant shop scheduling problem, where each job is operated in the preset sequence $M_1, M_2, M_3, ..., M_m$, and then back to the primary machine M_1. The authors prove some properties of optimal solutions for the two-machine case.

Chen (2006) study a reentrant permutation flow shop scheduling problem, in which each job is planned to

[1]Ming Liu is with Faculty of Economics and Management, Tongji University, Shanghai 200092, China mingliu@tongji.edu.cn

[2]Xin Liu is with the Department of Economics and Management, Tongji University, Shanghai 200092, China liuxin9735@126.com

[3]FeiFeng Zheng is with Glorious Sun School of Business & Management, Donghua University, Shanghai 200051, China ffzheng@dhu.edu.cn

[4]Feng Chu is with the Laboratory IIBISC, Univ Évry, University of Paris-Saclay, 91025 Évry, France, and is with Management Engineering Research Center Xihua University, Chengdu 610039, China feng.chu@ibisc.univ-evry.fr

be operated on all machines in the following sequence, $M_1, M_2, ..., M_m, M_1, M_2, ..., M_m,$ and $M_1, M_2, ..., M_m$. The objective is to minimize the makespan. They propose a branch and bound algorithm for the problem. Based on that, Chen et al. (2008) develop a tabu search algorithm and a hybrid genetic algorithm for this problem.

Kim (2005) investigates the reentrant hybrid flow shop scheduling problem with the aim of minimizing the total tardiness. The author develops a heuristic algorithm for the problem and evaluates it by conducting numerical experiments. Based on that, Choi and Kim (2009) study the two-machine reentrant flow shop scheduling problem with the objective of minimizing the total tardiness. They propose a branch and cut algorithm for the problem.

Desprez et al. (2009) address a industrial production scheduling problem in reality, which is a hybrid reentrant flow shop scheduling problem with the objective of minimizing the total weighted number of tardy jobs. The authors develop a genetic algorithm according to the characteristics of the studied problem.

Chu et al. (2010) study a reentrant flow shop scheduling problem, which can be regarded as a special case of the problem investigated by Wang et al. (1997). The authors propose an optimal schedule for the problem with the aim of minimizing the makespan. For the problem that minimize the total flow time, they divide the problem into many subproblems according to the relationship between the processing time in different procedure, and propose algorithms and some dominant properties for the different subproblems.

Huang et al. (2014) propose a farness particle swarm optimization algorithm for the reentrant shop problem, which is a hybrid flow shop scheduling problem considering due dates and time windows. The objective of the studied problem is minimizing the total earliness and tardiness. Xu et al. (2014) develop a memetic algorithm for the reentrant permutation flow shop scheduling problem with the objective of minimizing the makespan.

Sangsawang et al. (2015) address a reentrant two-stage hybrid flow shop scheduling problem considering the blocking constraint. The objective of the problem is minimizing the makespan. They proposed a hybrid particle swarm optimization and an adaptive auto-tuning hybrid genetic algorithm, which is based on the idea of relaxation and modification.

Zhou et al. (2016) investigate the reentrant hybrid flow shop scheduling problem with the objective of minimizing the total weighted completion time. They take inspection and repair operation into consideration, and these operations are conducted at the end of each cycle of processing. The authors propose a mathematical formulation for the problem and a differential evolution algorithm for the problem. Other studies on reentrant shop scheduling may include Bertel and Billaut (2004), Pan and Chen (2005), Dugardin et al. (2010), Jing et al. (2011), Yalaoui et al. (2014), Ying et al. (2014), Shen et al. (2016), Rifai et al. (2016), etc.

B. shop scheduling with time lags

The idea of time lags has been studied in various scheduling problems, in particular in the two-machine flow shop and the single machine case with the name of coupled-task problem. In the two-machine flow shop problem, the notation of time lag of a job represents that the time difference between the start time of the job on the second machine and the completion time of it on the first machine. Yu et al. (2004) prove the NP-hardness in the strong sense of the two-machine flow shop scheduling problem. For the single machine case with the name of couple-task problem, the process of each job contains two individual operations with required delay time between them. Ahr et al. (2004) present an exact algorithm based on dynamic programming for this problem.

Amrouche and Boudhar (2013) introduce the notion of time lags to the two-machine reentrant shop scheduling problem with the objectiveof minimizing the total completion time. In the problem, each job must be processed following the sequence M_1, M_2 and then back to M_1 with an identical exact time lag between the completion time of the first operation and the start time of the third operation. Based on that, Amrouche and Boudhar (2016) prove the NP-hardness of a special case and given some special subproblems that can be solved in polynomial time.

Amrouche et al. (2016) thoroughly study a chain-reentrant shop scheduling problem with an exact time lag in order to minimize the makespan. They propose heuristic algorithms for the problem with identical time lags, and establish a new NP-hardness result and some polynomial cases. However, the authors have not given the mathematical formulation for the problem. Based on this work, we extend the problem by considering energy consumption cost as one of the objectives.

C. Sustainability considerations in shop scheduling

Sustainability considerations in shop scheduling have been attracting more and more attention. Moreover, electricity is the main form of energy consumption in manufacturing industry (Sun and Li, 2013). Mouzon et al. (2007) develop operational methods for the minimization of the energy consumption of manufacturing equipment by turning machines off at idle times. Mouzon and Yildirim (2008) propose a metaheuristic to decide the timing and length of turn off operation and turn on operation to minimize total tardiness and total energy consumption simultaneously. Che et al. (2016) address an energy-conscious single machine scheduling problem under time-of-use (TOU) electricity tariffs, where the prices of electricity are various in different moment of a day. They propose a MILP formulation and a greedy insertion heuristic algorithm for the problem.

Demir et al. (2014) study a bi-objective Pollution-Routing problem, and the two objectives are minimizing the fuel consumption and the driving time of vehicles. Ding et al. (2015) consider a permutation flow shop scheduling problem with the objective of minimizing the carbon emissions and the makespan simultaneously. Mansouri et al. (2016)

investigate a two-machine flowshop scheduling problem by analyzing the trade-off between minimizing the makespan and total energy consumption. They propose a mixed integer bi-objective optimization formulation and a heuristic. Wang et al. (2016) investigate a single machine batch scheduling problem considering energy cost with the objective of minimizing the makespan and total energy cost simultaneously. A bi-objective discrete-time mathematical formulation and two heuristic algorithms are proposed. Other studies on shop scheduling problem considering energy consumption may include Luo et al. (2013), Shrouf et al. (2014), etc.

The framework presented in this paper extends the literature on two-machine reentrant flow shop scheduling by considering two objectives of minimizing the total energy consumption cost and the makespan.

III. PROBLEM DESCRIPTION AND FORMULATION

In the problem we discussed here, there are n jobs planed to be processed on two machines. Every job is operated on the machines following a preset sequence. A job j should be processed in the order M_1, M_2 and then back to machine M_1, and there is an exact time lag l_j between the two operations on the first machine.

The aim of this problem is to find a set of sequences of the jobs and determine the starting time for each job in different procedure. The two objectives are minimizing the total energy consumption and minimizing the makespan. We study the problem by considering non-identical job sizes and the time-of-use (TOU) electricity prices. Besides, we consider the reentrance by transforming the second procedure on the primary machine into a virtual machine, as shown in Fig. 1.

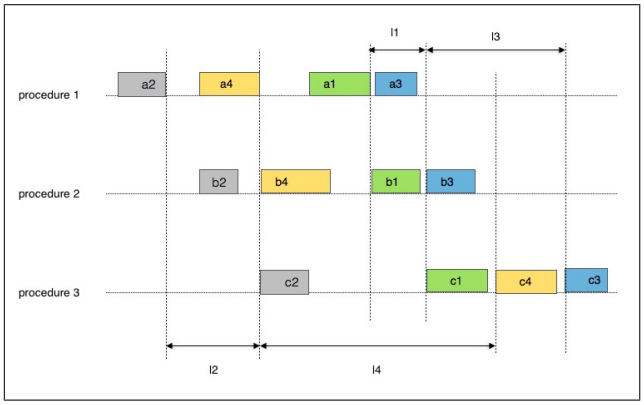


Fig. 1. Framework of initial solution generation

To illustrate the problem, a mixed integer programming formulation is established below.

A. Notation

Indices:

- i, j: indices of jobs, $i \neq j$.
- h: index of procedures.
- t: index of time.
- k: index of machines.

Parameters:

- N: set of jobs, i.e. $N = \{1, ..., n\}$.
- M: set of machines, i.e. $M = \{1, ..., m\}$.
- H: set of procedures.
- P_k: set of procedures on machine k.
- l_i: the exact time lag of job i.
- p_{hi}: the processing time of job i in procedure h.
- c_t: the unit price of energy consumption at time t.
- L: a large enough number.

Variables:

- x_{ij}^h: equal to 1 if job i is processed before job j in procedure h, 0 otherwise.
- z_{hi}^t: equal to 1 if job i is being processed in procedure h at time t, 0 otherwise.
- S_{hi}: positive integer, the start time of job i in procedure h.
- C_{hi}: positive integer, the completion time of job i in procedure h.
- C_{max}: positive integer, the makespan.

B. A mathematical formulation

We describe the two objective functions and the main constraints below. The first objective f_1 is to minimize the makespan, and the second one f_2 represents the total energy cost.

$$\min f_1 = C_{max} \tag{1}$$

In the minimization of item f_2, as the total energy cost is the function of the time of energy consumption and the price of electricity, we calculate it based on cumulative energy cost.

$$\min f_2 = \sum_{i \in N} \sum_{t \in T} \sum_{h \in H} z_{hi}^t c_t \tag{2}$$

In the problem of reentrant flow shop scheduling with exact time lag, some constraints must be satisfied and we describe them below. For two jobs i and j processed in a certain procedure h, job i is either processed before job j or after it.

$$\text{s.t.} \quad x_{ij}^h + x_{ji}^h = 1, \quad h \in H, i, j \in N, i \neq j \tag{3}$$

Each job must be processed in a procedure without interruption, which means that there is no preemption allowed in the problem.

$$C_{hi} = S_{hi} + p_{hi}, \quad h \in H, i \in N \tag{4}$$

The start time of a job i in procedure h must be after its completion time in the previous procedure $h-1$.

$$S_{hi} \geq C_{h-1,i}, \quad h \in H, i \in N \tag{5}$$

Each job is planned to be processed in the order M_1, M_2 and then return back to M_1 after an exact time lag l_i.

$$S_{3i} = C_{1i} + l_i, \quad i \in N \tag{6}$$

The start time of a job i must be ensured to be after the completion time of jobs prior to it in procedure h.

$$S_{hj} + (1 - x_{ij}^h)L \geq C_{hi}, \quad h \in H, i, j \in N, i \neq j \tag{7}$$

Item z_{hi}^t represents whether job i is being processed in procedure h at time t.

$$C_{hi} \geq (z_{hi}^t - 1)L + t, \quad h \in H, i \in N, t \in T \tag{8}$$

$$S_{hi} \leq (1 - z_{hi}^t)L + t - 1, \quad h \in H, i \in N, t \in T \tag{9}$$

$$\sum_{t \in T} z_{hi}^t = p_{hi}, \quad h \in H, i \in N \tag{10}$$

Since procedure 1 and procedure 3 are operated on the machine M_1, constraint below ensures that there cannot be more than one job being processed on the machine M_1 at a certain time t.

$$\sum_{i \in N} z_{1i}^t + \sum_{i \in N} z_{3i}^t \leq 1, \quad t \in T \tag{11}$$

The ranges of decision variables are given below.

$$x_{ij}^h \in \{0, 1\}, \quad h \in H, i, j \in N, i \neq j \tag{12}$$

$$z_{hi}^t \in \{0, 1\}, \quad h \in P, i \in N, t \in T \tag{13}$$

$$S_{hi}, C_{hi} \geq 0, \quad i \in N, h \in H \tag{14}$$

IV. SOLUTION APPROACH

For bi-objective optimization problems, the relationship between the two objectives is usually negative correlation. There is no single optimal solution, and the desired output is a solution set (i.e. Pareto front) instead. In this section, to solve the two-machine reentrant flow shop scheduling problem with an exact time lag, we attempt to obtain the Pareto solutions with ϵ-constraint method and NSGA-II approach. In the following, we present the frameworks of the proposed methods.

A. ϵ-constraint method

The ϵ-constraint method is widely used to obtain the optimal Pareto solutions. It is based on the the idea of focusing on a single objective with restricting remaining objectives. We solve a multi-objective solution with ϵ-constraint method by setting one of the objectives, i.e. f_2, as a single objective and transforming the other objective into a constraint with a gradual reduction value ϵ in each iteration. The upper bound and lower bound of the two objectives need to be obtained and considered as a constraint. For the bi-objective two-machine reentrant flow shop problem with an exact time lag, the following concepts are needed (Bérubé et al. (2009)).

- **Ideal point**: let $\mathbf{f}^I = (f_1^I, f_2^I)$ with $f_1^I = min\{f_1(\mathbf{X})\}$ and $f_2^I = min\{f_2(\mathbf{X})\}$, $\mathbf{X} \in \mathbf{Z^V}$;
- **Nadir point**: let $\mathbf{f}^N = (f_1^N, f_2^N)$ with $f_1^N = min\{f_1(\mathbf{X}) : f_2(\mathbf{X}) = f_2^I\}$ and $f_2^N = min\{f_2(\mathbf{X}) : f_1(\mathbf{X}) = f_1^I\}$, $\mathbf{X} \in \mathbf{Z^V}$;
- **Extreme point**: $\mathbf{f}^E = (f_1^I, f_2^N)$ and $\mathbf{f}^E = (f_1^I, f_2^N)$ are two extreme points on the Pareto front.

In this paper, the two objectives are the total energy consumption cost and makespan. The makespan is regarded as a constraint in the method. The ϵ-constraint method is detailed below:

Algorithm 1 Framework of ϵ-constraint method

Require: *Parameters for proposed problem*
$f_1^I = min\{f_1(\mathbf{X})\}$;
$f_2^I = min\{f_2(\mathbf{X})\}$;
$f_1^N = min\{f_1(\mathbf{X}) : f_2(\mathbf{X}) = f_2^I\}$;
$f_2^N = min\{f_2(\mathbf{X}) : f_1(\mathbf{X}) = f_1^I\}$;
$F_{apo} = [f_1^I, f_2^N]$;
$\Delta = 1$;
$\epsilon = f_2^N - \Delta$;
while $\epsilon \geq f_2^I$ **do**
  $f_1 = min\{f_1(\mathbf{X}) : f_2(\mathbf{X}) \leq \epsilon\}$;
  $f_2 = min\{f_2(\mathbf{X}) : f_1(\mathbf{X}) = f_1\}$;
  $F_{apo} = [F_{apo}; f_1, f_2]$;
  $\epsilon = f_2 - \Delta$;
end while
$F = non_dominated_sort(F_{apo})$;

B. Optimal solutions by CPLEX

We adopt ϵ-constraint method to obtain optimal solutions for the proposed mathematical model. The parameters are generated randomly, as shown below.

- Number of jobs, N = 4;
- Number of machines, M = 2;
- Number of procedures, H = 3;
- Planning time horizon, T = 80;
- Processing time of jobs in different procedure:

$$\begin{pmatrix} 3 & 4 & 5 & 2 \\ 2 & 3 & 4 & 3 \\ 3 & 2 & 4 & 5 \end{pmatrix}$$

- Time lags. l_j = $[5, 5, 6, 7]$;
- The prices of electricity in different time of a day, which is randomly generated from 1 to 4.

We use ϵ-constraint method to obtain the optimal Pareto solutions for the instance above, as shown in the Fig.2. The computational time limit is 300 seconds. Table 1 represents the solution combinations of the results obtained.

C. NSGA-II approach

In this section, we present NSGA-II method in order to solve the large scale cases efficiently. NSGA-II has been widely used to solve the multi-objective problems, which is first proposed by Deb et al. (2002). Details of NSGA-II

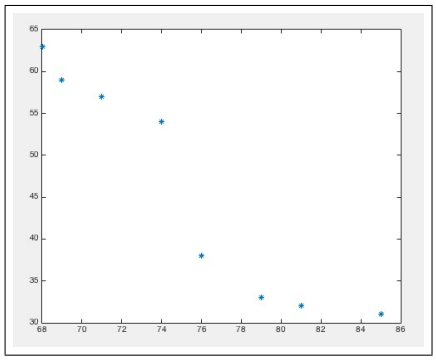


Fig. 2. Framework of initial solution generation

TABLE I

RESULTS OF ϵ-CONSTRAINT METHOD

Pareto frontier	total energy cost	makespan
1	68	63
2	69	59
3	71	57
4	74	54
5	76	38
6	79	33
7	81	32
8	85	31

are presented as follows, the $initialize_variables()$ procedure generates a set of feasible solutions for the problem, the initial solutions are sorted by $non_domination_sort()$. For each iteration, the offspring solutions are generated by $genetic_operator()$ procedure. Then the solutions are sorted by $non_domination_sort()$ procedure. K members from the next non_selected level F_i are needed to meet the requirement of pop.

Algorithm 2 Framework of NSGA-II method

```
Data: Parameters for proposed problem.
Result: Pareto front solution set.
  pop (% Population size);
  gen (% Generation number);
  chromosome = initialize_variables();
  P = non_domination_sort(chromosome);
  for i = 1: gen do
     pool = gen/ 2 (% Matching pool size);
     tour = 2 (% Tournament size);
     parent_P = tournament_selection(P, pool, tour);
     offspring_P = genetic_operator(parent_P);
     intermediate_P = parent_P ∪ offspring_P;
     F_i = non_domination_sort(intermediate_P);
     P = ∅;
     i = 1 (%non_domination level index);
     while |P| + |F_i| ≤ pop do
        P = P ∪ F_i;
        i = i + 1;
     end while
     K = pop - |P| (% Number of individuals to be further added);
     P = P ∪ selection(P, F_i, K)
  end for
```

1) Solution representation: A solution S is composed of four parts: (i) the sequence of jobs processed in the first procedure on the primary machine, (ii) the starting time of jobs in the first procedure, (iii) the starting time of jobs in the second procedure, and (iv) the starting time of jobs in the third procedure. The above four sets of decision variables are all integral.

Observe that once the sequence of jobs and the starting time and completion time of jobs in different procedures are determined, the total energy consumption cost and makespan can be calculated, and the feasibility of the solution can be checked.

2) Population initialization: To ensure the effectiveness of the NSGA-II approach, the quality of initial solutions should be as high as possible, i.e. the initial solutions are all feasible. The framework of initial solutions generation is illustrated below.

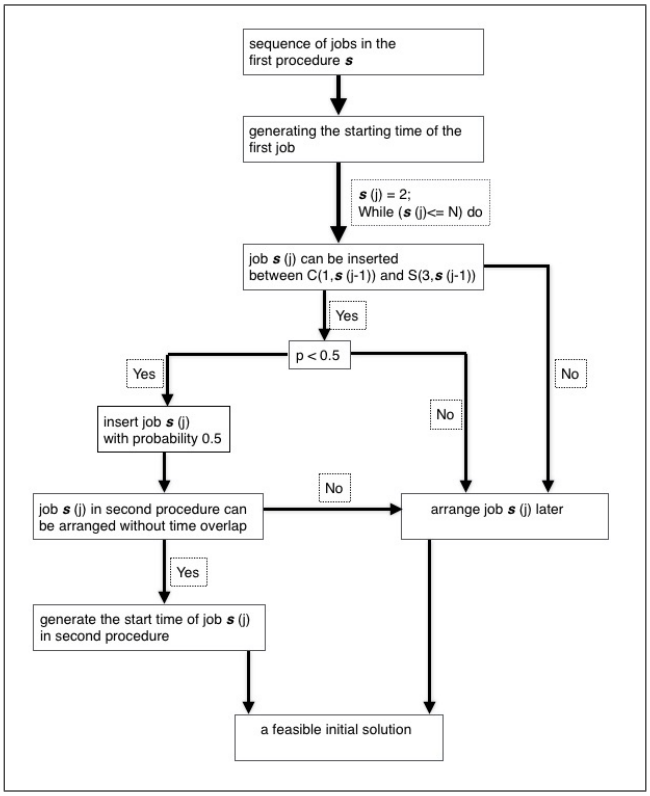


Fig. 3. Framework of initial solution generation

V. CONCLUSIONS

This paper investigates a bi-objective two-machine reentrant flow shop scheduling problem with an exact time lag, which aims to minimize the total energy consumption cost and makespan simultaneously. A bi-objective mathematical formulation is established to illustrate the problem. To obtain Pareto solutions of the problem, an exact ϵ-constraint method and NSGA-II approach are proposed.

In future research, we should pay more attention on (i) more efficient bi-objective algorithm and (ii) some other factors that may affect the objectives, such as the speed of machines and the salary of operators.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (NSFC) under Grants 71428002, 71432007, 71531011, 71571134 and 71571135. The work was supported by "the Fundamental Research Funds for the Central Universities".

REFERENCES

[1] A. Che, Y. Zeng, L. Ke, *An efficient greedy insertion heuristic for energy-conscious single machine scheduling problem under time-of-use electricity tariffs*. Journal of Cleaner Production, 2016, vol. 129, pp. 565-577.

[2] A. P. Rifai, H. T. Nguyen, S. Z. M. Dawal, *Multi-objective adaptive large neighborhood search for distributed reentrant permutation flow shop scheduling*. Applied Soft Computing, 2016, vol. 40, pp. 42-57.

[3] B. H. Zhou, L. M. Hu, Z. Y. Zhong, *A hybrid differential evolution algorithm with estimation of distribution algorithm for reentrant hybrid flow shop scheduling problem*. Neural Computing & Applications, 2016, pp.1-17.

[4] C. C. Wu, W. C. Lee, *Single-machine and flowshop scheduling with a general learning effect model*. Computers & Industrial Engineering, 2009, vol. 56, pp. 1553-1558.

[5] C. Desprez, F. Chu, C. Chu. *Minimising the weighted number of tardy jobs in a hybrid flow shop with genetic algorithm*. International Journal of Computer Integrated Manufacturing, 2009, vol. 22, pp. 745-757.

[6] C. H. Pan, J. S. Chen, *Mixed binary integer programming formulations for the reentrant job shop scheduling problem*. Computers & Operations Research, 2005, vol. 32, pp. 1197-1212.

[7] C. Jing, W. Huang, G. Tang, *Minimizing total completion time for re-entrant flow shop scheduling problems*. Theoretical Computer Science, 2011, vol. 412, pp. 6712-6719.

[8] C. Sangsawang, K. Sethanan, T. Fujimoto, M. Gen, *Metaheuristics optimization approaches for two-stage reentrant flexible flow shop with blocking constraint*. Expert Systems with Applications An International Journal, 2015, vol. 42, pp. 2395-2410.

[9] C. Jing, W. Huang, G. Tang. *Minimizing total completion time for re-entrant flow shop scheduling problems*. Theoretical Computer Science, 2011, vol. 412, pp. 6712-6719.

[10] D. Ahr, J. Békési, G. Galambos, M. Oswald, G. Reinelt. *An exact algorithm for scheduling identical coupled tasks*. Mathematical Methods of Operations Research, 2004, vol. 59, pp. 193-203.

[11] D. L. Yang, M. S. Chern, *Two-machine flowshop group scheduling problem*. Computers & Operations Research, 2000, vol. 27, pp. 975-985.

[12] D. L. Yang, C. J. Hsu, W. H. Kuo, *A two-machine flowshop scheduling problem with a separated maintenance constraint*. Computers & Operations Research, 2008, vol. 35, pp. 876-883.

[13] E. Demir, T. Bektaş, G. Laporte, *The bi-objective pollution-routing problem*. European Journal of Operational Research, 2014, vol. 232, pp. 464-478.

[14] F. Chu, C. Chu, C Desprez. *Series production in a basic re-entrant shop to minimize makespan or total flow time*. Computers & Industrial Engineering, 2010, vol. 58, pp. 257-268.

[15] F. Dugardin, F. Yalaoui F, L. Amodeo, *New multi-objective method to solve reentrant hybrid flow shop scheduling problem*. European Journal of Operational Research, 2010, vol. 203, pp. 22-31.

[16] F. Chu, C. Chu, C Desprez, *Series production in a basic re-entrant shop to minimize makespan or total flow time*. Computers & Industrial Engineering, 2010, vol. 58, pp. 257-268.

[17] F. Shrouf, J. Ordieres-Meré, A. García-Sánchez, M. Ortega-Mier, *Optimizing the production scheduling of a single machine to minimize total energy consumption costs*. Journal of Cleaner Production, 2014, vol. 67, pp. 197-207.

[18] Gilles Mouzon, Mehmet B. Yildirim, Janet Twomey. *Operational methods for minimization of energy consumption of manufacturing equipment*. International Journal of Production Research, 2007, vol. 45, pp. 4247-4271.

[19] Gilles Mouzon, Mehmet B. Yildirim. *A framework to minimise total energy consumption and total tardiness on a single machine*. International Journal of Sustainable Engineering, 2008, vol. 1, pp. 105-116.

[20] H. Hwang, J. U. Sun, *Production sequencing problem with re-entrant work flows and sequence dependent setup times*. International Journal of Production Research, 1998, vol. 33, pp. 2435-2450.

[21] H. Luo, B. Du, G. Q. Huang, H. Chen, X. Li, *Hybrid flow shop scheduling considering machine electricity consumption cost*. International Journal of Production Economics, 2013, vol. 146, pp. 423-439.

[22] J. F. Bérubé, M. Gendreau, J. Y. Potvin, *An exact ϵ-constraint method for bi-objective combinatorial optimization problems: Application to the Traveling Salesman Problem with Profits*. European Journal of Operational Research, 2011, vol. 194, pp. 39-50.

[23] J. Shen, L. Wang, J. Deng J, X. Zheng, *A Pareto-Based Discrete Harmony Search Algorithm for Bi-objective Reentrant Hybrid Flowshop Scheduling Problem*. Harmony Search Algorithm, 2016, vol. 382, pp. 435-445.

[24] J. Shen, L. Wang, H. Zheng, *A modified teachingClearning-based optimisation algorithm for bi-objective re-entrant hybrid flowshop scheduling*. International Journal of Production Research, 2015, vol. 54, pp. 1-18.

[25] J. Xu, Y. Yin, T. C. E. Cheng, C. C. Wu, S. Gu. *A memetic algorithm for the re-entrant permutation flowshop scheduling problem to minimize the makespan*. Applied Soft Computing, 2014, vol. 24, pp. 277-283.

[26] J. Y. Ding, S. Song, C. Wu, *Carbon-efficient scheduling of flow shops by multi-objective optimization*. European Journal of Operational Research, 2015, vol. 248, pp. 758-771.

[27] J. Zoghby, J. W. Barnes, J. J. Hasenbein, *Modeling the reentrant job shop scheduling problem with setups for metaheuristic searches*. European Journal of Operational Research, 2005, vol. 167, pp. 336-348.

[28] K. Amrouche, M. Boudhar, M. Bendraouche, F. Yalaoui, *Chain-reentrant shop with an exact time lag: new results*. International Journal of Production Research, 2016, vol. 55, pp. 1-11.

[29] K. C. Ying, S. W. Lin, S. Y. Wan, *Bi-objective reentrant hybrid flowshop scheduling: an iterated Pareto greedy algorithm*. International Journal of Production Research, 2014, vol. 52, pp. 5735-5747.

[30] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, *A fast and elitist multiobjective genetic algorithm: NSGA-II*. IEEE Transactions on Evolutionary Computation, 2002, vol. 6, pp. 182-197.

[31] M. Boudhar, K. Amrouche, *Two machines flow shop with reentrance and exact time lag*. Multidisciplinary International Scheduling Conference, 2013, vol. 50, pp. 223-232.

[32] N. Yalaoui, L. Amodeo, F. Yalaoui, H. Mahdi, *Efficient methods to schedule reentrant flowshop system*. Journal of Intelligent & Fuzzy Systems, 2014, vol. 26, pp. 1113-1121.

[33] R. H. Huang, S. C. Yu, C. W. Kuo. *Reentrant two-stage multiprocessor flow shop scheduling with due windows*. The International Journal of Advanced Manufacturing Technology, 2014, vol. 71, pp. 1263-1276.

[34] S. A. Mansouri, E. Aktas, U. Besikci, *Green scheduling of a two-machine flowshop: trade-off between makespan and energy consumption*. European Journal of Operational Research, 2016, vol. 248, pp. 772-788.

[35] S. Bertel, J. C. Billaut, *A genetic algorithm for an industrial multiprocessor flow shop scheduling problem with recirculation*. European Journal of Operational Research,2004, vol. 159, pp. 651-662.

[36] S. C. Graves, H. C. Meal, D. Stefek, A. H. Zeghmi, *Scheduling of re-entrant flow shops*. Journal of Operations Management. 1983, vol. 3, pp. 197-207.

[37] S. Wang, M. Liu, C. Chu, *A branch-and-bound algorithm for two-stage no-wait hybrid flow-shop scheduling*. International Journal of Production Research, 2015, vol. 53, pp. 1143-1167.

[38] S. Wang, M. Liu, C. Chu, *A branch-and-bound algorithm for two-stage no-wait hybrid flow-shop scheduling*. International Journal of Production Research, 2015, vol. 53, pp. 1143-1167.

[39] S. Wang, M. Liu, F. Chu, C. Chu, *Bi-objective optimization of a single machine batch scheduling problem with energy cost consideration*. Journal of Cleaner Production, 2016, vol. 137, pp. 1205-1215.

[40] S. W. Choi, Y. D. Kim, *Minimizing total tardiness on a two-machine re-entrant flowshop*. European Journal of Operational Research, 2009, vol. 199, pp. 375-384.

[41] V. Lev, I. Adiri, *V-shop scheduling*. European Journal of Operational Research, 1984, vol. 18, pp. 51-56.

[42] Y. D. Kim, *Minimizing total tardiness of orders with reentrant lots in a hybrid flowshop* . International Journal of Production Research, 2005, vol. 43, pp. 2149-2167.

[43] Z. Sun, L. Li, *Opportunity estimation for real-time energy control of sustainable manufacturing systems*. IEEE Transactions on Automation Science & Engineering, 2013, vol. 10, pp. 38-44.

Scheduling reentrant jobs on parallel machines with a remote server to minimize makespan, revisited

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Yang Sui†, Feifeng Zheng‡, Chengbin Chu*§
*School of Economics & Management, Tongji University, China
Email:mingliu@tongji.edu.cn
†Glorious Sun School of Business & Management, Donghua University, China
Email:suisuiyy2012@163.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email:ffzheng@dhu.edu.cn
*School of Economics & Management, Tongji University, China
§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
Email:chengbin.chu@centralesupelec.fr

***Abstract*—In this study, we revisit the problem of scheduling reentrant jobs on parallel primary machines with a remote server machine. For each job, it has three operations. The first operation of a job is performed on the primary machine, the second operation is processed on the single server machine and the last operation is also performed by the same primary machine, i.e, we consider reentrant job operations. The order of jobs is predetermined in our context. The objective is minimizing the makespan through assigning jobs onto the primary machines and splitting jobs into several batches. This problem was first proposed by Chakhlevitch and Glass (2009). We are the first to present a mixed integer programming model for the considered problem, and solve it by CPLEX for small instances, to the best of our knowledge. For large instances with more jobs, we propose a genetic algorithm which produces feasible solutions, in which the generation of initial chromosomes, batches are initialized by alternative machines.**

I. Introduction

In recent decades, there has been a great interest in the problem of scheduling parallel machines with a common server environment. A series of problems has been considered in the literature and obtained complexity results for different objective functions (Hall et al., 2000; Brucker et al., 2002; Kravchenko et al., 1997; Hasani et al., 2014). Abdekhodaee and Wirth (2002) presented an exact algorithm for special case, and proposed a metaheuristic algorithm to solve the general case in scheduling parallel machines with a single server problem. Abdekhodaee et al. (2006) addressed a greedy heuristic and a metaheuristic algorithm to minimize makespan with scheduling two-operation jobs on two parallel machines with a single server. Zhang et al. (2009) proposed two heuristics asymptotic competitive ratios in the on-line scheduling with two parallel identical machines sharing a single server. Huang et al. (2010) discussed the scheduling problem of parallel machines with sequence-dependent setup times and a remote server. They established the hybird genetic algorithm for the general case and showed that a special case can be solved in polynomial time. Kim and Lee (2012) developed a hybird heuristic algorithm for the identical parallel machines scheduling problem (PMSP) with a single server in charge of job setups. Gan et al. (2012) applied a column generation method to the problem of scheduling parallel machines with a single server. Jiang et al. (2013) addressed a preemptive scheduling problem on two parallel machines with a single server, in which each job had to be set up by the server before being processed on the machines. They presented an algorithm with a tight bound for the general case of the scheduling problem. Hasani et al. (2014) proposed a mixed integer linear programming formulation based on the decomposing a schedule into a set of blocks, for scheduling a set of non-preemptable jobs on two identical parallel machines. They considered the objective to minimize the forced idle time by scheduling jobs on two parallel machines with setup times.

Notice that the above articles studied the scheduling problem which a server visits the machines to perform a setup operation, they did not consider the reentrant operations of jobs when there is a server machine. In this paper, the setup operation is performed by a remote server after a job was processed on a primary machine, and finally back to the same primary machine for processing the main process of a job, implying the reentrant process. Similar problems are discussed by Huang et al. (2014). However, they did not involve the reentrant process. Wang et al. (1997) studied the problem of scheduling a chain-reentrant shop, in which each job went for first processing to a primary machine, and went to other machines in a fixed sequence, finally back to the primary machine for the last operation. Pan and Chen (2005) presented two binary integer programming optimization formulations for the reentrant job shop scheduling problem. Chen et al. (2008) developed a scheduling algorithm which included two major modules for the job shop by parallel machines and reentrant process. The scheduling problem studied in this paper was first discussed by Chakhlevitch and Glass (2009). They discussed a combinatorial problem relating to reentrant jobs on parallel primary machines with a remote server machine, which was inspired by the logistics of a semi-automated microbiology

laboratory. The laboratory performed microbiological testing on samples of processed food to ensure been safe to eat at the end of its shell-life. Each diluted test sample arrived to the pipetting stations by scanning the barcode label on the plastic bag of the test sample. Then the labelling operation collected the required number of test plates of different types in accordance with the test suit of a sample from the plate storage desk automatically, put a barcode label containing information about individual test onto each plate. The labelled plates were delivered to the corresponding pipetting station for pipet operation. Each pipetting station and the plate labelling machine can deal with only one job at any one time. The plate labelling machine processed incoming requests on the first-come-first-served basis. They referred to this problem as the initialization-setup-processing (ISP) problem, established the NP-hard nature of the problem and developed a heuristic algorithm called coordinated scheduling (CS). However, no mathematical model was presented. Wang et al. (2013) discussed the same scheduling problem by Chakhlevitch and Glass (2009). They developed genetic algorithm to minimize the makespan.

However, the initial populations in genetic algorithm often contain some solutions which have greater distance from optimal solution, so we present an improved heuristic algorithm to generate initial populations which are closer to optimal solution. Then the initial chromosomes can construct a high quality solution. This paper revisits the ISP problem. Though the CS has a good performance, the investigation shows that the optimal schedules obtained by the branch and bound algorithm often contains a few idle intervals on the server machine. On the other hand, the algorithm CS only produces schedules with either no idle time or a single idle interval on the server machine at the end of the schedule. Therefore, Chakhlevitch and Glass (2009) studied the schedule which generated by the algorithm CS can be improved by splitting large batches. In this paper, the objective is minimizing the latest completion time through assigning jobs onto parallel machines and splitting jobs into different batches. According to this, we establish a mixed integer programming model to obtain the objective, and solve the problem with small instances by CPLEX. For large instances, an improved genetic algorithm is designed in this paper. The remaining of this paper is organized as follows. In Section **II**, we describes the problem under consideration. Section **III** proposes the mathematical formulation. A optimal solution for a small instance of the model solved by CPLEX is provided in Section **IV**. In Section **V**, we present an improved genetic algorithm to solve the problem with large instances, in which the generation of initial chromosomes, batches are initialized by alternative machines. Finally, the conclusion and future work are given in Section **VI**.

II. Problem Statement and Mathematical Model

A. Problem Statement

There is a job shop with n jobs and three machines, in which two parallel machines called primary machines are identical and a single machine S servers the primary machines. Each job i consists of three operations. The first and third operations are performed on the same primary machine and the second operation is carried out on the server machine. The order of jobs and of each of their operations is predetermined in our context. The challenge is that of assigning jobs onto the primary machines and splitting jobs into different batches. The job can be processed on arbitrary parallel primary machine. The first operation called initialization requires a small amount of time δ; setup operation which performed by server machine S during P_2^i time units. The main processing operation which is the third operation is completed by the same machine that has initialized the job, needs P_3^i time units, implying the reentrant.

All the remaining of this paper is based on the following preliminary assumptions.

Assumptions:

(1) All the jobs can be processed by at most one machine at one time.
(2) Once an operation is started, another cannot be started on this machine until the operation completes.
(3) The order of jobs is predetermined. For example, for any pair of jobs i and j, if $i < j$, each operation of job i is performed before that of job j.
(4) The planning horizon is started with time 0.
(5) The processing time of initialization δ is not bigger than any setup time and main processing time, i.e., $\delta \leq \min_{i \in N}\{P_2^i\}$ and $\delta \leq \min_{i \in N}\{P_3^i\}$.
(6) Jobs can be processed in a consistent batch.

Because of the initialization time is small and the order of jobs is predetermined, it is possible to initialize a number of jobs in succession on the same machine, thus creating a batch of jobs. Batch processing means that the jobs with consecutive number can be processed on the primary machine, and the third operation of the first job in the batch can be started after all the first operations of jobs are completed. Another batch of jobs cannot be processed until all the operations of the current batch of jobs are completed on the same machine. In this paper, we focus on the scheduling problem with two primary machines. It is convenient to refer to the machines as A and B, respectively. The challenge is to assign the jobs onto the primary machines and to split jobs into different batches. There is always an optimal schedule for the two-machines ISP problem with the properties. Chakhlevitch and Glass (2009) discussed the optimized conditions which batches are initialized by alternative machines, and the property for each batch is that initialization, setup and main processing operations of jobs are sequenced in the same order.

B. Model formulation

Indices

i, j : indices of jobs;
k : index of primary machines.

Parameters

N : set of jobs, i.e. $N = \{1, 2, \ldots, n\}$;

K : set of primary machines, i.e. $K = \{1, \ldots, k\}$;
δ : the initialization time of jobs;
P_2^i : the setup time of job i by server machine, where $i \in N$;
P_3^i : the main processing time of job i, where $i \in N$;
M : a sufficiently large integer.

Decision Variables

I_i : the start time of initialization operation of job i, which the planning horizon is started with time 0, where $2 \leq i \leq n$;
S_i : the start time of setup operation of job i, where $i \in N$;
R_i : the start time of the main processing operation of job i, where $i \in N$;
C_{max} : the latest processing time of all jobs;
$x_{i,k}$: a binary variable equal to 1 if job i is assigned to machine k, 0 otherwise, where $i \in N$, $k \in K$;
β_i : a binary variable equal to 1 if job i is the first processed job in batch 2, 0 otherwise, where $2 \leq i \leq n$;
$\gamma_{i,j}$: a binary variable equal to 1 if job i and job j are the first processed jobs of contiguous batches on the same primary machine except batch 1, 0 otherwise. For example, the contiguous batches indicate the batch 2 and batch 4, where $2 \leq i \leq n-1$, $\{i < j\} \in N$;
α_i : a binary variable equal to 1 if the initialized completion time of job $i-1$, which job i is the first processed job of a batch, except batch 1 and batch 2, is not less than the main processing completion time of the previous job, which has been assigned to the same machine with job i, 0 otherwise. For example, job i is the first processed job in the batch 3, the previous job of job i indicate the last processed job in batch 1, where $3 \leq i \leq n$;
ϵ_i : a binary variable equal to 1 if the setup completion time of job $i-1$ is not less than the initialized completion time of job i, 0 otherwise, where $i \in N$ and $i \neq 1$;
ζ_i : a binary variable equal to 1 if the setup completion time of job i which is not the first processed job of a batch, when the batch has at least two or more jobs, is not less than the main processing completion time of job $i-1$, 0 otherwise. For example, a batch includes job 3, job 4 and job 5, job i is job 4 or job 5, where $i \in N$ and $i \neq 1$;
ν : a binary variable equal to 1 if the setup completion time of job 1 which batch 1 has at least two or more jobs, is not less than the initialized completion time of the last processed job of the batch 1, 0 otherwise. For example, batch 1 includes job 1, job 2 and job 3, the last processed job is job 3.
μ_i : a binary variable equal to 1 if the setup completion time of job i which is the first processed job in a batch that has at least two or more jobs, except batch 1 and batch 2, is not less than the initialized completion time of the last processed job of the batch, 0 otherwise. For example, a batch includes job 5, job 6 and job 7, the first and last processed jobs are job 5 and job 7, where $2 \leq i \leq n-1$;
$\xi_{i,j}$: a binary variable equal to 1 if job i and job j are the first and last processed jobs in the batch which includes the last indexed job, 0 otherwise. For example, a batch includes job $n-2$, job $n-1$ and job n, the first and last processed jobs are job n and job n, where $2 \leq i \leq n-1$, $j = n$;
ρ_i : a binary variable equal to 1 if the setup completion time of job i, which is the first processed job in a batch that includes the last indexed job and has at least two or more jobs, is not less than the initialized completion time of the last indexed job, 0 otherwise. For example, the last indexed job is job n, where $2 \leq i \leq n-1$.

Mathematical Model: The MIP model of the ISP problem can be formulated as follows.

$$\min C_{max}$$

$$I_i \geq I_{i-1} + \delta - M(2 - x_{i,k} - x_{i-1,k}) \quad 2 \leq i \leq n, k \in K \tag{1}$$

$$I_i \leq I_{i-1} + \delta + M(2 - x_{i,k} - x_{i-1,k}) \quad 2 \leq i \leq n, k \in K \tag{2}$$

$$\sum_{a=1}^{i} x_{a,k} \geq (i-1) - M(2 - x_{1,k} + x_{i,k} - \beta_i) \quad k \in K, 2 \leq i \leq n \tag{3}$$

$$\sum_{a=1}^{i} x_{a,k} \leq (i-1) + M(2 - x_{1,k} + x_{i,k} - \beta_i) \quad k \in K, 2 \leq i \leq n \tag{4}$$

$$\sum_{a=1}^{i} x_{a,k} \leq (i-2) + M(1 - x_{1,k} + x_{i,k} + \beta_i) \quad k \in K, 2 \leq i \leq n \tag{5}$$

$$\beta_i \leq M(2 - x_{1,k} - x_{i,k}), \quad k \in K, 2 \leq i \leq n \tag{6}$$

$$I_i \geq I_{i-1} + \delta - M(1 - \beta_i), \quad 2 \leq i \leq n \tag{7}$$

$$I_i \leq I_{i-1} + \delta + M(1 - \beta_i), \quad 2 \leq i \leq n \tag{8}$$

$$\sum_{a=i}^{j} x_{a,k} \geq (j-i) - M(2 - x_{i,k} + x_{i-1,k} + x_{j,k} - \gamma_{i,j}) \quad 2 \leq i \leq n-1, k \in K, \{j > i\} \in N \tag{9}$$

$$\sum_{a=i}^{j} x_{a,k} \leq (j-i) + M(2 - x_{i,k} + x_{i-1,k} + x_{j,k} - \gamma_{i,j}) \quad 2 \leq i \leq n-1, k \in K, \{j > i\} \in N \tag{10}$$

$$\sum_{a=i}^{j} x_{a,k} \leq (j-i-1) + M(1 - x_{i,k} + x_{i-1,k} + x_{j,k} + \gamma_{i,j}) \quad 2 \leq i \leq n-1, k \in K, \{j > i\} \in N \tag{11}$$

$$I_{j-1} + \delta \geq R_{i-1} + P_3^{i-1} - M(2 - \gamma_{i,j} - \alpha_j) \quad 2 \leq i \leq n-1, \{j > i\} \in N \tag{12}$$

$$I_{j-1} + \delta \leq R_{i-1} + P_3^{i-1} - 1 + M(2 - \gamma_{i,j} - \alpha_j)$$

$$2 \le i \le n-1, \{j > i\} \in N \quad (13)$$
$$I_j \ge I_{j-1} + \delta - M(2 - \gamma_{i,j} - \alpha_j)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (14)$$
$$I_j \le I_{j-1} + \delta + M(2 - \gamma_{i,j} - \alpha_j)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (15)$$
$$I_j \ge R_{i-1} + P_3^{i-1} - M(1 - \gamma_{i,j} + \alpha_j)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (16)$$
$$I_j \le R_{j-1} + P_3^{i-1} + M(1 - \gamma_{i,j} - \alpha_j)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (17)$$
$$S_{i-1} + P_2^{i-1} \ge I_i + \delta - M(1 - \epsilon_i), \quad 2 \le i \le n \quad (18)$$
$$S_{i-1} + P_2^{i-1} \le I_i + \delta - 1 + M\epsilon_i, \quad 2 \le i \le n \quad (19)$$
$$S_i \ge S_{i-1} + P_2^{i-1} - M(1 - \epsilon_i), \quad 2 \le i \le n \quad (20)$$
$$S_i \le S_{i-1} + P_2^{i-1} + M(1 - \epsilon_i), \quad 2 \le i \le n \quad (21)$$
$$S_i \ge I_i + \delta - M\epsilon_i, \quad 2 \le i \le n \quad (22)$$
$$S_i \le I_i + \delta + M\epsilon_i, \quad 2 \le i \le n \quad (23)$$
$$S_i + P_2^i \ge R_{i-1} + P_3^{i-1} - M(3 - x_{i,k} - x_{i-1,k} - \zeta_i)$$
$$2 \le i \le n, k \in K \quad (24)$$
$$S_i + P_2^i \le R_{i-1} + P_3^{i-1} - 1 + M(2 - x_{i,k} - x_{i-1,k} + \zeta_i)$$
$$2 \le i \le n, k \in K \quad (25)$$
$$R_i \ge S_i + P_2^i - M(3 - x_{i,k} - x_{i-1,k} - \zeta_i)$$
$$2 \le i \le n, k \in K \quad (26)$$
$$R_i \le S_i + P_2^i + M(3 - x_{i,k} - x_{i-1,k} - \zeta_i)$$
$$2 \le i \le n, k \in K \quad (27)$$
$$R_i \ge R_{i-1} + P_3^{i-1} - M(2 - x_{i,k} - x_{i-1,k} + \zeta_i)$$
$$2 \le i \le n, k \in K \quad (28)$$
$$R_i \le R_{i-1} + P_3^{i-1} + M(2 - x_{i,k} - x_{i-1,k} + \zeta_i)$$
$$2 \le i \le n, k \in K \quad (29)$$
$$R_1 \ge S_1 + P_2^1 - M(1 + x_{2,k} - x_{1,k}), \quad k \in K \quad (30)$$
$$R_1 \le S_1 + P_2^1 + M(1 + x_{2,k} - x_{1,k}), \quad k \in K \quad (31)$$
$$S_1 + P_2^1 \ge I_1 + (i-1) \times \delta - M(4 - x_{2,k} - x_{1,k} - \beta_i - \nu)$$
$$3 \le i \le n, k \in K \quad (32)$$
$$S_1 + P_2^1 \le I_1 + (i-1) \times \delta - 1 + M(3 - x_{2,k} - x_{1,k} - \beta_i + \nu)$$
0
$$3 \le i \le n, k \in K \quad (33)$$
$$R_1 \ge S_1 + P_2^1 - M(1 - \nu) \quad (34)$$
$$R_1 \le S_1 + P_2^1 + M(1 - \nu) \quad (35)$$
$$R_1 \ge I_1 + (i-1) \times \delta - M(1 - \beta_{1,i} + \nu)$$
$$3 \le i \le n \quad (36)$$
$$R_1 \le I_1 + (i-1) \times \delta + M(1 - \beta_{1,i} + \nu)$$
$$3 \le i \le n \quad (37)$$
$$R_i \ge S_i + P_i^2 - M(1 + x_{i-1,k} - x_{i,k} + x_{i+1,k})$$
$$2 \le i \le n-1, k \in K \quad (38)$$
$$R_i \le S_i + P_i^2 + M(1 + x_{i-1,k} - x_{i,k} + x_{i+1,k})$$
$$2 \le i \le n-1, k \in K \quad (39)$$
$$S_i + P_2^i \ge I_i + (j-i) \times \delta - M(2 - \gamma_{i,j} - \mu_i)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (40)$$
$$S_i + P_2^i \le I_i + (j-i) \times \delta - 1 + M(1 - \gamma_{i,j} + \mu_i)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (41)$$
$$\mu_i \le M \times (1 - \beta_i), \quad 2 \le i \le n-1 \quad (42)$$
$$R_i \ge S_i + P_2^i - M(1 - \mu_i), \quad 2 \le i \le n-1 \quad (43)$$
$$R_i \le S_i + P_2^i + M(1 - \mu_i), \quad 2 \le i \le n-1 \quad (44)$$
$$R_i \ge I_i + (j-i) \times \delta - M(1 - \gamma_{i,j} + \mu_i)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (45)$$
$$R_i \le I_i + (j-i) \times \delta + M(1 - \gamma_{i,j} + \mu_i)$$
$$2 \le i \le n-1, \{j > i\} \in N \quad (46)$$
$$R_n \ge S_n + P_2^n - M(1 - x_{n,k} + x_{n-1,k}), \quad k \in K \quad (47)$$
$$R_n \le S_n + P_2^n + M(1 - x_{n,k} + x_{n-1,k}), \quad k \in K \quad (48)$$
$$\sum_{a=i}^{n} x_{a,k} \ge (n - i + 1) - M(3 - \xi_{i,n} - x_{i,k} - x_{n,k} + x_{i-1,k})$$
$$2 \le i \le n-1, k \in K \quad (49)$$
$$\sum_{a=i}^{n} x_{a,k} \le (n - i + 1) + M(3 - \xi_{i,n} - x_{i,k} - x_{n,k} + x_{i-1,k})$$
$$2 \le i \le n-1, k \in K \quad (50)$$
$$\sum_{a=i}^{n} x_{a,k} \ge (n - i) + M(2 + \xi_{i,n} - x_{i,k} - x_{n,k} + x_{i-1,k})$$
$$2 \le i \le n-1, k \in K \quad (51)$$
$$S_i + P_2^i \ge I_i + (n - i + 1) \times \delta - M(2 - \rho_i - \xi_{i,n})$$
$$2 \le i \le n-1 \quad (52)$$
$$S_i + P_2^i \le I_i + (n - i + 1) \times \delta - 1 + M(1 + \rho_i - \xi_{i,n})$$
$$2 \le i \le n-1 \quad (53)$$
$$R_i \ge S_i + P_2^i - M(1 - \rho_i), \quad 2 \le i \le n-1 \quad (54)$$
$$R_i \le S_i + P_2^i + M(1 - \rho_i), \quad 2 \le i \le n-1 \quad (55)$$
$$R_i \ge I_i + (n - i + 1) \times \delta - M(1 + \rho_i - \xi_{i,n})$$
$$2 \le i \le n-1 \quad (56)$$

$$R_i \le I_i + (n-i+1) \times \delta + M(1+\rho_i - \xi_{i,n})$$

$$2 \le i \le n-1 \tag{57}$$

$$C_{max} \ge R_i + P_3^i, \quad i \in N \tag{58}$$

$$\sum_{k \in K} x_{i,k} = 1, \quad i \in N \tag{59}$$

$$S_1 = I_1 + P_2^1 \tag{60}$$

$$I_i, S_i, R_i, C_{max} \in Z^+, \quad 2 \le i \le n \tag{61}$$

$$\beta_i, \epsilon_i, \zeta_i \in \{0,1\}, \quad 2 \le i \le n \tag{62}$$

$$\mu_i, \rho_i \in \{0,1\}, \quad 2 \le i \le n-1 \tag{63}$$

$$x_{i,k} \in \{0,1\}, \quad i \in N, k \in K \tag{64}$$

$$\alpha_i \in \{0,1\}, \quad 3 \le i \le n \tag{65}$$

$$\gamma_{i,j} \in \{0,1\}, \quad 2 \le i \le n-1, \{i<j\} \in N \tag{66}$$

$$\xi_{i,n} \in \{0,1\}, \quad 2 \le i \le n-1 \tag{67}$$

The objective is minimizing the makespan. Constraint $(1)-(2)$ determine the initialized start time of job i, which is not the first processed job of a batch that has at least two or more jobs. Constraint $(3)-(6)$ determine the job which is the first processed job in batch 2. Constraint $(7)-(8)$ determine the initialized start time of the first processed job in batch 2. Constraints $(9)-(11)$ indicate the first processed jobs of the contiguous batches which not include job 1. Constraints $(12)-(13)$ make a comparison between the initialized completion time of the job $i-1$ and the main process completion time of the previous job. The job which has been assigned to the same machine with job i, that is the first processed job of a batch except batch 1 and batch 2. Constraint $(14)-(17)$ determine the initialized start time of job j, which is the first processed job of a batch except batch 1 and batch 2. Constraints $(18)-(19)$ make a comparison between the setup completion time of job $i-1$ and the initialized completion time for job i. Constraint $(20)-(23)$ determine the setup start time of job i except job 1. Constraint $(24)-(25)$ make a comparison between the setup completion time and the main process completion time of job $i-1$, which is not the first processed job in a batch that has at least two or more jobs. Constraint $(26)-(29)$ determine the processing start time of job i, which is not the first processed job in a batch that has two or more jobs at least. Constraint $(30)-(31)$ indicate the main processing start time of job 1 when a batch has only job 1. Constraint $(32)-(33)$ make a comparison between the setup completion time of job 1 and the initialized completion time of the last processed job of the batch, which has at least two or more jobs and includes job 1. Constraint $(34)-(37)$ determine the main processing start time of job 1 in batch 1, which has at least two or more jobs. Constraint $(38)-(39)$ determine the main processing start time of the first processed job in a batch, which includes only one job except job 1. Constraint $(40)-(42)$ make a comparison between the setup completion time of the first processed job in a batch except batch 1 and batch 2, and the initialized completion time of the last processed job in the batch which has at least two or more jobs. Constraint $(43)-(46)$ determine the main processing start time of the first processed job in a batch except batch 1 and batch 2. Constraints $(47)-(48)$ determine the main processing start time of job n which is assigned to a batch only itself. Constraint $(49)-(51)$ determine the number of jobs in a batch, which includes job n and has at least two or more jobs. Constraint $(52)-(53)$ make a comparison between the setup completion time of the first processed job in a batch, which includes job n and has at least two or more jobs, and the initialized completion time of job n. Constraint $(54)-(57)$ determine the main processing start time of the first processed job in a batch, which includes job n and has at least two or more jobs. Constraints (58) indicates the makespan is not less than the main processing completion time of any jobs. Constraint (59) ensures a job can only be assigned on one machine. Constraint (60) determines the setup start time of job 1.

III. Optimal Solutions for small instances

We first consider a small instance where there are ten jobs to be processed by two primary machines and a single server machine. The processing time of initialization stage is randomly generated. The processing time of setup stage and reentrant stage is randomly generated and not less than the initialized processing time. The processing time of three operations with ten jobs is as follows in Table I.

TABLE I
INPUT DATA OF THE PROCESSING TIME OF THREE OPERATIONS WITH TEN JOBS

Job	1	2	3	4	5	6	7	8	9	10
δ	2	2	2	2	2	2	2	2	2	2
P_2^i	5	3	2	4	5	2	3	5	4	3
P_3^i	5	4	6	7	6	7	4	5	8	5

Figure 1 illustrates the optimal solution for the ISP problem. In the figure, ten jobs are assigned into three batches respectively. The start time of three operations with ten jobs is followed in the Table II.

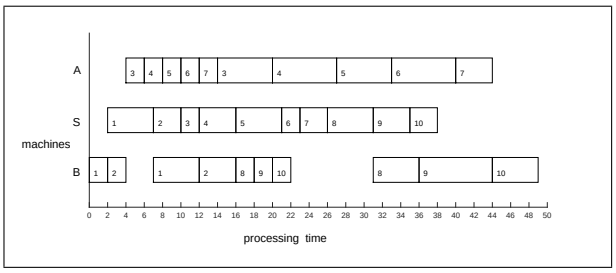


Fig. 1. An optimal solution for ten jobs

IV. Genetic algorithm

In this section, we develop an improved GA algorithm to resolve the considered problem. As is described in the context,

TABLE II
OUTPUT DATA OF THE START TIME OF THREE OPERATIONS WITH TEN JOBS

Job	1	2	3	4	5	6	7	8	9	10
I_i	0	2	4	6	8	10	12	16	18	20
S_i	2	7	10	12	16	21	23	26	31	35
R_i	7	12	14	20	27	33	40	31	36	44

the challenge of this paper is assigning jobs onto different primary machines and jobs are split into several batches. The chromosome is represented by assigning jobs onto different primary machines. They are generated randomly. For example, the number one job is assigned onto primary machine A marks 1, and the number two job is assigned onto primary machine B marks 2 (see Fig.2).

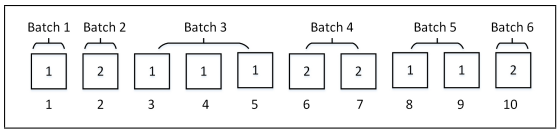


Fig. 2. Chromosome representation

In the generation of initial chromosomes, batches are initialized by alternative machines. We design a heuristic algorithm to generate initial populations which are closer to the optimal solution. Then the initial chromosomes construct a high quality solution. The objective of the considered problem aims at minimizing the makespan. We adopt the corresponding objective value to indicate the fitness of a solution. Two chromosomes are selected to act as parents and generate new individuals for crossover by exchanging some genes of the two chromosomes. The crossover genes are selected randomly. Mutation operation is performed with prespecified rate. A computational experiment need less time to obtain optimal solutions and improve the optimal makespan through the improved GA algorithm.

V. CONCLUSION

This paper considers the reentrant jobs scheduling for ISP problem. The challenge is that assigning jobs onto different primary machines and splitting jobs into servals batches. The objective is minimizing the makespan. According to this problem, we establish a mixed integer programming model and obtain the objective for small instances of the model by CPLEX. We propose an improved genetic algorithm for different scales of instances, in which the generation of initial chromosomes, batches are initialized by alternative machines. We design a heuristic algorithm to generate initial populations which are closer to the optimal solution. Then the initial chromosomes construct a high quality solution. Further research directions may include (1) the order of jobs is unknown; (2) the non-zero release time of jobs is given; (3) development of more efficient algorithms, such as Lagrangian relaxation algorithm.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China (71432007, 71531011, 71571134, 71571135, and 71428002), and the Cai Yuanpei Program between the French Ministries of Foreign and European Affairs and the Higher Education and Research and the Chinese Ministry of Education (27927VE). The work was also supported by the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program.

REFERENCES

[1] A. H. Abdekhodaee, A. Wirth, *Scheduling parallel machines with a single server: some solvable cases and heuristics*. Computers and Operations Research, 29(3), 295-315, 2002.
[2] A. H. Abdekhodaee, A. Wirth, H. S. Gan, *Scheduling two parallel machines with a single server: the general case*. Computers and Operations Research, 33(4), 994-1009, 2006.
[3] P. Brucker, C. Dhaenens-Flipo, S. Knust, S. A. Kravchenko, F. Werner, *Complexity results for parallel machine problems with a single server*. Journal of Scheduling, 5(6), 429-457, 2002.
[4] K. Chakhlevitch, C. A. Glass, *Scheduling reentrant jobs on parallel machines with a remote server*. Computers and Operations Research, 36(9), 2580-2589, 2009.
[5] J. C. Chen, K. H. Chen, J. J. Wu, C. W. Chen, *A study of the flexible job shop scheduling problem with parallel machines and reentrant process*. The International Journal of Advanced Manufacturing Technology, 39(3), 344-354, 2008.
[6] H. S. Gan, A. Wirth, A. Abdekhodaee, *A branch-and-price algorithm for the general case of scheduling parallel machines with a single server*. Computers and Operations Research, 39(9), 2242-2247, 2012.
[7] N. G. Hall, C. N. Potts, C. Sriskandarajah, *Parallel machine scheduling with a common server*. Discrete Applied Mathematics, 102(3), 223-243, 2000.
[8] K. Hasani, S. A. Kravchenko, F. Werner, *Block models for scheduling jobs on two parallel machines with a single server*. Computers and Operations Research, 41(1), 94-97, 2014.
[9] S. M. Huang, L. N. Cai, X. Y. Zhang, *Parallel dedicated machine scheduling problem with sequence-dependent setups and a single server*. Computers and Industrial Engineering, 58(1), 165-174, 2010.
[10] R. H. Huang, S. C, Yu, C. W. Kuo, *Reentrant two-stage multiprocessor flow shop scheduling with due windows*. The International Journal of Advanced Manufacturing Technology, 71(5), 1263-1276, 2014.
[11] Y. Jiang, J. Dong, M. Ji, *Preemptive scheduling on two parallel machines with a single server*. Computers and Industrial Engineering, 66(2), 514-518, 2013.
[12] M. Y. Kim, Y. H. Lee, *MIP models and hybrid algorithm for minimizing the makespan of parallel machines scheduling problem with a single server*. Computers and Operations Research, 39(11), 2457-2468, 2012.
[13] S. A. Kravchenko, F. Werner, *Parallel machine scheduling problems with a single server*. Mathematical and Computer Modelling An International Journal, 26(12), 1-11, 1997.
[14] C. H. Pan, J. S. Chen, *Mixed binary integer programming formulations for the reentrant job shop scheduling problem*. Computers and Operations Research, 32(5), 1197-1212, 2005.
[15] M. Y. Wang, S. P. Sethi, S. L. V. D. Velde, *Minimizing Makespan in a Class of Reentrant Shops*. Operations Research, 45(5), 702-712, 1997.
[16] H. Wang, H. J. Li, Y. Zhao, D. Lin, J. W. Li, *Genetic Algorithm for Scheduling Reentrant Jobs on Parallel Machines with a Remote Server*. Transactions of Tianjin University, 19(6), 463-469, 2013.
[17] L. Zhang, A. Wirth, *On-line scheduling of two parallel machines with a single server*. Computers and Operations Research, 36(36), 1529-1553, 2009.

A Prelude to Anarchic Manufacturing Systems

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Aydin Nassehi
Department of Mechanical Engineering,
University of Bristol,
Bristol, BS8 1TR, United Kingdom
Email: aydin.nassehi@bristol.ac.uk

Andrew Ma
Department of Mechanical Engineering,
University of Bath,
Bath, BA2 7AY, United Kingdom
Email:andrew.ma62@googlemail.com

***Abstract*—Industrie 4.0 envisages smart factories with autonomous manufacturing units and products that communicate with one another wirelessly to fulfil production goals. This is, in part, the reaction to modern customer requirements for a wider variety of products and reduced batch sizes, forcing manufacturers to become flexible and adaptable to continuously changing market demands. In order to take advantage of the enhancements offered by smart factories, new approaches to production planning and scheduling would be necessary as the traditional hierarchical methods are too rigid. In the research reported in this paper, the distributed heterarchical approach to scheduling which is generally coupled with centralised monitoring, is taken further to construct a planning framework with no central monitoring or control. This framework, dubbed the "anarchic" manufacturing system is comprised of a loose collection of machines as service providers and jobs as service consumers with a local negotiation mechanism. It is shown that despite having no central control, and only through the use of local variables, the desired behaviour of effective production throughput emerges in the system as the result of interaction between the individual elements. This provides a novel foundation for production planning in future connected smart factories.**

***Keywords*—Distributed systems, Multi agent systems, Simulation**

I. Introduction

The global manufacturing shift from mass production to mass customisation is driven by demand for increased product variants and smaller batches of high product complexity [1]. This forces manufacturers to be flexible, dynamic and responsive to frequently changing customer demands for quality and cost-effectiveness [2]. Modern manufacturing facilities increasingly deal with dynamic and volatile environments [3]; particularly prevalent for high variety manufacturers such as job shops and flow shops. Flexibility is becoming a key competitive priority [4]; enabling a manufacturer to deal with uncertain environments.

The manufacturing scheduling problem, particularly for manufacturers with high product turnover, is increasingly significant and researched [5]. Hierarchical methods are traditionally used for optimal scheduling; however, these are based on static well-defined problems and adapt poorly to real-time disruptions.

Emergent synthesis methods, where the aggregate effect of many elements pursuing their individual goals creates a global emergent direction, are identified to best solve problems with incomplete knowledge areas [6]. Multi-agent distributed systems with a flat heterarchical structure are underpinned by the emergent synthesis concept. Autonomous and flexible multi-agent systems have been demonstrated to successfully solve complex problems [5]. Scholz-Reiter et al. state the next research steps should evaluate distributed systems in general, unforeseen disruptions and create general problem formulations such as job shop problems [3]. This paper aims to explore an extension of the distributed system by removing central control and monitoring from the factory architecture. In this context, "anarchy" is defined as a distributed heterarchical structure, where decision making authority and autonomy is at the lowest level. The proposed structure is then tested against benchmark job shop scheduling problems [7] to determine its effectiveness.

II. Background

Flexible Manufacturing Systems (FMS) and Reconfigurable Manufacturing Systems (RMS) defined paradigms for introducing increasing flexibility in manufacturing [8], [9] with the emerging cloud based manufacturing (CBM) continuing the trend [10]. In an FMS, mechanisation and automation of interdependent and interactive elements improve a machine tool's (MT) utilisation [2]. Typically MTs are autonomously fed jobs using pallets, transported by an autonomous rail guided vehicle. High software complexity, lack of reconfigurability and high investment and maintenance costs are attributed to the low FMS uptake [9].

RMS have reconfigurable and modular elements, these are rapidly re-integrated reliably with low ramp-up time. Koren compares dedicated manufacturing lines, FMS and RMS identifying their relative limitations and benefits. The necessary enabling technologies and future improvements for RMS are discussed; most significantly integration of reconfigurable system design coupled with open-architecture reconfigurable controllers [8]. Advances in interrelated technologies and design for reconfigurability are required to realise RMS; enabling manufacturers to achieve the new economic objective of manufacturing responsiveness. Job shop manufacturing systems are usually small and manufacture customised one-off or small batches of jobs. Jobs are not necessarily constrained to individual MTs; job shops typically have high flexibility.

CBM is a service oriented and customer focused demand driven manufacturing model, where customers have on-demand access to a shared collection of diversified and distributed resources. This vision utilises distributed resources to create reconfigurable and temporary production systems

to improve efficiencies and resource utilisation, by assigning manufacturing jobs and operations to underutilised resources regardless of their geographical location [10].

Allwood et al. observe that increasing variety in a manufacturers product mix results in a decrease in productivity, and the predominate response to managing complexity has been to design products in 'families' or 'platforms' [11]. During their study, they found increased variety dramatically increased production time due to conflicting demands.

There is also a paradigm shift towards decentralised control of 'intelligent' items, from centralised control in hierarchical structures [12]. Distributed intelligent systems have autonomous control, using a flat heterarchical structure, where decision making authority is distributed locally [13]. The multi-agent 'society' is an example where autonomous agents live and interact with each other aiming to solve a given problem [14]. Intelligent agents are capable of flexible autonomous action to meet their design objectives, whilst having social ability, reactivity and pro-activeness [15].

Scholz-Retier et al. compares the performance of autonomous control methods against scheduling heuristics using a flexible flow-shop scenario. They found centralised production planning is superior for static scenarios regardless of their complexity; however, multi-stage and dynamic environments favour autonomously controlled systems. They conclude central strategies try to satisfy all constraints at once, often over constraining the system in changing environments; however, distributed and autonomous strategies divide the problem into sub-problems and successively solve these [3].

This paper focuses on flat autonomous intelligent multi-agent systems for job shop scenarios, and incorporates a free market bidding and negotiation architecture that has a low level cost factor negotiation mechanism variant. The physical implementation of such a system is shown in Fig 1 where resources (human and machine) work together without a predefined structure to maximise flexibility. It aims to characterise the inherent flexibility that distributed intelligent agents offer against a centralised scheduling system. The individual agents do not know the system's overall aim, and follow their individual local goals to achieve a global objective; emergent synthesis underpins the system structure to solve unforeseen or unquantifiable problems.

A. Modelling Platform

For this study a multi-agent system with integrated discrete-event processes within agents is required to model the behaviour of individual agents (e.g. MTs, jobs). This can be achieved through programming intelligent agents using a general programming language such as Java [18]; however, a software platform eliminates the need to create agents from scratch. AnyLogic 7 was chosen as a suitable multi-method modelling platform that can incorporate multi-agent, system dynamics and discrete event modelling [19]. Synchronous vs asynchronous communication, deadlocks and simultaneous event firing are issues faced during multi-agent modelling, the way that AnyLogic 7 processes commands has been incorporated and adjusted for to obtain expected model functionality.

III. Agent Based Model of an Anarchic Manufacturing System

Agent based models use a population of agents that interact with each other; each agent has predetermined behaviour patterns and decision making logic. Agents are used to represent elements of a system, such as jobs, MTs, operators or transporters. Agents predominately use discrete event state charts to model behaviour and actions; additionally, agents can interact with each other by sending messages and observe and change each other's parameters. More advanced agent based models have non-deterministic behaviour, and can simulate complex scenarios. Fig. 2 shows the graphical representation of two basic agent sets that are defined to represent jobs (J) and MTs (M). The machines are shown as rectangular boxes and jobs as cylinders that change colour according to their status. Communicating elements are shown by a solid line between the communicating agents.

An important aspect of agent based models is the asynchronous nature of messages sent between the agents. This would mean that there is no guarantee that if messages are sent in a sequence, they would necessarily be received in the same sequence. The synchronisation issues that could ensue are well recognised as a potential challenge in agent based modelling of distributed systems. In the framework of this paper, a set of *operation handlers* O has been defined to avoid simultaneous event firing issues and clashes. The anarchic manufacturing system S is thus represented by the ternary (J, M, O).

A. Job agents

In the distributed system each job is defined as an agent that seeks to complete a sequence of operations. At each stage, jobs only plan their next operation, rather than a complete route through the manufacturing system for every operation. Assuming that the planning is concerned with n jobs, the job set is composed as:

$$J = \{j_1, j_2, \ldots, j_n\} \tag{1}$$

with j_i representing job i. The individual jobs are given budgets and tasked to complete their production sequence whilst minimising the amount that they spend against their

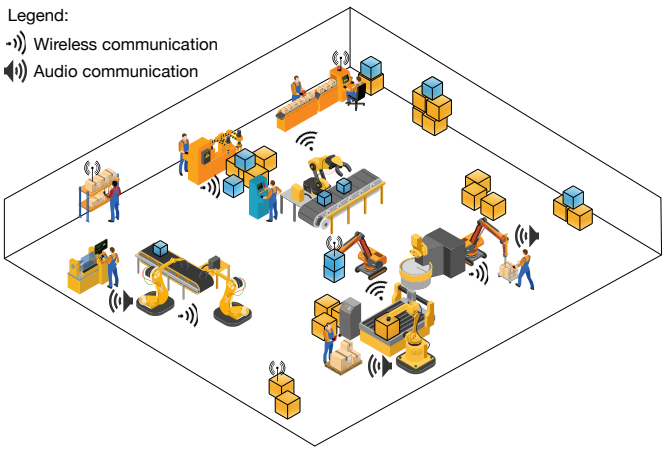


Fig. 1. An anarchic manufacturing system

budget. In the proposed agent based model, jobs may have multiple budgets; for example, operational, emissions and environmental. Operational budget and emissions budget are associated with an operational cost, and environment budget is associated with time in system; it is continuously taken away from a job's environment budget. γ_{ijt} denotes the remaining budget in category j for job i at time t. Jobs can only connect and communicate to MTs and jobs within a specified range; this is used to limit negotiations. The communication range for job i is denoted by ϕ_i. ψ_{it} denotes the remaining operations for job i at time t. To simplify, the job allocates the same budget to remaining operations at any given time. So if λ_{ijt} is the allocated budget for each operation by job i in category j at time t, the following equation relates it to the overall budget:

$$\lambda_{ijt} = \frac{\gamma_{ijt}}{\psi_{it}} \tag{2}$$

The state chart governing the behaviour of job agents is shown in Fig. 3. When a job enters the system it starts the process by evaluating the possible operations and then tries to join the queues for machines. If the machine costs are too high, it reconsiders the operations and tries to find machines anew. When the job is allocated to a machine, it waits in the queue until the operation is carried out. If all operations are complete the job leaves the system, otherwise it repeats the process. Whilst the job is waiting for assignment, some budget categories are increased regularly.

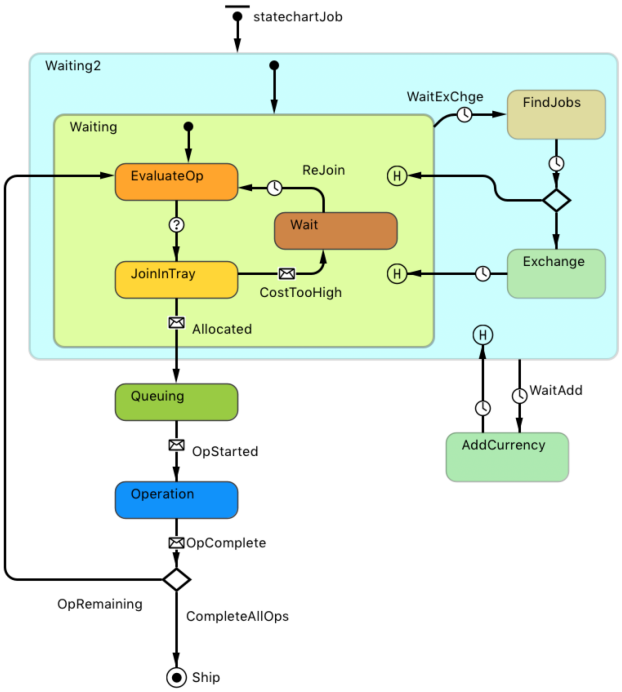


Fig. 3. The state chart for the job agent

B. Machine tool agents

MTs are defined as agents that provide production services to job agents on a chargeable basis. This is to allow machines to perform tasks that are most valuable to the company (and hence reflected to the budget that is assigned to each job). With the assumption that there are k machines in the system, the set of machines is defined as:

$$M = \{m_1, m_2, \ldots, m_k\} \tag{3}$$

The range of communication for machine i is denoted by Φ_i. The cost of processing on machine i in budget category j at time t is denoted by β_{ijt}.

Fig. 4 shows the two state charts that govern machine behaviour. The chart in Fig. 4(a) governs the bidding behaviour of the machine where it evaluates its costs when a call for bids is received, sends the bid and re-evaluates costs and rebids until the bidding is finished. The chart in Fig. 4(b) models the machine's working status. In this implementation, machine failures are not modelled for simplicity.

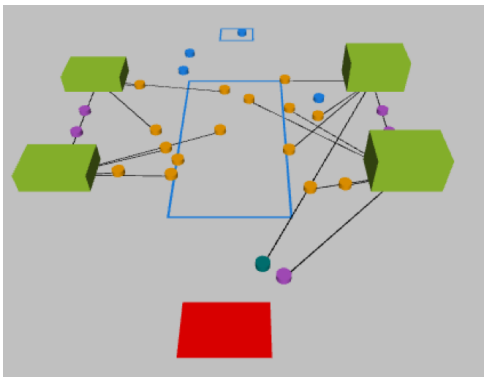


Fig. 2. An agent based representation of the anarchic manufacturing system

C. Operation handler agents

The operation handler agents are modelled for two purposes: (i) to model synchronised aspects of communication between the agents in an unsynchronised environment; and, (ii) to have the ability to assign message priorities when communication bandwidth is limited. The state chart governing these agents is shown in Fig. 5. The agent ensures that the bids are received and evaluated at each round and resets the environment in anticipation of a new round of bidding.

IV. Agent Negotiation Framework

The negotiation mechanisms are a combination and adaptation of the contract net protocol with cost factor adaptation [14], free market architecture for distributed systems [20] with

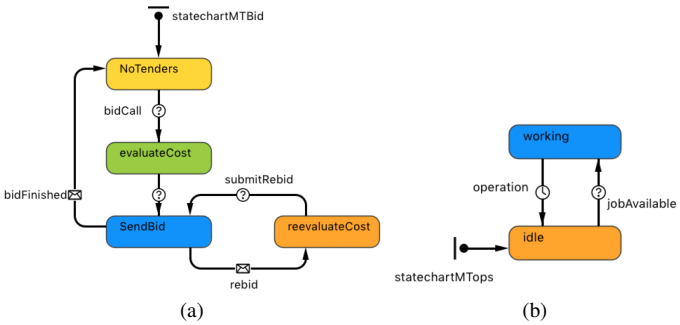


Fig. 4. The state chart for the machine tool agent

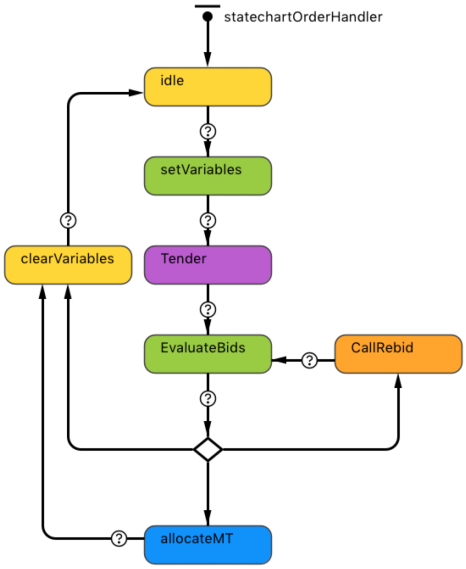


Fig. 5. The state chart for the operation handler agent

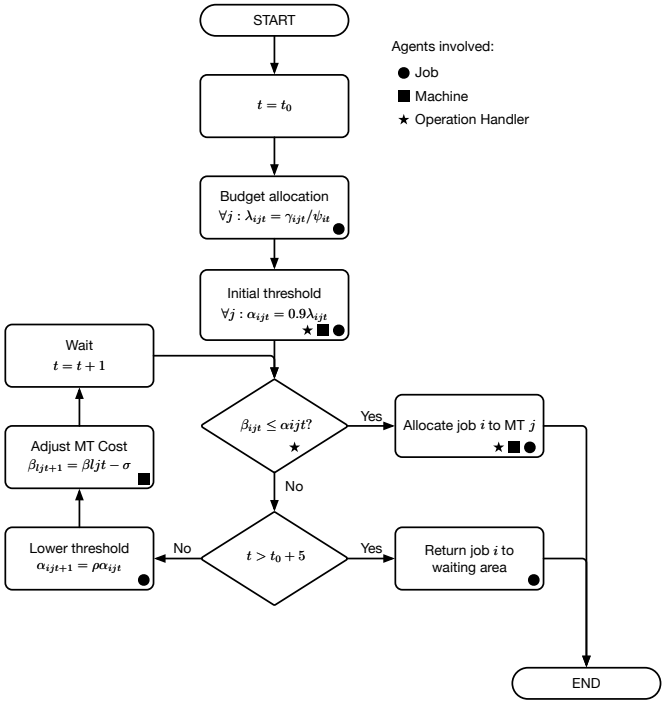


Fig. 6. The state chart for the operation handler agent

local information methods [3]; whilst being underpinned by emergent synthesis where local agents pursue their objectives to globally solve unclear problems [6]. The negotiation system has a free market structure and is a tool to obtain an improved distributed scheduling method; low level individual negotiations use a variant of the contract net protocol with cost factor adaptation. The system aims to take advantage of flexibility and high adaptability traits in free markets to gain a highly productive society created from pursuing individual interests [20].

Initially each job is prepared to pay a pre-set fraction of the allocated budget for each operation to a machine tool. MTs fluctuate their costs according to their utilisation and bid success within certain bounds, this is a permutation of Kádár's negotiation method [14]. The part is assigned to the machine if the machine cost is below the threshold that the part sets for the operation. The part reconsiders its bid if it is unsuccessful in bidding and gives up after a certain number of tries restarting the process with other machines. Fig.6 shows the negotiation framework in the form of a flowchart where α is the job threshold, n the bid round, ρ a risk factor that changes over time, γ the overall budget, λ the operation budget, ψ number of operations remaining, β the machine tool cost and σ the machine tool cost reduction.. To ensure that the flow takes place in sequence, the operation handler agent maintains the order. In practice, there are several potential solutions (including the use of timecodes, electronic kanbans and so on) to ensure the execution order in a distributed manner.

V. Simulation Results and Discussion

The framework presented in previous sections is implemented in the Anylogic modelling and multi-method simulation environment. The benchmark problems presented by Taillard [7] were used as the framework for the simulation model with the total makespan as the basis for comparison.

The negotiation method in the multi agent framework was adapted to use a single currency, the initial bid from each machine is calculated by:

$$\beta_0 = \frac{\omega}{2} + \frac{\kappa}{2} \tag{4}$$

where ω is the utilisation of the machine and κ is the duration of the operation being considered. The cost reduction factor is selected as:

$$\sigma = \frac{2}{\nu} + 1 \tag{5}$$

where ν is the historical success rate of the machine tool in the bidding process. ϕ and Φ are chosen as $15m$ in the model space which corresponds to about 25% coverage as the range of the wireless signal between the agents. This means that effectively only agents that are within this distance of each other can communicate and participate in bidding with each other.

Four scenarios were run from the Taillard benchmarks as shown in Table I and the overall makespan was recorded. In addition, a simple FIFO scheduling system was set up as the basis for comparison with the anarchic system.

Fig. 7 shows a run of the anarchic manufacturing system simulation where some jobs are complete, some are waiting and bidding for machines and some are being processed. The

TABLE I. Benchmark scenarios for the job shop

Scenario	Jobs	Machine Tools
A	15	15
B	30	15
C	50	20
D	50	15

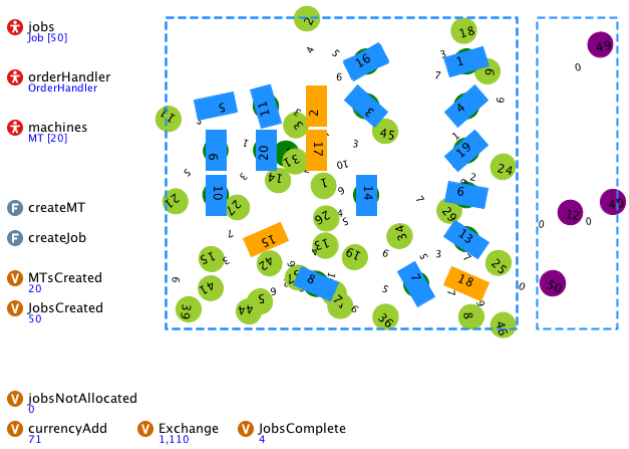


Fig. 7. Simulation of an anarchic job shop

jobs are shown by circles and the machines by rectangles. The number in the circle is the job number and the number lower left of the circle is the number of operations remaining for the job to be complete. The machine number is shown in the rectangles representing the machines.

The results for the four scenarios are reported in tables 2 to 5 below. For each scenario the makespan for the anarchic system is compared to the simple FIFO scheduling and the theoretical bound as reported by Taillard. Overall the results show that the anarchic system achieves a makespan about 6% shorter than a simple FIFO rule but about 25% longer than the lower bound established by Taillard.

TABLE II. SCENARIO A (15 JOBS, 15 MTS)

Run	Makespan (Anarchic)	Makespan (FIFO)	Lower Bound
1	1609	1709	1231
2	1513	1583	1252
3	1498	1583	1223
4	1525	1726	1181
5	1512	1843	1234
6	1568	1659	1243
7	1635	1465	1228
8	1483	1472	1221
9	1591	1632	1289
10	1545	1740	1261

Average improvement over FIFO = 5%
Average above bound= 25%

VI. CONCLUSIONS AND FUTURE WORK

In this paper the notion of distributed manufacturing systems was modified to remove central or system-wide communications and control and the feasibility of creating a functional manufacturing system solely based on local communications was explored. The proposed 'anarchic' framework was defined using intelligent agents and the system was defined by specifying the state charts that govern agent behaviour. The system was then implemented on an agent based simulation platform and a set of established benchmarks for job shop scheduling

TABLE III. SCENARIO B (30 JOBS, 15 MTS)

Run	Makespan (Anarchic)	Makespan (FIFO)	Lower Bound
1	2296	2678	1770
2	2328	2566	1853
3	2490	2848	1855
4	2487	2392	1851
5	2392	2464	2007
6	2407	2640	1844
7	2406	2343	1822
8	2179	2470	1714
9	2391	2552	1824
10	2116	2388	1723

Average improvement over FIFO = 7%
Average above bound= 29%

TABLE IV. SCENARIO C (50 JOBS, 20 MTS)

Run	Makespan (Anarchic)	Makespan (FIFO)	Lower Bound
1	3588	3978	2921
2	3501	3732	3002
3	3396	3593	2835
4	3320	3723	2775
5	3408	3568	2800
6	3584	3619	2914
7	3641	3720	2895
8	3309	3682	2835
9	3681	4033	3097
10	3710	3775	3075

Average improvement over FIFO = 6%
Average above bound= 21%

TABLE V. SCENARIO D (50 JOBS, 15 MTS)

Run	Makespan (Anarchic)	Makespan (FIFO)	Lower Bound
1	3456	3444	2760
2	3501	3472	2756
3	3085	3300	2717
4	3341	3364	2839
5	3417	3424	2689
6	3332	3546	2781
7	3418	3626	2943
8	3429	3681	2885
9	3430	3679	2655
10	3271	3553	2723

Average improvement over FIFO = 4%
Average above bound= 21%

was used to assess the performance through comparing the achieved overall makespan of the jobs. The system showed that it performs better than a rule of thumb based system (First In First Out) but with limited communication, the performance achieved is significantly lower than the established bounds.

It should, however, be noted that the Taillard problems used for benchmarking have fixed processing times and exclude variables such set-up times, due dates or release dates. This 'static' test scenario is useful to compare relative performance of the distributed anarchic negotiation method. The scenario, however, has its limitations in real world application, where in a more dynamic environment release dates, due dates and disruptions occur. The anarchic system is suited to the dynamic

environment, as the system has minimal forward planning and proceeds through a scenario one step at a time.

To continue the work in this paper, the authors will thus consider including such aspects of a real-world industrial system in the implementation. The anarchic framework allows multiple budget currencies to be defined and used by the agents, these may be used to model emissions, energy expenditure and use of rare resources in production. This would allow the production system to dynamically adapt its behaviour according to the rarity of resources or emerging legislation. Such multiple currency models will be further studied and compared to traditional multi-objective optimisation methods in production planning. Another potentially powerful mechanism that will be introduced together with multiple currencies, is giving the job agents the capability to trade their budgets with each other (similar to the environmental cap and trade policies); this leads to dynamically self-optimising resource concious manufacturing systems.

In addition, the authors will test the sensitivity of performance against changes in system parameters and compare this to traditional systems. An interesting venue for research would be to use a formal definition of flexibility and resilience against external disturbances to compare the anarchic system with traditional centralised and distributed production systems as any benefits of not relying on system wide communication should start to become important in such circumstance.

REFERENCES

[1] Windt, K., Böse, F. and Philipp, T., 2008. Autonomy in production logistics: Identification, characterisation and application. Robotics and Computer-Integrated Manufacturing, 24(4), pp.572-578.

[2] Boer, H., 1994. New Wave Manufacturing Strategies: Organizational and Human Resource Management Dimensions, Chapter 5, pp. 80-103.

[3] Scholz-Reiter, B., Rekersbrink, H. and Görges, M., 2010. Dynamic flexible flow shop problemsScheduling heuristics vs. autonomous control. CIRP Annals-Manufacturing Technology, 59(1), pp.465-468.

[4] Wang, R.C. and Chuu, S.J., 2004. Group decision-making using a fuzzy linguistic approach for evaluating the flexibility in a manufacturing system. European Journal of Operational Research, 154(3), pp.563-572.

[5] Ouelhadj, D. 2003. A multi-agent system for the integrated dynamic scheduling of steel production. Diss. The University of Nottingham.

[6] Ueda, K., Markus, A., Monostori, L., Kals, H.J.J. and Arai, T., 2001. Emergent synthesis methodologies for manufacturing. CIRP Annals-Manufacturing Technology, 50(2), pp.535-551.

[7] Taillard, E., 1993. Benchmarks for basic scheduling problems. European Journal of Operational Research, 64(2), pp.278-285.

[8] Koren, Y., Heisel, U., Jovane, F., Moriwaki, T., Pritschow, G., Ulsoy, G. and Van Brussel, H., 1999. Reconfigurable manufacturing systems. CIRP Annals-Manufacturing Technology, 48(2), pp.527-540.

[9] Mehrabi, M.G., Ulsoy, A.G. and Koren, Y., 2000. Reconfigurable manufacturing systems and their enabling technologies. International Journal of Manufacturing Technology and Management, 1(1), pp.114-131.

[10] Wu, D., Greer, M.J., Rosen, D.W. and Schaefer, D., 2013. Cloud manufacturing: Strategic vision and state-of-the-art. Journal of Manufacturing Systems, 32(4), pp.564-579.

[11] Allwood, J.M., Childs, T.H., Clare, A.T., De Silva, A.K., Dhokia, V., Hutchings, I.M., Leach, R.K., Leal-Ayala, D.R., Lowth, S., Majewski, C.E. and Marzano, A., 2016. Manufacturing at double the speed. Journal of Materials Processing Technology, 229, pp.729-757.

[12] Scholz-Reiter, B., Windt, K. and Freitag, M., 2004, May. Autonomous logistic processes: New demands and first approaches. In Proceedings of the 37th CIRP international seminar on manufacturing systems (pp. 357-362). Budapest.

[13] Hülsmann, M. and Windt, K. eds., 2007. Understanding autonomous cooperation and control in logistics: the impact of autonomy on management, information, communication and material flow. Springer Science & Business Media.

[14] Kádár, B. and Monostori, L., 2001. Approaches to increase the performance of agent-based production systems. Engineering of intelligent systems, pp.612-621.

[15] Wooldridge, M. and Jennings, N.R., 1995. Intelligent agents: Theory and practice. The knowledge engineering review, 10(02), pp.115-152.

[16] Sethi, A.K. and Sethi, S.P., 1990. Flexibility in manufacturing: a survey. International journal of flexible manufacturing systems, 2(4), pp.289-328

[17] Golden, W. and Powell, P., 2000. Towards a definition of flexibility: in search of the Holy Grail?. Omega, 28(4), pp.373-384.

[18] Bigus, J.P. and Bigus, J., 2001. Constructing intelligent agents using Java.

[19] Borshchev, A., 2013. The big book of simulation modeling: multimethod modeling with AnyLogic 6. Anylogic North America.

[20] Dias, M.B. and Stentz, A., 2000. A free market architecture for distributed control of a multirobot system. In 6th International Conference on Intelligent Autonomous Systems (IAS-6) (pp. 115-122).

Heuristics for hybrid flowshop problem solving with non-identical parallel machines

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Adriana Pacheco, Christophe Sauvey, Nathalie Sauer, Wajdi Trabelsi
Laboratoire de Génie Industriel, Production et Maintenance
Université de Lorraine, UFR MIM, Ile du Saulcy
57045 Metz Cedex, France

{christophe.sauvey; nathalie.sauer; wajdi.trabelsi}@univ-lorraine.fr

Abstract— In this paper, we propose eight heuristics for solving hybrid flow-shop scheduling (HFS) problems, with non-identical parallel machines, taking into account mixed blocking constraints among which *RSb*, *RCb** and *RCb*, often encountered in industrial environments. Objective function is makespan minimization. HFS problems and different types of constraints are described. Then, the proposed heuristics are precisely described. The results of the experiments performed on benchmark problems are then presented, and heuristics relative effectiveness are discussed in function of problems configurations. Some further research perspectives are finally proposed to address completely this problem, with a new kind of combined constraints.

Keywords— Hybrid flow-shop; Scheduling; Heuristics; Non-identical parallel machines.

I. Introduction

Flow-shop scheduling problem has been widely studied over past decades. Meanwhile, it is difficult to find references dealing with hybrid flow-shop (HFS) problem with non-identical parallel machines, taking into account mixed blocking constraints. Most of works dealing with blocking constraints, consider either the identical parallel machines [1,2] or non-identical parallel machines problems with only one stage [3,4].

Authors who studied hybrid flow-shop problems usually consider 2-stage problems, with parallel machines at first stage and only one machine at the second, or inversely, one at first stage and more than one at second stage [5]. Some others take into account three stages with several parallel machines [6, 7].

From the heuristic point of view, a big part of HFS problems are addressed or improved by evolutionist algorithms means. The principle of such algorithms is inspired from evolution theory. They use mechanisms inspired on biological evolution, like reproduction, mutation, recombination and selection. Some examples, among a lot of others, are given hereafter: Genetic algorithms (GA) [8], Biogeography-based Optimization (BBO) [6], Imperialist Competitive Algorithms (ICA) [6] or Simulated Annealing (SA) [6].

More, famous methods have been employed and modified, like NEH method [9], or easier ones like SPT (Shortest Processing Time first), LPT (Longer Processing Time first) and RAND (Random order). This later is the less efficient than metaheuristics, but gives an idea of an upper bound to evaluate the heuristics. Another method (NH) has proved its efficiency, particularly on uniform HFS problems with *RSb* blocking constraint [10]. A combination of aforementioned heuristics is often efficient and can give good results within a rather short computing time.

Results obtained in [10] indicate that average computing times are shorter when a mixed blocking constraints problem is considered than with problems where no blocking constraint is taken into account. To solve problems with a high number of jobs and machines, computing time considerably increases, and we must use different approached solving methods. The author also indicates that the Linear Integer Mathematical Model becomes quickly insufficient, and suggests to develop Branch and Bound methods, for example, to improve computation speed.

In this paper, we present the hybrid flow-shop problem with non-identical parallel machines we propose to address. Then, we present the heuristics we developed to solve this class of problems, taking into account non-identical parallel machines. The results obtained with those heuristics are then presented and discussed. To conclude, we detail research directions in which we will make progress this work in next months.

II. Problem description

The scheduling problem on which we take interest here is the hybrid flow-shop problem with mixed blocking constraints and non-identical parallel machines. This kind of problem is encountered when n jobs are composed of maximum K operations, which must be treated in the same order (flow-shop) on K stages. On each stage k, there are m_k machines, which can be non-identical. When two machines of a same stage are non-identical, the processing time of a given operation is different from a machine to the other. The objective function is the minimization of the maximal total time after which all the operations of all the jobs are completed, also called the makespan (C_{max}).

In a production line, the aim of being more productive stimulates the investment in new faster machines. This creates a difference between the old ones, slower but often multi-purpose, and the new ones, faster but sometimes less flexible. To increase flexibility, the number of parallel machines of each stage, able to process the same operation, is increased, and its machines are considered as non-identical.

Blocking constraints occur when stocking space between machines is limited or null. This can come from the process itself, from the wish to lessen at minimum the number of products in the stock or to gain place for other activities. Several blocking types exist. In the case of a classical blocking, a job remains blocked on a machine of a stage, as long as the next machine in the routing is not yet available, or no more place does exist in the stock. This blocking is called Release when Starting blocking (*RSb*). In the case of a *RCb** blocking constraint (Release when Completing blocking), a machine is available to treat its next operation just after its just treated operation is completed on the following machine in the routing. Moreover, for the *RCb* blocking constraint case, this operation must necessarily have quitted its machine to start its operation on further machine [10].

We have already studied flow-shop type scheduling problems with mixed blocking in the case of identical parallel machines. Particularly, we have developed exact methods, linear mathematical models, Branch and Bound type algorithms, and heuristics and meta-heuristics ([2], [8], [10]). Other researchers have also worked on these kinds of problems, some of which having already considered the problem of unrelated machines. We can cite [6] who have developed a mixed non-linear model and heuristics in the case where waiting time is limited between the machines. Others have worked on the particular No-Wait blocking case [11] but, as far as we know, no one has already addressed blocking cases encountered in industry (*RSb*, *RCb** and *RCb*) with non-identical machines.

To describe this kind of problem, we consider a sequence of constraints, in a vector *V*, adapted in function of the number and the arrangement of the machines in stages. This vector represents the different blocking constraints between two successive stages in the case of a hybrid flow-shop. Fig. 1 illustrates an example of a mixed blocking flow-shop production system with 4 jobs et 5 machines, where *V* = (*RCb*, *RSb*, *RCb**, *Wb*). *RCb* is the blocking constraint between machines M_1 and M_2, *RSb* is the blocking constraint between machines M_2 and M_3, etc. Wb refers to non-blocking conditions (Without blocking).

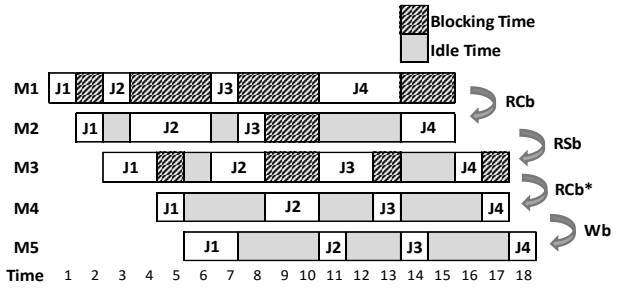


Fig. 1. Flow-shop with mixed blocking constraints *V* = (*RCb*, *RSb*, *RCb**, *Wb*).

III. Proposed heuristics

For this kind of scheduling problems (HFS with non-identical parallel machines and mixed blocking constraints), required computing times are too long to obtain an optimal solution with an exact method. Therefore, obtaining a feasible solution by using an approximate method becomes more interesting for a NP-hard optimization problem due to its ease of programming and its smaller computing times. Seeking to obtain a quick, feasible, but not necessarily optimal solution, we propose eight heuristics, based on known and simple rules, such as Shortest Processing Time (SPT) and Longest Processing Time (LPT).

In the following paragraphs, we describe these heuristics in detail, on a particular example, and illustrate them with Gantt charts. This example is a HFS scheduling problem, composed of four jobs (J_1 to J_4) and three stages (S_1 to S_3), with respectively 2, 2, and 1 machine. The blocking constraints considered between stages are *V* = (*RCb**, *RSb*). The processing times of this example are given in Table I. We give them for each of the 4 jobs, on each of the 5 machines of the 3 stages (2-2-1) HFS problem. We can note that jobs processing times on the machines of a same stage are different. Thus, this example illustrates a non-identical parallel machine case we deal with on this paper.

TABLE I. Processing times of the example

	S_1		S_2		S_3
Jobs	M_1	M_2	M_1	M_2	M_1
J_1	3	2	2	2	2
J_2	3	2	2	1	1
J_3	2	1	2	1	2
J_4	1	3	1	2	1

The four first heuristics (A to D) are precisely described with a pseudo-code, and are based on the SPT heuristic. Since the four others are based on LPT heuristic rule and follow the same structure, their presentation is summed up in a single paragraph (E).

A. First heuristic: SPT-based (Partial Processing Times), First Fit strategy.

The SPT rule sequences a vector composed of each jobs sum of minimal processing times on each stage with more than one machine, in an increasing order. Then, this first heuristic places, at each stage, each job in the first machine available (First Fit strategy).

Let $LJ = \{J_1, J_2, \ldots, J_n\}$ be the list of jobs ordered by increasing sum of minimal processing time among all the stages where $m_k > 1$ (m_k is the number of machines at stage k). With this heuristic, only the stages with several machines are taken into consideration. It may be particularly useful in problems with one machine at the last stage (a common studied HFS problem). In this case, the aim of this method is to eliminate from the vector SPT, the last stage processing time which, in case of being too long, might affect the SPT scheduling of the jobs before the last machine. We note:

- RT_{mk}: Release Time of m^{th} machine of stage k, once a job has been processed.
- BT_{mk}: Blocking Time of m^{th} machine of stage k (when a machine is released).
- V = (B_1, B_2, …, B_{k-1}): Blocking Vector, where B_i is the blocking constraint between stages i and i+1.
- C_{max}: Makespan.

Algorithm 1

Initialisation of RT_{mk}= 0, BT_{mk} = 0
For i = 1 to n
  Select Job J_i from LJ
  For k = 1 to K
    Schedule Job J_i in machine M_m of S_k with minimal RT_{mk}
    For m = 1 to m_k
      Calculate RT_{mk}
      Calculate BT_{mk} for blocking type B_k
      RT_{mk} = max (RT_{mk}, BT_{mk})
    End for
  End for
End for
C_{max} = max (RT_{mK})

Example: To obtain the SPT sequence for processing the jobs, we calculate the sum of the minimal processing times (given in Table I) among the first two stages, where m_k > 1. The considered blocking vector is V = (*RCb**, *RSb*). We obtain 4, 3, 2 and 2 ut (units of time) respectively. Therefore, the SPT sequence can be: LJ = {J_3, J_4, J_2, J_1} or {J_4, J_3, J_2, J_1}. The Gantt chart on Fig. 2 presents the first option.

For the example, we take the sequence {J_3, J_4, J_2, J_1}. We choose to schedule the first job, J_3, in machine M_1 at the first stage, because it's the first machine available at stage 1 at t=0. Similarly, for the following stages, we schedule J_3 in M_1. In the same way, we schedule the job J_4 in machine M_2 at first stage because it is available at t=0, while M_1 remains unavailable until t=2. Similarly, for the following stage and for the final one, there is only one option in M_1. At first stage, both machines remain blocked until the respective jobs that were being processed finish their treatment in the following stage in t=2 (*RCb** blocking constraint). At second stage, the machine M_1 remains blocked until the job J_4, that was being processed, starts its treatment at next stage at t=6 (*RSb* blocking constraint). Similarly, we schedule the next jobs in the sequence in the first machine available per stage. The corresponding Gantt chart is given on Fig. 2.

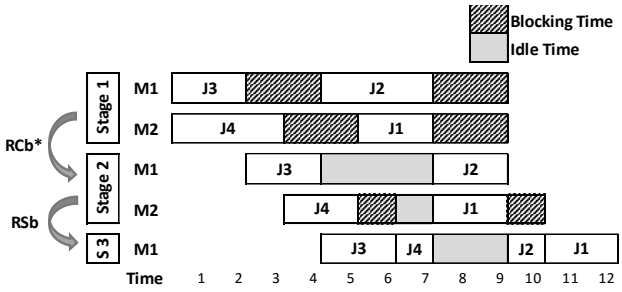


Fig. 2. Gantt chart for solution given by Heuristic 1.

B. Second heuristic: SPT-based (Total Processing Times), First Fit strategy.

The previous heuristic is modified to include all the stages, not only the stages with more than one machine, in the vector on which the SPT rule is applied.

Let LJ = {J_1, J_2, …, J_n} be the list of jobs ordered by increasing sum of minimum processing time in all the stages ($m_k \geq 1$). This means that all stages are taken into consideration for obtaining the SPT sequence of jobs. Excepted the aforementioned consideration, the algorithm for this second heuristic is the same as the previous one (Algorithm 1).

Example: To obtain the SPT sequence for processing the jobs, we calculate the sum of the minimal processing times (given in Table I) among all the stages. We obtain 6, 4, 4 and 3 ut respectively. Therefore, the SPT sequence can be: LJ = {J_4, J_3, J_2, J_1} or {J_4, J_2, J_3, J_1}. For the example, we take the first option. The corresponding Gantt chart is given on Fig. 3.

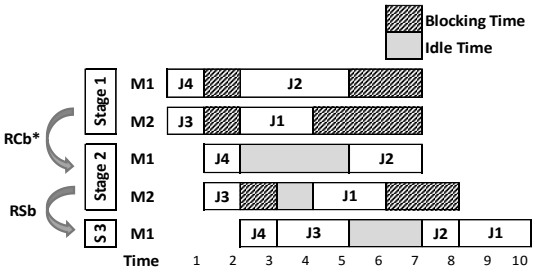


Fig. 3. Gantt chart for solution given by Heuristic 2.

C. Third heuristic: SPT-based (Partial Processing Times), Best Fit strategy.

This heuristic places each job in the machine where its completion time would be smaller (Best Fit strategy). With this heuristic, we try to foresee when an operation could finish, if it is placed on each machine of a stage. With this information, we place the operation on the machine which gives the smaller completing time. Here, we try to benefit from the particularity addressed in this paper, to build a hopefully better heuristic than the previous one. Even if its computing time is a little higher, we hope better results. The pseudo-code of this heuristic is presented in Algorithm 2.

Let LJ = {J_1, J_2, …, J_n} be the list of jobs ordered by increasing sum of minimum processing time in the stages where m_k > 1. The algorithm 2 presents this heuristic, keeps the same notations as for algorithm 1, and adds the following one: CT_{imk}, completion time of J_i in m^{th} machine of stage k.

Algorithm 2

Initialisation of RT_{mk}= 0, BT_{mk} = *0*
Initialisation of CT_{imk} = 0
For i=1 to n
  Select Job J_i from LJ
  For k=1 to K
    For m=1 to m_k
      Calculate CT_{imk} = RT_{mk} + P_{im}
      Schedule J_i in machine M_m of S_k with minimal CT_{imk}
      Calculate RT_{mk}
      Calculate BT_{mk} for blocking type B_k
      RT_{mk} = max (RT_{mk}, BT_{mk})
    End for
  End for
End for
C_{max} = max (RT_{mK})

Example: As in the first heuristic, the SPT sequence can be: $LJ = \{J_3, J_4, J_2, J_1\}$ or $\{J_4, J_3, J_2, J_1\}$, but for scheduling the jobs, the aim is to do it in the machine on which its treatment ends earlier. For the example, we take the sequence $\{J_3, J_4, J_2, J_1\}$. We choose to schedule the first job, J_3, in machine M_2 at the first stage, because its processing ends at t=1. Otherwise, in machine M_1, its processing would end at t=2. Similarly, for the following stage, we schedule J_3 in M_2, and for the final one, there is only one option in M_1. In the same way, we schedule the job J_4. At first stage, both machines remain blocked until the respective jobs that were being processed finish their treatment in the following stage in t=2 (*RCb** blocking constraint). At second stage, the machine M_1 remains blocked until the job J_4, that was being processed, starts its treatment at next stage in t=4 (*RSb* blocking constraint). Similarly, we schedule the next jobs in the sequence. The corresponding Gantt chart is given on Fig. 4.

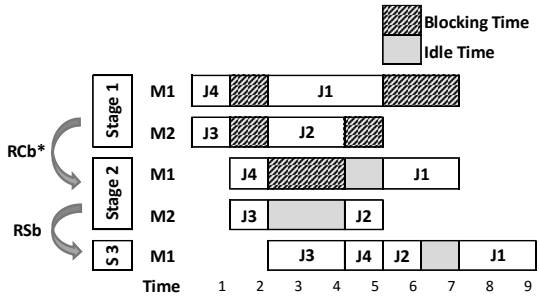


Fig. 4. Gantt chart for solution given by Heuristic 3.

D. Fourth heuristic: SPT-based (Total Processing Times), Best Fit strategy.

The previous heuristic was modified to include all the stages. Let $LJ = \{J_1, J_2, \ldots, J_n\}$ the list of jobs ordered by increasing sum of minimum processing time in all the stages ($m_k \geq 1$). This means that all stages are taken into consideration for obtaining the SPT sequence of jobs. Excepted the aforementioned consideration, the algorithm for this fourth heuristic is the same as the previous one (Algorithm 2).

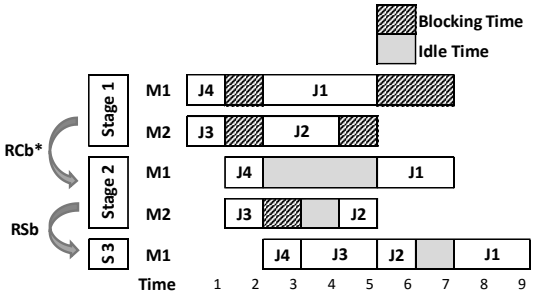


Fig. 5. Gantt chart for solution given by Heuristic 4.

Example: To obtain the SPT sequence for processing the jobs, we calculate the sum of the minimal processing times (given in Table I), as we did in the second heuristic, among all the stages. We obtain 6, 4, 4 and 3 ut respectively. Therefore, the SPT sequence can be $LJ = \{J_4, J_3, J_2, J_1\}$ or $\{J_4, J_2, J_3, J_1\}$. For the example, we take the first option. The corresponding Gantt chart is given on Fig. 5.

E. LPT-based heuristics

Other four LPT-based heuristics were programmed, replacing SPT dispatching rule by LPT in the previous ones. The LPT rule sequence the minimal processing times vector in a decreasing order. Let $LJ = \{J_1, J_2, \ldots, J_n\}$ be the list of jobs ordered by decreasing sum of minimum processing time in all the stages ($m_k \geq 1$).

Heuristic 5: This algorithm is the same as Heuristic 1 (Partial Processing Times), but scheduling according to an LPT vector. In example, we obtain $LJ = \{J_1, J_2, J_3, J_4\}$ or $\{J_1, J_2, J_4, J_3\}$. We take the first option for the example of which corresponding Gantt chart is given on Fig. 6. Even if job J_2 could be processed at t=3, since the heuristic's LPT order is $\{J_1, J_2, J_3, J_4\}$, jobs are sequenced in this order on stage 3.

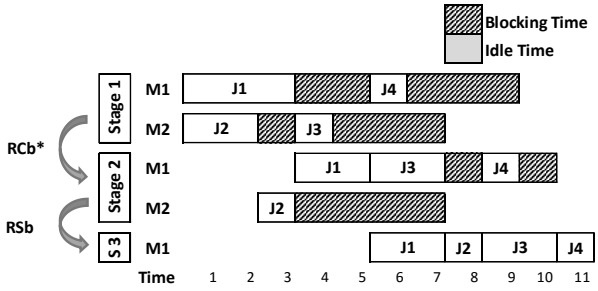


Fig. 6. Gantt chart for solution given by Heuristics 5 and 6.

Heuristic 6: This algorithm is the same as Heuristic 2 (Total Processing Times), but scheduling according to an LPT vector. In example, we obtain $LJ = \{J_1, J_2, J_3, J_4\}$ or $\{J_1, J_3, J_2, J_4\}$. We take the first option for the example, resulting in the same LPT sequence as previous. Therefore, the corresponding Gantt chart is already given in Fig. 6.

Heuristic 7: This algorithm is the same as Heuristic 3 (Partial Processing Times), but scheduling according to an LPT vector that can be $LJ = \{J_1, J_2, J_3, J_4\}$ or $\{J_1, J_2, J_4, J_3\}$ in the example. We take the first option for the example. The corresponding Gantt chart is given on Fig. 7.

Heuristic 8: This algorithm is the same as Heuristic 4 (Total Processing Times), but scheduling according to an LPT vector that can be $LJ = \{J_1, J_2, J_3, J_4\}$ or $\{J_1, J_3, J_2, J_4\}$ in the example. We take the first option for the example, resulting in the same LPT sequence as previous. Therefore, the corresponding Gantt chart is already given in Fig. 7.

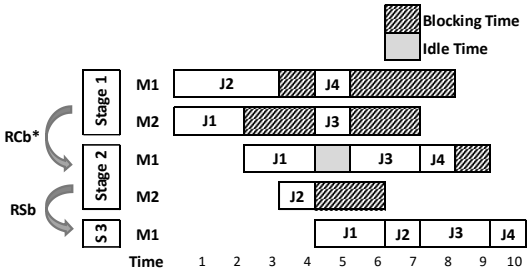


Fig. 7. Gantt chart for solution given by Heuristics 7 and 8.

IV. Results and discussions

We have tested the heuristics on problems composed of 10 jobs and 3 stages. The benchmark on which we have worked contains 10 problem instances per stage configuration. The results obtained with our heuristics are presented in error percentages, reported to best solution obtained with one of our heuristics. The values of processing times on each machine have been randomly filled with integer numbers, following a uniform law [1, 99]. The problem instances treated have the same values for the problem with mixed blocking constraints as for the problem without blocking constraint. Three different stage configurations have also been tested, one with only one machine at the last stage, another with only one machine at the first stage and the last with 2 machines at each stage. Experiments have been performed on a Core i7 2.50GHz laptop. Heuristics have been programmed in Java, using the free and open-source Eclipse IDE software. All results presented on Tables II and III have been obtained, for each instance, within a time inferior to one second. H1 to H4 are the heuristics based on SPT rule, and H5 to H8 are based on LPT rule. In the literature, classically, SPT rule performs correctly on problems with one machine at last stage, because it is worth quickly arriving to throughput last stage. For symmetrical problems, contrarily, LPT rule performs well [10].

For the first experiment, the blocking vector V = (*RCb**, *RSb*) has been taken into consideration and the average relative error percentages and number of best solutions found are presented in Table II. As can be seen on the two first lines, classical literature results are globally verified. For the first line, reporting results with (2-2-1) stages configuration, we can remark that H3 is the best performing heuristic. H4 performed better on only 3 instances. The second line validates the superiority of H7 for (1-2-2) problems, even if H8 performed better only once. On the third line, results of twin heuristics are strictly equal because, with the (2-2-2) stages configuration, all stages verify the condition $m_k > 1$. Again, H7 clearly appears to be best performing. H3 better performed only once.

TABLE II. Average relative error percentages (number of best solutions found) with 10 jobs, 3 stages HFS benchmark, with *V* = (*RCb**, *RSb*) blocking vector.

(m_1-m_2-m_3)	H1	H2	H3	H4	H5	H6	H7	H8
(2-2-1)	21	24	**2,1 (7)**	5,5 (3)	26	19	16	9,3
(1-2-2)	31	31	6,7	5,6	20	22	**0,1 (9)**	1,0 (5)
(2-2-2)	38	38	9,9 (1)	9,9 (1)	28	28	**0,1 (9)**	**0,1 (9)**

On the second experiment, the blocking vector V = (*Wb*, *Wb*), has been taken into consideration. It proves that our heuristics are able to treat non-identical hybrid flow-shop in the general case, with no blocking constraint. Average relative error percentages are presented in Table III. Classical literature results are globally verified on the two first lines, but here, H3 and H7 have respectively not been outperformed. On the third line, with the (2-2-2) stages configuration, results of H3 and H7 are here balanced. Even if H7 seems to have a slight numerical advantage, 5 instances were solved with H3 and 5 others with H7.

TABLE III. Average relative error percentages (number of best solutions found) with 10 jobs, 3 stages HFS benchmark, with *V* = (*Wb*, *Wb*) blocking vector.

(m_1-m_2-m_3)	H1	H2	H3	H4	H5	H6	H7	H8
(2-2-1)	13	23	**0,0 (10)**	5,7 (2)	26	22	22	19
(1-2-2)	28	22	18	15	12	21	**0,0 (10)**	6,3
(2-2-2)	31	31	8,2 **(5)**	8,2 **(5)**	30	30	**5,7 (5)**	**5,7 (5)**

In tables IV, V and VI, we present the results of another experiment with 10 jobs and 5 stages, respectively obtained with blocking vectors *V = (RSb, Wb, RSb, Wb)*, *V = (Wb, RCb*, Wb, RSb)* and *V = (Wb, Wb, Wb, Wb)*.

TABLE IV. Average relative error percentages (number of best solutions found) with 10 jobs, 5 stages HFS benchmark, with V = *(RSb, Wb, RSb, Wb)* blocking vector.

(m_1-…-m_5)	H1	H2	H3	H4	H5	H6	H7	H8
(1-2-2-2-2)	40	35	27	26	19 (1)	24	**0,9 (7)**	2,3 (2)
(1-2-2-3-3)	27	26	18	16	23	27	**0,0 (9)**	5,2 (1)
(2-2-2-2-2)	27	27	**2,7** (5)	**2,7** (5)	25	25	5,4 **(6)**	5,4 **(6)**
(3-3-3-3-3)	81	81	**12 (5)**	**12 (5)**	66	66	**4,2 (5)**	**4,2 (5)**
(3-3-2-2-1)	22	30	**0,0 (10)**	5,3 (2)	36	35	25	23
(2-2-2-2-1)	25	28	**0,3 (9)**	5,3 (3)	39	32	29	24

TABLE V. Average relative error percentages (number of best solutions found) with 10 jobs, 5 stages HFS benchmark, with V = *(Wb, RCb*, Wb, RSb)* blocking vector.

(m_1-…-m_5)	H1	H2	H3	H4	H5	H6	H7	H8
(1-2-2-2-2)	48	40	26	25	19	33	**0,0 (10)**	6,9 (1)
(1-2-2-3-3)	35	28	19	15	25	35	**0,1 (9)**	6,5 (1)
(2-2-2-2-2)	36	36	**5,5 (5)**	**5,5 (5)**	28	28	6,5 **(5)**	6,5 **(5)**
(3-3-3-3-3)	76	76	6,7 **(6)**	6,7 **(6)**	76	76	**4,8** (4)	**4,8** (4)
(3-3-2-2-1)	23	25	**2,3 (5)**	2,4 **(5)**	34	31	22	19
(2-2-2-2-1)	24	31	**0,0 (10)**	4,8 (1)	38	33	28	21

Globally, these experiments validate three main ideas. It is always better to dispose of several heuristics, because there is not only one heuristic well performing at all kinds of problems. Then, in the case addressed here, with non-identical parallel machines, it is clearly worth scheduling with the times at which operations are expected to finish on a stage rather than at their starting times on the same given stage. All the results presented here prove this affirmation, because there was only one case where H1, H2, H5 or H6 did perform better than one of the four other heuristics. Finally, summing processing times on no-

throughput stages (where $m_k > 1$) seems to globally give better results than summing on all stages, when it remains possible.

TABLE VI. Average relative error percentages (number of best solutions found) with 10 jobs, 5 stages HFS benchmark, with V = *(Wb, Wb, Wb, Wb*) blocking vector.

(m_1-…m_k …-m_5)	H1	H2	H3	H4	H5	H6	H7	H8
(1-2-2-2-2)	43	35	28	26	21	27	**0,2 (9)**	3,2 (2)
(1-2-2-3-3)	27	26	19	16	22	29	**0,2 (9)**	5,7 (1)
(2-2-2-2-2)	31	31	**4,5 (5)**	**4,5 (5)**	32	32	8,2 **(5)**	8,2 **(5)**
(3-3-3-3-3)	79	79	8,7 **(5)**	8,7 **(5)**	67	67	**6,2 (5)**	**6,2 (5)**
(3-3-2-2-1)	23	28	**0,0 (10)**	5,1 (2)	35	36	25	23
(2-2-2-2-1)	25	31	**0,0 (10)**	5,2 (2)	40	33	30	25

On this particular experiment, we can remark that both third lines of Tables II and III give rather different results. Indeed, results are balanced with no-blocking constraints, and quite completely in favour of H7 with V = (*RCb**, *RSb*). This could mean that mixed blocking constraints change classical performing fields of both SPT and LPT rules. This trend, slightly perceived on a 10 jobs, 3 stages problem benchmark, seems to be infirmed on a 5 stage problem, with 6 different machine configurations and 3 blocking vectors. Nevertheless, this point will be dug in the next future.

V. Conclusion and Perspectives

In this paper, we propose eight constructive heuristics to solve the hybrid flow-shop scheduling problem, under both mixed blocking constraints and non-identical parallel machines. We have tested three main ideas, LPT or SPT rules, scheduling with the time at which an operation arrives at a stage or at the date it leaves the stage, and taking into consideration processing times of all stages or only those concerning multi-machines stages ($m_k > 1$). The combination of these ideas gave birth to the eight heuristics, presented in details in this paper, with a numerical example and Gantt charts to illustrate respective solutions.

The experiments reported in this paper proved both the effectiveness of LPT and SPT-based heuristics and the superiority of scheduling at the dates at which operations are supposed to leave a stage (Best Fit strategy). On the third point, taking into account only the stages where $m_k > 1$ seems to be the best choice, even if, on some instances, this strategy is outperformed. This research work will be completed with some further research investigations.

Our future research works on this subject will deal first with a linear integer mathematical model development, so that we will be able to undoubtedly validate the quality of the heuristics we have proposed in this paper. Then, we will propose other heuristics to further dig the good-working principle identified in this paper. A genetic algorithm will also be developed to test different meta-heuristics performance on this particular kind of problems.

Another field of research which can be launched could be the analysis of the problem instances solved with H3 and those solved with H7, with the aim to identify the reasons of their correct resolution by a heuristic or the other. An accurate observation could lead us to both better forecast which heuristics are likely to solve correctly a problem, and catch new ideas to develop even better heuristics.

Acknowledgment

This work has received a fund granted by the French Region Lorraine, which supports this regional interest research project, under the number AAP-004-107.

References

[1] Sangsawang, C., Sethanan, K., Fujimoto, T., & Gen, M. (2015). Metaheuristics optimization approaches for two-stage reentrant flexible flow shop with blocking constraint. Expert Systems with Applications, 42(5), pp. 2395–2410.

[2] Trabelsi, W., Sauvey, C., Sauer, N. (2011a). Mathematical Model and Lower Bound for Flowshop Problem With Mixed Blocking Constraints. International Conference on Industrial Engineering and Systems Management (IESM'2011). Metz, France.

[3] Alcan, P., Basligil, H. (2012). A genetic algorithm application using fuzzy processing times in non-identical parallel machine scheduling problem. Advances in engineering software 45, pp. 272-280.

[4] Balin, S. (2011). Non-identical parallel machine scheduling using genetic algorithm. Expert systems with applications 38, pp. 6814-21.

[5] Figielska, E. (2014). A heuristic for scheduling in a two-stage hybrid flowshop with renewable resources shared among the stages. *European Journal of Operational Research, 236*, pp.433–444.

[6] Attar, S.F., Mohammadi, M., Tavakkoli-Moghaddam, R. (2013). Hybrid flexible flowshop scheduling problem with unrelated parallel machines and limited waiting times. International Journal of Advanced Manufacturing Technology 68, pp. 1583-99.

[7] Bozorgirad, M. A., & Logendran, R. (2016). A comparison of local search algorithms with population-based algorithms in hybrid flow shop scheduling problems with realistic characteristics. *International Journal of Advanced Manufacturing Technology, 83*, pp. pp. 1135–1151.

[8] Sauvey, C., Sauer, N. (2012). A genetic algorithm with genes-association recognition for flowshop scheduling problems, Journal of Intelligent Manufacturing 23, pp.1167-1177.

[9] Nawaz, M., Enscore, E. E., Ham, I. (1983). A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem. Omega, 11(1), pp. 91-95.

[10] Trabelsi, W., Sauvey, C., Sauer, N. (2012). Heuristics and meta heuristics for mixed blocking constraints flowshop scheduling problems, Computers & Operations Research 39, pp.2520–2527.

[11] Rabiee, M., Sadeghi Rad, R., Mazinani, M., Shafaei, R. (2014). An intelligent hybrid meta-heuristic for solving a case of no-wait two-storage flexible flowshop scheduling problem with unrelated parallel machines. International Journal of Advanced Manufacturing Technology 71, pp. 1229-45.

Exact Solution of the Single Machine Total Tardiness Problem : the Power of Memorization

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Lei Shang
Université François-Rabelais de Tours,
LI (EA 6300), ERL CNRS OC 6305
64 avenue Jean Portalis
37200 Tours, France
Email: shang@univ-tours.fr

Vincent T'Kindt
Université François-Rabelais de Tours,
LI (EA 6300), ERL CNRS OC 6305
64 avenue Jean Portalis
37200 Tours, France
Email: tkindt@univ-tours.fr

Federico Della Croce
Politecnico di Torino, DAUIN
Corso Duca degli Abruzzi 24
10129 Torino, Italy
Email: federico.dellacroce@polito.it

***Abstract*—This paper provides new experimental results on the exact solution of the single machine total tardiness problem (SMTT). The currently fastest algorithm reported in the literature is due to Szwarc et al [12], which solves instances with up to 500 jobs in size. The key elements of their algorithm are the use of *Memorization* and the exploitation of an "Algorithmic Paradox". Recently, we revisited the problem and proposed a *Branch & Merge* algorithm that runs in $\mathcal{O}^*((2+\epsilon)^n)$ in time, with ϵ arbitrarily small, and in polynomial space. During the study of the computational performance of this algorithm, the authors occasionally found that the *Memorization* technique used in [12] can be further boosted and their "Paradox" can be enhanced. The resulting algorithm is able to solve all instances with up to 1200 jobs in size. We also discuss the worst-case running time of the algorithm.**

I. Introduction

The design of exact algorithms for NP-hard problems has always been a challenging issue. We consider here a classic scheduling problem: the single machine total tardiness ($1||\sum T_j$) problem which is known to be NP-hard in the ordinary sense [1]. The problem asks to schedule a set of n jobs $N = \{1, 2, \ldots, n\}$ on a single machine. For each job j, a processing time p_j and a due date d_j are given and the objective is to arrange the jobs into a sequence $S = (a_1, ..., a_n)$ so as to minimize $T(N, S) = \sum_{j=1}^{n} \max\{\sum_{i=1}^{j} p_{a_i} - d_{a_j}, 0\}$. The problem has been extensively studied in the literature. The conventional *Dynamic Programming* algorithm across all subsets can solve this problem in $\mathcal{O}^*(2^n)$ in time and space, which is shown to be impractical on large instances. On the experimental aspect, the current state-of-the-art exact method in practice is a *Branch & Bound* algorithm which solves to optimality problems with up to 500 jobs in size [12]. Latest theoretical developments for the problem can be found in the survey of Koulamas [7].

Recently, we have discovered some interesting structural properties of the problem and an exact algorithm called *Branch & Merge* has been proposed and proved to have a worst-case time complexity converging to $\mathcal{O}^*(2^n)$ with a polynomial space complexity [3]. During the study of the experimental performance of this algorithm, we found that the *Memorization* technique used in the algorithm of [12] is in fact a time efficient way to implement the merging technique introduced in [3] and it can be further boosted. Besides, the "Algorithmic Paradox" stated in [12] can also be enhanced. With all improvements integrated, our algorithm solves the problem with up to 1200 jobs.

The main properties used to solve the problem are recapitulated in section II. Our original *Branch & Merge* algorithm is briefly recalled in section III with a discussion on its relationship with *Memorization*. Section IV presents the main experimental results with an enhanced "Algorithmic Paradox" stated.

II. Preliminaries

Let $(1, 2, \ldots, n)$ be a LPT (Longest Processing Time first) sequence and $([1], [2], \ldots, [n])$ be an EDD (Earliest Due Date first) sequence of all jobs.

Property 1. *[6](Emmons' conditions)*
There exists an optimal sequence where for each pair of jobs i, j, $p_i < p_j$:

1) *i precedes j if $d_i < \max(E_j, d_j)$.*
2) *j precedes i if $d_i + p_i > L_j$ and $d_i > \max(E_j, d_j)$.*

Emmons' conditions allow to derive precedence relations between jobs. The following property allows to decompose the problem.

Decomposition 1. *[8] (Lawler's decomposition) Let job 1 in LPT sequence correspond to job $[k]$ in EDD sequence. Then, job 1 can be set only in positions $h \geq k$ and the jobs preceding and following job 1 are uniquely determined as $B_1(h) = \{[1], [2], \ldots, [k-1], [k+1], \ldots, [h]\}$ and $A_1(h) = \{[h+1], \ldots, [n]\}$.*

Moreover, it is shown that not all positions after k are candidate positions for h. Some well established rules can be used to eliminate dominated positions.

Property 2. *(Elimination Rules) Consider an EDD sequence where job 1 is moved from the original position k to position $i > k$. Let $T(1, i)$ be the tardiness of the resulting sequence and $C_1(i)$ be the completion time of job 1 that was moved to position i. Assume job 1 is completed as soon as possible then the following rules hold. Delete r from the initial possible position list $\{k, k+1, ..., n\}$ of job 1 if:*

1) $C_1(r) > d_{[r+1]}$, $k \leq r \leq n-1$. *[8]*
2) $C_1(r) \leq d_{[r]} + p_{[r]}$, $s+1 \leq r \leq n$. *[10]*
3) $C_1(r) \leq d[i] + p[i]$ *for some* $s+1 \leq i \leq r-1$. *[11]*
4) $T(1,r) > T(1,r+1)$ *or* $T(1,r) \geq T(1,r-1)$ *or* $T(1,r) \geq T(1,i)$ *for some* i, $s \leq i < r-1$. *[13]*

Another decompostion scheme (Decomposition 2) has a similar structure but the job to branch on is the smallest-due-date job. The Elimination Rules can be also adapted for it.

Decomposition 2. *[9] Let job k in LPT sequence correspond to job $[1]$ in EDD sequence. Then, job k can be set only in positions $h \leq (n-k+1)$ and the jobs preceding job k are uniquely determined as $B_k(h)$, where $B_k(h) \subseteq \{k+1, k+2, \ldots, n\}$ and $\forall i \in B_k(h), j \in \{n, n-1, \ldots, k+1\} \setminus B_k(h)$, $d_i \leq d_j$*

Property 3 allows to split a problem into several subproblems based on precedence relations between jobs. It is clear that if all jobs from set B precedes set A, $A \bigcup B = N$, then the problem naturally decomposes into two subproblems of jobset B and A. We call (B, A) an optimal block sequence of the initial problem. Let E_j and L_j be the earliest and latest completion times of job j, that is if B_j (resp. A_j) is the currently known jobset that should precede (resp. follow) job j, then $E_j = p(B_j) + p_j$, $L_j = p(N \setminus A_j)$.

Property 3. *[9] (Split)*
(B, A) corresponds to an optimal block sequence if $\max_{i \in B} L_i \leq \min_{j \in A} E_j$. Notice that there could be up to n blocks in a block sequence.

The value of E_i and L_i of each job i can be obtained by applying Emmons' conditions.

From now on we refer to the *Branch & Bound* algorithm of [12] as BB2001. The above properties are essential building blocks of BB2001, nevertheless the results in [12] can not be obtained without the strong catalyst: *Memorization*. The idea is to "memorize" the solution of solved subproblems so that when that subproblem is encountered again, its solution is retrieved directly from the "memory" instead of solving it again. The "memory" is implemented as a hash table, which allows a fast access to stored solutions.

While using the *Memorization* technique, Szwarc et al [12] found the following "Algorithmic Paradox".

Paradox 1. *"Based on the Branch & Bound algorithm in [9], deleting a lower bound drastically improves the performance of the algorithm, while adding a stronger component, like a better decomposition rule, negatively affects its performance."*

Paradox 1 is simply due to the fact that a lot of identical subproblems are encountered during the solution of the problem. The computation time used by lower bounding procedures to cut these identical problems are much longer than simply solve that subproblem once, memorize the solution and reuse the solution whenever the subproblem appears again. Besides, pruning nodes by the lower bound may prevent memorization from working thus lead to a loss of efficiency of the algorithm.

BB2001 has the following main structure:

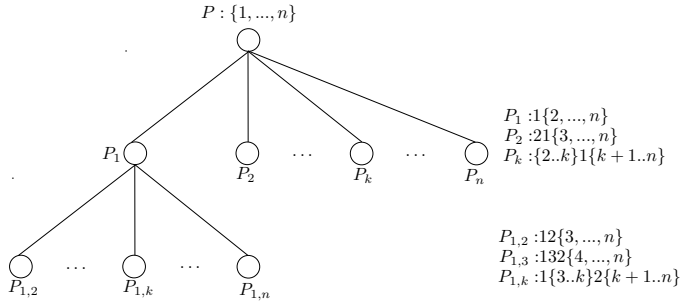
1) Search the problem in "memory", return the solution if found, otherwise go to 2.
2) Use Property 3 to split the problem and solve each subproblem recursively. If no split can be done, go to 3.
3) Apply Property 2 to get the list of possible branching positons for the longest job and the smallest-due-date job.
4) Combine Decomposition 1 and 2 to branch on the longest job and the smallest-due-date job to every candidate position given by step 3. For each branching case, solve subproblems recursively then finally store the best solution among all branching cases into the memory and return it.

Note that due to Paradox 1, all lower bounding procedures are removed, which makes the *Branch & Bound* algorithm a *Branch* algorithm.

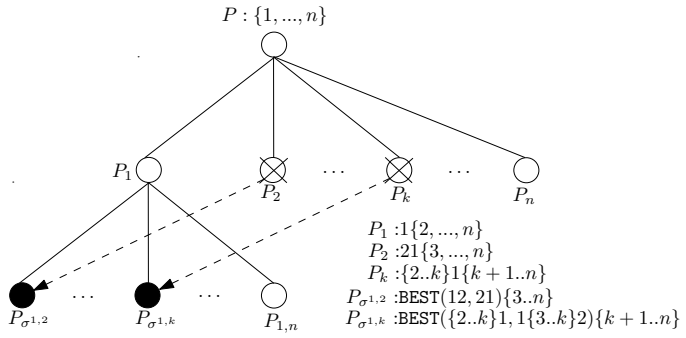
III. Branch & Merge and Memorization

Previous works aimed at providing an algorithm with worst-case complexity guarantee on the problem, by exploiting the fact that a lot of nodes in the branching tree relate to identical subproblems. The resulting algorithm is called *Branch & Merge* and is reported in [2], [3]. For the sake of completeness we recall the main idea of *Branch & Merge*, knowing that the technical details are not essential for this paper and they can be found in [2].

Examining the branching tree based on Decomposition 1 and focusing on the worst-case instances where LPT sequence equals to EDD sequence, we can see that by developing a polynomial number of subnodes, many idential subproblems are showing up: branching job 1 on position 1 then job 2 on position 2 gives a subproblem of the form $12\{3, ..., n\}$, consisting in scheduling jobs $\{3, ..., n\}$ at time point $p_1 + p_2$. This is the same problem as $21\{3, ..., n\}$, given by branching job 1 on position 2, with job 2 forced to the first position due to Decomposition 1. The node $12\{3, ..., n\}$ dominates $21\{3, ..., n\}$ if the partial sequence $(1, 2)$ yields smaller total tardiness than $(2, 1)$. Based on this observation, at each node, *Branch & Merge* open a polynomial number of nodes, *merge* nodes related to identical problems to avoid solving them twice and then continue in a recursive way. The algorithm runs in $\mathcal{O}^*((2+\epsilon)^n)$ time, with ϵ arbitrarily small, and polynomial space. An illustrative example is shown in Figure 1.



(a) Before *Merge*



(b) After *Merge*

Fig. 1: example of *Branch & Merge* on some left branches

The whole mechanism of *Branch & Merge* has been implemented but without a practical success. After some investigation, we found the main reasons as follows:

1) Decomposition 3 and the Elimination Rules are so strong that the number of subnodes on each branching is very limited. Therefore the structural merging operation is not happening so often.
2) When memory usage is not limited, *Memorization* covers merging in the sense that all nodes that are cut by merging can also be cut by *Memorization*. Moreover, some cases can only be treated by *Memorization*. As an example consider node A as $12\{4,5,6\}3$ and node B $13\{4,5,6\}2$ with $p_2 = p_3$. Merging can not be performed because the prefixed jobs (12 vs 13) are not on same jobs. However since $p_2 = p_3$, A and B contain the same problem: scheduling $\{4,5,6\}$ at time point $p_1 + p_2$.

Therefore, the conclusion here is that if there is no limit on memory usage, *Memorization* is preferable to *Branch & Merge*, and since it covers *Branch & Merge* the time complexity of *Memorization* also converges to $\mathcal{O}^*(2^n)$. However, the memory usage was declared as a bottleneck in BB2001, and now the question is, whether the computer memory of today can make a difference in practice. If yes, how much?

With these questions, we retest BB2001 on our machine, then investigate and enhance it to obtain better results.

IV. Experimental results

Our tests are performed on a HP Z400 work station with 3.07GHz CPU and 8Gb RAM. 200 instances are generated randomly for each problem size using the same generation scheme as in [10]. Processing times are integers generated from a uniform distribution in the range $[1, 100]$ and due dates d_i are integers from a uniform distribution in the range $[p_i u, p_i v]$ where $u = 1 - T - R/2$ and $v = 1 - T + R/2$. Each due date is set to zero whenever its generated value is negative. Twenty combinations (R, T) are considered where $R \in \{0.2, 0.4, 0.6, 0.8, 1\}$, and $T \in \{0.2, 0.4, 0.6, 0.8\}$. Ten instances are generated for each combination and the combination $(R = 0.2, T = 0.6)$ yields the hardest instances as reported in the literature (see [9]) and confirmed by our experiments. We first test our implementation of BB2001. It solves instances with up to 900 jobs in size, with an average time of 764s and a maximum time of 9403s, knowing that the original program in 2001 was limited to instances with up to 500 jobs due to memory size limit. Their tests are done on a Sun Ultra-Entreprise Station with limited CPU frequency (<450MHz) and RAM size (not stated). It is anyway interesting to see that with just the computer hardware evolution we can now solve instances with 400 jobs more.

The program is out of time (>4h) at 1000 jobs, hence the memory seems no longer to be the bottleneck. However we were not aware that it is possible to take more advantage of the remaining memory to accelerate the solution, until we found the following "paradox".

Paradox 2. *Remove Split procedure (Property 3) from BB2001 drastically accelerate the solution.*

The effect of Paradox 2 is astonishing: the resulting algorithm *NoSplit* solves instances of 700 jobs with an average time 24 times faster (Table I). In fact *Split* is performed based on the precedence relation between jobs. The computation of there precedence relations is time consuming. Moreover, as already described in Paradox 1, many identical problems appear in the search tree and the *Split* procedure is re-run each time. When *Split* is removed, nodes are solved with more time at their first occurrence, but then never solved twice thanks to *Memorization*. However, the disadvantage is also clear: more memory is needed than before and that is why *NoSplit* encounters memory problems on instances with 800 jobs.

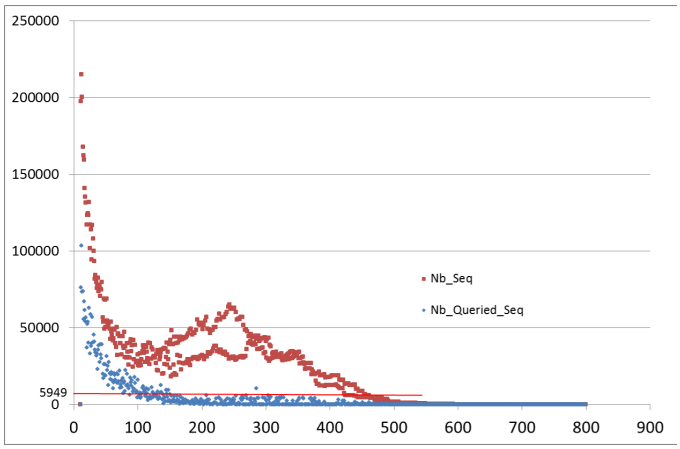
	TMin	TAvg	TMax	#Nodes	SizeMem
BB2001	0.0	192.81	2963.0	880268	111175
NoSplit	0.0	8.0	114.0	3053648	1262895

TABLE I: Results for instances of size 700

SizeMem : number of solutions in memory when the algorithm terminates.

At this point, we have a better understanding of the power of *Memorization* on this problem and we become curious on the effectiveness of memorized solutions. In other words, what are the stored solutions that are effectively used? To answer this question, we test the following three memory cleaning strategies, applied to remove unuseful solutions when the memory is "full".

FIFO: First In First Out
This is one of the most common database cleaning strategy:



Nb_Seq: number of sequences of a given size, stored in the memory.
Nb_Queried_Seq: number of sequences of a given size, that are used to avoid solving twice identical problems.

Fig. 2: Number of solutions and useful solutions in memory for an instance of 800 jobs

when the memory is full, we first remove the first added solutions.

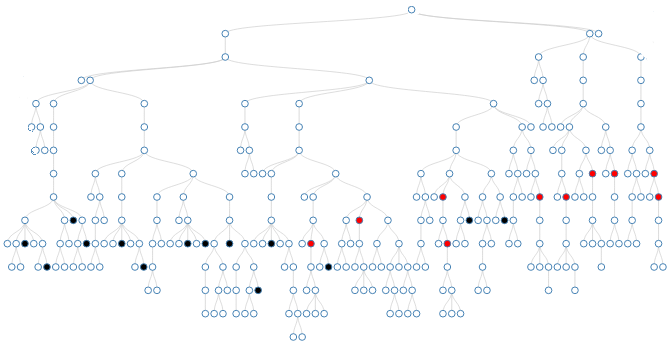
BEFO: Biggest Entry First Out
Big entries in the memory correspond to solution sequences of big problems that are on higher levels in the search tree. Since there are fewer nodes at a higher level, the possibility that a big entry be reused is reduced. Moreover, even if the big subproblem appears again and we have no choice but branch again on it, we may expect to only branch for a small number of times since then *small entries* from the memory can be used to solve the problem quickly.

At the implementation level, the memory is organized in a way such that solution sequences of different sizes are stored separately. This allows us to implement the strategy easily. At each cleaning, we also tend to clean up a large amount of space in order to decrease the time cost induced by the cleaning operation.

LUFO: Least Used First Out
In order to verify if memorized solutions are queried in a homogeneous way, we analyze the usage of memorized sequences when the input instance is solved. The result shows that most of entries stored in the memory are never queried at all, though some solutions are queried many times.

Figure 2 shows the statistical result on an instance of 800 jobs (the results on instances of other sizes basically express the same message). The horizontal axis represents the size of memorized solutions and the vertical axis is the number of solutions. More intuitively, we visualized the branching tree of an instance of 100 jobs. A snapshot of a representative part of the tree is shown in Figure 3. Both figures confirms the fact that only a small portion of memorized sequences are useful.

These observation suggest the LUFO strategy. For each stored solution, keep a counter *NbUsed* indicating the number of times it is used to prune nodes. When the memory is full, do



All drawn nodes are stored into the memory, but only red nodes are used to solve identical subproblems which are colored in black.

Fig. 3: Part of the branching tree of an instance of 100 jobs

the cleaning by removing all sequences whose counter is zero, at the same time decrease the counter by one for remaining sequences.

Now we are ready to present the final results with different cleaning strategies. Note that *Split* is removed. Strategies FIFO and BEFO allow to solve instances with up to 800 jobs (Table II). But they are completely dominated by LUFO, which allows to solve up to 1200 jobs, for which it solves all 200 instances with an average time of 192 seconds. The detailed results with different value of (R, T) is presented in Table III.

	TMin	TAvg	TMax	#Nodes	SizeMem
FIFO	0.0	60.0	3144.0	16161758	1727397
BEFO	0.0	59.0	4828.0	6356245	2006948
LUFO	0.0	19.0	275.0	5408511	1354477

TABLE II: Results for instances of size 800

R	T	TMin	TAvg	TMax	#Nodes	SizeMem
0.2	0.2	3	5	9	3015983	1266906
0.2	0.4	279	426	610	81282222	5416691
0.2	0.6	949	2460	3763	312887225	4361022
0.2	0.8	331	416	494	63333939	3157289
0.4	0.2	0	0	0	503	190
0.4	0.4	33	48	69	13978451	5865902
0.4	0.6	264	402	592	63852328	3400014
0.4	0.8	0	1	7	707379	150800
0.6	0.2	0	0	0	505	191
0.6	0.4	0	2	6	1525355	489848
0.6	0.6	59	90	138	22924336	4201467
0.6	0.8	0	0	0	2321	612
0.8	0.2	0	0	0	497	190
0.8	0.4	0	0	0	516	193
0.8	0.6	0	1	6	943904	175988
0.8	0.8	0	0	0	1158	408
1.0	0.2	0	0	0	501	189
1.0	0.4	0	0	0	502	189
1.0	0.6	0	0	0	16636	3783
1.0	0.8	0	0	0	1052	380
Global		0	192	3763	28223765	1424612

TABLE III: Result of LUFO for instances of size 1200

V. Conclusion

The power of *Memorization* is well illustrated on this problem. By taking advantage of Paradox 2 and by a careful design of memory cleaning strategy (LUFO), our algorithm is able to solve instances with up to 1200 jobs while the state-of-the-art exact algorithm was limited to 900 jobs. This exciting result suggests us to always consider *Memorization* as an essential building block when designing branching algorithm for NP-hard problems.

References

[1] J. Du and J. Y. T. Leung (1990), "Minimizing total tardiness on one machine is NP–hard", *Mathematics of Operations Research* 15, 483–495.

[2] M. Garraffa, L. Shang, F. Della Croce, V. T'Kindt (2017). An Exact Exponential Branch-and-Merge Algorithm for the Single Machine Total Tardiness Problem. HAL. https://hal.archives-ouvertes.fr/hal-01477835.

[3] L. Shang, M. Garraffa, F. Della Croce and V. T'Kindt (2016). "An Exact Exponential Branch-and-Merge Algorithm for the Single Machine Total Tardiness Problem", In Proc. PMS'16, 182-185,Valencia (Espagne).

[4] F. Della Croce, M. Garraffa, L. Shang and V. T'kindt (2015), "A branch-and-reduce exact algorithm for the single machine total tardiness problem", In Proc. MISTA2015, 879-881, Prague, Czech Republic.

[5] F. Della Croce, R. Tadei, P. Baracco and A. Grosso (1998), "A new decomposition approach for the single machine total tardiness scheduling problem", *Journal of the Operational Research Society* 49, 1101–1106.

[6] H. Emmons (1969), "One-machine sequencing to minimize certain functions of job tardiness", *Operations Research* 17, 701–715.

[7] C. Koulamas (2010),"The single-machine total tardiness scheduling problem: review and extensions", *European Journal of Operational Research*, 202, 1-7.

[8] E. L. Lawler (1977), "A pseudopolynomial algorithm for sequencing jobs to minimize total tardiness", *Annals of Discrete Mathematics* 1, 331–342.

[9] W. Szwarc, F. Della Croce, A. Grosso (1999). Solution of the single-machine total tardiness problem. Journal of Scheduling. 2:55-71.

[10] C. N. Potts and L. N. Van Wassenhove (1982), "A decomposition algorithm for the single machine total tardiness problem", *Operations Research Letters* 5, 177–181.

[11] W. Szwarc (1993), "Single machine total tardiness problem revisited", Y. Ijiri (ed.), Creative and Innovative Approaches to the Science of Management, Quorum Books, Westport, Connecticut (USA), 407–419.

[12] W. Szwarc, A. Grosso and F. Della Croce (2001), "Algorithmic paradoxes of the single machine total tardiness problem", *Journal of Scheduling* 4, 93–104.

[13] S. Chang, Q. Lu, G. Tang and W. Yu (1995). On decomposition of the total tardiness problem. Operations Research Letters, 17(5), 221-229.

New objective functions based on jobs positions for single machine scheduling with deadlines

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Thanh Thuy Tien TA, Kivin RINGARD, Jean-Charles BILLAUT
Laboratoire d'Informatique, ERL CNRS 6305,
64 avenue Jean Portalis
F-37200, Tours
Email: thanhthuytien.ta@etu.univ-tours.fr, jean-charles.billaut@univ-tours.fr

***Abstract*—In this paper, we study new single machine scheduling problems where the objective functions are based on jobs positions. The original idea is to characterize easily a huge set of optimal solutions, without enumerating them. The (polynomially solvable) scheduling problems that can be addressed by this approach are such that a simple rule exists, giving a priority between two consecutive jobs. Assuming that the solution obtained by applying the rule is sequence $Id = (J_1, J_2, ..., J_n)$ (correct after a renumbering of jobs), we search for a feasible solution (i.e. also optimal), at a maximum distance from sequence Id. Such considerations are also usefull for some robustness considerations. We identify some new objective functions and present first theoretical results and first computational experiments. Some new problems remain open.**

Key words: Scheduling, Single machine, Complexity, MILP, branch-and-bound

I. Introduction

Many scheduling problems can be solved in polynomial time by applying a simple sequencing rule, also called "priority rule" [10]. The application of such a rule indicates whether it is preferable to place a job J_i before a job J_j, or the contrary. For example, it is well known that problem $1||\sum C_j$ is solved optimally by sorting the jobs according to their nondecreasing processing time order (SPT rule [10]), the problem $1||L_{max}$ is solved optimally by sorting the jobs according to their non decreasing due dates order (EDD rule, [7]), the problem $F2||C_{max}$ is solved by the Johnson's rule ([8]), etc. (see [5] for a generalization to other scheduling problems). It is also well known that each scheduling problem generally has a lot of optimal solutions (potentially an exponential number).

The original goal is to characterize a set of optimal solutions without enumerating them. Such an approach is useful in a dynamic environment, to react in real time to unexpected events, or to data uncertainty. Some preliminary studies concerning the search of the characteristics of the optimal solutions (characteristics but not the list) have been conducted in this direction, using the lattice of permutations as support, for problems $1||L_{max}$ and $F2||C_{max}$ in [2] and for problem $F2||C_{max}$ only in [3].

In this paper, we only consider problem $1||L_{max}$. For any instance, we apply the following pre-treatment in polynomial time: (1) we apply EDD rule and we obtain L^*_{max}, the optimal maximum lateness value. Then, (2) we modify the due dates in order to obtain deadlines: $\tilde{d}_j = d_j + L^*_{max}$, for any $j \in \{1, 2, ..., n\}$, limiting the deadlines to $\sum p_j$. Finally, (3) we renumber the jobs in EDD order. At the end we know that sequence $Id = (J_1, J_2, ..., J_n)$ is feasible and that jobs are numbered in EDD order. Suppose for example that we have $n = 4$ jobs and that – after the pre-treatment – it is found a sequence $\sigma = (J_3, J_1, J_4, J_2)$ which is also a feasible sequence. Then, because of the EDD rule, it is possible by simple pairwise exchanges, to modify this sequence iteratively and to reach sequence Id. All the sequences obtained during these changes are also feasible sequences: sequence (J_1, J_3, J_4, J_2) is feasible, sequence (J_3, J_1, J_2, J_4) is feasible, etc. So, all the "predecessors" of sequence σ are feasible. These sequences are exactly those such that $J_1 \prec J_4$ and $J_1 \prec J_2$ and $J_3 \prec J_4$, i.e. are characterized by the conjunction of three simple precedence relations. Therefore, finding a feasible sequence, *as far as possible* from sequence Id, allows immediately to characterize a *big* set of feasible solutions.

Such an approach – described in details in [3] is valid for any scheduling problem for which such a rule exists such as problems $1|r_j|C_{max}$, $F2|s_{nsd}|C_{max}$, etc.

The paper is organized as follows. We describe the problem more formally in Section II, where we define some new objective functions and we give first results. In Section III we focuse on the minimization of the sum of weighted positions. We show that this problem is strongly NP-hard and we propose some exact and approximated resolution methods. These methods are evaluated by randomly generated instances in Section IV. Finally, in the section V, we present some ideas for the future research directions.

II. Problems definitions

We consider a set of n jobs $\{J_1, J_2, ..., J_n\}$ to schedule on a single machine. Preemption is not allowed and the machine can perform only one job at a time. To each job J_j is associated a processing time p_j and a deadline $\tilde{d}_j$. It is assumed that the sequence is feasible (maximum lateness equal to 0) and that the jobs are numbered in EDD order (notice that it can be done in $O(n \log n)$ time for any instance).

Let consider a sequence σ. We denote by N_j the number of jobs after J_j in σ with an index greater than j. For example, consider $\sigma = (J_3, J_1, J_4, J_2)$, we have $N_1 = 2$, $N_2 = 0$, $N_3 = 1$ and $N_4 = 0$. The sum of N_j, $\sum_{j=1}^{n} N_j$ is exactly equal to the number of simple precedence constraints in the

conjunction (here $\sum_{j=1}^{n} N_j = 3$). It also corresponds to the level of sequence σ in the lattice of permutations (see [3]).

Finding a sequence as deep as possible in this lattice, or finding a sequence as far as possible from sequence Id is the same, and is equivalent to minimise the objective function $\sum N_j$. The problem can be denoted by $1||Lex(L_{\max}, \sum N_j)$ with a notation of a bi-objective problem [11], which is equivalent to $1|\tilde{d}_j, \tilde{d}_1 \leq ... \leq \tilde{d}_n, \tilde{d}_n = \sum p_j| \sum N_j$ for the instance after pre-treatment.

A. Expression of N_j

Let us define the boolean variables $x_{j,k} = 1$ if job J_j is in position k and 0 otherwise. The expression of $\sum N_j$ is the following:

$$\sum N_j = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{i=j+1}^{n} \sum_{h=k+1}^{n} x_{i,h} x_{j,k}$$

This expression is quadratic, it can be linearized for a linear programming formulation [1], [3].

Let us define by $y_{j,k} = \sum_{i=j+1}^{n} \sum_{h=k+1}^{n} x_{i,h}$. We have:

$$\sum N_j = \sum_{j=1}^{n} \sum_{k=1}^{n} y_{j,k} x_{j,k}$$

Proposition 1: $\sum_{j=1}^{n} \sum_{k=1}^{n} y_{j,k}$ is a function related to the positions of the jobs in the sequence.

Proof. We have:

$$\sum_{j=1}^{n} \sum_{k=1}^{n} y_{j,k} = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{i=j+1}^{n} \sum_{h=k+1}^{n} x_{i,h}$$

(notice that this expression is related to, but different than the expression of $\sum N_j$).

We can show that

$$\begin{aligned} \sum_{j=1}^{n} \sum_{k=1}^{n} y_{j,k} &= -n^2 + \sum_{j=1}^{n} \sum_{k=1}^{n} jk x_{j,k} \\ &= -n^2 + \sum_{j=1}^{n} j \sum_{k=1}^{n} k x_{j,k} \end{aligned}$$

In this expression, $P_j = \sum_{k=1}^{n} k x_{j,k}$ is exactly the position of J_j in the sequence. So, minimizing $\sum_{j=1}^{n} \sum_{k=1}^{n} y_{j,k}$ is equivalent to minimize $\sum_{j=1}^{n} j P_j$.

This new indicator, P_j, and this new objective function $\sum jP_j$ never appeared in the scheduling literature before, to the best of our knowledge. Notice that this objective function is not related to the jobs completion time, which is also not frequent in the literature.

B. New objective functions and first results

Even if $\sum N_j$ and $\sum jP_j$ are not the same objective functions, we can notice that these functions are related in some sense.

We introduce the following new objective functions:

- $P_{\max} = \max_{1 \leq j \leq n} P_j$, the maximum position of a job in the sequence,
- $\sum P_j = \sum_{j=1}^{n} P_j$, the sum of positions of the jobs in the sequence,
- $\sum w_j P_j = \sum_{j=1}^{n} w_j P_j$, the weighted sum of positions of jobs in the sequence

Let us consider only single machine problems, without deadlines. We can exhibit the first following results:

- Problem $1||P_{\max}$ is trivial (similar to $1||C_{\max}$) since for any solution, we have $P_{\max} = n$. It is the same result for any sequencing problem.
- Problem $1||\sum P_j$ is trivial since in any solution we have $\sum P_j = n(n+1)/2$. Again, it is the same result for any sequencing problem.
- Problem $1||\sum w_j P_j$ is equivalent to problem $1|p_j = 1|\sum w_j C_j$, and therefore it can be solved in $O(n \log n)$ by sorting the jobs in w_j nonincreasing order.

Let us consider now the case with deadlines. We can exhibit the first following results:

- Problem $1|\tilde{d}_j|P_{max}$ is trivial because if EDD is a feasible solution, this solution has $P_{\max} = n$, which is optimal. Otherwise, there is no solution.
- Problem $1|\tilde{d}_j|\sum P_j$ is trivial for the same reasons, and EDD is such that $\sum P_j = n(n+1)/2$, which is optimal.
- Problem $1|\tilde{d}_j, p_j = 1|\sum w_j P_j$ is equivalent to problem $1|\tilde{d}_j, p_j = 1|\sum w_j C_j$, which can be solved in polynomial time [6].

We focuse in the rest of the paper on problem $1|\tilde{d}_j|\sum w_j P_j$.

III. Problem $1|\tilde{d}_j|\sum w_j P_j$

In this section, we assume that jobs are numbered in EDD order and that sequence $Id = (J_1, J_2, ..., J_n)$ is feasible and has a maximum lateness of 0. We search to minimise the weighted sum of jobs positions.

A. Complexity

Proposition 2: Problem $1|\tilde{d}_j|\sum w_j P_j$ is strongly NP-hard.

Proof. It is clear that the problem is in NP. Then, we proceed by reduction from 3-PARTITION problem.

3-PARTITION problem:
Instance: $A = \{a_i\}_{1 \leq i \leq 3m}$ a set of $3m$ elements such that $B/4 < a_i < B/2$ and $\sum_A a_i = mB$, with a bound $B \in N$.

Question: Can A be partitioned into m disjont sets $A_1, A_2, ..., A_m$ such that $\sum_{A_k} a_i = B, \forall k \in \{1, 2, ..., m\}$ (note that each A_k must contain exactly 3 elements of A)

Clearly, problem $1|\tilde{d}_j|\sum w_j P_j$ is in NP. We define $\overline{p} = |A| \times \max(A)$. We build an instance to our problem with m dummy jobs $J_i, 1 \leq i \leq m$ such that $p_i = 1, w_i = 0, \widetilde{d}_i = i(B + 3\overline{p} + 1)$. We add $3m$ regular jobs $J_{m+j}, 1 \leq j \leq 3m$ with $p_{m+j} = w_{m+j} = a_j + \overline{p}$ and $\widetilde{d}_{m+j} = m(B + 3\overline{p} + 1)$.

We first show that in an optimal solution, each job J_i is precisely at position $4i$, $1 \leq i \leq m$ and then we show that it completes exactly at its deadline if and only if the answer to 3-PARTITION is yes.

(a) We consider an optimal sequence S^*. We prove by induction that in any optimal sequence, each job $J_i, 1 \leq i \leq m$, has a position equal to $4i$.

First, we have to check it for $i = 1$.
Suppose that J_1 is in position k, and suppose that it completes at time $C_1 < \tilde{d}_1$. We denote by D the set of the first $k-1$ jobs (before J_1). We denote by J_l the job in position $k+1$, and we know that this job completes after $\tilde{d}_1$ (since otherwise swapping J_1 and J_l would lead to a better solution). We denote by Δ the difference between the starting time of J_1 and $\widetilde{d}_1$.

We have:

$$\begin{aligned}\Delta \geq 1 &\Rightarrow \widetilde{d}_1 - \sum_{J_j \in D} p_j \geq 1 \\ &\Rightarrow \widetilde{d}_1 - \sum_{J_j \in D} a_j - (k-1)\overline{p} \geq 1 \\ &\Rightarrow B + 3\overline{p} + 1 - \sum_{J_j \in D} a_j - k\overline{p} + \overline{p} \geq 1 \\ &\Rightarrow B + 4\overline{p} - \sum_{J_j \in D} a_j - k\overline{p} \geq 0\end{aligned}$$

So we have:

$$k \leq 4 + \frac{1}{\overline{p}}(B - \sum_{J_j \in D} a_j) \quad (1)$$

Because $C_l > \widetilde{d}_1$, we have:

$$\sum_{J_j \in D} a_j + (k-1)\overline{p} + 1 + a_l + \overline{p} > B + 3\overline{p} + 1$$

$$\sum_{J_j \in D} a_j + k\overline{p} + a_l > B + 3\overline{p}$$

So we have:

$$k > 3 + \frac{1}{\overline{p}}(B - \sum_{J_j \in D} a_j - a_l) \quad (2)$$

We deduce from (1), (2) that:

$$3 + \frac{1}{\overline{p}}(B - \sum_{J_j \in D} a_j - a_l) < k \leq 4 + \frac{1}{\overline{p}}(B - \sum_{J_j \in D} a_j)$$
$$\Leftrightarrow 3 + \varepsilon_1 < k \leq 4 + \varepsilon_2$$

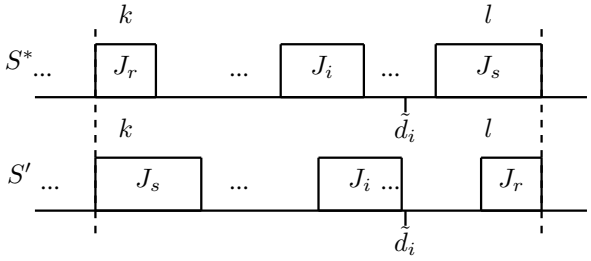


Fig. 1. Case where J_i does not complete at time $\tilde{d}_i$ $(1 \leq i \leq m)$

It shows that J_1 is necessarilly in position 4.

Suppose now that job J_i is in position $4i, \forall i, 1 \leq i \leq m-1$. We want to show that job J_{i+1} is in position $4(i+1)$. Suppose that J_{i+1} is in position k. We denote by D the jobs preceding J_{i+1}. There are $(k-1-i)$ jobs in D that are regular jobs. We have: $C_{i+1} \leq \widetilde{d}_{i+1}$

$$\Rightarrow (i+1)(B + 3\overline{p} + 1) - (\sum_{J_j \in D} a_j + (k-1-i)\overline{p}) - i - 1 \geq 0$$

$$\Rightarrow k \leq 4(i+1) + \varepsilon \quad (3)$$

Let us denote by J_l the job just after J_{i+1}. We have:

$$C_l > \tilde{d}_{i+1} \Rightarrow \sum_{J_j \in D} a_j + |D|\overline{p} + i + 1 + \overline{p} + a_l > (i+1)(B + 3\overline{p} + 1)$$

If J_{i+1} is in position k, the number of jobs before J_{i+1} is $k-1$, and before this job, there are $|D|$ regular jobs and i dummy jobs. So we have $k - 1 = |D| + i$. Hence:

$$\sum_{J_j \in D} a_j + (k-1-i)\overline{p} + i + 1 + \overline{p} + a_l > (i+1)(B + 3\overline{p} + 1) \quad (4)$$

$$\Rightarrow k > (4i+3) + \varepsilon \quad (5)$$

We deduce from (3) and (5) that $k = 4i + 4$.

(b) We must show now that there is an answer YES to 3-PARTITION problem, if and only if in an optimal solution to our problem, each dummy job completes at its deadline.

Suppose there is an answer YES to 3-PARTITION problem. In this case, we can take the partition of the jobs and put a dummy job between each partition. Because of the deadlines definition, each dummy jobs completes exactly at its deadline. Breaking the partition would either violate a deadline, or generate a weaker solution.

Suppose that we do not follow the 3-PARTITION solution and that in S^*, we have $C_i < \widetilde{d}_i$ for one job J_i with $1 \leq i \leq m$ (i.e. we have an optimal solution that is not a 3-partition). Suppose there are two jobs: J_r before J_i in position k and J_s after J_i in position l such that if the jobs are swapped, then $C_i = \widetilde{d}_i$. In this case, illustrated in Fig. 1 we must have $a_r < a_s$ and $k < l$.

We denote by S' the solution S^* except that J_r and J_s are swapped in S'. We have:

$$\begin{aligned}Z(S') - Z(S^*) &= k(a_s + \overline{p}) + l(a_r + \overline{p}) - k(a_r + \overline{p}) - l(a_s + \overline{p}) \\ &= k(a_s - a_r) + l(a_r - a_s) \\ &= (l-k)(a_r - a_s) < 0\end{aligned}$$

This quantity is negative, which means that S^* is not optimal, so this case is not possible.

Suppose now that there are two jobs J_{r_1} and J_{r_2} before J_i and two jobs J_{s_1} and J_{s_2} after J_i so that if the jobs are swapped, then $C_i = \widetilde{d}_i$. In this case, we have $a_{s_1} + a_{s_2} > a_{r_1} + a_{r_2}$ and :
$Z(S') - Z(S^*) = k_1 a_{s_1} + k_2 a_{s_2} + l_1 a_{r_1} + l_2 a_{r_2} - k_1 a_{r_1} + k_2 a_{r_2} + l_1 a_{s_1} + l_2 a_{s_2} = (k_1 - l_1)(a_{s_1} - a_{r_1}) + (k_2 - l_2)(a_{s_2} - a_{r_2})$

We know that $k_1 - l_1 < 0, k_2 - l_2 < 0$
$\Rightarrow \max(k_1 - l_1, k_2 - l_2) < 0$ and we have
$Z(S') - Z(S^*) = (k_1 - l_1)(a_{s_1} - a_{r_1}) + (k_2 - l_2)(a_{s_2} - a_{r_2})$
$< (a_{s_1} - a_{r_1}) \max(k_1 - l_1, k_2 - l_2) + (a_{s_2} - a_{r_2}) \max(k_1 - l_1, k_2 - l_2) < (a_{s_1} - a_{r_1} - a_{s_2} - a_{r_2})) \max(k_1 - l_1, k_2 - l_2) < 0$

For the same reasons, this case is not possible because S^* is optimal. By extending this reasoning to any number of jobs to exchange (same number before and after J_1 due to the position of this job), we prove that if the answer to 3-PARTITION problem is YES, we must follow the 3-PARTITION solution and therefore we have a feasible solution that is optimal.

B. Two exact resolution methods

We propose two ways to solve the problem to optimality: a linear programming formulation and a branch-and-bound algorithm.

1) MILP formulation: An MILP formulation of the problem with variables $x_{j,k}$ is the following.

$$\text{Minimize} \sum_{j=1}^{n} \sum_{k=1}^{n} w_j k x_{j,k}$$

$$\sum_{j=1}^{n} x_{j,k} = 1, \forall k \in \{1, 2, ...n\}$$

$$\sum_{k=1}^{n} x_{j,k} = 1, \forall j \in \{1, 2, ...n\}$$

$$\sum_{l=1}^{k} \sum_{j=1}^{n} p_j x_{j,l} \leq \sum_{j=1}^{n} \tilde{d}_j x_{j,k}, \forall k \in \{1, 2, ...n\}$$

This model contains n^2 binary variables and $3n$ constraints. This model will be solved by CPLEX solver.

2) Branch-and-bound: A branch-and-bound procedure is an implicite enumeration method, characterized by two elements: the branching and the bounds. Here, a node is characterized by a partial ending sequence σ composed by k jobs (from position $n - k + 1$ to n), a set S of unscheduled jobs and a lower bound. The branching consists in adding a job of S at position $n - k$ in σ.

The initial upper bound is given by the (simple) Backward Algorithm presented in Alg. 1.

The lower bound is evaluated as follows. The partial sequence σ is evaluated, and the jobs of S are sequenced in w_j non-increasing order. If the partial sequence is not feasible, the evaluation is put to HV (high value).

Let consider the following instance with $n = 5$ jobs.

Algorithm 1 BW algorithm

$\mathcal{J} \leftarrow \{J_1, J_2, ..., J_n\}$
$t \leftarrow \sum_{j=1}^{n} p_j$; $k \leftarrow n$
while $\mathcal{J} \neq \emptyset$ **do**
  Let $J_r \in \mathcal{J}$ such that $w_r = \min_{j \in \mathcal{J}/\tilde{d}_j \geq t} w_j$
  Put J_r in position k
  $t \leftarrow t - p_r$; $k \leftarrow k - 1$
  $\mathcal{J} \leftarrow \mathcal{J} \setminus \{J_r\}$
end while

j	1	2	3	4	5
p_j	1	2	5	2	4
w_j	1	2	4	2	3
$\tilde{d}_j$	7	7	12	14	14

The BW algorithm gives sequence $(J_2, J_1, J_3, J_5, J_4)$. The evaluation is equal to 38.

Let consider node $\sigma = (J_3, J_4)$ and $S = \{J_1, J_2, J_5\}$. The partial sequence σ has an evaluation equal to 16+10=26. The unscheduled jobs are sequenced in an 'optimal' (but not necesserally feasible) way, leading to sequence (J_5, J_2, J_1) evaluated by 3+4+3=10. The evaluation of this node is equal to 36.

C. Heuristic algorithms

We propose two heuristic algorithms. The first one is the previous Backward algorithm. This algorithm is important because it is very intuitive (remember that such an algorithm is optimal for problems $1|prec|f_{\max}$ [9]). We also propose a Forward algorithm described in Alg. 2 ($J_{[si]}$ is the job in position si).

Algorithm 2 FW algorithm

$\sigma = (J_1, J_2, ..., J_n)$
$si = 1$ // si stands for "smallest index"
while $P_{si} < n$ **do**
  $S' \leftarrow \{J_j / P_j > si\}$
  while $S' \neq \emptyset$ **do**
    Let $J_k \in S'$ such that $w_k = \max_{J_j \in S'} w_j$
    // σ is a sequence of type $\sigma_1 // J_{[si]} // \sigma_2 // J_k // \sigma_3$
    $\sigma' \leftarrow \sigma_1 // J_k // J_{[si]} // \sigma_2 // \sigma_3$
    if σ' is not feasible **then**
      $S' \leftarrow S' \setminus \{J_k\}$
    else
      $\sigma \leftarrow \sigma'$
      $si \leftarrow si + 1$
    end if
  end while
  $S' \leftarrow \{J_j / P_j > si\}$
  $si \leftarrow$ smallest index in S'
end while

Let us consider the previous example. We start with $\sigma = (J_1, J_2, J_3, J_4, J_5)$. We have $si = 1$ Job J_3 cannot be put before J_1 (otherwise J_2 is late). Job J_5 can be put before J_1. We obtain sequence $\sigma = (J_5, J_1, J_2, J_3, J_4)$. Then, J_2 can be put before J_1. We obtain sequence $\sigma = (J_5, J_2, J_1, J_3, J_4)$. Then, we have $si = 3$. It is not possible to put J_4 before J_3. Then, we have $si = 4$ and the procedure stops because $P_4 = 5$. The evaluation of this solution is equal to 36.

A global lower bound LB can also be easily computed by relaxing the MILP.

IV. Computational experiments

A. Data generation

The computational experiments have been run on a MacBook Air Intel Core i5 1,6 GHz, using IBM ILOG CPLEX 12.6. Two types of instances have been generated. Instances of type 1 are classical random data sets, instances of type 2 are difficult random data sets. For each type, 30 instances have been generated for each value of n, with $n \in \{10, 20, ..., 100\}$. For instances of type 1, random data sets have been generated as follows:

- $p_j \in [1, 100]$, $w_j \in [1, 100]$
- $d_j \in [(\alpha - \beta/2)P, (\alpha + \beta/2)P]$ with $P = \sum p_j$, $\alpha = 0.75$ and $\beta = 0.25$

Then, these instances receive the pre-treatment.

For instances of type 2, random data sets have been generated as follows:

- for $n' = \lfloor n/4 \rfloor$ jobs:
  - $p_j = 1$, $w_j = 0$
  - $\tilde{d}_j = 4jP/n$
- for the $(n - n')$ remaining jobs:
  - $p_j \in [1, 100]$, $w_j = w_{0j} + P$, with $w_{0j} \in [1, 100]$ and $P = \sum p_j$
  - $\tilde{d}_j = P + \lfloor n/4 \rfloor$

These instances do not need the pre-treatment.

The CPU time to solve each instance has been limited to 180 seconds.

B. Results

The computational results are reported in Tables I to IV. Tables I and II are related to instances of type I, Tables III and IV are related to instances of type II. Tables I and III give the results of the exact methods. Columns cpu and opt indicate respectively the CPU time in seconds and the number of times the method found the optimal solution in less than 180 seconds. Tables II and IV give the results of the heuristic methods. Columns best, DeltaBest and DeltaLB indicate respectively the number of times the method returns the best solution among all the methods (including the exact methods), the relative deviation to the best solution, and the relative deviation to the lower bound. Column Δ indicates the relative deviation between BW and FW:

$$\Delta = \frac{BW - FW}{BW}$$

About the exact methods (Tables I and III), one can see that the B&B algorithm is competitive with CPLEX for type I instances, both methods solving 29 problems over 30 with up to 70 jobs, and less with more jobs. However, for instances of type II, CPLEX is much more better than the B&B, solving instances with up to 40 jobs, where the B&B cannot solve instances with more than 10 jobs. It is clear that it is possible to improving the B&B algorithm, by pruning some nodes by using some dominance conditions.

	B&B		CPLEX	
n	cpu	opt	cpu	opt
10	0,01	30	0,11	30
20	0,03	30	0,15	30
30	0,09	30	0,42	30
40	0,18	30	0,99	30
50	0,54	30	3,22	30
60	0,81	30	7,22	30
70	11,87	29	26,45	29
80	53,11	23	52,17	26
90	87,15	18	86,94	23
100	139,17	11	145,93	10

TABLE I. Results of the exact methods for Type I instances

	BW			FW			Δ
n	best	DeltaBest	DeltaLB	best	DeltaBest	DeltaLB	BW/FW
10	28	0,08%	2,28 %	28	0,04%	2,24%	0,04%
20	23	0,08%	1,37 %	28	0,02%	1,31%	0,06%
30	18	0,19%	1,51 %	24	0,03%	1,35%	0,16%
40	6	0,44%	1,52 %	16	0,04%	1,09%	0,41%
50	3	0,24%	1,72 %	11	0,11%	1,59%	0,13%
60	6	0,15%	1,46 %	8	0,08%	1,38%	0,07%
70	3	0,24%	1,22 %	4	0,1%	1,08%	0,14%
80	3	0,28%	1,2 %	2	0,17%	1,08%	0,12%
90	3	0,35%	1,22 %	3	0,15%	1,01%	0,20%
100	0	0,29%	1,08 %	6	0,09%	0,87%	0,20%

TABLE II. Results of the heuristic methods for Type I instances

About the heuristic methods (Tables II and IV), FW algorithm returns better solution than BW algorithm. For instances of type I, BW returns solution that are not so bad, with a relative deviation to the best solution under 0.44%, whereas FW is under 0.17%. Note that the relative deviation between BW and FW is in favor of FW, but with a percentage less than 0.41%. However, for instances of type II, BW is dominated by FW. BW returns solutions that are far from the optimal solution, and the relative deviation between BW and FW is around 10%.

V. Conclusion and perspectives

We identify in this paper a new category of scheduling problems, with the definition of new objective functions, based on jobs positions. Some trivial problems are identified, and a problem with weights is proved strongly NP-hard. Some problems remain open, such as the $1|\tilde{d}_j|\sum jP_j$ problem and the original $1|\tilde{d}_j|\sum N_j$ problem.

For the difficult problem, heuristic algorithms and exact methods are proposed and evaluated on randomly generated instances. The results of the exact methods and of the heuristic algorithms are relatively good for the first type instances, where data are generated 'as usual' in the scheduling literature. However, for more difficult instances, the limits of the exact methods are reached, and the difference between the heuristic algorithms is more clear.

The perspectives consist in continuing the investigations for the open problems, and in the improvment of the exact methods – introducing cuts and dominance conditions – and of the heuristic methods, by using metaheuristic algorithms.

	B&B		CPLEX	
n	cpu	opt	cpu	opt
10	0,12	30	0,18	30
20	178,18	1	1,57	30
30	180	0	9,85	30
40	180	0	50,87	28
50	180	0	174,78	5
60	180	0	180,06	0
70	180	0	180,06	0
80	180	0	180,08	0
90	180	0	180,11	0

TABLE III. RESULTS OF THE EXACT METHODS FOR TYPE II INSTANCES

	BW			FW			Δ
n	best	DeltaBest	DeltaLB	best	DeltaBest	DeltaLB	BW/FW
10	0	8,44%	21,5 %	1	1,26%	12,62%	7,26%
20	0	7,84%	20,8 %	0	0,31%	11,66%	7,56%
30	0	10,5%	23,43 %	0	0,34%	10,85%	10,19%
40	0	10,55%	23,55 %	0	0,39%	10,95%	10,20%
50	0	9,74%	22,39 %	0	0,48%	11%	9,31%
60	0	10,66%	23,35 %	0	0,35%	10,59%	10,34%
70	0	10,49%	23,47 %	0	0,46%	11,04%	10,07%
80	0	13,6%	27,22 %	0	0,52%	10,5%	13,14%
90	0	8,85%	21,88 %	0	0,25%	11,38%	8,62%

TABLE IV. RESULTS OF THE HEURISTIC METHODS FOR TYPE II INSTANCES

REFERENCES

[1] Billaut J-C., Lopez P. and Takouti L, *A method for characterizing the set of optimal sequences for some scheduling problems*. 12th Congress of the French OR Society (ROADEF'2011), Saint-Etienne, March 2011.

[2] Billaut J-C., Lopez P., *Characterization of all $\rho-$approximated sequences for some scheduling problems*. IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, art. no. 6059026, 2011.

[3] Billaut J-C., Hebrard E. and Lopez P., *Complete Characterization of Near-Optimal Sequences for the Two-Machine Flow Shop Scheduling Problem*, Ninth International Conference on Integration of Articial Intelligence and Operations Research Techniques in Constraint Programming (CPAIOR'2012), Nantes, June 2012.

[4] Blazewicz J, Ecker K.H., Pesch E., Schmidt G., Weglarz J., *Scheduling Computer And Manufacturing Processes*. Springer, 2nd edition, 2007.

[5] Bouquard J-L., Lente C, Billaut J-C., *Application of an optimization problem in Max-Plus algebra to scheduling problems*, Discrete Applied Mathematics, 154(15), pp. 2041-2238, 2006.

[6] Chen C-L., Bulfin R.L., *Scheduling unit processing time jobs on a single machine with multiple criteria*, Computers & Operations Research, 17(1), pp. 1-7, 1990.

[7] Jackson J-R., *Scheduling a production line to minimize maximum tardiness*. Research report 43, Management Science Research Project, University of California, Los Angeles, 1955.

[8] Johnson S.M., *Optimal two- and three-stage production schedules with setup times included*. Naval Research Logistics Quarterly, 1, pp. 61-68, 1954.

[9] Lawler E., emphOptimal sequencing of a single machine subject to precedence constraints. Management Science, 19(8), 544–546, 1973.

[10] Smith W.E, *Various optimizers for single stage production*. Naval Research Logistics Quarterly, 3(1-2):59-66, 1956.

[11] T'Kindt V., Billaut J-C., *Multicriteria scheduling. Theory, models and algorithms*, Springer, 2nd edition, 2006.

Robust parallel machine scheduling with job splitting under uncertain processing times

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Widad Naji, Marie-Laure Espinouse and Van-Dat Cung
Univ. Grenoble Alpes, G-SCOP, F-38000 Grenoble, France
CNRS, G-SCOP, F-38000 Grenoble, France
Email: firstname.lastname@grenoble-inp.fr

***Abstract*—The uncertainty of processing times is prevalent in scheduling applications. In this paper, we focus on the makespan minimization on parallel machines (identical, uniform and unrelated) with splitting under uncertain processing times. The splitting is common to many real-life applications. Under uncertainty, the classical scheduling algorithms lead to solutions that are infeasible or suboptimal under the actual realizations of processing times. Thus, we will investigate the problem under discrete scenarios representation of uncertain processing times. We propose to use the min-max objective to generate robust schedules minimizing the worst-case makespan over all the scenarios. For the unrelated parallel machine case, we show that the direct application of robust optimization techniques leads to no solution. To overcome this difficulty, we propose a formulation in which the uncertainty appears in the objective instead of the assignment constraints. Thus, we provide robust solutions in polynomial time and we extend these results to the special cases of uniform and identical parallel machines. Some numerical experiments will be carry out to illustrate these new proposals.**

I. Introduction

In this paper, we focus on the makespan minimization on parallel machines with splitting under uncertain processing times. Parallel Machine Scheduling (PMS) consists in sequencing n jobs on m same function machines in order to optimize the objective. The processing time of job j on machine i represents the duration required by machine i to process job j and is denoted p_{ij}. PMS is a very common problem in many manufacturing industries such as textile industry, semiconductors manufacturing, in multiprocessors computing application, *etc.* [1] proposed a three-field notation $\alpha/\beta/\gamma$ where α describes the machine environment, β describes job characteristics and γ describes optimality criteria. In parallel machine scheduling, we distinguish three environment machines. In the parallel identical machines ($\alpha = P$), the job processing time is independent of the machine where it is processed. In the uniform parallel machines ($\alpha = Q$), each machine has a different speed. The processing time is equal to the processing requirement divided by the speed of the machine. In unrelated parallel machines ($\alpha = R$), no particular relationship exists between the processing times on the different machines, i.e., there is no proportionality between the processing time of a job on a given machine and the time of the same job on another machine. When the processing requirement of a job is considered as its total demand, we can split the jobs into sub-jobs that can be processed independently on the machines. We accept to overlap the sub-jobs so as to finish the processing requirement as soon as possible. The splitting takes often place in parallel machine environment. [2] referred to this process as lot-streaming or lot-sizing and denote the job characteristic as $\beta = Split$ to distinguish the splitting assumptions from the non overlapping preemptive assumptions where $\beta = pmtn$). The makespan minimization is denoted $\gamma = C_{max}$.

From a mathematical point of view, the unrelated parallel machines represent the general case of PMS. The makespan minimization on m unrelated parallel machine with splitting $Rm/Split/C_{max}$ is a relaxed version of the preemptive problem $Rm/pmtn/C_{max}$. [3] have shown that the problem $Rm/pmtn/C_{max}$ can be solved in polynomial time. Likewise, the problem $Rm/Split/C_{max}$ is polynomial as shown in [2]. However, a nominal solution calculated according to the deterministic model will have poor performance under uncertain processing times.

In a lot of activities, experts are able to describe uncertainty realizations by the construction of a set of discrete scenarios that represent several futures reflecting economic and/or technological possibilities. Therefore, we will investigate the problem $Rm/Split/C_{max}$ under discrete scenario representation of uncertain processing times. We propose to use the min-max objective to generate robust schedules minimizing the worst-case makespan.

The main contribution of this paper is to provide a formulation to compute the robust schedule of $Rm/Split/C_{max}$ under discrete processing time scenarios in a polynomial time. We provide also robust solutions for the special cases of uniform and identical parallel machines. In section II, we review the existing approaches to deal with uncertainty in scheduling. We focus on the proactive scheduling approach based on Robust Optimization (RO) models under discrete representation of uncertain data. We report some complexity results of robust scheduling problems under discrete processing time scenarios. In section III, we present the static formulation of $Rm/Split/C_{max}$ and we formulate its robust counterpart by the direct application of RO techniques. We show that this leads to no feasible solution. Thus, we propose a formulation to generate the robust schedule that minimizes the worst-case makespan. Based on this formulation, we show that the robust makespan minimization on unrelated parallel machine with jobs splitting can be solved in polynomial time. We draw the same results for special cases. Section IV describes the computational experiments and results in which we focus on the gap between the robust makespan and the nominal deterministic makespan. Section V summarizes the main contributions and provides some guidelines for future studies.

II. Robust scheduling: Literature Review

There exists a variety of approaches to deal with data uncertainty in scheduling problems. [4] distinguish mainly three scheduling approaches: Proactive approach, reactive approach, and proactive-reactive approach. The proactive approach is one of the most extensively studied. Its particular strengths are the anticipation of the uncertainty, and the generation of baseline schedules that does not need to be repaired to meet the realization of uncertain processing times. This proactive baseline schedule is of crucial importance for decisions visibility in uncertain environment since it gives a starting point for communication and coordination between the production, suppliers and customers.

The proactive approach includes both robust and stochastic models. Stochastic models aim to provide schedules that perform in average when probability distributions of uncertain processing times are known and can be modelled while robust models aim to provide solutions with the best performance in the worst-case under any potential realization of uncertain data. Because producing probability distributions is not trivial for many decision-makers, due to deficient data and interdependencies between uncertainty factors, robust models represent a good alternative to stochastic models. Robust models use deterministic set-based representations of uncertainty, and robust objectives as min-max or min-max regret to provide robust solutions. Min-max and min-max regret were originally defined in the strategic decision-making literature and introduced later in operations research problems by [5]. A large part of robust optimization literature was developed since the publication of the works of [6], [7] in the late 1990's. Those works have developed many min-max and min-max regret models along with tractable solving techniques for both discrete and continuous uncertainty set representations, especially in linear cases. This has made robust models computationally more attractive than stochastic models (see [8]).

In robust scheduling applications under uncertain processing times, the use of discrete scenarios to represent uncertain data is very practical. Under this uncertainty representation, min-max objectives generate solutions minimizing the worst-case performance over all the scenarios. The worst-case performance measures the maximal cost of each solution and it is used when protection measures are indispensable. For many polynomial solvable scheduling problems, their corresponding robust versions (min-max or min max regret) are weakly or strongly NP-hard under discrete processing time scenario representation even in the case of single machine. For instance, the single machine scheduling problem with a total flow time criterion denoted $1//\sum c_j$ is studied in [9], the authors have shown that its min-max version is NP-hard even in the case of 2 scenarios. These results are confirmed in [10] where the authors propose a dynamic algorithm, as well as polynomial heuristics to solve it. But, the complexity of these methods increases with the number of discrete scenarios. In the more general case, a cutting-plane algorithm for solving the min-max version of $(1//\sum w_j c_j)$ is proposed in [11]. The computational complexity for several single machine scheduling problems where some job characteristics are uncertain is investigated in [12]. Many cases were shown to be polynomial for general maximum cost functions f_{max}. Under discrete processing time scenarios, the authors have shown that min max version of $(1|prec|f_{max})$ are solved to optimality in polynomial time by a MinMax-Lawler algorithm. The two machine permutation flow shop problem with makespan minimization denoted $F2//C_{max}$ is considered in [7] and [13]. $F2//Cmax$ is known to be polynomially solvable by using Johnson's algorithm but, it is shown that its min-max version is strongly NP-hard even for two scenarios ([13]). Other uncertainties were considered in scheduling problems as due dates (d_j) and release dates (r_j) uncertainties (see [12], [14], *etc.*). All these studies have focused either on single-machine or on flow-shop scheduling problems. However, the min-max parallel machine scheduling problems have not been investigated under the discrete scenarios. To the best of our knowledge, only [15] have addressed the robust makespan minimization on identical parallel machines $Pm//C_{max}$ with interval processing times, and [16] have considered the robust total flow time minimization on uniform parallel machine $Qm//\sum c_j$ with interval processing times, they have proposed a model to identify a robust schedule that minimizes the maximum deviation from the optimal total flow time.

Parallel machine with split jobs are used to model various applications in environments where the uncertainty of processing times is significant. Nevertheless, the generation of robust solutions in a proactive way under such constraints is rarely addressed in the literature. In the next section, we will apply RO techniques to compute these robust solutions.

III. Robust makespan minimization on parallel machines with job splitting under uncertain processing times

In this section, we will address the problem of robust makespan minimization under uncertain processing times on parallel machines. We consider the special case in which jobs can be split into numerous sub-jobs. We consider first the general parallel machine environment i.e. the unrelated parallel machines and then we address the special cases of uniform and identical parallel machines.

A. Robust makespan minimization on unrelated parallel machines

The splitting scheduling problem on unrelated parallel machines with makespan minimization is defined as follows. Given n independent jobs allowed to be processed with any of the m unrelated parallel machines as :

- Each machine i has a job dependent speed v_{ij} to accomplish the processing requirement p_j of job j. The processing times p_{ij} depends on job j and machine i so as: $p_{ij} = p_j/v_{ij}$. They are arbitrary for all i and j.
- Each job can be split into continuous sub-jobs and processed independently on the m machines with allowed parallelism so as to finish the processing of all demands as soon as possible.

For a high utilization of the machines, the objective is to determine the optimal schedule so as to minimize the makespan C_{max}.

The most common sources of uncertainty in the aforementioned scheduling problem are the demand of product (cus-

tomer order quantity) represented by the processing requirement p_j of job j and, the capability of machine represented by speed v_{ij}. One or both of these sources of uncertainty leads to the uncertainty of processing times p_{ij}.

The deterministic formulation of the problem $Rm/Split/C_{max}$ is based on the linear formulation of $Rm/pmtn/C_{max}$ given in [3]. We adapt their formulation to the split problem by relaxing the constraints that prohibit the parallelism of same job parts. The resulting LP is as follows:

Minimize C_{max}

Subject to

$$\sum_{i=1}^{m} \frac{t_{ij}}{p_{ij}} = 1, \qquad j = 1, \ldots, n, \ s = s_1, \ldots, s_k \quad (1)$$

$$\sum_{j=1}^{n} t_{ij} - C_{max} \leq 0, \qquad i = 1, \ldots, m \quad (2)$$

$$0 \leq t_{ij}, \qquad i = 1, \ldots, m, \ j = 1, \ldots, n \quad (3)$$

where each temporal variable t_{ij} is the total amount of time spent by job j on machine i and C_{max} is the makespan which is the maximum job completion time over the m machines.

The objective function minimizes the makespan. Constraints (1) ensure that a job j is completed, while constraints (2) ensure that the total time of running machine i never exceeds C_{max}. Lastly, constraints (3) set the non-negativity of temporal variables t_{ij}.

This problem is indeed an assignment problem where all sequences that respect the optimal assignment are optimal schedules. The goal is to determine the size of each sub-lot and the machine to be assigned to so as to complete all the jobs as soon as possible. This problem can be solved in polynomial time. But, in environments with significant processing time uncertainty, the deterministic model will not respect the constraints of the actual realizations of processing times. To insure the feasibly, the predictive solution can yield poor performance. Indeed, the deterministic problem $Rm/Split/C_{max}$ admits an LP formulation with only assignment constraints. Under uncertain processing times, those constraints are uncertain. Consequently, the nominal solution has poor performance. Therefore, we can apply some techniques developed by Robust Optimization (RO) literature to generate schedules which are insensitive to the uncertainty of processing times.

When no probability distribution is available and multiple futures are possible, the discrete scenario representation is a practical representation of the uncertain processing times. Then, we assume that the uncertainty of processing times is represented by a set of discrete scenarios $S = \{s_1, s_2, \ldots, s_k\}$. Each scenario $s \in S$ represents a potential realization of the uncertain processing times and captures a situation that might occur. We denote by p_{ij}^s the processing time of job j on machine i under scenario s, and C_{max}^{s*} the optimal makespan under scenario s. Based on the deterministic formulation of $Rm/Split/C_{max}$, the uncertainty of processing times is affecting the data in the equality constraints (1) while the objective and the inequality constraints are not uncertain. Thus, the direct application of RO techniques leads to a constraint feasibility problem where the objective is to find the schedule that is feasible under any scenario as expressed in the robust constraints (4).

$$\sum_{i=1}^{m} \frac{t_{ij}}{p_{ij}^s} = 1, \quad j = 1, \ldots, n, \ s = s_1, \ldots, s_k, \qquad (4)$$

In this formulation, the robust problem has no solution since this equality cannot be insured for different scenarios. We are not able to find a schedule that respect the requirement of all scenarios. To overcome this difficulty, we propose to change the decision variables in the deterministic problem. We will consider x_{ij} as the ratio of participation of machine i in the processing of job j: x_{ij}=$\frac{t_{ij}}{p_{ij}}$.

We will denote M the set of m machines, N the set of n jobs, S the set of k discrete scenarios, and F the set of all feasible solutions under the constraints of the deterministic LP_2 *i.e*

$$F = \{(x_{ij})_{i \in M, j \in N} \in [0,1]^{mn} / \sum_{i=1}^{m} x_{ij} = 1, \forall j \in N\}.$$

Let X be an element of F the set of feasible solutions. The completion time of machine i when the solution X is applied to the scenario s is defined as

$$C_i^s(X) = \sum_{j=1}^{n} p_{ij}^s x_{ij},$$

and the makespan is defined as

$$C_{max}^s(X) = \max_{i \in M} \{C_i^s(X)\}.$$

The worst-case makespan of solution X over all the scenarios of S, denoted by $Z_a(X)$, is then :

$$Z_a(X) = \max_{s \in S} \{C_{max}^s(X)\}.$$

The discrete min-max version of $Rm/Split/C_{max}$ under uncertain processing times can be formulated as the problem of finding X the robust optimal solution in absolute sense, i.e the solution minimizing the worst-case makespan among all feasible solutions $X \in F$. It can be stated as

$$\min_{X \in F} \max_{s \in S} \max_{i \in M} \{C_i^s(X)\}.$$

Thus, the discrete min-max version of $Rm/Split/C_{max}$ under uncertain processing times could be formulated as the following mathematical model:

Minimize $\max_{s \in S} \max_{i \in M} \{C_i^s(X)\}$

Subject to

$$\sum_{i=1}^{m} x_{ij} = 1, \qquad j = 1, \ldots, n$$
$$0 \leq x_{ij} \leq 1, \qquad i = 1, \ldots, m, \ j = 1, \ldots, n.$$

This formulation is nonlinear. We linearize $\max_{i \in M} \{C_i^s(X)\}$ by introducing a variable C_{max}^s such as

$$C_i^s(X) \leq C_{max}^s, \quad i = 1, \ldots, m, \quad s = s_1, \ldots, s_k,$$

which leads to the following (still nonlinear) formulation:

$$\text{Minimize } \max_{s\in S}\{C^s_{max}\}$$

Subject to

$$\begin{aligned}
\sum_{i=1}^{m} x_{ij} &= 1, && j = 1,\dots,n\\
\sum_{j=1}^{n} p^s_{ij}x_{ij} - C^s_{max} &\le 0, && i = 1,\dots,m,\ s = s_1,\dots,s_k\\
0 \le x_{ij} &\le 1, && i = 1,\dots,m,\ j = 1,\dots,n
\end{aligned}$$

We linearize now $\max_{s\in S}\{C^s_{max}\}$ by introducing a variable y such as

$$C^s_{max} \le y, \quad s = s_1,\dots,s_k.$$

Hence, we obtain the following formulation:

$$\text{Minimize } y$$

Subject to

$$\begin{aligned}
\sum_{i=1}^{m} x_{ij} &= 1, && j = 1,\dots,n\\
C^s_{max} &\le y, && s = s_1,\dots,s_k\\
\sum_{j=1}^{n} p^s_{ij}x_{ij} - C^s_{max} &\le 0, && i = 1,\dots,m,\ s = s_1,\dots,s_k\\
0 \le x_{ij} &\le 1, && i = 1,\dots,m,\ j = 1,\dots,n.
\end{aligned}$$

C^s_{max} is an intermediate variable that could be eliminated, and then discrete min-max version of $Rm/Split/C_{max}$ under uncertain processing times can be solved to optimality by the linear program bellow:

$$\text{Minimize } y$$

Subject to

$$\begin{aligned}
\sum_{i=1}^{m} x_{ij} &= 1, && j = 1,\dots,n\\
\sum_{j=1}^{n} {p_{ij}}^s x_{ij} &\le y, && i = 1,\dots,m,\ s = s_1,\dots,s_k\\
0 \le x_{ij} &\le 1, && i = 1,\dots,m, j = 1,\dots,n.
\end{aligned}$$

The uncertain processing times appear in the inequality constraints instead of equality constraints which make the model solvable. X^* is the optimal robust assignment solution and y^* is the optimal worst case makespan.

The robust discrete min-max version of $(Rm/Split/C_{max})$ remains an LP. To get the robust schedule duration of each scenario s, we should compute the matrix T^s as: t^s_{ij}=$p^s_{ij}x^*_{ij}$.

B. Robust makespan minimization on uniform and identical parallel machines with splitting

When the parallel machines are uniform, the machine speeds are not job dependent:

$$p_{ij} = p_j/v_i, \forall i \in M, \forall j \in N.$$

where v_i is the speed of machine i and p_j the processing requirement of job j.

If the uncertainty of processing times is due to the uncertainty of the processing requirements (p_j), we have the following results:

The min-max version of $(Qm/Split/C_{max})$ under discrete processing requirement scenarios could be solved to optimality in $O(n|S|)$ by mean of the optimal deterministic solution X^* where $x^*_{ij} = \frac{v_i}{\sum_{i=1}^{m} v_i}$, applied to the worst-case scenario s_w where the total of job processing requirements is maximal.

Indeed, in the deterministic case, there exists an optimal solution that consists in splitting each job into m sub-jobs, the size of each sub-job of job j to assign to machine i is equal to the ratio $\frac{v_i}{\sum_{i=1}^{m} v_i}$. The optimal assignment solution X^* of $Qm/Split/C_{max}$ in the deterministic case depends only on machine speeds.

$$x^*_{ij} = \frac{v_i}{\sum_{i=1}^{m} v_i}, \forall i \in M, \forall j \in N.$$

Consequently, the optimal deterministic makespan is such as:

$$\begin{aligned}
C^*_{max}(X^*) &= \sum_{j=1}^{n} p_{ij}x^*_{ij}\\
&= \sum_{j=1}^{n} (\frac{p_j}{v_i})\frac{v_i}{\sum_{i=1}^{m} v_i}\\
&= \frac{\sum_{j=1}^{n} p_j}{\sum_{i=1}^{m} v_i}
\end{aligned}$$

The deterministic makespan calculation is done in $O(n)$.

If the processing times are uncertain due to the uncertainty of job requirements, we will denote:

$$p^s_{ij} = (p^s_j)/v_i$$

where p^s_j is the processing requirement of job j under scenario s.

Given that the optimal solution X^* depends only on machine speeds, X^* is an optimal solution for any scenario from the set of scenarios S. The worst-case makespan of X^* is then:

$$\begin{aligned}
\max_{s_k\in S} C^{s_k}_{max}(X) &= \max_{s_k\in S} \frac{\sum_{j=1}^{n} p^{s_k}_j}{\sum_{i=1}^{m} v_i}\\
&= \frac{1}{\sum_{i=1}^{m} v_i} \max_{s_k\in S} \sum_{j=1}^{n} p^{s_k}_j.
\end{aligned}$$

Based on this equality, the worst-case scenario s_w is the scenario where the total quantity of demand over jobs is maximal. The computation of s_w is done in $O(n|S|)$.

The same results apply for identical parallel machines where $p_{ij} = p_j$. We could only consider that for each i: v_i=1 to produce the same proof.

TABLE I. CPU TIME FOR THE WORST-CASE MAKESPAN COMPUTATION UNDER DIFFERENT PROBLEM SIZE

		CPU(ms)							
m	n	$\|S\|=3$	$\|S\|=6$	$\|S\|=12$	$\|S\|=24$	$\|S\|=48$	$\|S\|=64$	$\|S\|=96$	$\|S\|=128$
7	10	2.1	3.4	3.7	5.8	7.3	9.6	37.9	57.9
	50	10.0	11.5	12.7	46.5	66.0	93.0	154.65	172.65
	200	35.2	45.7	119.3	155.9	468.0	591.2	791.2	1287.8
15	15	3.1	5.4	7.5	9.2	9.3	11.6	56.2	89.9
	50	30.0	45.5	62.7	96.2	116	154	278.5	472.6
	200	85.2	208.9	225.9	334.5	881.4	1324.9	1748.5	3400.8

IV. COMPUTATIONAL EXPERIMENTS AND RESULTS

In the numerical experiments, we focus on the unrelated parallel machines. We consider different problem sizes m=3,7,15, n=10,50,200 and $|S|$=3,6,12,24,48,96. The number of scenarios was increased by completing progressively the previous list. Processing times were randomly generated for each scenario from a uniform distribution in the interval I=$[5, 20]$. We have generated 100 replications for each combination of m, n, $|S|$.

First, we report the results regarding the computational times for varying problem sizes in Table I. The CPU times illustrate clearly the effect of the problem size on the CPU times. We notice that the computational time increases as the size of the deterministic problem increases and as the number of considered scenarios increases. We can also deduce from the CPU time table that the proposed approach is able to provide robust solution for small-sized, medium-sized and high-sized instances with modest computational efforts.

Secondly, we will measure the cost of robustness of a given solution X. For this purpose, we measure the relative deviation between its worst-case makespan $Z_a(X)$ and the nominal makespan $C_{max}^{nominal}$. We call this relative deviation "the worst-case makespan gap". We denoted as $GAP(\%)$ and we computed as follows:

$$Gap(\%) = \frac{Z_a(X) - C_{max}^{nominal}}{C_{max}^{nominal}}$$

The curves in figures (Fig. 1) show that the optimal worst-case makespan gap increases as the number of discrete scenarios increases (the curves being increasing functions). It is intuitively true that the more we need to be protected against uncertainty, the more we need to pay a robustness cost. However, we can observe that the cost of robustness measured through the Gap(%) is not a linear function in the number of discrete scenarios. This shows that we can cover more uncertainty while we pay only a marginal cost. We observe also that the curve evolution according to the number of discrete scenarios conserve the same trends for different problem sizes. Likewise, the optimal worst-case makespan gap increases as the number of machines increases when the number of scenarios is important. For example, when we schedule 50 jobs on 3 machines, the cost of robustness is comprised between 4% and 12%. When we schedule the 50 jobs on 7 machines, the cost of robustness is comprised between 4% and 15%. Moreover, on 15 machines, the cost of robustness is comprised between 4% and 16%. In contrary, the optimal worst-case makespan gap decreases as the number

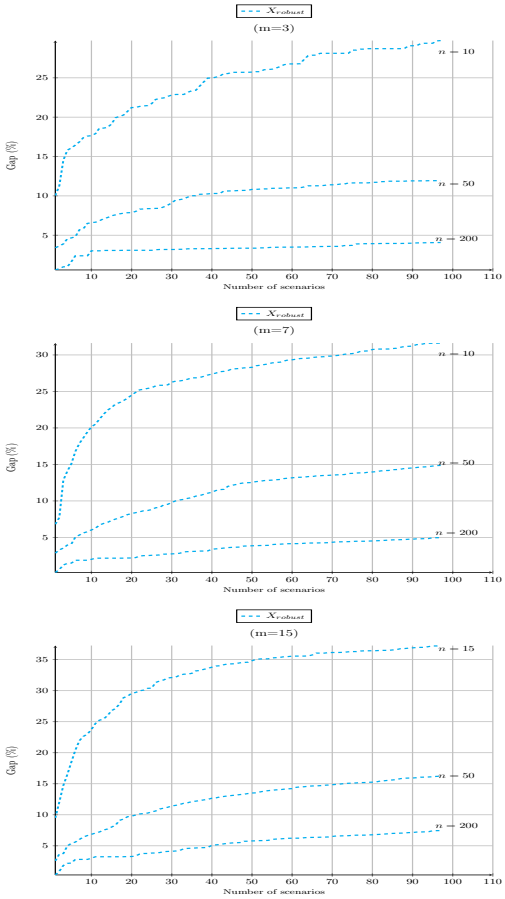


Fig. 1. The optimal robust solution cost of robustness $(GAP\%)$ evolution under different problem sizes

of jobs increases. For 3 machines, the robustness cost of the optimal robust solution under 10 jobs is comprised between 10% and 30% while it is comprised between 4% and 12% under 50 jobs, and its less than 5% for 200 jobs.

Lastly, a computational experiment is performed to compare the robustness cost of the robust optimal solution and some common industrial strategy solutions. In this experiment, we construct the artificial scenario s_{max} and the artificial scenario $s_{average}$ as described in [17]. We compute respectively their optimal assignment solutions and their worst-case makespan. The gaps of X_{max}, $X_{average}$ and X_{robust} are represented in Fig. 2. The curves show that the robustness cost

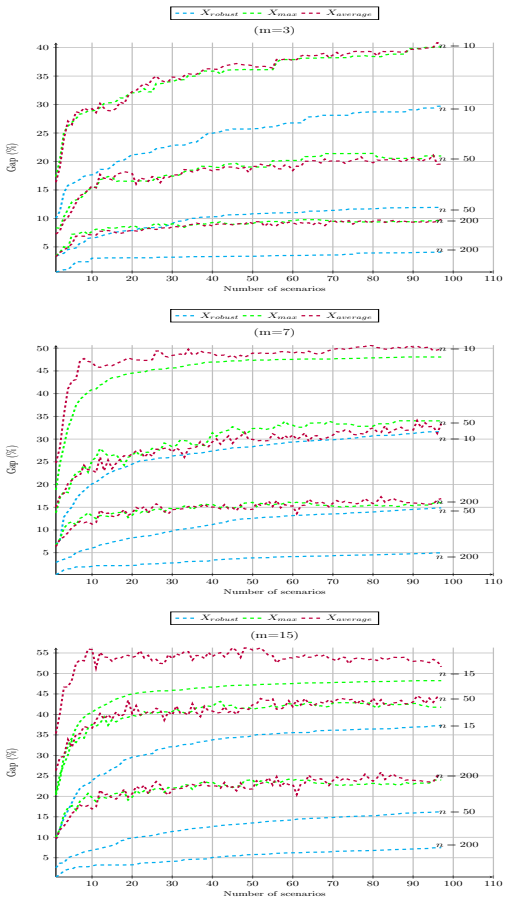


Fig. 2. Solution comparison according to the cost of robustness ($GAP\%$) for different problem sizes

of optimal robust solution is the most interesting. Compared with $X_{average}$ and X_{max}, the optimal robust solution X_{robust} provides protection against the uncertain processing times with less expensive cost of robustness for different problem sizes and under a wide range of scenarios (under 7 machines and 15 machines, the deviation between the cost of robustness when we apply X_{robust} a the cost of robustness when we apply the other strategies is very significant).

V. Conclusion

In this work, we proposed an approach to address parallel machine scheduling with splitting under uncertain processing times for $\alpha \in \{Pm, Qm, Rm\}$. We investigate first the most general problem of unrelated parallel machines $Rm/Split/C_{max}$ under discrete processing time scenarios and we propose to generate robust solutions by minimizing the worst-case makespan. The robust solutions are immune against uncertainties. This result is extended to the special cases of uniform (Qm) and identical (Pm) parallel machines. Our computational results show that this approach provides an effective way to address these scheduling problems under uncertain processing times, to produce robust schedules and to solve real-world problems with a large number of scenarios. Furthermore, the proposed approach can address many uncertain scheduling problems. It can be applied to address other scheduling problems with uncertain processing times, uncertain due dates, and/or uncertain ready times, *etc.*

References

[1] R. L. Graham, E. L. Lawler, J. K. Lenstra, and A. R. Kan, "Optimization and approximation in deterministic sequencing and scheduling: a survey," *Annals of discrete mathematics*, vol. 5, pp. 287–326, 1979.

[2] W. Xing and J. Zhang, "Parallel machine scheduling with splitting jobs," *Discrete Applied Mathematics*, vol. 103, no. 1, pp. 259–269, 2000.

[3] E. L. Lawler and J. Labetoulle, "On preemptive scheduling of unrelated parallel processors by linear programming," *Journal of the ACM (JACM)*, vol. 25, no. 4, pp. 612–619, 1978.

[4] J.-C. Billaut, A. Moukrim, and E. Sanlaville, *Flexibility and Robustness in Scheduling*, J. Wiley and sons, Eds. Wiley Online Library, 2008.

[5] M. J. Rosenblatt and H. L. Lee, "A robustness approach to facilities design," *International Journal of Production Research*, vol. 25, no. 4, pp. 479–486, 1987.

[6] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations research*, vol. 43, no. 2, pp. 264–281, 1995.

[7] P. Kouvelis, R. L. Daniels, and G. Vairaktarakis, "Robust scheduling of a two-machine flow shop with uncertain processing times," *IIE Transactions*, vol. 32, no. 5, pp. 421–432, 2000.

[8] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*, ser. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[9] R. L. Daniels and P. Kouvelis, "Robust scheduling to hedge against processing time uncertainty in single-stage production," *Management Science*, vol. 41, no. 2, pp. 363–376, 1995.

[10] J. Yang and G. Yu, "On the robust single machine scheduling problem," *Journal of Combinatorial Optimization*, vol. 6, no. 1, pp. 17–33, 2002.

[11] I. R. de Farias, H. Zhao, and M. Zhao, "A family of inequalities valid for the robust single machine scheduling polyhedron," *Computers & Operations Research*, vol. 37, no. 9, pp. 1610–1614, 2010.

[12] M. A. Aloulou and F. Della Croce, "Complexity of single machine scheduling problems under scenario-based uncertainty," *Operations Research Letters*, vol. 36, no. 3, pp. 338–342, 2008.

[13] A. Kasperski, A. Kurpisz, and P. Zieliński, "Approximating a two-machine flow shop scheduling under discrete scenario uncertainty," *European Journal of Operational Research*, vol. 217, no. 1, pp. 36–43, 2012.

[14] H. Aissi, M. A. Aloulou, and M. Y. Kovalyov, "Minimizing the number of late jobs on a single machine under due date uncertainty," *Journal of Scheduling*, vol. 14, no. 4, pp. 351–360, 2011.

[15] X. Xu, W. Cui, J. Lin, and Y. Qian, "Robust makespan minimisation in identical parallel machine scheduling problem with interval data," *International Journal of Production Research*, vol. 51, no. 12, pp. 3532–3548, 2013.

[16] X. Xu, J. Lin, and W. Cui, "Hedge against total flow time uncertainty of the uniform parallel machine scheduling problem with interval data," *International Journal of Production Research*, vol. 52, no. 19, pp. 5611–5625, 2014.

[17] W. Naji, M.-L. Espinouse, and V.-D. Cung, "Towards a robust scheduling on unrelated parallel machines: A scenarios based approach," in *Modelling, Computation and Optimization in Information Systems and Management Sciences*. Springer, 2015, pp. 343 – 355.

Analysis of integrated scheduling and intermediate transportation systems in production lines with variability-related measures

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Miriam Zacharias, Rainer Leisten
Dep. of Bus. Admin. and Operations Management
Faculty of Engineering, University of Duisburg-Essen
Duisburg, Germany

Chandrasekharan Rajendran
Department of Management Studies
Indian Institute of Technology Madras
Chennai, India

Abstract— This paper builds upon the work on minimizing the variability-related objective function *squared variability of completion times* (CTDVB2) [1]. CTDVB2 and modifications of CTDVB2 are investigated in serial production systems. As we are building upon a makespan-optimized schedule, we are investigating how altering the core delay will help to improve CTDVB2 and the overall performance. The influence of variability in one sub-system on the performance of the combined system is investigated via computational experiments. First, a single line or a flow shop is evaluated from a scheduling perspective. Then, two flow shops in series with and without intra-shop logistics are analysed from scheduling and queuing perspective. Looking at the overall performance can help to analyse and close the gap between static-deterministic scheduling and dynamic-stochastic queuing theory.

Keywords—CTDVB2, CTV, variability, scheduling, queuing, flow shop

I. Introduction

There are three regular types of objective functions considered in job scheduling. These are utilization-orientated ones (e.g., makespan), material-flow-orientated ones (e.g., flowtime) and due-date-orientated ones (e.g., earliness and tardiness) [1]. Extensive research has been done since the seminal paper of Johnson (1954) to optimize these kinds of objective functions [2], and many exact and heuristic algorithms as well as meta-heuristics have been developed to solve problems with regular objective functions ([1] and the quoted references). Generally, uniformity-related objectives and specifically CTDVB2 have not been investigated into detail as it is the case with regular objectives like makespan and flowtime. With CTDVB2, Leisten and Rajendran (2015), proposed a new objective function which "aims at the smoothness of a permutation flow shop's output, i.e. at the similarity of the differences of completion times of each two consecutively scheduled and finished jobs. Ideally in this case, the *inter-completion* times are equal for all consecutive pairs of jobs in a sequence." This equals a fixed rhythm of job completion in the flow shop [1] and falls under the category of non-regular objectives.

We are using following notation:

i	job index, $i=1,...,n$
j	machine index, $j=1,...,m$
$p_{i,j}$	processing time of job i on machine j
$C_{[i],j}$	completion time of job in position i on machine j
seq	given job sequence (permutation /solution) of the problem
$[i]$	job in position i of sequence seq
$C_{[i],j}$	completion time of job $[i]$ on machine j, $C_{[1],1} = p_{[1],1}$; $C_{[1],j} = C_{[1],j-1} + p_{[1],j}$ for $j = 2,\ldots,m$; $C_{[i],1} = C_{[i-1],1} + p_{[i],1}$ for $i = 2,\ldots,n$; $C_{[i],j} = max(C_{[i],j-1}, C_{[i-1],j}) + p_{[i],j}$, all indices defined appropriately.
$C_mod_{[i],j}$	modified completion time of job in position i on machine j for a given sequence
$CT_q(S)$	average waiting time of jobs in front of System S
$FTS(S)$	average flowtime of jobs in scheduled sequence in system S
$FTA(S)$	estimated flowtime of jobs in system S

The mathematical formulation of the proposed objective function is:

$$\text{CTDVB2}_{seq} = \frac{(n-1)\sum_{i=1}^{n-1}(C_{[i+1],m} - C_{[i],m})^2}{(C_{[n],m} - C_{[1],m})^2} - 1 \quad (1)$$

Another objective function addressing variability of completion times is CTV. It is defined as the variance of the completion time vector (2nd statistical moment of the distribution of jobs' completion times) [1]:

$$\text{CTV}_{seq} = \frac{1}{n}\sum_{i=1}^{n}(C_{[i],m})^2 - \left(\frac{1}{n}\sum_{i=1}^{n} C_{[i],m}\right)^2 \quad (2.1)$$

or [2]:

$$\text{CTV}_{seq} = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}(C_{[i],m} - C_{[k],m})^2 = \frac{1}{n}\sum_{i=1}^{n-1}(C_{[i],m} - C_{[i+1],m})^2 \quad (2.2)$$

Both objective functions are looking at variability aspects of jobs' completion times, but the underlying idea of optimization is different. CTDVB2 aims at a smooth output in terms of a fixed rhythm between completion of consecutively scheduled jobs. This might be useful if a downstream system is fed, which requires a smooth input at the next stage to execute another

production step, or if a transport unit with a fixed routing has to be considered between two systems. CTV instead ideally aims at finishing all jobs about the same time. This is helpful if all jobs of the upstream system have to be ready before being processed at the next stage, which is mostly the case in assembly systems with parallel machines on stage 1. In a flow shop system, a CTV-value of zero can never be achieved as all jobs have to be processed one-by-one on the last machine. Therefore, minimizing CTV will only lead to less idle time of the last machine between consecutively scheduled jobs of a sequence and the jobs will be finished as closely as possible to the average flowtime.

$$FT = \frac{1}{n}\sum_{i=1}^{n} C_{[i],m} \quad (3)$$

In a first step, we take a given sequence with optimized makespan as input to our optimization model. We assume that the sequence and makespan are given by a scheduler and that we can only alter the core idle time in a sub-system. Front and back idle time cannot be changed as this would influence the schedules of other linked systems. Therefore, we keep $C_{[n],m}$ and $C_{[1],m}$ fixed. We will delay operations on the last machine and therefore modify CTDVB2 and core idle time, aiming at a reduced variation of resource consumption and work in process by smoothing the flow shop's output. As only right-shifting is allowed and no job can be finished earlier than in the original makespan solution, minimizing CTDVB2 can lead to a reduction of CTV and vice versa. In the special case, where processing times on the last machine of a flow shop are deterministic and equal for all jobs, the best CTDVB2 solution also gives the best CTV value for a given sequence (see Fig. 1).

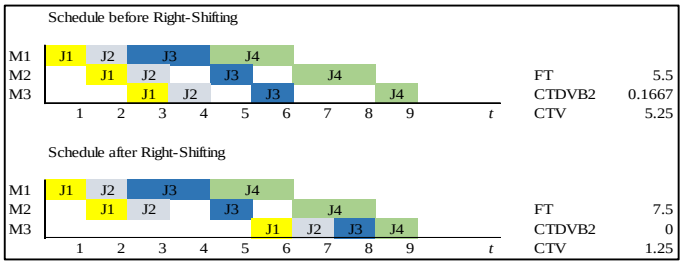


Fig. 1: Example of CTDVB2-optimized schedule for a Flow Shop with four jobs and three machines.

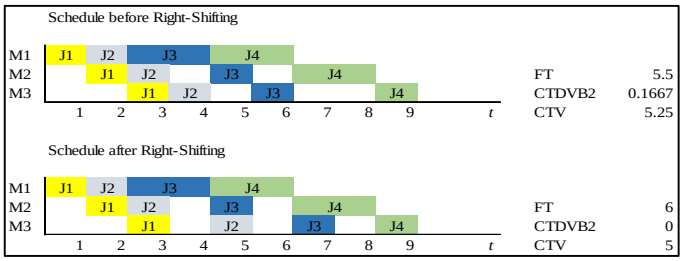


Fig. 2: Example of CTDVB2-optimized schedule for a Flow Shop with four jobs and three machines keeping first job's completion times fixed.

In this work, we keep the first and last jobs' completion times fixed while optimizing the numerator of CTDVB2. In this setting, we can achieve an improvement of CTV by minimizing CTDVB2 (see Fig. 2), because in this special case both objectives will adjust completion times of the remaining jobs closer towards the average flowtime, which is bound by the first and the last jobs' completion times and has a tendency to move towards the mean of the first and last jobs' completion times by minimizing CTDVB2 (if the first and last jobs' completion times are constant).

Many real life decisions in production planning have to be taken under uncertainty. While classic scheduling theory deals with problems from the static-deterministic point of view, dynamic-stochastic queuing theory considers uncertainty, addressing variability aspects by applying the mathematical tools of stochastic analysis. The actual process time is one of the most important random variables in manufacturing environ-ments. One fact that should be kept in mind is that the actual process times make up often only around 5 to 10 percent of the total cycle time in a manufacturing system. The remaining time is mostly spent waiting, in queues, for resources such as work-stations, transportation vehicles, operators, etc. [3]. This is the reason why our present work intends to link the static-deterministic approaches of scheduling theory with the ideas and methods of deterministic-stochastic queuing theory. Leisten and Rajendran combined the two approaches by looking at a combined system from scheduling and queuing perspectives. The setting of the combined system (i.e., two sub-systems in series), is given in Fig. 3.

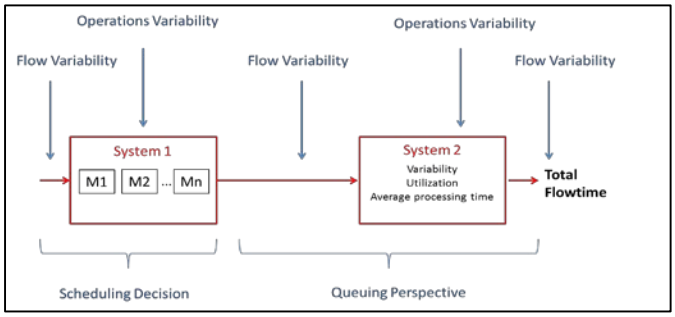


Fig. 3: Combined system from scheduling and queuing perspectives [1].

This system can give specific insights into the influence of scheduling decisions on other systems linked to it [1]. In this approach, only certain parameters of system 2 have to be known for the scheduling decision in system 1. For a detailed description of the queuing model, see [1]. It is important that system 1 is defined as a permutation flow shop [4], while system 2 is not looked at in detail but interpreted in terms of queuing indicators. The waiting time of a job in front of system 2 is approximated by the Variability-Utilization-Time-formula (VUT-formula or Kingman equation) and depends on the variability of departures from system 1, the variability of processing times on system 2, system 2's utilization and the processing times on system 2 [3]. For the overall system, the total (expected) flowtime (FT) per job is given by:

$$FT_{total,\ per\ job} = FTS(1) + CT_q(2) + FTA(2) \quad (4)$$

$FTS(1)$ is derived from the detailed makespan-optimized scheduling solution in system 1, whereas $CT_q(2)$ is calculated by a queuing-theoretical approach and $FTA(2)$ is estimated. This implies that the scheduling decision in system 1 does not only influence the time spent in system 1 (including waiting

time before being processed plus processing time) but also the overall performance of the joined system. Different schedules in system 1 alter the departure variability of system 1 (which is supposed to equal the arrival variability of system 2) and the waiting time in front of system 2 [1]. By varying the departure variability of system 1 (which is covered by the new objective function CTDVB2), the total flowtime of a job in the overall system can be affected as well. Several numerical examples of a single flow shop and flow shop production lines are presented including intra-shop logistics. In section II, a mathematical model, which aims at minimizing CTDVB2, is introduced. The example in section III.A will detect interdependencies between variability reduction and the system's performance in a combined system of two flow shops. In III.B, a transport unit is considered in the system. To determine the influence of different objectives on linked systems, a production line consisting of two flow shops in line with an intermediate transport system is evaluated in the sequel. The variability-utilization-time-formula, originating in queuing theory, is used to derive the overall performance of the joined system, which can help to link scheduling and queuing approaches.

II. Single Flow Shop without Outbound Logistics

As CTDVB2 should not be considered as a single objective (as it would give impractical solutions with very high makespan and flowtime values), we will follow a two-step approach. For the numerical examples in this section the test-bed of Taillard (1993) was used [5],[6]. First, the solutions of the well-known constructive NEH-heuristic [7] for makespan minimization were calculated for all Taillard instances up to the number of jobs n = 200 (there are 10 instances in every test-bed size). The NEH-heuristic is based on the assumption that there is a higher priority for jobs with higher total processing time than jobs with less total processing time [7]. In the next step, we introduce a mathematical model *"Min CTDVB2"*, that leads to right-shifting of operations on the last machine in the flow shop. A reduction of CTDVB2 can be achieved by postponing some or all operations on the last machine while the makespan is maintained. Otherwise, a CTDVB2 value of zero can be always achieved simply by right-shifting all operations on the last machine which worsens the values of makespan and flowtime. The core idle time of the input schedule is modified, while keeping front and back idle time unchanged.

Constraints:

$$C_mod_{[i],j} \geq C_mod_{[i],j-1} + p_{i,j}, \text{ for } i = 2,\ldots,n,\ j = 2,\ldots,m \quad (5)$$

$$C_mod_{[i],j} \geq C_mod_{[i-1],j} + p_{i,j}, \text{ for } i = 2,\ldots,n,\ j = 2,\ldots,m \quad (6)$$

$$C_mod_{[i],1} \geq C_mod_{[i-1],1} + p_{1,i}, \text{ for } i = 2,\ldots,n \quad (7)$$

Constraints (4) to (6) ensure that a job can only start being processed on a machine after it has been finished on the previous one and that a machine can only start processing a job after its predecessor has been finished on this machine.

$$C_mod_{[1],j} = C_{[1],j}, \text{ for } j = 1,\ldots,m \quad (8)$$

$$C_mod_{[n],j} = C_{[n],j}, \text{ for } j = 1,\ldots,m \quad (9)$$

$$C_mod_{[i],j} \geq C_{[i],j}, \text{ for } i = 1,\ldots,n,\ j = 1,\ldots,m \quad (10)$$

With these constraints the makespan and the completion times of the first job on all machines are kept from the initial NEH-makespan solution (for a given sequence) and only right-shifting of operations is allowed in any case. As we are not changing the first and last jobs' completion times, we will only influence the core idle and waiting time for a given makespan-optimized sequence which we assume to be given. Front and back delay remain untouched. By this, schedules of previous and succeeding systems remain undisturbed by the new time-table. Therefore, we can derive a convex quadratic function which can be solved in CPLEX.

Objective function:

$$\text{Min } Z1 = \sum_{i=1}^{n-1} \left(C_mod_{[i+1],m} - C_mod_{[i],m} \right)^2 \quad (11)$$

Min CTDVB2 was solved in IBM ILOG CPLEX Optimization Studio (version 12.5). It can be stated that although CTDVB2 is only indirectly minimized by the model (in the following called *Min CTDVB2*), the impact of the model on CTDVB2 is relatively higher than the impact on the actual objective function. A comparison of the initial values of the NEH-solution and the variability-optimized solution of *Min CTDVB2* for the test-bed (50x10) are given in Table 1. On an average for all test instances, the model achieves an improvement of CTDVB2 of 25.7 % with an average improvement of *Z1* of 7.89%.

TABLE 1 : Variability-optimized results for instances of size (50x10) by Taillard.

Tailard-Instance	NEH-solution		Min CTDVB2-solution		Improvement of CTDVB2
Number	CTDVB2	Z1	CTDVB2	Z1	[%]
1	0.4119	185685	0.4036	184399.8	2.0
2	0.3775	183985	0.2992	172835.5	20.7
3	0.3589	189123	0.3064	177614.4	14.6
4	0.4523	234210	0.2257	194103.7	50.1
5	0.3087	193829	0.2527	184560.1	18.1
6	0.4759	238219	0.2067	185283.3	56.6
7	0.3282	233692	0.1573	198126.3	52.1
8	0.2820	169988	0.2436	164708.4	13.6
9	0.7367	197199	0.4818	155731.1	34.6
10	0.3106	188057	0.2499	177572.9	19.5

III. Queuing Theory-based Analysis of Two Serial Flow Shops with and without a Transport Vehicle

A. Two Serial Flow Shops without a Transport Vehicle

In this chapter, the combined system of two flow shops in a sequential line is evaluated from a queuing-related point of view. While there is a detailed timetable determined for system 1 by optimizing different scheduling objectives, system 2 is considered from a queuing perspective, so that only certain parameters have to be known in advance for the evaluation of the overall system.

For a combined system without a transport vehicle, the basic relations (considered from a queuing perspective) are shown in Fig. 4. As mentioned before, the waiting time of a job in the queue in front of system 2 ($CTq(2)$) is approximated by the VUT-formula or Kingman equation and depends on the variability of departures from system 1 ($c_d^2(1) = c_a^2(2)$), the variability of processing times on system 2 ($c_e^2(2)$), system 2's

utilization ($u(2)$) and the processing times on system 2 ($t_e(2)$) [3] and [8].

$$CT_q(2) \approx \left(\frac{c_a^2(2)+ c_e^2(2)}{2}\right) \cdot \left(\frac{u(2)}{1- u(2)}\right) \cdot t_e(2) \qquad (12)$$

$c_d{}^2(1)$ is derived from the deterministic scheduling solution, whereas $c_e^2(2)$, $u(2)$ and $t_e(2)$ are estimated. In deterministic scheduling there is a natural variability e.g. of departures, even before including stochastic aspects such as stochastic processing times. Most flow shop papers deal with deterministic problems while job shop literature more often includes stochastic approaches. Here, we want to give a first idea of an approach linking scheduling and queuing theory via the VUT-formula, so that we did not look into stochastic scheduling problems for system 1. This might be interesting scope for future work. For the overall system, the total flowtime per job is given by formula 4.

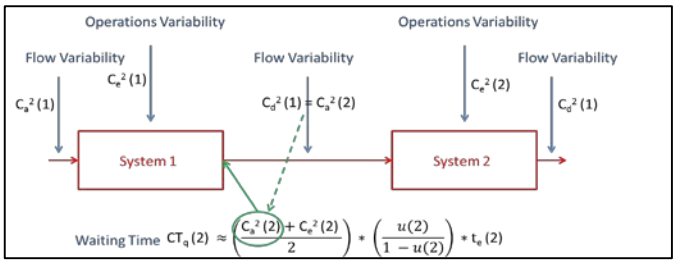


Fig. 4: Assembly line without transport vehicle (modified according to [1]).

As before, the time spent in system 1 and 2 include processing times and internal waiting times within the systems. The two objectives makespan and $Z1$ have been considered for the scheduling decision in system 1. The parameters of system 2 are set exemplarily and kept constant for both cases, with and without a transport vehicle. Regarding the utilization of system 2, there are three different cases investigated separately, namely $u(2)$ set to {0.80, 0.90, 0.95}. By taking the timetable obtained by the NEH-heuristic for makespan objective and the solution of *Min CTDVB2* as schedules in system 1, the different objectives are compared in terms of the overall performance of the joined system. In general, the variability of completion times increases by applying *Min CTDVB2*, while the variability of inter-departure times is minimized (as it is the objective of *Min CTDVB2*).

TABLE 2 : Results of queuing-based approach without the transport vehicle for instance 1 of size (50x10) by Taillard.

Schedule of System 1	**NEH-Makespan**	**Min CTDVB2**
Flowtime of System 1	2026.0	2030.8
CTq(2) for $u(2) = 0.8$	71.2	70.4
CTq(2) for $u(2) = 0.9$	160.2	158.3
CTq(2) for $u(2) = 0.95$	338.1	334.2
Total Flowtime per job for $u(2) = 0.8$	2147.2	2151.1
Total Flowtime per job for $u(2) = 0.9$	2236.2	2239.1
Total Flowtime per job for $u(2) = 0.95$	2414.1	2415.0

While the results are reported next for one instance of size (50x10), all results can be provided upon request. Table 2 contains the values for both objectives of system 1 and all cases of utilization of system 2. The NEH makespan values in column two are derived by applying the NEH heuristic with objective makespan minimization. Keeping the makespan of this initial solution fixed, we allow right-shifting of jobs on the last machine by applying our model *Min CTDVB2* (column 3). As completion time can be defined as a continuous variable, we release the integer constraint for completion times in this step, which leads to very good solutions regarding the new objective CTDVB2. The results show that reducing the variability of inter-departure times (by applying *Min CTDVB2*) leads to less waiting time in front of the next system. In some cases, this reduction of waiting time can even compensate a higher initial flowtime in system 1, and therefore improve the overall performance of the combined system. Table 3 gives an overview of the advantage of $Z1$ regarding the overall flowtime of the combined system. For each Taillard instance size there are 10 instances. Table 3 shows the number of cases, where the overall performance (total flow-time) improved by applying *Min CTDVB2*. A major outcome is that with higher utilization of downstream systems, the variability of inter-departure times of preceding systems has a higher impact on the overall performance. Therefore, *Min CTDVB2* performs well in the case of high utilization of system 1.

TABLE 3 : Cases (out of ten) with a reduction of total flowtime by applying model *Min CTDVB2* for the system without transport vehicle.

Utilization/Instance	$u(2) = 0.8$	$u(2) = 0.9$	$u(2) = 0.95$
(20x5)	3	7	9
(20x10)	6	8	10
(20x20)	4	10	10
(50x5)	5	8	9
(50x10)	2	7	9
(50x20)	6	9	10
(100x5)	5	7	9
(100x10)	1	6	9
(100x20)	2	10	10

B. Two Serial Flow Shops with a Transport Vehicle

Now an additional unit is included in the combined system, namely, a transport vehicle between the two flow shops of the serial system (see Fig. 5).

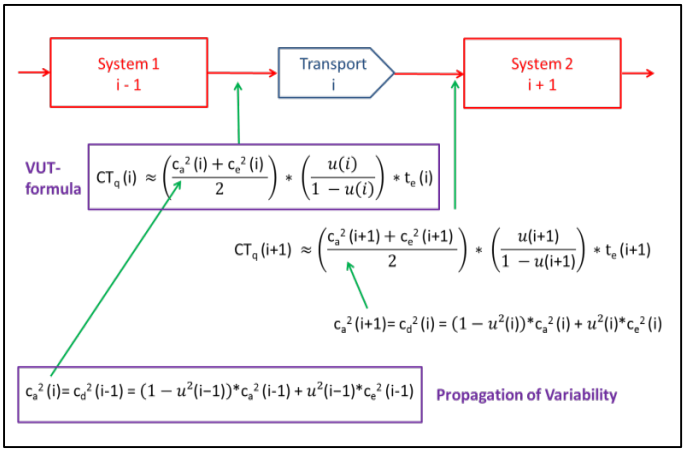


Figure 5: Assembly line with intermediate transport vehicle (modified according to [3]).

With the interpretation of the combined system from a queuing perspective, it can be extended to any number of sub-systems, as long as the basic parameters such as average processing time ($t_e(i)$), variability ($c_e^2(i)$) and utilization ($u(i)$) are given. The variability of the inter-departure times of the added sub-system can be calculated using formula 13 for the propagation of variability [3].

$$c_a^2(i) = c_d^2(i\text{-}1) = (1\text{-}u^2(i\text{-}1))*c_a^2(i\text{-}1)+u^2(i\text{-}1)*c_e^2(i\text{-}1) \quad (13)$$

We assume that the transport vehicle arrives in fixed time-intervals, as it would e.g., be the case in a milk-run setting. A finished job has to wait in the (uncapacitated) output buffer of system 1 till the vehicle arrives. If there is no finished job in the output buffer, the vehicle leaves empty. Also, jobs have to wait in the (uncapacitated) input buffer of system 2 till the first machine of system 2 is available. The total flowtime per job has to be extended by the waiting time in front of the transport system and by the processing time (i.e. transportation time) on this system.

$$FT'_{total,\ per\ job} = FTS(1) + CT_q(T) + t_e(T) + CT_q(2) + FTA(2) \quad (14)$$

The parameters of system 1 and 2 remain the same as explained for the case without transport vehicle. For the case including the transport vehicle, the following assumptions are made. The vehicle's utilization is calculated by the number of utilized runs (given by n = number of jobs) in relation to the number of total runs. The vehicle takes the same amount of time for delivering a job from system 1 to system 2 and returning to system 1. The transportation times are deterministic and constant for all jobs:

$$u(T) = \frac{n}{n + empty\ runs} \quad (15)$$

$$t_e(T) = 0.5 \text{ x } speed \quad (16)$$

$$speed = 2 \text{ x } \left\lceil \frac{C_{[last\ job],last\ machine} - C_{[first\ job],last\ machine}}{no.\ of\ jobs - 2} \right\rceil \quad (17)$$

$$c_e^2 = 0 \quad (18)$$

Given the set of assumptions, the queuing times in the second case can be calculated using the VUT-formula approach ([3] and [8]) as follows:

$$CT_q(T) = \frac{C_d^2(1) + 0}{2} \cdot \frac{u(T)}{1 - u(T)} \cdot t_e(T) \quad (19)$$

$$CT_q(2) = \frac{C_d^2(T) + C_e^2(2)}{2} \cdot \frac{u(2)}{1 - u(2)} \cdot t_e(2) \quad (20)$$

$$c_d^2(T) = \left(1 - u(T)\right) \cdot c_d^2(1) + 0 \quad (21)$$

Finally, the total flowtime for the combined system including the transport vehicle can be obtained by adding up all waiting and processing times in the combined system as it can be seen for the example instance in Table 4. An important outcome of this approach is that less variability of departures in system 1 can compensate a higher initial flowtime, when other systems (which are linked to system 1 in some way) are taken into consideration. In this example it can be seen that the total flowtime of *Min CTDVB2* is initially, and in the case without considering the transport vehicle, higher than that of the NEH makespan solution. If the transport unit is linked to the system, the total flowtime decreases and becomes lower than the one of the initial NEH solution.

TABLE 4 : Results of queuing-based approach with a transport vehicle for instance 1 of size (50x10) of Taillard.

Schedule of System 1	**NEH-Makespan**	**Min CTDVB2**
Flowtime of System 1	2026.0	2030.8
$CTq(T)$	42.7	41.8
$CTq(2)$ for $u(2) = 0.8$	39.5	39.3
$CTq(2)$ for $u(2) = 0.9$	88.9	88.4
$CTq(2)$ for $u(2) = 0.95$	187.6	186.7
Total Flowtime per job for $u(2) = 0.8$	2352.4	2351.3
Total Flowtime per job for $u(2) = 0.9$	2401.8	2400.5
Total Flowtime per job for $u(2) = 0.95$	2500.5	2498.7

It can be stated that model *Min CTDVB2* is generally advantageous in terms of total flowtime, although the initial flowtime in system 1 itself is generally higher due to the postponement of operations on the last machine, keeping the first and last jobs' completion times fixed. A comparison of the performance of both objectives in the linked system including the transport vehicle is given in Table 5.

TABLE 5 : Cases (out of ten) with a reduction of total flowtime of combined system by applying model *Min CTDVB2* for the system with a transport vehicle.

Utilization/Instance	***u*(2) = 0.8**	***u*(2) = 0.9**	***u*(2) = 0.95**
(20x5)	8	8	7
(20x10)	8	8	8
(20x20)	8	8	10
(50x5)	9	9	9
(50x10)	8	9	9
(50x20)	10	10	10
(100x5)	8	8	8
(100x10)	6	6	7
(100x20)	7	8	10

It is observed that a higher utilization of downstream systems increases the job waiting time in front of those systems and therefore results in higher total flowtime. Another finding is that the reduction of variability to zero is not only hardly to achieve in real-life examples but also not desirable, as the initial increase of makespan and flowtime can never be compensated by reduced waiting time (at least this can be stated for the two stage flow shop setting which has been investigated in this paper).

TABLE 6 : Job waiting times for pick-up and vehicle idle time for instance of size (50x10) by Taillard.

Taillard-Instance	**Job Waiting Times for Pick-up (System 1)**		
(50x10)	**NEH- Makespan**	**Min CTDVB2**	**Improvement [%]**
1	6813	6574.2	3.5
2	4873	4728.5	3.0
3	6229	6793.2	-9.1
4	5306	4692.0	11.6
5	7113	6117.5	14.0
6	3375	4080.6	-20.9
7	6755	5927.0	12.3
8	6681	6314.8	5.5
9	8951	7773.3	13.2
10	4296	4116.3	4.2

Another finding is that smoothing the output of system 1 by applying *Min CTDVB2* mostly leads to less waiting time of jobs in front of the transport vehicle, whereas the utilization of the

transport vehicle remained mostly the same as in the NEH makespan solution. In our setting we assume that the transport vehicle will not wait till an unfinished job is finished but continue its routine. Therefore, jobs will have to wait for nearly two transportation terms if they are not finished when the empty vehicle arrives. This explains increased waiting times and shows scope for future work. Table 6 shows the results for instances of size (50x10). In total, the sum of job waiting times for pick-up and vehicle idle time can be reduced for the majority of cases (see Table 7) by applying *Min CTDVB2*.

TABLE 7: Cases (out of ten) with a reduction of sum of Job Waiting Time and Vehicle Idle Time by applying model *Min CTDVB2* for the system with a transport vehicle.

Instance Size	50x5	50x10	50x20	100x5	100x10	100x20	200x10	200x20
Advantage CTDVB2	7	8	10	6	8	9	7	9
Average Improvement [%]	5.2	3.3	10.1	2.9	4.9	6.3	1.6	5.0

What can be seen in all examples is that the best performance in one sub-system with regards to one objective function does not always lead to the best overall results regarding the same objective. Therefore, we come to the conclusion that it is generally advantageous to not only control variability of processing times and other internal parameters of a production system, but also to look at variability aspects of interfaces with other sub-systems to enhance the overall performance of the production system.

IV. Conclusions and Outlook

This paper dealt with variability-related objective functions in different kinds of production systems and has intended to find a link between static-deterministic scheduling and dynamic-stochastic queuing theory. The first phase has analyzed a single flow shop with regard to different objective functions. A new mathematical model is introduced which aims at minimizing the variability of inter-departure times of jobs by minimizing the numerator of CTDVB2. Model *Min CTDVB2* gives compromise or trade-off solutions regarding makespan optimization and CTDVB2. What can be stated is that the impact on CTDVB2 is much higher than the impact on the objective function itself. This is appreciated as it means good results (on an average around 26% improvement of CTDVB2 by model *Min CTDVB2*) with less computational effort. Hence, *Min CTDVB2* is an adequate tool from a multi-objective perspective, as the makespan of the NEH solution is kept. A major part of this work is the interpretation of a production line of two serial flow shops with and without an intermediate transport unit from a queuing theory-related point of view. Detailed static-deterministic scheduling decisions in system 1 are combined with basic queuing-related parameters of system 2 and the transport vehicle. One main conclusion of this experimental evaluation is that an initially higher flowtime in flow shop 1 (resulting from applying model *Min CTDVB2*) can be compensated by less variability, which leads to less overall flowtime per job in the combined system. The experimental finding proves the advantage of reducing the variability of inter-departure times via right-shifting of operations, while keeping the makespan fixed, as it is done via model *Min CTDVB2*. While reaching zero variability is not realistic in practice, aiming at reducing (inter-departure times) variability has a positive effect on the overall performance of production systems. Finally, all experimental evaluations pointed out that reducing variability in job schedules leads to an improvement of the performance of combined systems, and is essential in multiple-shop settings. Our work was supposed to be a first step towards linking scheduling and queuing, so that there is a lot of scope for future work. First, one could start with including stochastic aspects (e.g. stochastic processing times, transportation times) in the scheduling problem. Stochastic scenarios and simulation studies could provide a deeper insight on the relations between different objectives in sub-systems. Also, the newly proposed non-regular objective function CTDVB2 could be investigated further in terms of practical applications as well as theoretical comparisons to CTV, including a formal proof of the relationship between CTDVB2 and CTV in different settings. It could also be interesting to adapt CTDVB2 to job shops or inter- and intra-job logistics. We followed a basic two-step approach while minimizing CTDVB2. Future work should look at real multi-objective optimization and investigate the relations between altering core idle time by minimizing CTDVB2 and resource consumption in the system.

Acknowledgment

First we would like to thank the four reviewers for their valuable comments and effort to improve the paper. We would also like to thank DAAD, IIT Madras and University of Duisburg-Essen for the support of this research work.

References

[1] Leisten, R. and Rajendran, C., "Variability of completion time differences in permutation flow shop scheduling", in: *Computers & Operations Research*, vol. 54, 2015, pp. 155-167.

[2] Gowrishankar, K., Rajendran, C. and Srinivasan, G., "Flow shop scheduling algorithms for minimizing the completion time variance and the sum of squares of completion time deviations from a common due date", in: *European Journal of Operational Research*, vol. 132, issue 3, 2001, pp. 643-665.

[3] Hopp, W. J. and Spearman, M. L., Factory Physics, 2nd ed., Boston: *McGraw-Hill*, 2001.

[4] Pinedo, M. L., Scheduling – Theory, Algorithms, and Systems, 4th ed., New York: *Springer Science+Business Media*, 2012.

[5] Taillard, E., "Benchmarks for basic scheduling problems", in: *European Journal of Operations Research*, vol. 64, issue 2, 1993, pp. 278-285.

[6] Taillard, E. (n.d.): http://mistic.heig-vd.ch/taillard/problemes.dir/ordonnancement.dir/ordonnancement.html

[7] Nawaz, M., Enscore E. E. and Ham, I., "A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem", in: *Omega*, vol. 11, issue 1, 1983, pp. 91-95

[8] Kingman, J. F. C., "On the Algebra of Queues, in: *Journal of Applied Probability*, vol. 3, issue 2, 1966, pp. 285-326.

A batching and scheduling optimisation for a cutting work-center: Acta-Mobilier case study

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Emmanuel ZIMMERMANN[1, 2, 3], Hind BRIL EL HAOUZI[1, 2], Philippe THOMAS[1, 2]
André THOMAS[1, 2], Mélanie NOYEL[3]

[1]*Université de Lorraine, CRAN, UMR 7039, Campus Sciences, BP 70239, 54506 Vandœuvre-lès-Nancy cedex, France*
[2]*CNRS, CRAN, UMR7039, France*
(e-mail: {emmanuel.zimmermann ; hind.el-haouzi ; philippe.thomas ; andre.thomas }@univ-lorraine.fr)
[3]*Acta-Mobilier, parc d'activité Macherin Auxerre Nord 89270 MONETEAU*
(e-mail : mnoyel@acta-mobilier.fr)

Abstract—The purpose of this study is to investigate an approach to group lots in batches and to schedule these batches on Acta-Mobilier cutting work-center while taking into account numerous constraints and objectives. The specific batching method was proposed to handle the Acta-Mobilier problem and a mathematical formalisation and genetic algorithm were proposed to deal with the scheduling problem. The proposed algorithm has been embedded in software to optimise production costs and emphasis the visual management on the production line. The application is currently being used in Acta-Mobilier plant and shows significant results

Keywords— part family, scheduling, genetic algorithm, multi-criteria optimisation

I. Introduction

Acta-Mobilier, a high-quality product lacquering manufacturer, is faced to high standards of quality and technological requirements challenges which lead this company to have a reworks rate upper than 30% and reaching 80% for some products. Many critical consequences, including production cost increasing, products flow perturbations and steady deterioration of deliveries rate, are implied by this fact. Penalty costs and brand image damage may also be at stake.

In addition, a mass customisation business strategy has been adopted by the company, which leads to a significant diversified products panel and, at the same time, pushed to optimise its manufacturing processes.

To reduce the production cost and time and also answer a specific customer need, a particular clustering/batching method on two levels combined to a scheduling optimization has been applied. In section 2, the industrial context is set. The section 3 exposes the studied problem. Then, the proposal is described in section 4. The obtained results are presented in the last section. The way these works would be used in a more global project is also described.

II. Industrial context

The Acta-Mobilier company has two core business: front manufacturing for kitchen specialists (subcontracting) and the design and realization of shops and stands (layout). Both use the same human and material resources.

In this paper, we focused on the subcontracting business. To manage this activity the company has adopted a mass customization approach which leads to very customized products. Thus, each front has custom dimensions, colours and many other parameters like, for example, the handle dimensions and position. Consequently, the company have to deal with a big variability on customer demands and manufacturing processes: different rooting sheets, rework rate… while keeping manufacturing process as standardized as possible and minimizing production cost.

In order to raise a high quality of product and in particular by avoiding differences in the texture and the appearance of the fronts which have to form a coherent and harmonious finished product reference, the company tried to manage its production by finished product reference (a finished product reference represents all the fronts forming a kitchen). For example, each employee in charge of applying the cement finishing has his own way to do it: the sense of application, thickness, painter claw .... There is also a risk of tint difference if all the fronts are not processed on the same period and with the same preparation, the drying time can also impact on the tint.

Moreover, sending the demand orders already sorted by finished product reference leads to a significant time saving for customer. All these requirements are essential regarding the company quality and customer service levels. For all those reasons, it becomes important to manufacture all fronts belonging to the same finished product reference on the same time.

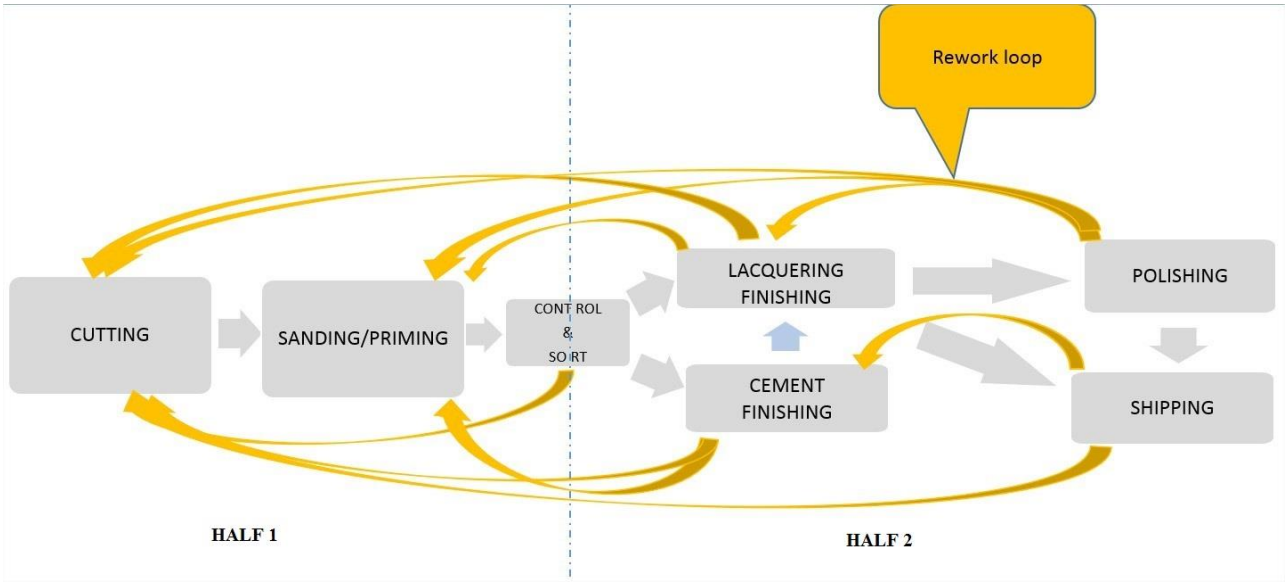


Fig. 1 Acta-Mobilier production line diagram

However, this policy requires to increase the surface area and implies many issues on material and time management (tool changes and organization in the cutting work-center). Consequently, the company needs a way to minimize time and production costs inferring the least possible functioning complication.

In previous works [15], the presence of indivisible batches, named "Kernels" has been highlighted. A Kernel is a batch of products of the same customer, with the same colour and having the same process (all the Kernel parts must stay together during the whole production line). These batches characteristics were particularly well adapted to the finishing process (the half 2 of Fig.1). The way to determine the Kernels composition is not the purpose of this work and should be seen as already known.

Working with this subdivision is irrelevant in the cutting work-center considering its process and objectives. That's why another batches construction regulation must be used to determine cutting batches: The cutting machines have dimensional limits lower than the ones available to customers and the fronts that exceed these limits are treated separately on another machine.

Moreover, the presence of a control step placed before the second part (Half 2 of Fig.1) of the manufacturing process will allow to split the cutting batches into the Kernels so that they could be worked one by one. To do so, it is, at a certain time, needed to physically achieve an unbundling and a re-bundling of the Kernels. However, the available floor space needed to achieve this sorting doesn't allow to have more than 6 or 7 pallets simultaneously. Regarding that, it is impossible to increase this space, the batching organization has to take this into account.

III. Problem statement and mathematical models

From the industrial context description, the cutting work-center optimizer problem can be seen as the combination of two well-known distinct problems classes:

- Making manufacturing batches (MBs) in order to optimise the production costs without making too complex batches (increase space area, adding no-added value tasks to sort those batches in the end of the production line).
- Propose a schedule of the MBs to optimise the control step. In this step, the human operator will be able to re-build the finished product references batches (FPRBs) in less than an hour on the available floor space.

A. Kernels batching

This problem aim is to find relevant MBs to minimise production and time wastes. There are at least two ways to interpret this problem:

- Take the whole customers' orders and split them into optimal sized lots which corresponds to the lot-sizing problems class [1], [2].
- Cluster the Kernels into bigger batches with common characteristics corresponding to the part family problems class [3], [4], [5].

In this study, the problem concerns the batches composition (Which distinct finished product references must be gathered? and how many batches?) and not the optimal number of products to put in batches. Moreover, as hypothesis, having the Kernel size already calculated, lead to put the studied problem in family part problems class. This class looks for grouping

products that have the same relevant characteristics (e.g. thickness and production range).

Numerous works about solving this kind of problems exist, most of them use clustering techniques: [6] treated this kind of problem in order to minimize sequence dependent changeover times. In the company case, the changeover times are sequence independent but there is a strong constraint on the number of finished product reference to put in a batch. [3], [4], [5] worked on clustering with multi-criteria. Those works show that batching problems can totally differ from one production to another. In our case, although, the number of Kernel in each batch is not a strong constraint, the standard deviation must stay weak to not create disproportionate batches and smooth the production load. Many method could be used to solve batching problem for example [7] used a genetic algorithm to determine batches.

B. Scheduling

The second problem is a batch scheduling problem on a single machine with multi criteria. The objective is to minimise the number of setup, but also to keep as close as possible the batches that are from the same finished product reference. The first objective is clearly to reduce production time and the second one guarantees that, on the control step, the collaborator will not wait more than one hour to have all the products he needs to rebuild the corresponding FPRBs with enough floor space to achieve it well.

Many works were done on the subject of scheduling problems with setup times, [8] expose them. According to the classification of scheduling problem with setup times proposed by [9] presented in fig.2, the studied problem belongs to batch sequence independent scheduling problem. This problem type is known as NP-hard.

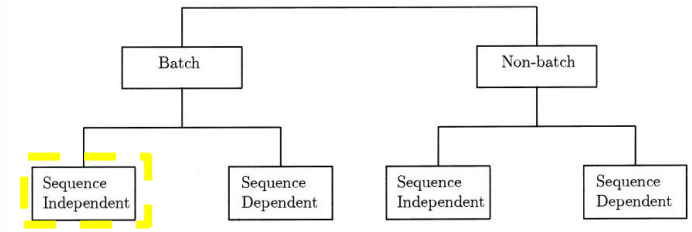


Fig. 2 The scheduling problem with setup time problem classification [9]

To be more specific, it belongs to the particular case of batch scheduling on a single machine, several solving methods were developed like branch and bound algorithm [10], Tabu search and also genetic algorithms [11],[12],[13].

A mathematical formulation of this specific problem is proposed below:

$N : Number\ of\ finished\ product\ references$
$M: Number\ of\ finished\ product\ references\ batches\ (FPRB)$
$F : Number\ of\ manufacturing\ batches\ (MB)$
$P_m : Number\ of\ MB\ in\ FPRB\ number\ m$

The first function F1 has as objective to optimise the distance between all the *MB* called $l_{e,m}$ of the same *FPRB* and characterised by its thickness *e* and its *FPRB* number *m*:

$$F1 = \min_i \left(\left(\sum_{j=1}^{F} a_{i,j} \ \max(x_j - x_i - 1, 0) \right) \right)$$

Where:

- $a_{i,j}$ expresses the fact that the two MB $i = l_{e,m}$ and $j = l'_{e',m'}$ belong to the same FPRB

  $a_{i,j} = 1$ if $m = m'$
  $0\ elsewhere.$
- x_i is the position of MB i in the schedule

The second function F2 has as objective to optimise the number of setup for the considered schedule:

$$F2 = \min_i \sum_{i=1}^{F} \sum_{j=1}^{F} y_{i,j}$$

Where:

- $y_{i,j} = 1$ if $(e \neq e' \ \&$ j is right after i
  0 elsewhere.

Consequently, the global function to minimise is:
$F = \alpha F1 + \beta F2$, with α, β mixing $ratios$
This optimisation has to be done under the following constraints:

$F1 \geq 0$
$F2 \geq M$
$\forall i, x_i \in (0, F)$
$\forall (i,j), \quad y_{i,j} \in [\![0,1]\!]$
$\forall i, \quad \sum a_{i,j} = P_k\ where\ k\ is\ FPRB\ number\ of\ MB\ i$
$\forall j, \quad \sum a_{i,j} = P_k where\ k\ is\ FPRB\ number\ of\ MB\ j$

IV. Proposal

A. Clustering

To answer the objective of quality requirement and to achieve a visual management standard for the human operators, the clustering has been made on two levels:

- On the first level, the different finished product references that have the more distinct thicknesses in common are gathered while also respecting the limit of grouping five finished references at most. This finished product references batches (FPRBs) are represented by an alphabetical character (e.g. batch A)
- On the second level, the Kernels of each FPRBs with the same thickness are grouped into manufacturing batches (MBs), represented by the FRPB character plus a randomly chosen integer (e.g. A1). The Fig.3 illustrated this explanation.

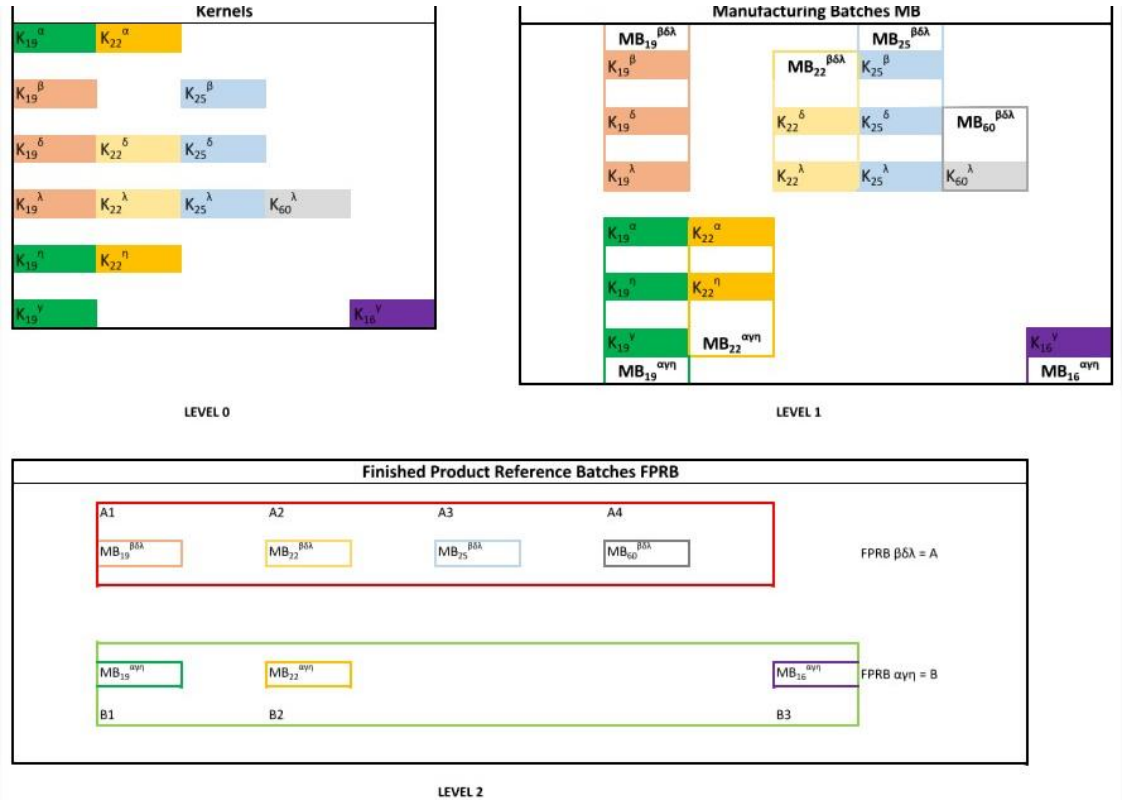


Fig. 3 : The two levels batching concept

With this two-levels batching, it is assured that the control & sort step will have enough floor space to reconstruct all the FPR Kernels without letting the pallets opened for a long time and without a lot no-added value tasks.

There are between 60 and 100 Kernels per week to cluster in MBs. This corresponds to a number of front which lies between 600 and 900.

B. Scheduling

The chosen solving approach is the genetic algorithm because of their capacity to provide good results in competitive computing times [8]. In the studied case, the computation time is more important than having an optimal solution, because the company suffers from its high reworks rate susceptible to happen anywhere in the production line. Furthermore, the probability to be able to follow exactly the schedule is weak and a wider objective will be to re-calculate the schedule each time a disturbance appears. The possibility to restart from a precedent result in case of reworks seems to be interesting and will be studied in further works. It will allow to measure the robustness of the algorithm in disturbed dynamic situation. In the output of the clustering phase there are between 16 and 40 MBs to schedule.

Here, it is proposed to apply a GA presented on Fig. 4 with the following parameters:

- An initial population of 1000 sequences of the batches randomly sorted.
- The natural selection with an elitism rule was chosen because even if, it is not the better rule to get closer to an optimal solution [14], it is known to be the fastest to converge and like already said velocity is more important than optimality.

The crossover process is similar to [12] and is functioning by iteration:

Following the rule presented on Fig 5 a new population is created by mixing the chromosomes of the remaining sequences of the last generation:

In the first loop, the two sequences that best suits to the optimisation criterium are chosen to make the crossover. Then, in the second loop, two of the three best sequences are randomly picked for the crossover. And, it keeps incrementing until the new population reaches one thousand inhabitants.

The mutation consists on swapping two chromosomes from a sequence. Two swapping methods were chosen: neighbour and foreign swapping explained on Fig. 6.

In a scheduling, it is non-sense to find the same batch at two distinct positions. That's why, in both mutation and crossover processes, the "abortion" phenomenon is important because it assumes that the same chromosome is not displayed more than a single time in a sequence. The algorithm stops as soon as the fitness function isn't better on ten consecutives runs.

The fitness mixing ratios were determined by experimentation and analysis of multiple runs. The tested schedules revealed that, to assure a fluent progress all along the production line, the first function is the one to be emphasized. The determined values of α and β corresponds to a ratio near from 2/3, 1/3.

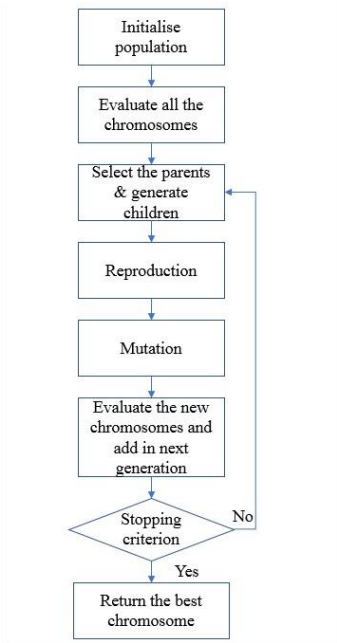


Fig. 4 A genetic algorithm flowchart

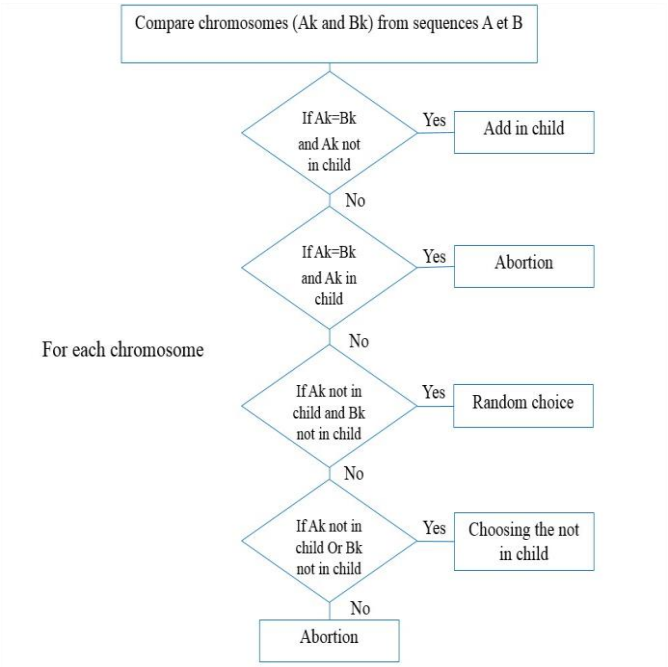


Fig. 5 Chromosome selection in crossover process

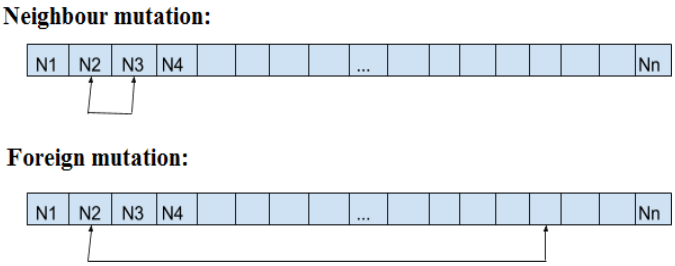


Fig. 6 Mutation principle

V. Results

The proposed method was implemented with VB.NET and used to manage one customer demands since last year. As results, its application allowed the company to improve the relationship and increase the volume of demands from this customer: it was able to satisfy his quality, delivery date requirements and also reduce total production time by decreasing the addition of a non-value-added operation. Figure 7, presents the comparison of the performances of the cutting work-center before and after the deployment of the proposed solution on the same period.

In comparison to the previous year at the same period, with a 32% ordered quantity bigger and similar conditions, the global production time is one day faster. The results are presented in Table 1.

TABLE 1: A comparison with and without optimisation

	2015 (without optimisation)	2016 (with optimisation)
Number of fronts ordered by week	W49: 456 pcs W50: 247 pcs W51 596 pcs	W49: 633 W50: 518 W51: 561
Number of days between first and last pallet opening in shipping workstation	W49: 5 days W50: 5 days W51: 5 days	W49: 4 days W50: 4 days W51: 4 days

Those results show tangible benefits to the company but not lead to direct link between results and the application of the proposed method. In order to validate the efficiency of the proposed method, we simulate one production week of the cutting work-center with a demand of 36 finished product references and 4 different thicknesses.

The results presented on Table 2 show the comparison between our method and two scheduling rules previously used by the company:

- Grouping all kernels from the same finished product reference into the same batch,
- Grouping all kernels with the same thickness into the same batch.

TABLE 2: A comparison between the proposed approach and the old used scheduling rules

	Proposed method	Finished product references	Thickness
Number of setups	11	4	22
Maximal W_{ip} between Kernels of the same FPR	5	14	4
Maximal number of pallets opened simultaneously	7	16	2

In addition, the developed application leads to a solution easily understandable by the co-workers thanks to its visual representation. Also, specifics documents are elaborated for the cutting WorkCentre and others for the rest of the process.

On the control step, a specific visual application has been developed to help the co-worker to simply split the manufacturing batches (named with the letters) and regroup together the Kernels by their FPR, the whole functioning using the coloured code shown in Fig.7.

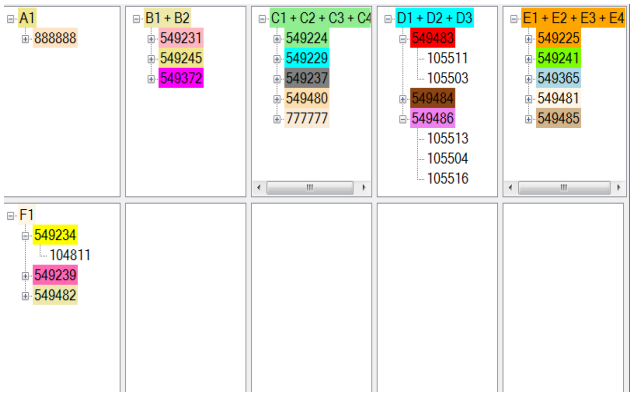


Fig. 7 An application screenshot presenting how to re-sort batches

VI. Outlooks

In this paper, an efficient method to batch and schedule the manufacturing orders for a specific problem induced by Acta-mobilier company was presented. Even if, significant results have been presented in this study, on-going work aims to use a CRAN laboratory platform named Tracilogis to determine the robustness of the generated solutions in case of reworks. Furthermore, this work can be seen as a part of wider project aimed to propose a hybrid manufacturing system which has been presented in [15]. The objective will be to use the developed solution as a local optimiser and make it collaborate with the other optimisers to provide the company a global dynamic system able to react to any disruption.

References

[1] N. O. Fernandes, M. J. Land, and S. Carmo-Silva, "Aligning workload control theory and practice: lot splitting and operation overlapping issues," *Int. J. Prod. Res.*, vol. 54, no. 10, pp. 2965–2975, 2016.

[2] R. Ramya, C. Rajendran, and H. Ziegler, "Capacitated lot-sizing problem with production carry-over and set-up splitting: mathematical models," *Int. J. Prod. Res.*, vol. 54, no. 8, pp. 2332–2344, 2016.

[3] H. A. Taboada and D. W. Coit, "Data clustering of solutions for multiple objective system reliability optimization problems," *Qual. Technol. Quant. Manag.*, vol. 4, no. 2, pp. 191–210, 2007.

[4] C. Han and I. Ham, "Multiobjective cluster analysis for part family formations," *J. Manuf. Syst.*, vol. 5, no. 4, pp. 223–230, 1986.

[5] H. C. Co, "Cluster analysis approach to batching in Flexible Manufacturing systems," *Comput. Integr. Manuf. Syst.*, vol. 3, no. 2, pp. 79–82, 1990.

[6] M. R. Bowers, K. Groom, W. M. Ng, and G. Zhang, "Cluster analysis to minimize sequence dependent changeover times," *Math. Comput. Model.*, vol. 21, no. 11, pp. 89–95, 1995.

[7] A. Lee-Post, "Part family identification using a simple genetic algorithm.," *Int. J. Prod. Res.*, vol. 38, no. 4, pp. 793–810, Mar. 2000.

[8] A. Allahverdi, C. T. Ng, T. C. E. Cheng, and M. Y. Kovalyov, "A survey of scheduling problems with setup times or costs," *Eur. J. Oper. Res.*, vol. 187, no. 3, pp. 985–1032, Jun. 2008.

[9] A. Allahverdi, J. N. Gupta, and T. Aldowaisan, "A review of scheduling research involving setup considerations," *Omega*, vol. 27, no. 2, pp. 219–239, 1999.

[10] M. Azizoglu and S. Webster, "Scheduling parallel machines to minimize weighted flowtime with family set-up times," *Int. J. Prod. Res.*, vol. 41, no. 6, pp. 1199–1215, Jan. 2003.

[11] Y. Yi, H.-Y. Chang, J. Wang, and J.-C. Bai, "Soft computing for parallel scheduling with setup times," in *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, 2004, vol. 4, pp. 2041–2046.

[12] V. A. Armentano and R. Mazzini, "A genetic algorithm for scheduling on a single machine with set-up times and due dates," *Prod. Plan. Control*, vol. 11, no. 7, pp. 713–720, Jan. 2000.

[13] H. Balasubramanian, L. Mönch, J. Fowler, and M. Pfund, "Genetic algorithm based scheduling of parallel batch machines with incompatible job families to minimize total weighted tardiness," *Int. J. Prod. Res.*, vol. 42, no. 8, pp. 1621–1638, Apr. 2004.

[14] D. Thierens and D. Goldberg, "Convergence models of genetic algorithm selection schemes," *Parallel Probl. Solving Nature—PPSN III*, pp. 119–129, 1994.

[15] E. Zimmermann, H. Bril, P. Thomas, A. Thomas and M. Noyel, "A hybrid manufacturing control based on smart lots in a disrupted industrial context," *IFAC WORLD CONGRESS*, 2017.

Machining elements sequence optimization of thin-walled and multi-cavity aircraft structural parts based on knowledge-based rules and genetic algorithm

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Dan Zhang [+], Tao Wei, Feng Xu, Dunwen Zuo
College of Mechanical and Electrical Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing, China
[+]zhd@nuaa.edu.cn

Abstract—Thin-walled and multi-cavity aircraft structural parts with a large quantity of machining features and easy-to-occur machining deformation are difficult to plan the numerical control (NC) machining process. One of the tasks of computer-aided process planning (CAPP) is to set the optimized sequence of machining element. Many previous researches on process planning take the geometric features as the optimization object which cannot be used by NC programming directly. To solve the problem, this paper proposed a method using both knowledge-based rules and genetic algorithm (GA) in the task of sequencing for machining elements with geometric and manufacturing informations. The knowledge-based rules are used to increase the effectiveness of chromosomes through adding a step to restructure chromosomes. The separating-face-relationship matrix and cavity-relationship matrix are built to express stepped symmetrically milling process. Machining quality and machining efficiency are used as optimal objectives of the genetic algorithm. At last, an instance part is used to verify the effectiveness of the proposed method.

Keywords—Machining elements; sequence optimization; knowledge-based; genetic algorithm

I. Introduction

Thin-walled and multi-cavity aircraft structural parts with masses of machining features have an easy-to-occur trait of deformation during machining. And the parts which are very common in aircraft structures are difficult and time-consuming to cutting. The efficiency of NC machining process can be improved effectively through intelligent and quick process planning. Therefore, CAPP of aircraft structural parts has received much attention [1-2]. To solve the sequencing problem of the manufacturing process, many approches are presented such as expert systems or knowledge-based systems, genetic algorithms, evolutionary algorithms [3], shuffled frog leaping algorithms [4], and etc.

The traditional process planning is driven by hierarchical constraints in stages. A typical method is the skeleton constraint matching approach [5] which is based on the part's hierarchy. It summarizes the commonness of each type of parts' process route to form rule sets. Then the reasonable order between each working procedure is determined by both the rule sets and several constraints like geometrical characteristics, requirements of processing technology, machining techniques, global optimization criteria(economy or productivity), and etc[6]. These constraints are considered hierarchically. Liu[7] uses knowledge-based rules to set processing priority. machining features are grouped into setups based on tool approaching directions, And Features in each setup are sequenced through geometric reasoning. Mokhtar[8] considers interacting features in a feature-based model for process planning through role-based system. For each feature, precedence information is generated considering both roughing and fishing operations. However, many technical experience and theory can be hardly understood as independent knowledge by inference engine due to different types of parts and various manufacturing environments. Actually, the staged decision method is very different from the thoughts of technologists that lead to the difficulty of extracting and expressing sequencing knowledge. Also, it's hard to use mathematical method or role match method to solve the problem. Some academics consider process planning as a topological sorting problem, Lin [9] studies the sequencing priority of interacting features considering processing quality and uses graph-search strategy in artificial intelligence. But the sequencing priority stays in feature level.

Many researches on process planning regard machining feature as optimization target which cannot be used as basis of NC programming directly. Meanwhile machining quality is always ignored or just create some blocks to do heuristic search. Machining deformation can be decreased significantly while using proper processing technology. This paper uses a hybrid method based on GA and knowledge with the target of machining elements. In order to increase effectiveness of chromosomes, a "chromosome recombination" step is added into genetic operations to improve the conventional genetic algorithm. The presented method not only considers machining efficiency but also takes stepped symmetrically milling process of thin-walled and multi-cavity structural parts into consideration so that less machining deformation can be achieved.

II. Description of sequencing problem

Machining elements are those information entities facing every machining operation which based on features and manufacturing resources. It includes machining feature of every operation, machining faces of every operation,

The research is supported by the Science and Technology Project of Jiangsu Province (No. BY2016003-12), and the Aeronautical Science Foundation of China (No. 20141652016).

processing method, processing parameters(cutting speed, cutting depth and feed) and some manufacturing resources like processing machines, tools and fixtures. The sequenced machining elements can be applied to NC programming directly. Each machining feature has its standard machining operation sequence according to machining precision, roughness, material, etc. For example, a cavity may have machining operations like rough milling webs, rough milling walls, finish milling webs, finish milling walls, finish milling corners, etc. A hole may contain machining operations like drilling, reaming, tapering, etc. The sequencing of machining elements is the combination of every processing operation satisfying process constraints.

Machining elements can be defined as follows:

$$ME = \{me_1, me_2, \dots, me_n\} \quad (1)$$

where ME is the set of all the machining elements and n is the number of machining elements.

$$me_j ::= < S_{feature-j} >< S_{faces-j} >< S_{resours-j} >< S_{machinfo-j} > \quad (2)$$

where j indicates the procedure number , the jth machining element is denoted by four-tuple me_j, in which $S_{feature-j}$ is the set of its related features, $S_{faces-j}$ is the set of its related machining faces, $S_{resours-j}$ is the set of its related manufacturing resources and $S_{machinfo-j}$ is the set of its related machining parameters.

According to the definition of machining elements, the process optimization problem is transferred into sequencing of machining elements with several constraints such as the selection of machine tools, fixtures and process constraints.

Assuming n machining elements in set $S=\{1,2,...,n\}$, let H_s, U_s, D_s, C_s show sets including tags of holes, faces, benchmarks, contours in the model respectively. F_s indicates the set including finish machining operations, R_s indicates the set including rough machining operations, $f(i)$ indicates the function to acquire tag of machining element i. Several constraints are formulated by the following knowledge-based rules:

Rule1 To describe that surface is machined before the hole:

$$\forall i, j \in S, M_{ij} = \begin{cases} 1 & (f(i) \in H_s) \&\& (f(j) \in U_s), i < j \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where value 1 means that the constraint is not met.

Rule2 To describe that finish machining is arranged after rough machining. The constraint can be modeled as follows:

$$\forall i, j \in S, K_{ij} = \begin{cases} 1 & (f(i) \in F_s) \&\& (f(j) \in R_s), i < j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Rule3 Similarly, to describe that benchmark is machined first:

$$\forall i, j \in S, B_{ij} = \begin{cases} 1 & (f(i) \notin D_s) \&\& (f(j) \in D_s), i < j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Rule4 To describe in order to guarantee theoretical dimensions of aerodynamic configurations, the contours should be processed first:

$$\forall i, j \in S, C_{ij} = \begin{cases} 1 & (f(i) \notin C_s) \&\& (f(j) \in C_s), i < j \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Considering the multi-cavity characteristic and particularity of intersecting hole, the Layer-based Rules are proposed to improve machining quality of intersecting hole.There are three types of layer in this paper:

(1) the layer of a step slot.

(2) the layer of machining features in an cavity.

(3) the layer of an cavity with rough machining.

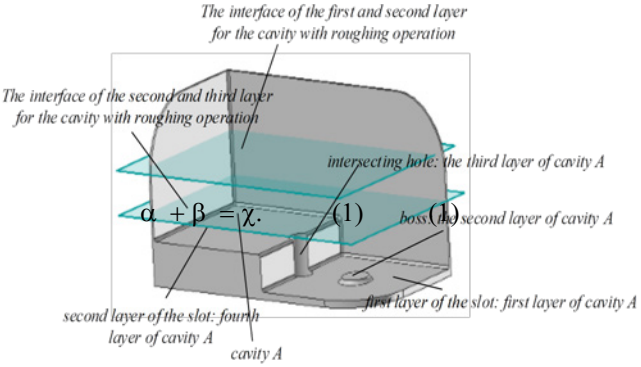
Rule5 To describe machining machinability of layer type 1 and 2, the equation is expressed as follows:

$$\forall i, j \in S, LBS_{ij} = \begin{cases} 1 & L(i) < L(j) (i < j) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $L(i)$ is the function calculating the layer number of machining element i.

To ensure the machinability of layer type 3, the finishing operations of an upper layer which is intersected with a hole should be machined before roughing operations of the hole while the lower layer which is intersected with the hole should be processed after finishing operations of the hole. The sketch is shown in Fig.1:

Fig. 1. Machining hierarchical diagram of a cavity.



Rule6 To disallow the sequence violate the compulsive constraints, the following equation is defined:

$$\sum_{i,j=0}^{n} M_{ij} + K_{ij} + B_{ij} + LBS_{ij} \le 0 \quad (8)$$

Rule7 Each machining feature under certain circumstance has only one standard machining operation sequence.

Rule8 Poorer rigidity element should be machined before larger ones. For example, a web of a cavity is processed before the wall and the web should be machined after the rib.

All the knowledge-based rules above are defined subjectively. There are also several flexible constraints which can not be defined subjectively.

III. Sequencing of machining elements by modified genetic algorithm

GA is an adaptive and global optimizing probability search method which simulates heredity and evolution of biology in environment, and it is widely used in process planning [10-12].

The fitness function in this paper is made up of two parts: (1) The fitness function of machining efficiency F_{efc}. (2) The fitness function of machining quality F_{qua}. It can be expressed by the following equation:

$$F = \omega_1 \times F_{efc} + \omega_2 \times F_{qua} \tag{9}$$

where ω_1 and ω_2 are the weight coefficients of machining efficiency and machining quality respectively.

A. Calculation of machining quality

Step-symmetry milling process can greatly reduce machining distortion and let the stress distribute uniformly for thin-walled and multi-cavity structural parts with roughing operations. However, there are two problems:

- Workload becomes heavier along with the increase of the number of cavities because the step-symmetry milling process is always obtained manually at present.
- The adjacent cavities may be staggered so that the impact on each adjacent cavity may be different.

Aiming at the problems above, this paper studies the expression of step-symmetry milling process.

Associated concepts and rules.

a) separating faces of a cavity: The derivative faces of outer-ring concave edges of each layer are separating faces of the cavity. The length of the concave edge is equal to the length of the separating face.

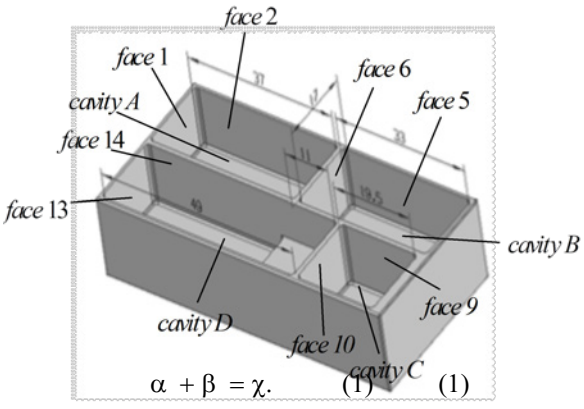
b) adjacent length of two cavities $L_{pub}(x,y)$: The length of adjacent part of cavity x and cavity y. It is regard as the indicator judging the extent of the influence.

c) **Rule 9:** If the minimum distance between two separating faces is within the limits of wall thickness and normal vectors of the separating faces are on the contrary approximately, then the two separating faces are adjacent. The limit of wall thickness depends on the real operation.

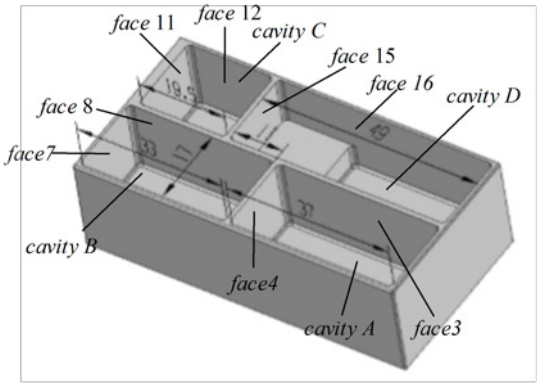
The separating-face-relationship matrix and cavity-relationship matrix are built to express whether the cavities are adjacent and the height of cavity at the current state. The calculation steps of two matrixes are listed as follows :

Step 1: Find the set of separating faces of each cavity and record the length of each separating face according to concept(a). The sample model is presented in Fig.2 from two directions. S_A={face1, face2, face3, face4} where S_A is the separating face set of cavity A.

Fig. 2. Instance model to calculate adjacent lengths of separating faces.



(a)



(b)

Fig. 3. Separating-face-relationship matrix.

	face1	face2	face3	face4	face5	face6	face7	face8	face9	face10	face11	face12	face13	face14	face15	face16
face1	17															
face2		37														
face3			37											1		
face4				17		1										
face5					33											
face6				1		17										
face7							17									
face8								33	1					1		
face9								1	19.5							
face10										18					1	
face11											18					
face12												19.5				
face13													18			
face14			1					1						49		
face15										1					18	
face16																49

(a)

	face1	face2	face3	face4	face5	face6	face7	face8	face9	face10	face11	face12	face13	face14	face15	face16
face1	17															
face2		37														
face3			37											37		
face4				17		17										
face5					33											
face6				17		17										
face7							17									
face8								33	19.5					11		
face9								19.5	19.5							
face10										18					18	
face11											18					
face12												19.5				
face13													18			
face14			37					11						49		
face15										18					18	
face16																49

(b)

Step 2: Judge whether the cavities are adjacent according to **Rule 9**. The separating-face-relationship matrix of the sample model is shown in Fig.3(a). The figures of diagonal line mean the length of the separating faces and figures "1" in non-diagonal line means the two separating faces are adjacent.

Step 3: Acquire $L_{pub}(x,y)$. The separating-face-relationship matrix of the sample model is shown in Fig.3(b) and the flow chart is shown in Fig.4.

Step 4: Convert the separating-face-relationship matrix into cavity- relationship matrix according to the cavities that the separating faces belong to. The figures of diagonal line mean the height of each cavity at current state and figures of non-diagonal line mean the adjacent length of cavities at the same state. The cavity- relationship matrix of the sample model at a certain state is presented in Fig.5.

Fig. 4. Creating Separating-face-relationship matrix.

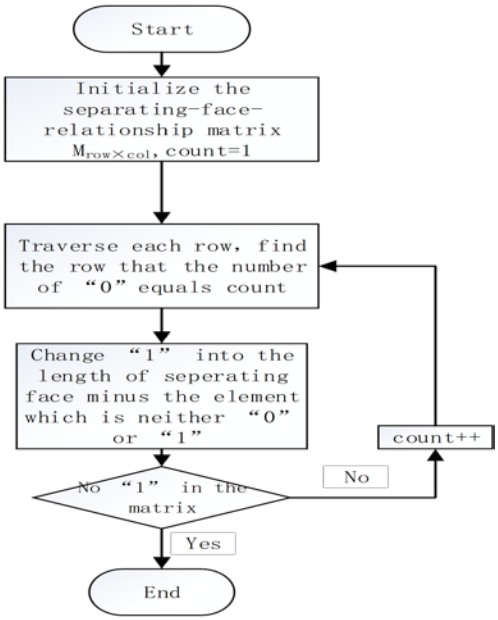


Fig. 5. Example of a Cavity-relationship matrix

	cavityA	cavityB	cavityC	cavityD
cavityA	15	17		37
cavityB	17	15	19.5	11
cavityC		19.5	15	18
cavityD	37	11	18	15

The figures of diagonal line mean height of each cavity and figures of non-diagonal line mean $L_{pub}(x,y)$ at the current state. The height between each cavity and its adjacent cavities is recorded before and after every operation to judge whether or not it satisfies the step-symmetry milling process. If the height is greater than a cutting depth a_p then it doesn't meet the requirement. The adjacent length of two cavities L_{pub} is regard as the indicator judging the extent of the influence. $F_{qua}=F_{qua}+L_{pub}$ if the demand of step-symmetry milling process doesn't meet.

B. Calculation of machining efficiency

The fitness function in this paper is made up of two parts: (1)The fitness function of machining efficiency F_{efc}. (2)The fitness function of machining quality F_{qua}. The switch time of machines, tools and fixtures is used to evaluate machining efficiency. The stepped symmetrically milling process is used to evaluate machining quality. It can be expressed by the following equations:

The changing times of machine tools, tools and fixtures are regard as the indicator to measure machining efficiency. The replacement is needed when two adjacent operations are machined on the different machine tools. The indicator can be given as follows:

$$F_{mac} = \omega_{mac} \times \sum_{i=1}^{n-1} \lambda(M_{i+1}, M_i) \tag{10}$$

where F_{mac} expresses the switch indicator of machine tools, $\omega_{mac}$is the transform factor of mahine tool, M_i represents the ID of machine tool at parameter i, $\lambda(M_{i+1}, M_i)$ can be computed by following equation:

$$\lambda(x,y) = \begin{cases} 1, & x \neq y \\ 0, & x = y \end{cases} \tag{11}$$

In the same way, the switch indicator of tools is given as follows:

$$F_{tol} = \omega_{tol} \times \sum_{i=1}^{n-1} \lambda(M_{i+1}, M_i) \tag{12}$$

where ω_{tol} is the transform factor of machine tool, which is also a constant.

$$F_{cla} = \omega_{cla} \times \sum_{i=1}^{n-1} \lambda(M_{i+1}, M_i) \tag{13}$$

where ω_{cla} is the transform factor of fixture.

The final indicator of machining efficiency F_{efc} can be therefor computed as follows:

$$F_{efc} = F_{tol} + F_{mac} + F_{cla} \tag{14}$$

where means the less value of F_{efc} the better machining efficiency will be.

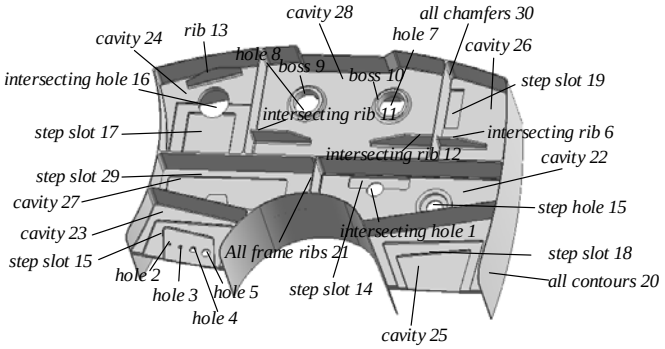
Z-Score standardization method is used in this paper. The disposed data accord with normal distribution. The conversion function can be written as follows: $x^* = \frac{x-\mu}{\sigma}$, where μ is mean value of sample data, σ is standard deviation of sample data. The value of fitness function represents the distance from the average level. The very top of the populations are reserved during iterations.

IV. Implementation and case study

A. Case of sequencing

The approach presented in this paper has been implemented as the Sequencing of thin-walled and multi-cavity aircraft structural parts prototype system. This module has been written in VS2010 and has done secondary development under UG8.5 environment. The instance model is shown in Fig.6. and the sequencing procedure of one side for the model is given.

Fig. 6. Sample model to sequence



There are 84 main machining elements in all. The number of the population is 100, the cross probability is 0.4 and the variation probability is 0.1. Given space limitation, part of the sequencing results are shown in table 1 after iterations of 500 times.

TABLE I. Optimized sequencing result of machining elements

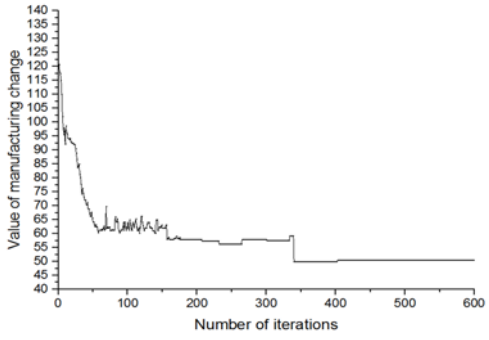
NO.	Feature	Opertion	Cavity	Machine	tool	fixture
1	21	rough milling (top face of rib 21)	/	M01,M03	T01	F01,F02
2	20	rough milling(contour 20)	/	M02	T02	F01
3	21	Semi-finish milling (top face of rib 21)	/	M01,M03	T03	F01
4	21	finish milling (top face of rib 21)	/	M01,M03	T04	F01,F02
5	20	finish milling(contour 20)	/	M02	T05,T04	F01
6	26	rough milling (5th layer of cavity 26)	26	M01	T02	F01
7	28	rough milling (4th layer of cavity 28)	28	M01	T02	F01
8	22	rough milling (3th layer of cavity 22)	22	M01	T02	F01
9	27	rough milling (4th layer of cavity 27)	27	M01	T02	F01
10	26	rough milling (4th layer of cavity 26)	26	M01	T02	F01
11	23	rough milling (4th layer of cavity 23)	23	M01	T02	F01
12	25	rough milling (3th layer of cavity 25)	25	M01	T02	F01
13	24	rough milling (3th layer of cavity 24)	24	M01	T02	F01
14	27	rough milling (3th layer of cavity 27)	27	M01	T02	F01
15	24	rough milling (2th layer of cavity 24)	25	M01	T02	F01
16	23	rough milling (3th layer of cavity 23)	23	M01	T02	F01
17	28	rough milling (3th layer of cavity 28)	28	M01	T02	F01
18	28	rough milling (2th layer of cavity 28)	28	M01	T02	F01
19	25	rough milling (2th layer of cavity 25)	25	M01	T02	F01
20	24	rough milling (1th layer of cavity 24)	24	M01	T02	F01
21	22	rough milling (2th layer of cavity 22)	22	M01	T02	F01
22	26	rough milling (3th layer of cavity 26)	26	M01	T02	F01
23	26	rough milling (2th layer of cavity 26)	26	M01	T02	F01
24	25	rough milling (1th layer of cavity 25)	25	M01	T02	F01
25	22	rough milling (2th layer of cavity 22)	22	M01	T02	F01
26	26	rough milling (1th layer of cavity 26)	26	M01	T02	F01
27	27	rough milling (2th layer of cavity 27)	27	M01	T02	F01
28	27	rough milling (1th layer of cavity 27)	27	M01	T02	F01
29	22	rough milling (1th layer of cavity 22)	22	M01	T02	F01
30	28	rough milling (1th layer of cavity 28)	28	M01	T02	F01
31	23	rough milling (1th layer of cavity 23)	23	M01	T02	F01
32	13	rough milling (top face of rib 13)	24	M01	T01	F01
33	6	rough milling (top face of intersecting rib 6)	26	M01,M03	T01	F01
34	12	rough milling(top face of intersecting rib 12)	28	M01,M03	T01	F01,F03
35	11	rough milling(top face of intersecting rib 11)	28	M01,M03	T01	F01,F03
36	6	finish milling (top face of intersecting rib 6)	26	M01,M03	T04	F01,F03
37	12	finish milling(top face of intersecting rib 12)	28	M01,M03	T04	F01,F03
38	13	finish milling(top face of rib 13)	24	M01,M03	T04	F01,F03
39	11	finish milling(top face of intersecting rib 11)	28	M01,M03	T04	F01,F03
40	16	rough milling(intersecting hole 16)	24	M01,M04	T06	F01,F04
41	16	boring (intersecting hole 16)	24	M01,M04	T08	F01,F04
42	17	rough milling(step slot 17)	24	M01	T07	F01
43	19	rough milling(step slot 19)	26	M01	T07	F01
44	1	rough milling(intersecting hole 1)	22	M01,M04	T06	F01,F04
45	1	boring (intersecting hole 1)	22	M01,M04	T08	F01,F04
46	14	rough milling(step slot 14)	22	M01	T07	F01
47	15	rough milling(step slot 15)	23	M01	T07	F01
48	18	rough milling(step slot 18)	25	M01	T07	F01
49	29	rough milling(step slot 29)	27	M01	T07	F01
50	10	rough milling(boss 10)	28	M01	T09	F01
51	9	rough milling(boss 9)	28	M01	T09	F01
52	10	finish milling(boss 10)	28	M01	T10	F01
53	9	finish milling(boss 9)	28	M01	T10	F01
54	28	finish milling(web of cavity 28)	28	M01,M05	T11,T12	F01
55	24,17	finish milling(web of cavity 24,setp slot 17)	24	M01,M05	T11,T12	F01
56	22,14	finish milling(web of cavity 22,setp slot 14)	22	M01,M05	T11,T12	F01
57	26,19	finish milling(web of cavity 26,setp slot 19)	26	M01,M05	T11,T12	F01
58	25,18	finish milling(web of cavity 25,setp slot 18)	25	M01,M05	T11,T12	F01
59	27,29	finish milling(web of cavity 27,setp slot 29)	27	M01,M05	T11,T12	F01
60	23,15	finish milling(web of cavity 23,setp slot 15)	23	M01,M05	T11,T12	F01
61	22,14	finish milling(web of cavity 22,setp slot 14)	22	M01,M05	T11,T12	F01
62	27,29	finish milling(wall of cavity 27, step slot 29)	27	M01,M05	T11,T12	F01
63	24,17	finish milling(wall of	24	M01,M05	T11,T12	F01

		cavity 24, step slot 17)				
64	28	finish milling(wall of cavity 28)	28	M01,M05	T11,T12	F01
65	22,14	finish milling(wall of cavity 22, step slot 14)	22	M01,M05	T11,T12	F01
66	26,19	finish milling(wall of cavity 26, step slot 19)	26	M01,M05	T11,T12	F01
67	23,15	finish milling(wall of cavity 23, step slot 15)	23	M01,M05	T11,T12	F01
68	25,18	finish milling(wall of cavity 25, step slot 18)	25	M01,M05	T11,T12	F01
69	30	finish milling (chamfer 30)	/	M01	T13	F01
70	30	finish milling (chamfer 30)	/	M01,M06	T14	F01
71	2	spot drilling(hole 2)	23	M01,M03	T07	F01,F04
72	3	spot drilling(hole 3)	23	M01,M03	T07	F01,F04
73	4	spot drilling(hole 4)	23	M01,M03	T07	F01,F04
74	5	spot drilling(hole 5)	23	M01,M03	T07	F01,F04
75	3	drilling(hole3)	23	M01,M03	T15	F01,F04
76	2	drilling(hole2)	23	M01,M03	T15T16	F01,F04
77	4	drilling(hole 4)	23	M01,M03	T15T16	F01,F04
78	5	drilling(hole 5)	23	M01,M03	T15T16	F01,F04
79	4	reaming(hole 4)	23	M01,M03	T15T16	F01,F04
80	5	reaming(hole 5)	23	M01,M03	T15T16	F01,F04
81	2	reaming(hole 2)	23	M01,M03	T15	F01,F04
82	3	reaming(hole 3)	23	M01,M03	T15	F01,F04
83	15	rough milling(step hole 15)	22	M01	T06	F01,F04
84	15	boring (step hole 15)	22	M01	T08	F01,F04

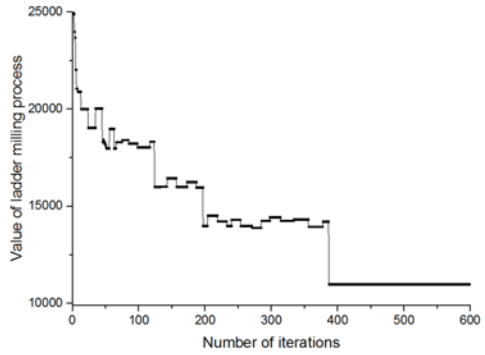
B. Analysis of convergence

The convergence is analyzed from aspects of machining effective and machining quality(Fig.8), plotted by the number of iterations on horizontal axis ,the value of manufacturing change on the vertical(Fig.8(a)) and the value of step-symmetry milling process on the vertical(Fig.8(b)). Due to the operations of Z-score standardization, the values of fitness function have minute variation within a certain range at some stage. It can be seen from Fig.8(a) that apparent declines occur at the 152 , 342 generation and the value begins to stabilize after the 400 generation. It is shown in Fig.8(b) that the values begin to fall significantly at the 120, 200, 391 generation and the value becomes stable after the 391generation.

Fig. 7.Evolution of the best individual



(a) Values of manufacturing changes



(b) Values of stepped symmetrically milling process

V. Conclusion

Compared to traditional sequencing method, this paper proposed a method using both knowledge-based rules and GA algorithm. The knowledge-based rules are used to meet the requirements of compulsory constraints. The separating-face-relationship matrix and cavity-relationship matrix are built to express step-symmetry milling process to meet the requirement of machining quality. The sequencing result of machining elements can be used to NC programming directly.

References

[1] Y. Yusof, and K. Latif, "Survey on computer-aided process planning", The International Journal of Advanced Manufacturing Technology , vol. 75, pp. 77-89, 2014.

[2] L. Wang, "Computer-aided process planning - A critical review of recent developments and future trends", International Journal of Computer Integrated Manufacturing, vol. 24, pp. 1-31, 2011.

[3] A. Nassehi, W. Essink and J. Barclay, "Evolutionary algorithms for generation and optimization of tool paths", CIRP Annals-Manufacturing Technology, vol. 64, pp. 455-458, 2015.

[4] A.M. Dalavi, P.J. Pawar and T.P. Singh, "Tool path planning of hole-making operations in ejector plate of injection mould using modified shuffled frog leaping algorithm", Journal of Computational Design and Engineering, vol. 3, pp. 266-273, 2016.

[5] T. C. Chang, "Expert process planning for manufacturing", Addison-Wesley Publishing Company, 1990.

[6] M. Salehi, and R. Tavakkoli-Moghaddam, "Application of genetic algorithm to computer-aided process planning in preliminary and detailed planning", Engineering Applications of Artificial Intelligence, vol. 22, pp. 1179-1187, 2009.

[7] Z. Liu, and L. Wang, "Sequencing of interacting prismatic machining features for process planning", Computers in Industry, vol. 58, pp. 295-303, 2007.

[8] A. Mokhtar, and X. Xu, "Machining precedence of 2½D interacting features in a feature-based data model", Journal of Intelligent Manufacturing, vol. 22, pp. 145-161, 2011.

[9] A. C. Lin, S. Y. Lin, D. Diganta, and et al, "An integrated approach to determining the sequence of machining operations for prismatic parts with interacting features", Journal of Materials Processing Technology, vol. 73, pp. 234-250, 1998.

[10] Y. Su, X. Chu, D. Chen, and et al, "A genetic algorithm for operation sequencing in CAPP using edge selection based encoding strategy", Journal of Intelligent Manufacturing, pp. 1-20, June, 2015.

[11] L. Liu, and L. H. Qiao, "Operation Sequencing Using Genetic Algorithm", Applied Mechanics & Materials, vol. 163, pp. 57-61, 2012.

[12] H. B. Ouyang, and B. Shen, "STEP-NC oriented process planning optimization based on hybrid genetic algorithm", Jisuanji Jicheng Zhizao Xitong/computer Integrated Manufacturing Systems Cims, vol. 18, pp. 66-75, 2012.

Maintenance Planning

Optimal joint production, maintenance and quality control for a single machine with random failures using Matlab/Simulink optimization-simulation

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Abdessamad AIT EL CADI[*], Ali GHARBI[†], Bassem BOUSLAH[†], Abdelhakim ARTIBA [*]
[*]LAMIH UMR CNRS 8201 – University of Valenciennes and Hainaut-Cambrésis, France
abdessamad.aitelcadi@univ-valenciennes.fr
[†]LC2SP – École de Technologie Supérieure, Montréal (Qc), Canada
ali.gharbi@etsmtl.ca

Abstract—Manufacturing systems subject to random outages are expensive to control. One should find the right maintenance strategy to implement and should cut quality impact on the final product to satisfy the customer's demand. In this paper, we consider the integration of the production, maintenance and quality control for a single machine. The system is subject to stochastic failures that are leading to more and more non-quality products. The inspection details the defective objects and sends them to a rectification phase, from where they become good. The objective is to find optimal production and maintenance control policies that minimize the long-term total cost. This cost includes cost of ownership, cost of lost sales, preventive and corrective maintenance costs, inspection costs and rectification costs. To solve this problem, we develop a simulation based optimization model with the framework Matlab / Simulink. The numerical results prove the effectiveness of the approach. The sensitivity analysis shows the complexity of the joint design of such systems and depicts the impact of its parameters, like maintenance costs, on the solutions.

Keywords— simulation; optimization; Simulink / Matlab; maintenance; quality; production control; preventive maintenance; inspection.

I. Introduction

Nowadays, for an intelligent asset management ones should take into account simultaneously the production, the maintenance and the quality control. Undeniably, the control of any of these three elements is dependent of the two others: The perfect production control, for example, could not be achieved without handling the maintenance and quality ones'. The quality and the reliability degradation impacts, indubitably, the planning and the control of the production. Also the operating speed impact the degradation of both reliability and quality [1].

Nonetheless traditionally, Production, maintenance and quality have been treated as separate problems [2]. In the last decade, there is a growing interest in integrating them. This trend may be motivated by the important findings that indicates that the integrated models outperform those dealing with only one separate aspect at a time [3, 4]. It could be, also, explained by the technology evolutions and the increasing computing performances. Integrated models fundamentally uses a holistic approach that takes into account the dependency, the interrelation, interaction and intersection between all these aspects. These models provide a best approach to analyzing, designing and optimizing manufacturing systems through by taking into consideration the substantial interrelations and interactions between production, quality and maintenance aspects. In the literature, the majority of the works handle the integration through the evaluation and the analysis of the performance of such systems under various configurations [5, 6 and 7]. While, little works aims the joint design of the three control policies [8, 9, 10 and 11].

The objective of this paper is to propose a simulation model for the joint production, maintenance and quality control in the case of a single machine that is subject to random failures and quality and reliability degradation. We integrate the three aspects within the same model.

The second objective of this paper is to introduce Simulink/Matlab as a simulation and a simulation-based optimization framework. We develop the model in Simulink and use Matlab Optimization Toolbox to find the optimum. We aim to prove the efficiency of this framework.

In-fact, Simulation and Simulation-based optimization are important for modeling and solving real problems that are generally more complex, of large size, dynamic and stochastic. The simulation-based optimization is widely used in a variety of applications including supply chain [12], manufacturing systems [13], project management [14], operations scheduling [15] and inventory systems [16]. Considering the mathematical limits of analytical resolution approaches [17], the development of empirical methods is necessary. For example, Gharbi et al. [18], to determine the optimal control policy parameters for an unreliable manufacturing cell with adjustable capacity, they combine a design of experiments, simulation modelling, and response surface methodology.

In the simulation-based optimization context, the simulation model should be fast, to allow many replication and evaluation. And the optimization tool should be robust enough to handle the randomness and uncertainty. We aim, in this work, to prove that

Matlab/Simulink is suitable for this task. Even if traditionally, in logistics and supply chain problem, researchers and practitioners use a simulation softwares like Arena - Arena popularity lies in its ability to handle very complex systems and in his optimization tool OptQuest [19].

As shown in our previous work [20], Simulink/Matlab is powerful plat-form that could be used to simulate manufacturing systems. In the present paper, we use this platform to tackle one of the hardest problem in the integration of production maintenance and quality control. No previous work dealt with the same problem under Simulink and our results shows again the power of this tools to handle more complex problem.

The present paper is organized as follow: In section two we present the studied problem. Section 3 introduces Simulink of Matlab, a simulation tool, and presents the simulation model. Section 4, is dedicated to the experiments and results. And section 5 concludes this work.

II. Problem description

In this section, we, first, present the studied manufacturing unit and its context then the used notations in the present paper.

A. The manufacturing unit

We consider a production unit with a single machine subject to stochastic failures as shown in Figure 1. The product quality and the machine reliability are degrading with time. The failure rate and the fraction of produced defects increase with time.

In Figure 1, the machine makes one unique product. The product is hold in a buffer before being inspected. The inspection phase, of non-negligible duration, is supposed perfect and check 100% of the items. After the inspection, the good parts are sent to the final buffer and the bad parts are rerouted to a rectification station for a reprocessing phase. This phase has a duration that affects the evolution of the final buffer level. The 100% inspection and the rectification ensure a defect-free production. The final products in the servicing stock are sent to the customer, to fulfill a constant demand rate. The unsatisfied sales are lost. The production control is done with the hedging-point policy (HPP), the maintenance control is completed through an age-based Preventive Maintenance (PM) and if failures occur a corrective maintenance (CM) actions remove them. After each maintenance action, PM or CM; the system is brought back to the state of as good as new.

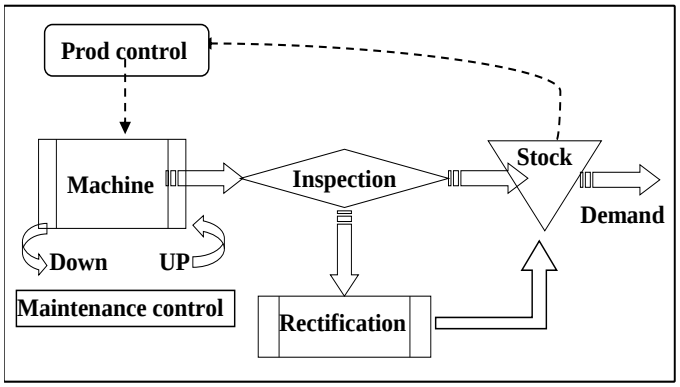


Fig. 1 Manufacturing system under study

The objective is to optimize jointly the production and maintenance control policies under the quality constraints, to minimize the total cost incurred over a long run. This cost includes the inventory holding cost, the lost sales cost, the CM and PM costs and the inspection and rectification costs.

B. Notations

The demand rate is fixe and equal to d and unfulfilled sales are lost.

The machine has two states:

- UP, the machine working and producing parts at a given rate u
- $DOWN$, the machine is not working and it is under preventive repair (or corrective one). The production rate is 0.

The buffer has the following characteristics:

- Z: Its capacity, the maximum level of storage.
- s: The current buffer level. s $\in]-\infty,$ Z]. It is also the state of the buffer.
- $s^+ = max(s, 0)$: The surplus.
- $s^- = max(-s, 0)$: The lost sales.

The inspection and rectification characteristics are:

- T_{insp}: Unit inspection duration.
- T_{rect}: Unit rectification duration.
- The quality inspection is error-free, so that all defective items are sorted and reprocessed before being transmitted to the serviceable stock.
- $p(a)$: The proportion of defect is depending on the age t of the machine. The aging of the machine leads to more defects i.e. $p(t)$ is increasing with the age.
- $q(a) = 1 - p(a)$: The proportion of good parts at the age a of the machine.

The machine characteristics are:

- U_{max}: The maximum production rate.
- a_f: Machine age at failure (working period with no failure). It is stochastic and has a given distribution. The failure rate is increasing with age of the machine.
- $MTTF$: is the mean time to failure.
- t_r: Time to repair (repair duration). It is stochastic and has a given distribution.
- $MTTR$: is the mean time to repair.
- M: Maintenance threshold (Period with no failure after which the machine is stopped for a preventive action).
- t_m: Time to maintain (maintenance duration). It is stochastic and has a given distribution.
- $MTTM$: is the mean time to maintain.

- Each repair, either after failure or preventive one, brings the machine to the state as good as new. These actions reset the quality level too.
- After each maintenance action, PM or CM, if the buffer size s is negative the machine starts immediately; otherwise we wait until the buffer is empty before restarting the production. Effectually, as the machine is good as new there is no need to build any safety stock.
- The production rate u of the machine is controlled through the so-called hedging point policy (HPP) presented in Akella and Kumar [21]. The production policy is given by the following equations:

$$u = \begin{cases} U_{max} & if\ s < Z\ \&\ machine\ UP \\ d & if\ s < Z\ \&\ machine\ UP \\ 0 & if\ machine\ DOWN \end{cases} \quad (1)$$

The costs structure:

- C_h: Unit inventory holding cost per unit time.
- C_l: Cost of one unit of lost sales.
- C_{pm}: Preventive maintenance cost.
- C_{cm}: Corrective maintenance cost.
- C_{insp}: Unit inspection cost.
- C_{rect}: Unit rectification cost of a defective item.

III. Modeling with Simulink / Matlab

Simulink, developed by MathWorks, is a graphical programming environment for modeling, simulating and analyzing multi-domain dynamic systems. Its primary interface is a graphical block diagramming tool and a customizable set of block libraries [22].

It is a modeling, simulation and analysis software for dynamic systems. It allows the study of linear and nonlinear systems modeled continuously (continuous-time), discrete (time steps) or hybrids. The system can be unicycle (single rate) or multi-cycle (multi-rate).

For modeling, Simulink uses a graphical interface GUI (Graphical User Interface) that enables, using predefined blocks or users' created ones and connectors, creating a model as quickly as possible without writing code. Models are hierarchical, i.e. each blocks can be itself a model composed of other blocks.

Unlike the majority of simulation softwares, who are entity-based discreet event simulation software, Simulink could be seeing as signal-based simulation software.

Simulink take advantage of the being fully integrated with Matlab. So, the user has access all the Matlab feature and toolboxes and could create its own scripts algorithm and model that he could use indefinitely.

Also Simulink has two important tools: SimEvents and Stateflow. SimEvents [23] has been designed to simulate Discrete Event Systems (DES) and it is fully embedded in Simulink, who is a traditional time-driven simulator, so that it is equipped with functionality that enables an effective co-existence of time-driven and event-driven components in complex hybrid systems. In addition, this design allows SimEvents to take advantage of a rich collection of visualization, data processing, and computation tools in both Simulink and MATLAB. Stateflow is an environment for modeling and simulating combinatorial and sequential decision logic based on state machines and flow charts. Stateflow lets you combine graphical and tabular representations, including state transition diagrams, flow charts, state transition tables, and truth tables, to model how your system reacts to events, time-based conditions, and external input signals.

With these feature Matlab/Simulink is a platform that offers a multi-paradigm tool for modeling complex systems.

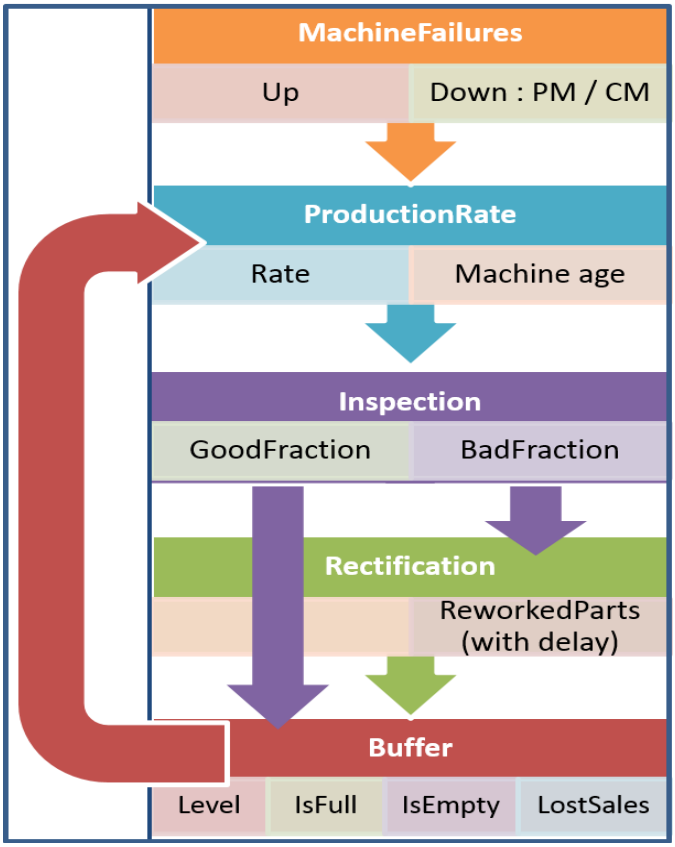


Fig. 2 The simulation model under Simulink

A. The simulation model

The modelling under Simulink is different from the one under classical manufacturing Simulation tools. Simulink is a signal-based simulation tools. We simulate signals that flow through different blocks. Each block has states, input and output ports. The states are used to store the block characteristics. The ports are used to receive input signals and broadcast output ones. The block uses the inputs to update its states and to compute outputs signals before broadcasting them. In this section. To build a model, in Simulink, we use blocks and connect them to each-other using their ports. There are some blocks that one could customize with his own code. The Simulink user guide [22] gives an overview and explains how to build a simulation model.

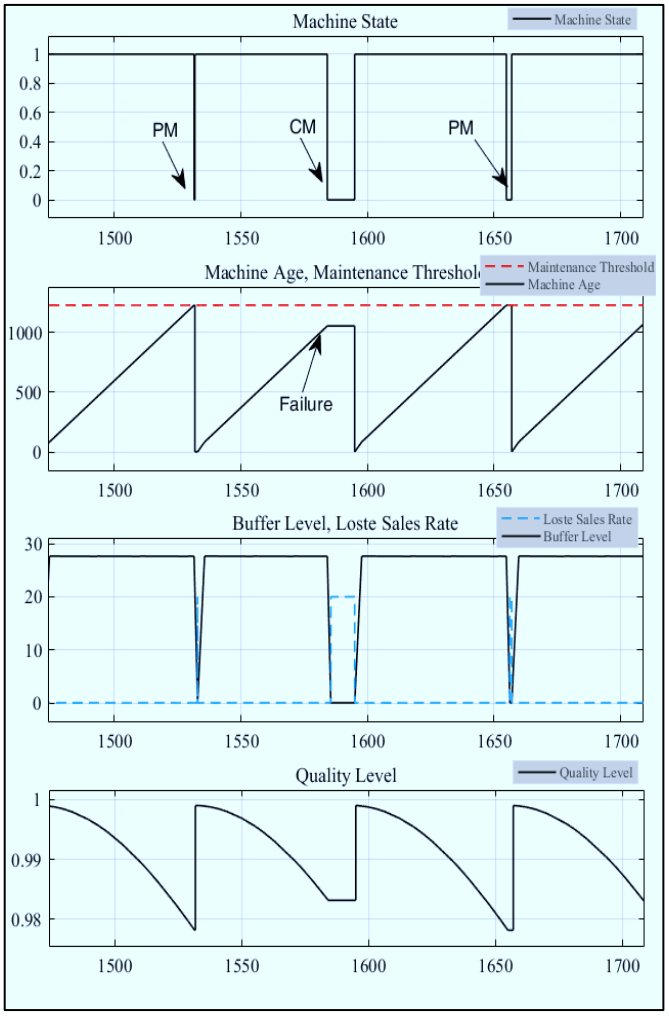


Fig. 3 Signals from the simulation: Machine state and age, buffer level, lost sales and the quality level.

Using the standard Simulink library, we build a simulation model to represent the manufacturing unit with the production control and maintenance control policies and the inspection and rectification phases. The Figure 2, illustrates the logic behind this Simulink simulation model. The model is built through five blocks: The failures generator "MachineFailures"; the production unit represented by its rate "ProductionRate"; the inspection phase "Inspection"; the rectification phase "Rectification"; the buffer block "Buffer". The arrows in figure 2 show the signals flow in the model.

First, the "MachineFailures" bock generate the machine's states (1: Up, 2: Stopped for PM action and 3: Down and under CM action). It uses the PM repair duration, the CM repair duration, the preventive maintenance threshold 'M' and the machine's lifespan distribution to compute its state (the machine state). The output signal - equal to the block current state - is sent to the "ProductionRate".

The "ProductionRate" block uses the input signals from the buffer level and the machine states to compute the current production rate. This rate is altered by the "Inspection block" to model the inspection phase. It splits the input signal into too signals, one for the good parts and the second for bad parts, and apply a delay on it. The delay is the inspection duration that depends on the rate itself. The resulting signals are sent, respectively, to the "Rectification" and "Buffer" blocks.

To implement the rectification phase, the "Rectification" block delay the input signal with a certain amount of time depending on the rate itself and the rectification duration. The good parts signal and the reworked part signal are gathered in the "Buffer" block to compute its state, the buffer level and the output signals like the lost sales rate.

The Figure 3 illustrate some output of the simulation model in Figure 2. We have, the machine state, the machine age, the buffer level, the lost sales rate and the quality level (percentage of good parts). The plots in this figure are used to validate the simulation model.

B. Optimum seeking

Once the simulation model is built, we use the Matlab/ Simulink framework to find the optimal policy, see Figure 4. For the optimum seeking, we use the simulation model as a black-box. A script in Matlab, wrap the simulation and define our fitness function. Then we use the optimisation toolbox of Matlab to run any of the heuristics and metaheuristics offered by the platform (see Fig. 4). All the optimisation algorithms are tunable. In Matlab we could interact with the solvers parameters, choose the right algorithm and adjust its ingredients.

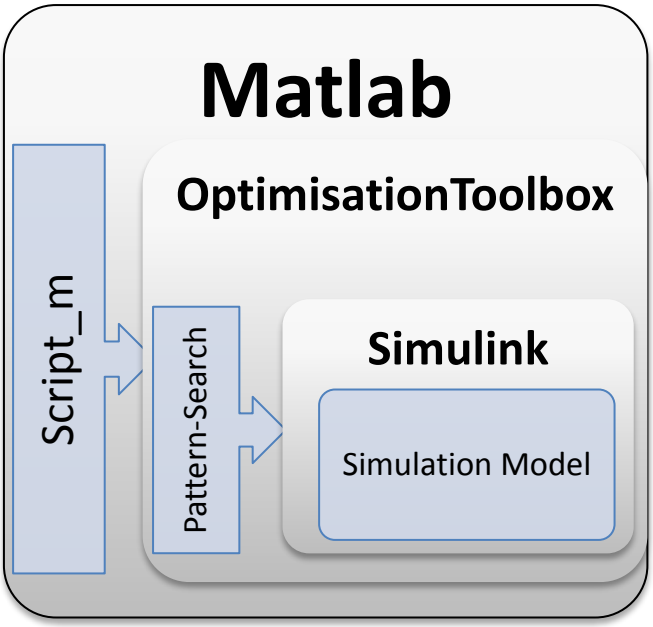


Fig. 4 Matlab optimization toolbox.

As the fitness the function, in our case, is not differentiable and the decision's variables, Z and M, are continues variables, we choose the most suitable algorithm in the Matlab optimisation toolbox, the "pattern-search". "Pattern-search" is derivative-free, or black-box search method. It does not require the gradient to find the extrema of a given function. It starts with an initial point in the search space. It uses a pattern to find a sequence of points X0, X1, X2 …, which approach an optimal point. The value of the objective function either decreases or remains the same from each point in the sequence to the next.

The explanation on how the Generalized Pattern Search (GPS) works could be find in [24].

IV. Experiments and results

This section provides an illustrative example for the proposed integrated model. We will present the used data for the experiments, second the results for the optimal solution and then a sensitivity analysis shows the impacts of the system parameters on the optimal solution. For all the results, including the sensitivity analysis, we provided a validation through the simulation and a confidence interval of 95% level.

All simulations are done over 100 000 units of time: From instant 0 to instant 101 000 with a transition period of length 1 000 units of time. All statistics are cleared at the instant 1 000. Thus, all the collected data are measured during the steady state and for 100 000 units of time.

For all the experiments, we used a PC with windows 10 as operating system and a processor "Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz". For the simulation, we used Simulink of MATLB R2016b.

A. The data

In this example, we consider the following parameters in the appropriate units:

- The time to failure has a Weibull distribution with a scale parameter $\lambda = \left(2/\sqrt{\pi}\right)10^4$ and a shape parameter $k = 2$.
  The failure rate is increasing with the machine age.
- The preventive maintenance duration has an exponential distribution with a mean: $MTTM = 1$.
- The correction maintenance duration has an exponential distribution with a mean: $MTTR = 5$.
- The quality, the proportion of defect, depends on the age and is increasing with age. We used:
- $p(a) = (0.1 + 9.9(1 - exp(-2\pi 10^{-9}a^2)))10^{-2}$
- The other parameters are:
  $U_{max} = 15;\ d = 10; T_{ins} = 1.10^{-2};\ T_{rect} = 2.10^{-2}$
  $C_h = 0.01;\ C_l = 1.5;\ C_{pm} = 400;\ C_{cm} = 800;$
  $C_{ins} = 0.05; C_{rect} = 0.5.$

The defect proportion is chosen such that: (1) When the machine is new, the proportion is *0.1%*; (2) When the machine is, somehow, aged, an age of 1 MTTF, the proportion is around *5%*, which is acceptable (3) and after an age of about 2 MTTF, the proportion drop drastically to the worst acceptable value, which around *10%*.

B. The results

With the Simulink model as a fitness function and Matlab "Pattern-Search", we were able to find the optimal controls for the system described above.

Table 1 shows the optimal control point *(Z*=122.1; M*=7586.7)* and the optimal cost *18.2*. The same table list all the key indicators related to the optimal solution. By order after the optimal control point, we have the cycle length followed by the average value of: stock in the two stations inspection and rectification, stock of final goods, lost sales, preventive maintenance frequency, corrective maintenance frequency, and number of inspected unit per unit of time and the number of rectified items per unit of time.

Table 1: The optimal solution

Element	Value
Z*	**122.1**
M*	**7586.7**
Cost*	**18.2**

After the determination of the optimal solution we conduct a sensitivity analysis to check the impacts of the maintenance cost and lost sales cost on the optimal solution. The Tables 2 and 3 shows respectively the results for the corrective maintenance cost and the lost sale cost.

Both Tables are shaped in the same way: In the first column we present the changes made to the concerned parameter; in the second and third column we have the optimal solution associated to the change. We added some figures (Fig. 5 – 8) to summarize the impacts.

In the case of the Table 3 and Fig. 5 and 6, we increases the corrective maintenance cost. In Fig. 5 we notice that *M** decreases: As the failures impacts are higher (corrective maintenance cost high), the system should do the necessary to avoid them! Then *M** should be decreased.

Table 2: The impact of the corrective mainetnance cost

Cost change	Optimum Z*	Optimum M*
$\boldsymbol{C_{cm} = 0.01 \times C_{cm0} \leq C_{pm}}$	160.9	17126.0
$\boldsymbol{C_{cm} = 0.1 \times C_{cm0} \leq C_{pm}}$	141.2	14029.0
$\boldsymbol{C_{cm} = 0.5 \times C_{cm0} \leq C_{pm}}$	121.2	9681.2
$\boldsymbol{C_{cm} = 0.9 \times C_{cm0}}$	121.3	7900.4
Baseline $\boldsymbol{C_{cm} = C_{cm0} = 800}$	122.1	7586.7
$\boldsymbol{C_{cm} = 2 \times C_{cm0}}$	132.4	5655.5
$\boldsymbol{C_{cm} = 5 \times C_{cm0}}$	159.4	3643.9
$\boldsymbol{C_{cm} = 10 \times C_{cm0}}$	188.8	2565.5
$\boldsymbol{C_{cm} = 20 \times C_{cm0}}$	229.7	1795.7
$\boldsymbol{C_{cm} = 30 \times C_{cm0}}$	270.3	1453.3
$\boldsymbol{C_{cm} = 35 \times C_{cm0}}$	310.2	1336.5
$\boldsymbol{C_{cm} = 37 \times C_{cm0}}$	427.6	1282.8
$\boldsymbol{C_{cm} = 40 \times C_{cm0}}$	412.5	1237.4
$\boldsymbol{C_{cm} = 50 \times C_{cm0}}$	371.6	1114.9
$\boldsymbol{C_{cm} = 100 \times C_{cm0}}$	266.6	799.8
$\boldsymbol{C_{cm} = 200 \times C_{cm0}}$	187.3	561.8

But in Fig. 6, the variation of *Z** are not obvious! It decreases, increases and decreases again. There is 3 part in the graph. In the first part, *Z** decreases, because as *M** increases, the quality is improved and there is no need to bigger safety stock. In the second part, *Z** increases; in this part, the system uses both, *M** and *Z** to protect itself against shortage and

defects. The third part is different, it correspond to the fact that the system plan preventive actions during the safety stock building phase. Otherwise, the preventive action arrives before the system reaches the maximum level of stock. This makes Z^*, bounded by the preventive action. As in the last (shaded) part of the Table 2, Z^* and M^* are proportional. This means that the preventive action is due during the building phase of the safety stock. This means also that in such case, the system is working all the time at full speed U_{max}.

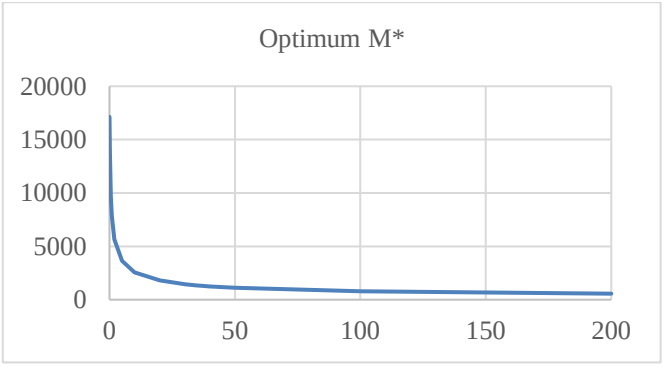


Fig. 5 Optimal M* with respect to C_{cm} change.

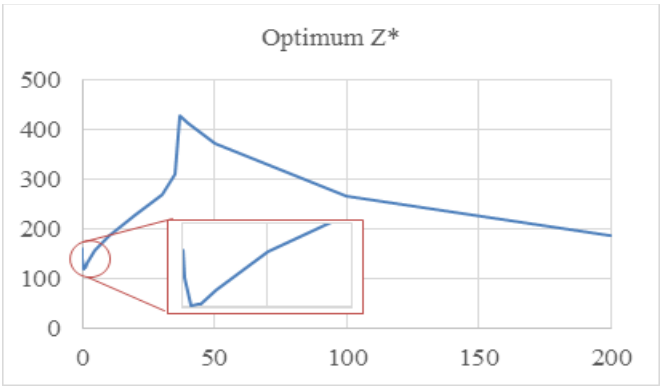


Fig. 6 Optimal Z* with respect to C_{cm} change.

In the first part (shaded one) of the Table 2, we notice that even with $C_{cm} \leq C_{pm}$, M^* is not infinite! Even when the preventive maintenance is too costly than the corrective one, we should keep the preventive actions at an optimal interval. The role of the preventive action here is to protect the systems from building too much defects.

Table 3: The impact of the lost sales cost

Cost change	Optimum Z*	Optimum M*
$\boldsymbol{C_l = 0.5 \times C_{l0}}$	0.0	8916.0
$\boldsymbol{C_l = 0.9 \times C_{l0}}$	96.6	7796.4
Base $\boldsymbol{C_l = C_{l0} = 1.5}$	122.1	7586.7
$\boldsymbol{C_l = 2 \times C_{l0}}$	317.9	6476.1
$\boldsymbol{C_l = 5 \times C_{l0}}$	760.7	5826.5
$\boldsymbol{C_l = 10 \times C_{l0}}$	1182.1	5703.3
$\boldsymbol{C_l = 20 \times C_{l0}}$	1630.1	6318.0
$\boldsymbol{C_l = 50 \times C_{l0}}$	2295.2	10179.9
$\boldsymbol{C_l = 100 \times C_{l0}}$	2813.7	18248.7

For the case of the Table 3 and Fig. 7 and 8, when the lost sales cost changes the levels of M* and Z* moves to keep the system performances. In the Fig. 8, Z* is increasing with the lost sales cost. This is normal, to protect the system from the high impact of the shortage the system build a larger safety stock.

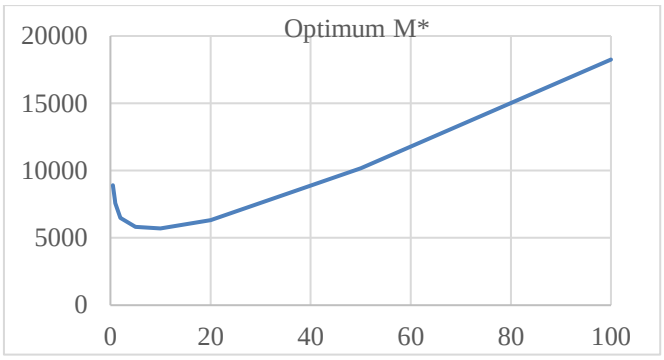


Fig. 7 Optimal M* with respect to C_l change.

But, in Fig. 7, the changes of M* is not monotonous! In the first part of the graph M* is decreasing to let help the system building good part, to make less defect. And in the second part M* increases so the system will have enough time to produce, to increase the availability of the system. If we go back to Fg.8, we have the two phases as in Fig. 7, but here the difference is in the slop of variation of Z*. In summary, these results show that the system should find the wright tradeoff between availability and safety stock.

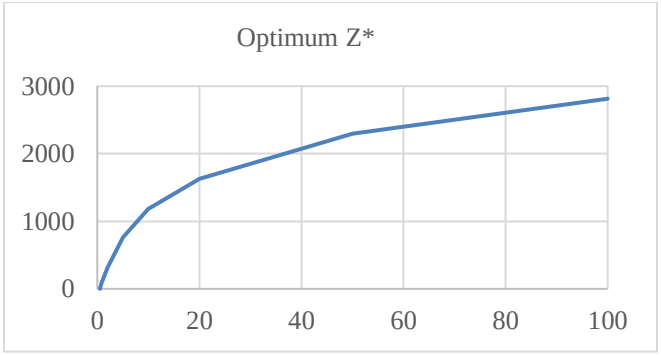


Fig. 8 Optimal Z* with respect to C_l change.

V. Conclusion

In this work we present the integration of the production, maintenance and quality control for a machine with random failures. We use Simulink/Matlab framework to simulate the system and to find the optimum.

The performances obtained are promising and prove the ability of the framework Matlab / Simulink to handle a complex system. The quality of the tools, especially Simulink, enable us to create a faster simulation model and the integration with Matlab help us to make a simulation optimization model that is quite performant.

The sensitivity analysis depicts the impact of the maintenance and lost sales cost on the control of the system. The impact are not obvious: the systems uses, jointly, the stock level and the maintenance threshold to mitigate the impact of quality and failure on the overall performance. The interaction between

these two control parameters is important. Without a joint design of the quality, maintenance and production we will not be able to see this interactions. Sometimes it is better to lower one control variable or both, sometimes to increase one and lower one and so on. There is an interaction and system should find the wright tradeoff between availability and safety stock in our case.

Acknowledgment

The present research work has been carried out in the context of the LIA (International Associate Laboratory) ROI-TML (Operational Research and Computer Science in Transportation, Mobility and Logistics) between the LAMIH UMR 8201 (France) and the CIRRELT (Canada). The authors gratefully acknowledge the support of these laboratories and relating institutions.

References

[1] Owen, J. H., & Blumenfeld, D. E. (2008). Effects of operating speed on production quality and throughput. International Journal of Production Research, 46(24), 7039-7056.

[2] Colledani, M., Ebrahimi, D., & Tolio, T. (2014). Integrated quality and production logistics modelling for the design of selective and adaptive assembly systems. CIRP Annals-Manufacturing Technology, 63(1), 453-456.

[3] Hadidi, L. A., Al-Turki, U. M., & Rahim, A. (2012). Integrated models in production planning and scheduling, maintenance and quality: a review. International Journal of Industrial and Systems Engineering, 10(1), 21-50.

[4] Inman, R. R., Blumenfeld, D. E., Huang, N., Li, J., & Li, J. (2013). Survey of recent advances on the interface between production system design and quality. IIE transactions, 45(6), 557-574.

[5] Kim, J., & Gershwin, S. B. (2008). Analysis of long flow lines with quality and operational failures. IIE transactions, 40(3), 284-296.

[6] Colledani, M., & Tolio, T. (2011a). Integrated analysis of quality and production logistics performance in manufacturing lines. International Journal of Production Research, 49(2), 485-518.

[7] Meerkov, S. M., & Zhang, L. (2010). Product quality inspection in Bernoulli lines: analysis, bottlenecks, and design. International Journal of Production Research, 48(16), 4745-4766.

[8] Rezg, N., Xie, X., & Mati, Y. (2004). Joint optimization of preventive maintenance and inventory control in a production line using simulation. International Journal of Production Research, 42(10), 2029-2046.

[9] Darwish, M., & Ben-Daya, M. (2007). Effect of inspection errors and preventive maintenance on a two-stage production inventory system. International Journal of Production Economics, 107(1), 301-313.

[10] Mhada, F. Z., Malhamé, R. P., & Pellerin, R. (2013). Joint assignment of buffer sizes and inspection points in unreliable transfer lines with scrapping of defective parts. Production & Manufacturing Research, 1(1), 79-101.

[11] Mourani, I., Hennequin, S., & Xie, X. (2007). Failure models and throughput rate of transfer lines. International Journal of Production Research, 45(8), 1835-1859.

[12] Hlioui, R., Gharbi, A., and Hajji, A. (2015). Replenishment, production and quality control strategies in three-stage supply chain. International Journal of Production Economics, 166, 90-102.

[13] Rivera-Gómez, H., Gharbi, A., Kenné, J. P., Montaño-Arango, O., and Hernandez-Gress, E. S. (2016). Production control problem integrating overhaul and subcontracting strategies for a quality deteriorating manufacturing system. International Journal of Production Economics, 171, 134-150.

[14] April, J., M. Better, F. Glover, and P.J. Kelly. 2004. New advances and applications for marrying simulation and optimization. In Proceedings of the 2004 Winter Simulation Conference, edited by R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 80-86. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

[15] Ait El Cadi, A., Benmansour, R., Serraj, F., and Artiba, A. (2015, October). A joint optimization-simulation model to minimize the makespan on a repairable machine. In Industrial Engineering and Systems Management (IESM), 2015 International Conference on (pp. 489-495). IEEE.

[16] Alferaei M.H., and A.H. Diabat. 2009. A Simulated Annealing Technique for multi-objective simulation optimization. Applied mathematics and computation 215: 3029-3035.

[17] Boukas, E.K., and Haurie, A. (1990). Manufacturing Flow Control and Preventive Maintenance: A Stochastic Control Approach. IEEE Transactions on Automatic Control, 35(9), p. 1024-1031.

[18] Gharbi, A., Hajji, A., and Dhouib, K., (2011). Production rate control of an unreliable manufacturing cell with adjustable capacity. International Journal of Production Research, 49(21), p. 6539-6557.

[19] Kelton, W. David, Randall P. Sadowski, and David T. Sturrock. 2007. Simulation with Arena. 4th edition. New

[20] Ait El Cadi, A., Gharbi, A., and Artiba, A. (2016, August). Matlab/Simulink -vs- Arena/Optquest: Optimal production control of unreliable manufacturing systems. In International Conference on Modeling, Optimization and Simulation - MOSIM'16.

[21] Akella, R., Kumar, P.R., 1986. Optimal control of production rate in a failure-prone manufacturing system. IEEE Transactions on Automatic Control AC 31, 116–126

[22] MathWorks. Simulink: A Program for Simulating Dynamic Systems, User Guide. The MathWorks, Inc., 2001

[23] MathWorks. SimEvents User's Guide. The MathWorks, Inc., 2005.

[24] Audet, C., & Dennis Jr, J. E. (2002). Analysis of generalized pattern searches. SIAM Journal on Optimization, 13(3), 889-903.

Opportunity study for the adoption of an extended warranty

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Zied BOUBAKER[1], Soumaya BOUGUERRA[2],
Anis CHELBI[3]
Centre de Recherche en Productique (CEREP)
Université de Tunis, Ecole Nationale Supérieure d'Ingénieurs de Tunis 5 Av. Taha Hussein,
Montfleury – Tunis, Tunisia
[1]Zied.ensit@gmail.com,
[2]soumaya.bouguerra@yahoo.fr,
[3]anis.chelbi@planet.tn

Nidhal REZG[4]
Laboratoire de génie industriel de production et de maintenance (LGIPM)
Université de Lorraine, UFR MIM, île du Saulcy,
Metz, France
[4]nidhal.rezg@univ-lorraine.fr

Abstract - In this paper, a mathematical model is developed to determine the minimum price at which a manufacturer should sell an extended warranty for a given product, as well as the maximum price the consumer should pay for the same extended warranty. The expressions of the expected total cost incurred by each party during the product's life cycle are developed considering a set of maintenance policies involving a random repair time.

Keywords — Extended warranty, random repair time, preventive maintenance.

I. Introduction

A great variety of products are sold with a warranty and with the possibility to extend it for a certain period when it expires. In fact, at the time of purchase, the consumer may be satisfied with the basic warranty period or may decide to pay an additional amount to have the benefit of an extended warranty period (generally shorter than the basic one).

The extended warranty helps the manufacturer to remain in touch with the consumer a certain period after the expiration of the basic warranty period. It also allows supplying more data on the performance of products during their lifecycle. Such data are particularly necessary for research activities and development of enhanced or new products.

Bouguerra S. and al. [1] developed a mathematical model to study the opportunity provided by the extended warranty for the buyer and for the manufacturer to determine the maximum additional cost that the consumer must pay and the minimum price for which the manufacturer should sell the extended warranty. They did this for different maintenance policies over the product's life cycle involving a negligible repair time after failure. The present work is an extension of Bouguerra and al [1] in the sense that it considers a more realistic framework in which the time to repair is random.

The rest of this paper is organized as follows. In section 2, we define the problem. Section 3 presents the working assumptions and the corresponding mathematical model considering different maintenance policies. Section 4 is dedicated to an illustrative numerical example with a discussion of the obtained results. Finally, the paper is concluded in section 5.

II. Problem statement

We study the opportunity provided by the extended warranty from the consumer and the manufacturer perspectives. We consider the total expected cost incurred by each side during the product's life cycle in order to determine, for any given situation, the maximum extra cost the consumer should pay for the extended warranty, and the minimum price at which the manufacturer should sell it. This is done while considering different maintenance policies involving a random repair time. The model allows looking, for each maintenance strategy, if there is a zone of possible compromise yielding a win-win relationship with respect to the extended warranty.

III. Mathematical model

A. Working assumptions and notations

We consider repairable products sold with a warranty period (W) with the possibility to be extended until instant (We) for an additional cost (Cw) paid by the consumer when purchasing the product. All failures during the basic and extended warranty periods are repaired by the manufacturer at no cost to the buyer. The repair actions are considered as minimal repairs with random durations keeping the product's failure rate at nearly the same level as before failure. The costs of repairs occurring during the post-warranty period are borne by the buyer.

We assume that during the product lifecycle, preventive maintenance actions are performed at instants $\tau_1, \tau_2, .. \tau_j$; with $\tau_0 = 0$. Preventive maintenance is modeled using the age reduction and virtual age considered by Kim et al. [2]. This age reduction depends on the effort of maintenance (m) included between 0 (no age reduction) and a maximal level (M) corresponding to reducing the system's age to the age just

after the last PM. We suppose that during the product life cycle all the preventive maintenance actions will have the same effort (m). The cost of preventive maintenance actions (Cm) is paid by the consumer during the basic and the extended warranty periods.
The following notation will be used:
$\mathbf{f(x)}, \mathbf{F(x)}, \bar{\mathbf{F}}(x)$: Pdf, cdf, reliability function of failure time X, respectively.

Y : Length of repair time with distribution G(_)

g(y), G(y) : Pdf, cdf of Y, respectively

Z(t) : C.d.f. of X + Y, convolution of the distributions F(t) and G(t)

$\bar{\mathbf{Z}}(\mathbf{t})$ =1- Z(L).

-log $\bar{Z}$() = E [N^M] : Average number of minimal repair during [0,t]

r(t): Failure rate function.

Cr: Cost of each minimal repair action.

Cm: Cost of each imperfect PM action with maintenance level m (paid by the consumer)

Cw: Cost of the extended warranty period paid by the consumer

L: Length of the product life cycle

W : The basic warranty period

We : The warranty period including the basic period W and the extension

m : PM level [0 to M].

δ (m) : Age reduction

$\boldsymbol{\tau_j}$ (j=1, 2,3,….) : PM instants

$\boldsymbol{v_j}$: Virtual age of the product following the j th PM action

$\mathbf{r}(\boldsymbol{v}(\boldsymbol{t}))$: Failure rate corresponding to virtual age.

$\boldsymbol{\Delta} = (\boldsymbol{\tau_j} - \boldsymbol{\tau_{j+1}})$(j=1,2,…) : PM actions period, $\Delta_0=\tau_1$.

$\mathbf{n_1}, \mathbf{n_2}, \ldots, \mathbf{n_6}$: Number of PM actions undertaken over each interval as shown in Table I.

TABLE I. Number of preventive maintenance actions by interval

Number	Interval	Value
n1	[0,W)	$\lfloor W/\Delta \rfloor$
n2	[0,L)	n2 = $\frac{L}{\Delta} - 1$ si $\frac{L}{\Delta} \in$ N
		n2=$\lfloor \frac{L}{\Delta} \rfloor$ si $\frac{L}{\Delta} \notin$ N
n3	[W,L)	n3=$\frac{(L-W)}{\Delta} - 1$ si $\frac{(L-W)}{\Delta} \in$ N
		n3=$\lfloor \frac{(L-W)}{\Delta} \rfloor$ si $\frac{(L-W)}{\Delta} \notin$ N
n4	[0,We)	$\lfloor We/\Delta \rfloor$
n5	[W,We)	$\lfloor (We-W)/\Delta \rfloor$
n6	[We,L)	n6=$\frac{(L-We)}{\Delta} - 1$ si $\frac{(L-We)}{\Delta} \in$ N
		n6=$\lfloor \frac{(L-We)}{\Delta} \rfloor$ si $\frac{(L-We)}{\Delta} \notin$ N

We consider the maintenance options as in the work of Bouguerra S. and al. (2012) defined as follows:

- Option I: no PM action performed over the life cycle of the system (0, L).
- Option II: periodic PM actions over (0,L), with $\tau_j = j\Delta$, (j=1,2,…… n2).

Option III: periodic PM actions during the post warranty period.

- Option III-1: PM actions are performed from the expiration of the basic warranty period on the interval [w, L], with $\tau_j = w + j\Delta$ (j = 1, 2,…. n3).
- Option III-2 : PM actions are performed after the expiry of the period of:

Basic warranty [0, w] if extended warranty is not purchased, with $\tau_j = w + j\Delta$ (j = 1,2, ..., n3)

Basic warranty and extended warranty at a time [0, we] if extended warranty is purchased, with $\tau_j = we + j\Delta$ (j = 1, 2,… n6)

Option IV: periodic PM actions performed during the warranty period.

- Option IV-1 : PM actions performed during the basic warranty period on period [0, w], with $\tau_j = j\Delta$ (j = 1,2, ..., n1)
- Option IV-2 : PM actions are performed during the period of:

Basic warranty [0, w] if the extended warranty is not purchased, with $\tau_j = j\Delta$ (j = 1,2, ..., n1)

Basic warranty and extended warranty at a time [0, we] if extended warranty is purchased, with $\tau_j = j\Delta$ (j = 1, 2, ..., n4)

B. Determining the average number of minimal repairs during a life cycle

In the literature, repair actions are considered as having a negligible duration. However, it is more realistic to consider that repair maintenance actions take a random duration. For such situations, we express the average number of minimal repairs during [0, t] as follows (Lim J.H. et al. [3]):

$$E[N^M(t)] = -\log\left[\int_t^\infty F(t-y)g(y)dy\right] \quad (4)$$

C. Determination of the maximum extra cost the consumer should pay for the extended warranty

Let C_{CSX} and C_{ASX} denote the total costs incurred by the consumer for maintenance option X respectively for the case

without the extended warranty period (*S*) and the case with the extended warranty period (*A*).
In what follows we state, for each option, the condition which should be satisfied so that buying the extended warranty period would be advantageous for the consumer. This condition is based on the fact that buying the extended warranty would be beneficial for the consumer if the total cost incurred to him during the product lifecycle, would be lower than what it would cost him in case he does not take it.

Option I:

C_{CAI} + Cw ≤ C_{CSI} → Cw ≤ C_{CSI} - C_{CAI}
→ Cw ≤ C_{CI}

$$C_{CI}=Cr\times \begin{bmatrix} +log\left(1-\int_0^{w}(1-exp\ (-\int_0^{(w-y)} r_0(t)dt))\,g(y)dy\right) \\ -log\left(1-\int_0^{we}(1-exp(-\int_0^{(we-y)} r_0(t)dt))\,g(y)dy\right) \end{bmatrix} \quad (5)$$

Option II :

C_{CAII} + $Cw \le C_{CSII}$ → $Cw \le C_{CSII}$ - C_{CAII}
→ $Cw \le C_{CII}$

$$C_{CII} = Cr \times \left[-\left[\begin{array}{c} \log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy)) \end{array} \right] + \left[\begin{array}{c} \log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_4-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_4}-y)}^{(we-y)} r_0(v_{n_4}+t-\tau_{n_4})dt)]\ g(y)dy)) \end{array} \right] \right] + Cm\,(n_3-n_5-n_6) \quad (6)$$

Option III-1 :

$C_{CAIII-1}$ + $Cw \le C_{CSIII-1}$ → $Cw \le C_{CSIII-1}$ - $C_{CAIII-1}$
→ $Cw \le C_{CIII-1}$

$$C_{CIII-1} = Cr \left[\left[\begin{array}{c} -\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_5-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_5}-y)}^{(we-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_{n_5}+t-\tau_{n_5})dt)]\ g(y)dy)) \end{array} \right] + \left[\begin{array}{c} +\log(1-(\int_0^{w}[1-exp-(\int_0^{(w-y)} r_0(t)dt)]\ g(y)dy)) \end{array} \right] \right] + Cm\,(n_3-n_5-n_6) \quad (7)$$

Option III-2 :

$C_{CAIII-2}$ + $Cw \le C_{CSIII-2}$ → $Cw \le C_{CSIII-2}$ - $C_{CAIII-2}$
→ $Cw \le C_{CIII-2}$

$$C_{CIII-2} = Cr \left[-\left[\left[\begin{array}{c} -\log(1-(\int_0^{L}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_3-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_3}-y)}^{(L-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_{n_3}+t-\tau_{n_3})dt)]\ g(y)dy)) \end{array} \right] + \left[\begin{array}{c} +\log(1-(\int_0^{w}[1-exp-(\int_0^{(w-y)} r_0(t)dt)]\ g(y)dy)) \end{array} \right] \right] + \left[\left[\begin{array}{c} -\log(1-(\int_0^{L}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_6-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_6}-y)}^{(L-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_{n_6}+t-\tau_{n_6})dt)]\ g(y)dy)) \end{array} \right] + \left[\begin{array}{c} +\log(1-(\int_0^{we}[1-exp-(\int_0^{(we-y)} r_0(t)dt)]\ g(y)dy)) \end{array} \right] \right] \right] + Cm(n_3-n_6)$$

Option IV-1:

$C_{CAIV-1} + Cw \leq C_{CSIV-1} \rightarrow Cw \leq C_{CSIV-1} - C_{CAIV-1}$
$\rightarrow Cw \leq C_{CIV-1}$

$$C_{CIV-1} = Cr\left[\begin{bmatrix}-\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(we-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy))\end{bmatrix} + \begin{bmatrix}\log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy))\end{bmatrix}\right] + Cm\,(n_1-n_1) \quad (9)$$

Option IV-2:

$C_{CAIV-2} + Cw \leq C_{CSIV-2} \rightarrow Cw \leq C_{CSIV-2} - C_{CAIV-2}$
$\rightarrow Cw \leq C_{CIV-2}$

$$C_{CIV-2} = Cr\left[\begin{bmatrix}\begin{bmatrix}-\log(1-(\int_0^{L}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(L-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy))\end{bmatrix} + \begin{bmatrix}\log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy))\end{bmatrix}\end{bmatrix} - \begin{bmatrix}\begin{bmatrix}-\log(1-(\int_0^{L}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_5-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_5}-y)}^{(L-y)} r_0(v_{n_5}+t-\tau_{n_5})dt)]\ g(y)dy))\end{bmatrix} + \begin{bmatrix}\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_5-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_5}-y)}^{(we-y)} r_0(v_5+t-\tau_{n_5})dt)]\ g(y)dy))\end{bmatrix}\end{bmatrix}\right] - Cm.n_5 \quad (10)$$

D. Determination of the minimum price at which the manufacturer should sell the extended warranty

Let C_{FAX} and C_{FSX} denote the total costs incurred to the manufacturer for maintenance option X respectively for the case without the extended warranty period (*S*) and the case with the extended warranty period (*A*).

We establish below, for each option, the condition which should be satisfied so that selling the extended warranty period would be profitable for the manufacturer. This condition is based on the fact that selling the extended warranty would be beneficial for the manufacturer if the total cost incurred to him during the product lifecycle, would be lower than what it would cost him in case he does not sell it.

Option I:

$C_{FAI} - Cw \leq C_{FSI} \rightarrow Cw \geq C_{FAI} - C_{FSI}$
$\rightarrow Cw \geq C_{FI}$

Avec : $C_{FI} = Cr\begin{bmatrix}-log\left(1-\int_0^{we}(1-exp\ (-\int_0^{(we-y)} r_0(t)dt))\,g(y)dy\right) \\ +log\left(1-\int_0^{w}(1-exp(-\int_0^{(w-y)} r_0(t)dt))\,g(y)dy\right)\end{bmatrix}$ (11)

Option II :

$C_{FAII} - Cw \leq C_{FSII} \rightarrow Cw \geq C_{FAII} - C_{FSII}$
$\rightarrow Cw \geq C_{FII}$

$$C_{FII} = Cr\left[\begin{bmatrix}-\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_4-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_4}-y)}^{(we-y)} r_0(v_{n_4}+t-\tau_{n_4})dt)]\ g(y)dy))\end{bmatrix} + \begin{bmatrix}\log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy))\end{bmatrix}\right] \quad (12)$$

Option III-1 :

$C_{FAIII-1} - Cw \leq C_{FSIII-1} \rightarrow Cw \geq C_{FAIII-1} - C_{FSIII-1}$
$\rightarrow Cw \geq C_{FIII-1}$

C_{FIII-1} =

$$Cr\left[\begin{array}{c} -\log(1-(\int_0^{we}[1-\exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_5-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_5}-y)}^{(we-y)}[r_0(\tau_1-y)-r_m(\tau_1-y)]+r_0(v_{n_5}+t-\tau_{n_5})dt)]\ g(y)dy)) \\ +\log(1-(\int_0^{w}[1-\exp-(\int_0^{(w-y)} r_0(t)dt)]g(y)dy)) \end{array}\right] \quad (13)$$

Option III-2 :

$C_{FBIII-2}$ - $Cw \le C_{FAIII-2}$ → $Cw \ge C_{FBIII-2} - C_{FAIII-2}$ → $Cw \ge C_{FIII-2}$

C_{FIII-2} =

$$Cr\left[\begin{array}{c} -\log(1-(\int_0^{we}[1-exp-(\int_0^{(we-y)} r_0(t)dt \\ +\log(1-(\int_0^{w}[1-exp-(\int_0^{(w-y)} r_0(t)dt)]\ g(y)dy)) \end{array}\right] \quad (14)$$

Option IV-1 :

C_{FAIV-1} - $Cw \le C_{FSIV-1}$ → $Cw \ge C_{FAIV-1} - C_{FSIV-1}$ → $Cw \ge C_{FIV-1}$

$$C_{FIV-1} = Cr\left[\begin{array}{c} \left[\begin{array}{c} -\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(we-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy)) \end{array}\right] \\ +\left[\begin{array}{c} \log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_{n_1}+t-\tau_{n_1})dt)]\ g(y)dy)) \end{array}\right] \end{array}\right] \quad (15)$$

Option IV-2 :

C_{FAIV-2} - $Cw \le C_{FSIV-2}$ → $Cw \ge C_{FAIV-2} - C_{FSIV-2}$ → $Cw \ge C_{FIV-2}$

$$C_{FIV-2} = Cr\left[\begin{array}{c} \left[\begin{array}{c} -\log(1-(\int_0^{we}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_5-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_5}-y)}^{(we-y)} r_0(v_{n_5}+t-\tau_{n_5})dt)]\ g(y)dy)) \end{array}\right] \\ +\left[\begin{array}{c} \log(1-(\int_0^{w}[1-exp-(\int_0^{(\tau_1-y)} r_0(t)dt \\ +\sum_{j=1}^{n_1-1}\int_{(\tau_j-y)}^{(\tau_{j+1}-y)} r_0(v_j+t-\tau_j)dt \\ +\int_{(\tau_{n_1}-y)}^{(w-y)} r_0(v_1+t-\tau_{n_1})dt)]\ g(y)dy)) \end{array}\right] \end{array}\right] \quad (16)$$

IV. Numerical exemple

We consider the example of systems with times between failures that follow a Weibull distribution with a shape parameter α = 2 and a scale parameter λ=1 with increasing failure rate function given by:

$$r_0(t) = (\alpha\ \lambda)(\lambda t)^{\alpha-1}$$

The following arbitrarily chosen input data are also considered:

- System life cycle length: L = 9 years.
- The product is sold with a basic warranty period: W = 2 years.
- The duration of the warranty period, including the extension: We = 3 years.
- The time between preventive maintenance actions: Δ = 0.33 year

We consider five preventive maintenance levels as δ(m) is defined by Kim et al. (2004):

$$\delta\ (m) = (1+m)e^{-m}\ ,\ m \ge 0 \quad (17)$$

For different values of the maintenance level (m), Table (II) presents preventive maintenance action cost and δ(m) values.

TABLE II. Maintenance level m, improvement factor δ (m) and the corresponding cost Cm.

Maintenance level (m)	δ (m)	Cm (monetary units)
0	1	0
1	0.74	10
2	0.41	30
3	0.2	60
4	0.09	100
5	0.04	160

We consider the case where repair time follows a log-normal distribution with mean time MTTR. Minimal repair cost (Cr) is proportional to the mean time to repair (MTTR) such that:

Cr = monthly cost* × MTTR

We performed calculations using the input data mentioned above. We have looked, for each maintenance option and each preventive maintenance level, for the existence of a compromise (win-win) interval [a,b] where 'a' stands for the minimum price at which the manufacturer should sell the extended warranty, and 'b' represents the maximum extra cost the consumer should pay for the extended warranty. The effect of varying the minimal repair duration and the monthly repair costs is felt in strategies III-2 and IV-2. The obtained results are grouped in Tables III and IV.

TABLE III. Compromise intervals for the extended warranty cost for different minimal repair costs and average durations for option III-2

MTTR	Cr	m=1	m=2	m=3	m=4	m=5
1	800	[0, 1386]∩ [1420,∞]= Ø	[0, 1392]∩ [1420,∞]= Ø	[1420,1542]	[1420, 1694]	[1420, 1884]
	1000	[0, 1725]∩ [1775,∞] = Ø	[0, 1717]∩[1775,∞]=Ø	[1775,1883]	[1775,2043]	[1775,2235]
	1200	[0, 2064]∩ [2130,∞] = Ø	[0, 2042]∩ [2130,∞] = Ø	[2130,2224]	[2130,2391]	[2130,2586]
3	800	[0, 3062]∩ [3204,∞] = Ø	[0, 3019]∩ [3204,∞] = Ø	[3204,3263]	[3204,3457]	[3204,3658]
	1000	[0, 3820]∩ [4005,∞] = Ø	[0, 3752]∩ [4005,∞] = Ø	[4005,4034]	[4005,4246]	[4005,4453]
	1200	[0, 4578]∩ [4806,∞] = Ø	[0, 4484]∩ [4806,∞] = Ø	[0, 4805]∩ [4806,∞] = Ø	[4806,5035]	[4806,5248]
6	800	[0, 5353]∩ [5892,∞] = Ø	[0, 5221]∩ [5892,∞] = Ø	[0, 5713]∩ [5892,∞] = Ø	[5892,6048]	[5892,6295]
	1000	[0, 6683]∩ [7365,∞] = Ø	[0, 6504]∩ [7365,∞] = Ø	[0, 7097]∩ [7365,∞] = Ø	[7365,7485]	[7365,7749]
	1200	[0, 8014]∩ [8838,∞] = Ø	[0, 7786]∩ [8838,∞] = Ø	[0, 8480]∩ [8838,∞] = Ø	[8838,8922]	[8838,9203]

TABLE IV. Compromise intervals for the extended warranty cost for different minimal repair costs and durations for option IV-2

MTTR	Cr	m=1	m=2	m=3	m=4	m=5
1	800	[0, 3615]∩ [3639,∞]= Ø	[3440, 3532]	[2999, 3840]	[2691, 3289]	[2528, 2560]
	1000	[0, 4527]∩ [4549,∞]= Ø	[4301, 4437]	[3749, 4845]	[3364, 4186]	[3160, 3320]
	1200	[0, 5438]∩ [5458,∞]= Ø	[5161, 5342]	[4499, 5851]	[4037, 5084]	[3793, 4080]
3	800	[0, 8305]∩ [8310,∞]= Ø	[7795, 8382]	[6542, 9739]	[5646, 8039]	[5167, 6163]
	1000	[10388, 10390]	[9744, 10500]	[8178, 12219]	[7058, 10124]	[6459, 7824]
	1200	[12466, 12474]	[11693, 12618]	[9814, 14698]	[8470, 12209]	[7750, 9485]
6	800	[13164, 13203]	[11756, 13448]	[9148, 16458]	[7331, 12304]	[6366, 8605]
	1000	[16456, 16511]	[14695, 16833]	[11436, 20617]	[9164, 15455]	[7957,10876]
	1200	[19747, 19820]	[17635, 20217]	[13723, 24776]	[10997, 18606]	[9549, 13148]

From Table (III) for Option III-2, we note that for preventive maintenance levels (m = 1 and m = 2) and regardless of the average minimal repair time and the corresponding cost, the compromise intervals for the extended warranty remains inexistent. However, for preventive maintenance levels m = 4 and m = 5, the compromise intervals exist even if the MTTR and the minimal repair cost are varied. For a preventive maintenance level m = 3, the compromise region doesn't exist for a minimal repair cost greater than (1200 monetary unit/ month) and a mean time to repair (3 months) thus we can conclude, that for an average repair time, the compromise interval doesn't exist .

For option IV-2, Table (IV) shows that The compromise interval exists for preventive maintenance levels m ≥2. Varying the duration of minimal repair and the different repair costs has only influenced the expanse of the compromise intervals. For a preventive maintenance level m = 1, the consumer and the manufacturer have a different threshold for which the extended warranty will be interesting and this is noticed for low mean time to repair (MTTR) and low values of minimal repair costs.

For example for MTTR = 1 month and Cr = 1000 monetary unit/month, the consumer will be interested in the extended warranty for any price less than $Cw \leq 4527$ monetary units, for the manufacturer it is necessary that the price of the extended warranty exceeds $Cw \geq 4549$ monetary units. The intervals begin to exist for important repair times and costs.

Conclusion

The problem we tackled is inspired of situations largely encountered in practice. When acquiring a product, the consumer is in a position of doubt when the manufacturer offers to pay an additional cost to enjoy an extended warranty period. In this article, we developed a mathematical model allowing the simultaneous determination the condition under which it would be in the interest of the consumer to buy an extended warranty, and the minimum price with which it would be profitable for the manufacturer to sell it. We considered several maintenance policies to be adopted during the life cycle of the product with minimal repair actions with random durations. The developed model was tested by performing numerical calculations. We determined for each maintenance policy, the win-win interval (when it exists) representing a compromise between the manufacturer and the consumer regarding the price of the extended warranty.

Among future research perspectives following this work, it would be interesting to study the opportunity provided by the extended warranty for non-repairable products. One can also explore the opportunity provided by the extended warranty considering a variable preventive maintenance level during the life cycle of the system.

References

[1] Bouguerra S., Chelbi A, Rezg N., (2012), A decision model for adopting an extended warranty under different maintenance policies, Int. J. Production Economics 135,840–849

[2] Kim, C.S., Djamaludin, I., Murthy, D.N.P., (2004). Warranty and discrete preventive maintenance. Reliability Engineering and System Safety 84, 301–309

[3] Lim J.H. , Kim D.K., Park D.H.,(2005), Cost evaluation for an imperfect-repair model with random repair time, International Journal of Systems ScienceVol. 36, No. 11, 15, 717–726

A joint selective maintenance and multiple repair-person assignment problem

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

C. Diallo*, A. Khatab†, U. Venkatadri‡, and E.-H. Aghezzaf§
*Department of Industrial Engineering, Dalhousie University, Halifax, Canada.
Email: Claver.Diallo@Dal.ca
†LGIPM. Lorraine University, National school of Engineering. Metz, France.
E-mail: abdelhakim.khatab@univ-lorraine.fr
‡Department of Industrial Engineering, Dalhousie University, Halifax, Canada.
Email: Uday.Venkatadri@Dal.ca
§Department of Industrial Systems Engineering and Product Design.
Faculty of Engineering and Architecture, Ghent University, Belgium.
E-mail: ElHoussaine.Aghezzaf@UGent.be

***Abstract*—It is common in many industrial settings to have large multicomponent systems perform consecutive missions interspersed with finite breaks during which a limited set of component repairs or replacements can be carried out due to limited time, budget, or resources. The decision maker then has to decide which components to repair in order to guarantee a pre-specified performance level. This is known as the selective maintenance problem. This paper introduces a novel variant of the selective maintenance problem for a multi-component system performing consecutive missions separated by scheduled finite breaks by specifically modelling the assignment of the repair tasks to multiple repair-persons. Current models in the literature usually assume that only one repair channel is available or that the assignment optimization can be done at a subsequent stage. A novel integrated non-linear programming formulation is proposed and optimally solved. Numerical experiments show the benefits of jointly carrying out the assignment of the tasks to repair-persons and the selection of the components to be repaired.**

I. Introduction

It is common in many industrial settings to have large and complex multicomponent systems perform consecutive missions interspersed with finite breaks during which a limited set of component repairs or replacements can be carried out due to limited time, budget, or resources. In the airline industry for example, aircraft are subjected to overnight maintenance at night during a break interval typically varying from 2 to 8 hours. To prepare the system to successfully complete its next mission, its components must be properly maintained during the scheduled intermission break. Because of the built-in redundancies, and due to the limited duration of the scheduled breaks and scarce maintenance resources, only a limited set of components have and can be maintained during the breaks. It is therefore necessary to identify an optimal subset of components to maintain to meet the predetermined reliability level required for the next mission. In the maintenance literature, this is known as the selective maintenance problem (SMP).

The selective maintenance problem was initially introduced by Rice at al. [1]. The authors considered a series-parallel system in which subsystems are made of independent and identically distributed (*i.i.d*) components with constant failure rate (CFR). The replacement of components at failure is the only maintenance option. To overcome the restrictive hypothesis of identical subsystem components in [1], Cassady et al. [2] developed a more general framework for selective maintenance formulation.

Cassady et al. [3] studied the SMP in a series-parallel system, where the components have Weibull distributed lifetimes. For each component, any of three potential maintenance actions can be carried out: minimal repair, corrective replacement of a failed component, and preventive replacement of a working component. The resulting SMP is solved using an enumeration method. However, enumeration solution methods become rapidly cumbersome when the number of the system components increases. To deal with the combinatorial complexity arising from large-size systems, four improved enumeration procedures are proposed in Rajagopalan and Cassady [4] to reduce the computation times. An exact method based on the branch-and-bound procedure and a Tabu search based algorithm are proposed in Lust et al. [5]. Khatab et al. [6] proposed two heuristic methods, adapted from those used to solve the redundancy allocation problem in [7], [8], [9]. Maaroufi et al. [10] studied the SMP for a system where some components are subject to both global failure propagation and isolation.

The development of SMP models has since increased and dealt with extensions such as imperfect maintenance, multistate systems. Imperfect maintenance in the selective maintenance setting is addressed by Liu and Huang [11], where the age reduction coefficient approach [12] is used to model imperfect maintenance. An imperfect selective

maintenance model was also developed by Zhu et al. [13] and applied to a machining line system. Panday et al. [14] also studied the selective maintenance problem for binary systems under imperfect maintenance using the hybrid hazard rate approach introduced in [15]. In [11] and [14], a set of maintenance levels, ranging from minimal repair to replacement, are used to improve the reliability of a system component. Liu and Zhang [11] and Zhu et al. [13], the only parameter that determines the improvement in the component's health is the age reduction coefficient. In [14], however, both the age reduction coefficient and the adjustment coefficient impact the component's health. A more recent work [16] studied the selective maintenance problem when the quality of imperfect maintenance is stochastic. A nonlinear and stochastic optimization problem was proposed and solved for a series-parallel system.

SMPs have also been investigated for multi-state systems (MSS) in Chen et al. [17] where system components and the overall system may be found in more than two possible states. Liu and Huang [11] considered MSS where components have two operating states and are subjected to imperfect maintenance that can bring the condition of a component to an intermediate degradation level. Panday et al. [18] investigated the SMP for a MSS where the functioning of each component is modeled as a continuous-time Markov chain with more than two states. Dao et al. [19], [20] studied the SMP under economic and stochastic dependencies among MSS components. In [21], several levels of imperfect maintenance are considered and the system is composed of multi-state components modeled as by Pandey et al. [18]. In a recent paper, Dao and Zuo [22] investigated the SMP in a MSS with structural relationships between components.

All papers surveyed above do not consider the very common case where multiple repair-persons are available to carry out the maintenance actions. In the airline industry, when an aircraft undergoes overnight maintenance, several repair-persons are usually available and will be utilized as needed to inspect and perform repairs. Furthermore, given the size of the systems it is possible for several members of the repair crew to work simultaneously on the plane. Up to two crew members can be in the cockpit to deal with the avionics, while several others can take care of the structural components, the wings, the engines, etc. It is therefore interesting to jointly optimize the selective maintenance decisions and the assignment of the repairs to the repair crews.

The remainder of the article is organized as follows. Section II describes the multicomponent system under consideration and the resources available to carry out the repairs. In this section, we also present the development of the mathematical formulation of the joint selective maintenance and multiple repair-person assignment problem. A solution procedure is proposed in Section III along with several numerical experiments and the discussion of their results. Conclusions are drawn and future extensions discussed in the last section.

II. System description and problem formulation

Without loss of generality, the selective maintenance problem addressed in the present work concerns a series-parallel system S composed of n series subsystems S_i $(i = 1, \ldots, n)$ each of which is composed of N_i s-independent, and possibly, non-identical components/parts P_{ij} $(j = 1, \ldots, N_i)$ arranged in parallel. The system is assumed to have just finished a mission and then, turned off during the scheduled break of finite length and becomes available for possible maintenance activities. The system is thereafter used to execute the next mission of a given duration. The duration of the scheduled break is denoted as D and the duration of the next mission is denoted by U.

At the end of the current mission (i.e., at the beginning of the current break), the state variable X_{ij} describes the status of the component P_{ij}. Similarly, the state variable z_{ij} describes the status P_{ij} at the end of the current break (i.e., at the beginning of the next mission). These two state variables are then formally defined as:

$$X_{ij} : \begin{cases} 1 & \text{if } P_{ij} \text{ is working at the start of the break} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$z_{ij} : \begin{cases} 1 & \text{if } P_{ij} \text{ is working at the end of the break} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The following assumptions are considered in this paper:

1) The system consists of multiple, repairable binary components (the components and the system are either functioning or failed).
2) During the break, system components do not age, i.e. the age of a component is operation-dependent.
3) No maintenance activity is allowed during the mission. Maintenance activities are allowed only during the break.
4) Multiple components can be worked on simultaneously without repairpersons colliding. This assumption is reasonable in large multicomponent systems and with modular design as previously discussed in the case of an aircraft.
5) The duration of the break is longer than the longest component repair time. This is reasonable as the contrary would automatically make the component non-eligible and out of consideration for optimization.

The following notation is used in the mathematical formulation of the RCSMP.

Indices

i : Index of subsystems, $\quad i = 1, ..., s$
j : Index of parts in subsystem i, $\quad j = 1, ..., N_i$
r : Index of potential repair-persons, $\quad r = 1, ..., m$

Parameters

D : Duration of the break
R_0 : Minimum required reliability level
k : Variable labour cost per unit of time
K_r : Fixed cost for hiring/utilizing repair-person r
t_{ijr} : Time to repair part P_{ij} by repair-person r
c_{ij} : Cost to repair/replace part to P_{ij}
X_{ij} : Status of P_{ij} at the start of the break
R_{ij} : Reliability of P_{ij} at the start of the break
m : Number of repair-persons available

Decision variables

x_{ijr} : Binary, = 1 if P_{ij} is repaired by repair-person r;
0, otherwise
z_{ij} : Binary, = 1 if P_{ij} is working at the end of the break;
0, otherwise
y_r : Binary, = 1 is repair-person r is hired/utilized;
0, otherwise.

The mathematical formulation for the minimization of the total cost (model 1) is as follows:

$$\text{Min } \mathcal{C} = \sum_{i=1}^{s}\sum_{j=1}^{N_i}\left[c_{ij}\,(z_{ij} - X_{ij}) + k\sum_{r=1}^{m} t_{ijr}x_{ijr}\right] + \sum_{r=1}^{m} K_r y_r$$

Subject to:

$$\sum_{i=1}^{s} ln\left[1 - \prod_{j=1}^{N_i}(1 - R_{ij}\,z_{ij})\right] \geq lnR_0 \qquad \forall i, j \qquad (3)$$

$$z_{ij} = X_{ij} + \sum_{r=1}^{m}(1 - X_{ij})\,x_{ijr} \qquad \forall i, j \qquad (4)$$

$$x_{ijr} \leq 1 - X_{ij} \qquad \forall i, j, r \qquad (5)$$

$$\sum_{i=1}^{s}\sum_{j=1}^{N_i} t_{ijr}\,x_{ijr} \leq D\,y_r \qquad \forall r \qquad (6)$$

$$\sum_{r=1}^{m} x_{ijr} \leq 1 \qquad \forall i, j \qquad (7)$$

$$\sum_{j=1}^{N_i} z_{ij} \geq 1 \qquad \forall i \qquad (8)$$

$$x_{ijr},\ z_{ij},\ y_r \in \{0, 1\} \qquad \forall i, j, r. \qquad (9)$$

The objective function is the sum of three terms. The first term is the total cost to repair/replace the components that have been selected to undergo maintenance. The second term is the related labor cost. The third term is the cost to hire the repair-persons needed to carry out the repairs.

Constraint (3) ensures that the system reliability is at least equal to the minimum required reliability R_0. Since the system under consideration is a series arrangement of subsystems, the system reliability during the following mission is given by

$$R_s(U) = \prod_{i=1}^{s}(1 - \prod_{j=1}^{N_i}(1 - R_{ij}\,z_{ij})).$$

Constraints (4) defines the status of $P_{i,j}$ at the end of the break, which is its status at the start of the break plus any change that occurred during the break. A working $P_{i,j}$ does not need replacement. A failed $P_{i,j}$ may or may not be replaced.

Constraint (5) ensures that only failed parts are candidates for replacement/repair.

Constraint (6) ensures that if a repair-person is utilized then their total repair times is no more than the break duration. If the repair-person is not hired, then they cannot perform any repair.

Constraint (7) guarantees that each component will receive at most one repair.

Constraint (8) ensures that each subsystem has at least one working component at the start of the next mission. This constraint is introduced to lift the commonly used assumption that requires all subsystems to be operating at the end of the previous mission.

The last constraint defines the binary decision variables used in the formulation.

The above mathematical formulation contains one nonlinear constraint (3). It is optimally solved using the KNitro 9 solver [28] in MPL 5.0 [29]. In the following section, numerical experiments are carried out to show that our model yields valid maintenance decisions.

Note that an alternative formulation (model 2), if a maintenance budget $\mathcal{C}_0$ is pre-specified, is:

$$\text{Max } R_s = \prod_{i=1}^{s}(1 - \prod_{j=1}^{N_i}(1 - R_{ij}\, z_{ij}))$$

Subject to:

$$\sum_{i=1}^{s}\sum_{j=1}^{N_i}\left[c_{ij}\,(z_{ij} - X_{ij}) + k\sum_{r=1}^{m} t_{ijr}x_{ijr}\right] + \sum_{r=1}^{m} K_r y_r \leq \mathcal{C}_0$$

$$z_{ij} = X_{ij} + \sum_{r=1}^{m}(1 - X_{ij})\, x_{ijr} \qquad \forall i, j$$

$$x_{ijr} \leq 1 - X_{ij} \qquad \forall i, j, r$$

$$\sum_{i=1}^{s}\sum_{j=1}^{N_i} t_{ijr}\, x_{ijr} \leq D\, y_r \qquad \forall r$$

$$\sum_{r=1}^{m} x_{ijr} \leq 1 \qquad \forall i, j$$

$$\sum_{j=1}^{N_i} z_{ij} \geq 1 \qquad \forall i$$

$$x_{ijr},\ z_{ij},\ y_r \in \{0, 1\} \qquad \forall i, j, r.$$

III. Numerical examples

In this section, three sets of numerical experiments are conducted. Table I lists all the parameters used. These are the same ones used by Cassady et al. [2]. Because, we are introducing multiple repair-persons instead of one as considered by [2], we have created 3 additional columns with repair times. Repair-persons 2 and 3 are considered to have standard skills and have the same repair times as in [2]. Repair-person 1 is more skilled, therefore their repair times are shorter but they are more expensive to hire and use. Repair-person 4 would be equivalent to a rookie or trainee. Thus, their repair times are longer than the standard ones but they are cheaper to hire and use.

TABLE I. Parameter values adapted from Cassady [2]

Part	**Reliability**	**Time to repair per repair-person**				**Unit cost**	**Status**
$P_{i,j}$	$R_{i,j}$	1	2	3	4	$c_{i,j}$	$S_{i,j}$
$P_{1,1}$	0.80	6	7	7	8	2	1
$P_{1,2}$	0.70	3	4	4	5	3	0
$P_{1,3}$	0.85	5	6	6	7	1	1
$P_{2,1}$	0.65	2	3	3	4	4	0
$P_{2,2}$	0.50	1	2	2	3	2	0
$P_{2,3}$	0.70	5	6	6	7	1	1
$P_{2,4}$	0.75	6	7	7	8	3	1
$P_{2,5}$	0.60	3	4	4	5	4	0
$P_{3,1}$	0.55	4	5	5	6	3	1
$P_{3,2}$	0.75	2	3	3	4	5	1
$P_{3,3}$	0.60	6	7	7	8	2	0
$P_{3,4}$	0.70	3	4	4	5	2	0

A. Experiment #1: Model 1 - All repair-persons have the same skills

For this experiment all repair-persons have the same standard skills and costs. The results in Table II are obtained for $K_r = 60$, $k_r = 4$, $t_{ijr} = t_{ij2}$, and $D = 11$. This is equivalent to the first experiment in Cassady et al. [2] when $R_0 = 0.9475$. Our results are identical to Cassady et al. [2] for that case: one repair-person has to be hired and the following components have to be replaced: $P_{1,2}$, $P_{2,1}$, $P_{2,5}$ and $P_{3,4}$. The results table also lists the computation times (CPUt) in seconds, which are negligible.

As the required minimum reliability increases, the number of components to be replaced increases and more repair-persons are needed.

TABLE II. Results for different values of R_0

R_0	m^*	CPUt	$\mathcal{C}$(\$)	Components replaced
0.8	1	0.14s	66	$P_{2,2}$
0.9	1	0.17s	80	$P_{2,2}$; $P_{3,4}$
0.9475	1	0.25s	117	$P_{1,2}$; $P_{2,1}$; $P_{2,5}$; $P_{3,4}$
0.97	2	1.88s	209	Rep.1: $P_{2,1}$; $P_{3,3}$; $P_{3,4}$ Rep.2: $P_{1,2}$; $P_{2,2}$; $P_{2,5}$

B. Experiment #2: Model 1 - Repair-persons have different skills

For this experiment, the repair-persons have different skills as listed in Table I. The results in Tables III and IV are obtained respectively for $R_0 = 0.9475$ and $R_0 = 0.97$. The duration of the break is varied. The other parameter values are: $K_1 = 60$, $K_2 = K_3 = 50$, $K_4 = 30$, $k_1 = 4$, $k_2 = k_3 = 2$, and $k_4 = 1$.

TABLE III. Results when varying D - Case of $R_0 = 0.9475$

D	m^*	CPUt	Repair-persons 1	2	3	4	$\mathcal{C}$(\$)	Parts replaced
11	1	0.25s			✓		117	Rep.3: $P_{1,2}$; $P_{2,1}$; $P_{2,5}$; $P_{3,4}$
9	2	0.74s			✓	✓	118	Rep.3: $P_{1,2}$; $P_{2,5}$ Rep.4: $P_{2,1}$; $P_{3,4}$
8	2	0.89s	✓			✓	140	Rep.1: $P_{1,2}$; $P_{2,1}$; $P_{2,5}$ Rep.4: $P_{3,4}$

TABLE IV. Results when varying D - Case of $R_0 = 0.97$

D	m^*	CPUt	Repair-persons 1	2	3	4	$\mathcal{C}$(\$)	Parts replaced
11	2	0.79s	✓			✓	162	Rep.1: $P_{1,2}$; $P_{2,1}$; $P_{2,5}$; $P_{3,4}$ Rep.4: $P_{2,2}$; $P_{3,3}$
8	3	1.13s	✓		✓	✓	205	Rep.1: $P_{2,1}$; $P_{2,2}$; $P_{2,5}$ Rep.3: $P_{1,2}$; $P_{3,4}$ Rep.4: $P_{3,3}$

The results in Tables III and IV show that when the duration of the break reduces, more repair-persons are needed. At first, the trainee is used to help the standard employee. At some point, the skilled repair-person is utilized to help speed up the repairs and make all required repairs fit in the interval of time available. Here again, we see that computation times are very small, showing that the joint selective maintenance and assignment problem is not more complex to solve with a proper formulation.

C. Experiment #3: Model 2 - Repair-persons have different skills

This experiment is designed to maximize the system reliability when a predetermined budget $\mathcal{C}_0$ is made available to the decision maker. Solving model 2 using the parameter values from Table I and varying the values of $\mathcal{C}_0$, the results in Table V are obtained for $D = 8$ units of time.

TABLE V. Results for varying maintenance budget $\mathcal{C}_0$

			Repair persons					
$\mathcal{C}_0$	m^*	CPUt	1	2	3	4	R^*	Parts replaced by
205	3	0.23s	✓	✓		✓	0.972	R.1: $P_{1,2}$; $P_{2,1}$; $P_{2,2}$ R.2: $P_{2,5}$; $P_{3,4}$ R.4: $P_{3,3}$
200	3	0.43s		✓	✓	✓	0.967	R.2: $P_{1,2}$; $P_{2,1}$ R.3: $P_{2,5}$; $P_{3,4}$ R.4: $P_{3,3}$
150	2	0.11s	✓			✓	0.952	R.1: $P_{1,2}$; $P_{2,1}$; $P_{3,4}$ R.4: $P_{3,3}$
125	2	0.26s			✓	✓	0.945	R.3: $P_{1,2}$; $P_{3,4}$ R.4: $P_{2,1}$; $P_{2,2}$
100	1	0.23s	✓				0.925	R.1: $P_{2,1}$; $P_{2,2}$; $P_{3,4}$
70	1	0.07s		✓			0.913	R.2: $P_{2,1}$; $P_{3,4}$
60	1	0.11s				✓	0.902	R.4: $P_{2,2}$; $P_{3,4}$

A clear trend can be seen from the above results. As the budget decreases, the number of repair-persons that can be hired and utilized decreases along with the number of components that can be replaced within the duration of the break. Thus, the maximum achievable reliability decreases with the budget.

When the budget allows it, the model will always give priority to the skilled repair-person as they are capable to repair more components in the fixed repair window. Additional repair-persons are added as permitted by the budget to complement the work of the skilled repair-person.

This model is capable of finding the best trade-offs between component costs, repair durations and labor costs to achieve the highest reliability possible while performing the assignment of repairs to workers.

IV. Conclusion

In this paper, we introduced a novel variant of the selective maintenance problem for a multi-component system performing consecutive missions separated by scheduled finite breaks by specifically modelling the assignment of the repair tasks to multiple repair-persons. Current models in the literature usually assume that only one repair channel is available or that the assignment optimization can be done at a subsequent stage. A novel integrated nonlinear programming formulation was proposed and optimally solved. Numerical experiments show the benefits of jointly carrying out the assignment of the tasks to repair-persons and the selection of the components to be repaired.

Future extensions that the authors are working on include a generalization to k-out-of-n subsystems with imperfect maintenance and repair-person eligibility constraints on the subsystems. Most models study the selective maintenance problem with reliability as the performance indicator. It would be of great value to consider system availability. Therefore, we are planning to study the trade-offs between system availability and the hiring of repair-persons using a bi-objective optimization model.

References

[1] W. F. Rice, C. R. Cassady, and J. Nachlas, "Optimal maintenance plans under limited maintenance time," in *Proceedings of Industrial Engineering Conference, Banff, BC, Canada*, 1998.

[2] C. R. Cassady, E. A. Pohl, and W. P. Murdock, "Selective maintenance modeling for industrial systems," *Journal of Quality in Maintenance Engineering*, vol. 7, no. 2, pp. 104–117, 2001.

[3] C. R. Cassady, W. P. Murdock, and E. A. Pohl, "Selective maintenance for support equipement involving multiple maintenance actions," *European Journal of Operational Research*, vol. 129, pp. 252–258, 2001.

[4] R. Rajagopalan and C. R. Cassady, "An improved selective maintenance solution approach," *Journal of Quality in Maintenance Engineering*, vol. 12, no. 2, pp. 172–185, 2006.

[5] T. Lust, O. Roux, and F. Riane, "Exact and heuristic methods for the selective maintenance problem," *European Journal of Operational Research*, vol. 197, pp. 1166–1177, 2009.

[6] A. Khatab, D. Ait-Kadi, and M. Nourelfath, "Heuristic-based methods for solving the selective maintenance problem for series-prallel systems," in *International Conference on Industrial Engineering and Systems Management, Beijing, China*, 2007.

[7] K. K. Aggrawal, "redundancy optimization in general systems," *IEEE Transactions on Reliability*, vol. R-25, no. 25, pp. 330–332, 1976.

[8] K. Gopal, K. K. Aggrawal, and J. S. Gupta, "An improved algorithm for reliability optimization," *IEEE Transactions on Reliability*, vol. R-27, no. 27, pp. 325–328, 1978.

[9] J. Sharma and K. V. Venkateswaran, "redundancy optimization in general systems," *IEEE Transactions on Reliability*, vol. R-20, no. 20, pp. 256–259, 1971.

[10] G. Maaroufi, A. Chelbi, and N. Rezg, "Optimal selective renewal for systems subject to propoagated failures with global effect and failure isolation phenomena," *Reliability Engineering and System Safety*, vol. 114, pp. 61–70, 2013.

[11] Y. Liu and H.-Z. Huang, "Optimal selective maintenance strategy for multi-state systems under imperfect maintenance," *IEEE Transactions on Reliability*, vol. 59, no. 2, pp. 356–367, 2010.

[12] M. Malik, "Reliable preventive maintenance scheduling," *AIIE transactions*, vol. 11, no. 3, pp. 221–228, 1979.

[13] H. Zhu, F. Liu, X. Shao, Q. Liu, and Y. Deng, "A cost-based selective maintenance decision-making method for machining line," *Quality and Reliability Engineering International*, vol. 27, pp. 191–201, 2011.

[14] M. Panday, M. J. Zuo, R. Moghaddass, and M. K. Tiwari, "Selective maintenance for binary systems under imperfect repair," *Reliability Engineering and System Safety*, vol. 113, pp. 42–51, 2013.

[15] D. Lin, M. J. Zuo, and R. C. M. Yam, "General sequential imperfect preventive maintenance models," *International Journal of Reliability, Quality and Safety Engineering*, vol. 7, no. 3, pp. 253–266, 2000.

[16] A. Khatab and E.-H. Aghezzaf, "Selective maintenance optimization when quality of imperfect maintenance actions are stochastic," *Reliability Engineering & System Safety,*, vol. 150, pp. 182–189, 2016.

[17] C. Chen, M. Q.-H. Mend, and M. J. Zuo, "Selective maintenance optimization for multi-state systems," in *In Proc. of the IEEE Canadian Conference on Electrical and Computer Engineering, Edmonton, Canada*, 1999.

[18] M. Pandey, M. J. Zuo, and R. Moghaddass, "Selective maintenance modeling for multistate system with multistate components under imperfect maintenance," *IIE Transcations*, vol. 45, pp. 1221–1234, 2013.

[19] C. D. Dao, M. J. Zuo, and M. Pandey, "Selective maintenance for multi-state series–parallel systems under economic dependence," *Reliability Engineering and System Safety*, vol. 121, pp. 240–249, 2014.

[20] C. D. Dao and M. J. Zuo, "Selective maintenance for multi-state series systems with s-dependent components," *IEEE Transactions on Reliability*, vol. 65, no. 2, pp. 525–539, 2016.

[21] Y. Zhou, T. R. Lin, Y. Sun, Y. Bian, and L. Ma, "An effective approach to reducing strategy space for maintenance optimisation of multistate series-parallel systems," *Reliability Engineering and System Safety*, vol. 138, pp. 40–53, 2015.

[22] C. D. Dao and M. J. Zuo, "Selective maintenance of multi-state systems with structural dependence," *Reliability Engineering and System Safety*, vol. 159, pp. 184–195, 2017.

[23] Y. Won and H. Chung, "Optimum replacement intervals with random time horizon," *Journal of Quality in Maintenance Engineering*, vol. 6, no. 4, pp. 269–274, 2000.

[24] A. Khatab, N. Rezg, and D. Ait-Kadi, "Optimal replacement with minimal repair policy for a system operating over a random time horizon," *Journal of Quality in Maintenance Engineering*, vol. 17, no. 4, pp. 415–423, 2011.

[25] ——, "Optimum block replacement policy over a random time horizon," *Journal of Intelligent Manufacturing*, vol. 22, no. 6, pp. 885–889, 2011.

[26] I. Djelloul, A. Khatab, E.-H. Aghezzaf, and Z. Sari, "Optimal selective maintenance policy for series-parallel systems operating missions of random durations," in *International conference on Computers & Industrial Engineering (CIE 45)*, 2015.

[27] A. Khatab, E.-H. Aghezzaf, I. Djelloul, and Z. Sari, "Selective maintenance for series-parallel systems when durations of missions and planned breaks are stochastic," in 8^{th} *IFAC Conference on Manufacturing Modelling, Management, and Control (MIM 2016)*, 2016.

[28] Artelys, "Artelys knitro," https://www.artelys.com/en/optimization-tools/knitro, 2017, retrieved Apr. 17, 2017.

[29] Maximal Software, "Mpl modeling system," http://www.maximalsoftware.com/mpl/, 2017, retrieved Apr. 17, 2017.

A study of variability induced by events dependency in microelectronic production

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Kean Dequeant, Pierre Lemaire, Marie-Laure Espinouse
Univ.Grenoble Alpes, CNRS, Grenoble INP, G-SCOP,
F-38000 Grenoble, France.
Email: kean.dequeant@grenoble-inp.fr,
pierre.lemaire@grenoble-inp.fr,
Marie-Laure.Espinouse@g-scop.grenoble-inp.fr

Philippe Vialletelle
STMicroelectronics,
F-38926, Crolles Cedex, FRANCE
Email: philippe.vialletelle@st.com

Abstract—Complex manufacturing systems are subject to high levels of variability that decrease productivity, increase cycle times and severely impact the systems tractability. As accurate modelling of the sources of variability is a cornerstone to intelligent decision making, we investigate the consequences of the assumption of independent and identically distributed variables that is often made when modelling sources of variability such as down-times, arrivals, or process-times. We first explain the experiment setting that allows, through simulations and statistical tests, to measure the variability potential stored in a specific sequence of data. We show from industrial data that dependent behaviors might actually be the rule with potentially considerable consequences in terms of cycle time. As complex industries require strong levers to allow their tractability, this work underlines the need for a richer and more accurate modelling of real systems.

Keywords—variability; cycle time; dependent events; simulation; complex manufacturing; industry 4.0

I. Accurate modelling of variability and the independence assumption

Industry 4.0 is said to be the next industrial revolution. The proper use of real-time information in complex manufacturing systems is expected to allow more customization of products in highly flexible production factories. Semiconductor High Mix Low Volume (HMLV) manufacturing facilities (called fabs) are one example of candidates for this transition towards "smart industries". However, because of the high levels of variability, the environment of a HMLV fab is highly stochastic and difficult to manage. The uncontrolled variability limits the predictability of the system and thus the ability to meet delivery requirements in terms of volumes, cycle times and due dates.

Typically, the HMLV STMicroelectronics Crolles 300 fab regularly experiences significant mix changes that result in unanticipated bottlenecks, leading to firefighting to meet commitment to customers. The overarching goal of our strategy is to improve the forecasting of future occurrences of bottlenecks and cycle time issues in order to anticipate them through allocation of the correct attention and resources. Our current finite capacity projection engine can effectively forecast bottlenecks, but it does not include reliable cycle time estimates. In order to enhance our projections, better forecast cycle time losses (queuing times), improve the tractability of our system and reduce our cycle times, we now need accurate dynamic cycle time predictions.

As increased cycle-time is the main reason workflow variability is studied (both by the scientific community and practitioners, see e.g. [1] and [2]), what follows concentrates on cycle times. Moreover, the "variability" we account for should be understood as the potential to create higher cycle times, even though "variability" may be understood in a broader meaning. This choice is made for the sake of clarity, but the methodology we propose and the discussion we lead can be applied to any other measurable indicator.

Sources of variability have been intensely investigated in both the literature and the industry, and tool down-times, arrivals variability as well as process-time variability are recognized as the major sources of variability in that sense that they create higher cycle times (see [3] for a review and discussion). As a consequence, these factors are widely integrated into queuing formulas and simulation models with the objective to better model the complex reality of manufacturing facilities. One commonly accepted assumption in the development of these models is that the variables (MTBF, MTTR, processing times, time between arrivals, etc.) are independent and identically distributed (i.i.d.) random variables. However, these assumptions might be the reason for models inaccuracies as [4] points out in a literature review on queuing theory. Several authors have studied the potential effects of dependencies, such as [5] who studied the potential effects of dependencies between arrivals and process-times or [6] who investigated dependent process times, [4] also gives further references for studies on dependencies effects. In a previous work [3], we pinpointed a few elements from industrial data that questioned the viability of this assumption in complex manufacturing systems.

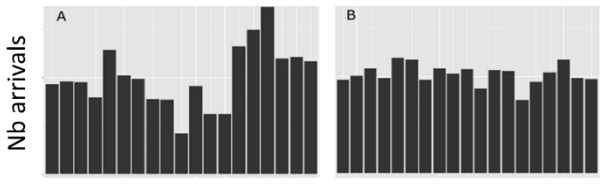


Figure 1: Number of arrivals per week from real data (A) and generated by removing dependencies (B)

For instance, Figure 1.A shows the number of arrivals per week at a toolset from STMicroelectronics Crolles300 fab, while Figure 1.B shows arrivals per week generated using the same inter-arrival rate as in Figure 1.A, but assuming independence of arrivals. The number of arrivals per week appears to be much more variable in reality than with the i.i.d. assumption.

From the (variability) modeling point-of-view, the i.i.d assumption means that all the information of many individual similar events (e.g., down-times) is contained within their distribution, and that the actual *sequence* of the events does not carry any additional information or, here, variability potential. That is why some events can be described by standard statistics, such as the mean and the variance, that completely define the statistical distribution. For simplicity of explanations, we will refer to sequences of i.i.d. random variables as i.i.d. sequences and the events they represent as i.i.d. events. The objective of this article is to investigate the impact that the i.i.d assumption may have on the quality of cycle time models and simulations by testing real world sequences of data. We therefore propose in the rest of our article an experiment framework for testing the variability potential of dependencies and show the results obtained on down-time distributions.

II. An experiment framework for testing the variability potential of dependencies

A. Objective: testing sources of variability for "dependency variability potential"

The objective we have is to investigate whether or not some information contained in the sequence of events might contribute to the variability. We therefore wish to investigate if there is any significant variability potential stored in the sequences of industrial data. Note that, from an operational perspective, if the events are not i.i.d but the result in terms of cycle time generated is similar, it can be considered that there is no useful information contained in the sequence and that the assumption is acceptable.

What we understand by the "variability potential" of a source of variability is its tendency to generate higher cycle times. For instance, queuing theory proposes using the coefficient of variation (the ratio between the standard deviation and the mean) for measuring the variability potential of arrivals (applied on inter-arrival times) and processing times (see e.g. Factory Physics [7]). To go further, we propose a framework to test if a specific *sequence* of data points carries a different variability potential than an i.i.d. sequence from the same data points. We will illustrate this approach by running a simulation model that integrates tool down-times, as they are the main source of variability in semiconductor manufacturing. Figure 1 shows a sequence of down-times and up-times of a specific tool from the Crolles 300 semiconductor fab recorded for a period of 1 year.

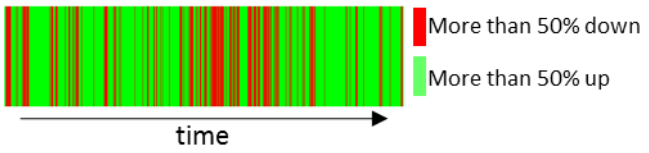


Figure 2: Historical down-times/up-times from a specific Crolles 300 tool over a year.

We call S_0 such a particular (real) sequence of events. Our fundamental question is whether the cycle time induced by the specific sequence S_0 is significantly different than that of an i.i.d. sequence from the same distribution, which we can create by applying a random permutation to the data points representing S_0. To answer this question, we propose an experiment framework that measures through simulations whether the effect of S_0 is statistically significantly different to that of i.i.d. sequences from the same distribution. The following subsections detail the different components of the proposed framework.

B. Setting the simulation

The framework we propose is quite generic and could be applied to many situations: the simulation model could be any system, as long as one can control its inputs and measure its ouputs according to the hypothesis one wants to test.

In our case, we want to challenge the i.i.d. hypothesis, and test its effects on cycle-times. We therefore settle for a minimalistic simulation model, so as to disregard all potential sources of variability other than the independence (or not) of the down-times. We therefore consider the following system: agents queue in an infinite FIFO queue, waiting to be processed by a single tool with constant process times that treats each agent one after another with no inefficiency whatsoever except for the down-time events. These down-time events are set to follow specific sequences, being on the one hand the reference (real) sequence S_0, and on the other hand (as a basis for comparison) many other i.i.d. sequences. We ran all our simulations with Anylogic Professional in an agent based simulation exported as a standalone application: this allowed us to control the simulation model with R, and manage the inputs/ouputs of thousands of simulations in an automated manner.

C. Measuring the effect of S_0

In our case, the effect of a sequence of down-times is measured as the average cycle-time of the agents in a simulation. This should be evaluated on an infinite horizon, that is until the measure is stable enough, but it is in practice impossible since S_0 is finite (as it is real, empirical data). As an alternative, to prevent any bias from a particular arrival sequence, we simply evaluate the long term average cycle-time by running many simulations on different scenarios, i.e. on different arrival sequences (following i.i.d. exponentially-distributed inter-arrival times). Thus, let $CT_{0,j}$ be the mean cycle time of agents from the simulation run on sequence S_0 of down-times and scenario j of arrivals; the effect of S_0 is then $\overline{CT}_0$, the mean of mean cycle times from simulations that used S_0 as an input. The number of different scenarios j is an important parameter and will be discussed in section II.F.

D. Measuring the effect of i.i.d. sequences

To measure the effect of i.i.d. sequences, we generate many such sequences from the same distribution; those sequences ($S_{i,i\neq 0}$) are generated as random uniform permutations of S_0. Using the same procedure as for S_0, we compute for each sequence $S_{i,i\neq 0}$ its effect: $\overline{CT}_{i,i\neq 0}$. Note that we use the exact same scenarios for arrivals as for S_0, this extra requirement insuring that any significant difference between the population

$\overline{CT}_{i,i\neq 0}$ and the value $\overline{CT}_0$ comes strictly from the difference between $S_{i,i\neq 0}$ and S_0.

The effect of a particular i.i.d. sequence is of virtually no interest as we want to measure the effect of *any* i.i.d. sequence. This could be measured as the mean value of the $\overline{CT}_{i,i\neq 0}$, but to get a better control on the comparison we prefer measuring the effect as a 95% confidence interval I_{95}.

We compute I_{95} as $\mu \mp 2\sigma$ where μ and σ are respectively the mean and the standard-deviation of $\overline{CT}_{i,i\neq 0}$. Doing such calculation, we assume that the $\overline{CT}_{i,i\neq 0}$ follow a Gaussian distribution. We justify this statement both by the central limit theorem as each point of our sample population is a mean and by the fact that the underlying population $CT_{i\neq 0,j}$ most likely follows a Gaussian distribution (since the cycle time of each distribution comes from the accumulation of many independent random effects).

As for the computation of $\overline{CT}_0$, the right number of sequences and of scenarios must be carefully chosen so as to get good estimates within reasonable running times. This is also discussed in section II.F.

E. Comparing the effects

Once $\overline{CT}_0$ and I_{95} have been computed accurately, they must/can be compared using a standard procedure for a statistical test: if $\overline{CT}_0$ falls within I_{95}, then it cannot be said that S_0 has an effect significantly different than that of an i.i.d. sequence; In the alternative, it can be assumed that S_0 has a different effect on cycle-time than an i.i.d. sequence. Figure 3 illustrates the three possible outcomes of the test. If $\overline{CT}_0$ is on the left side of I_{95} (Figure 3 A), S_0 carries a negative variability potential; If $\overline{CT}_0$ is within I_{95}(Figure 3.B), it cannot be said that S_0 carries any significant potential; If $\overline{CT}_0$ is on the right side of I_{95} (Figure 3.C), S_0 carries a positive variability potential.

Note that setting a 95% confidence interval actually means, from a skeptic point-of-view, that there is a 5% chance for any sequence tested that the results will turn out to be positive by pure randomness. It is therefore essential to test different S_0 sequences, and to compare the number of positives with the probability of getting such results by "luck". For instance, if we test 19 sequences, the chance of getting more than half results positive by luck is less than 10^{-5}.

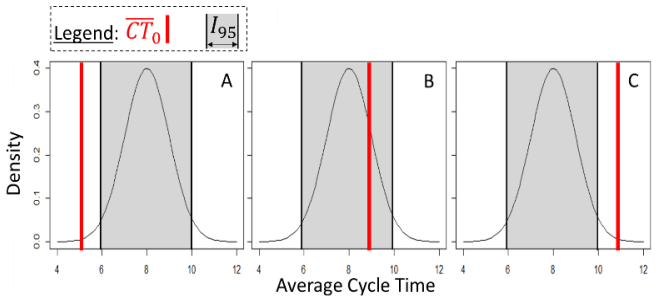


Figure 3: Three possible outcomes when comparing $\overline{CT}_0$ to the sample population $\overline{CT}_{i,i\neq 0}$

Note that a Student's t-test is not directly applicable here as it does not answer the question we are asking: should we perform a t-test between the populations $\overline{CT}_{0,j}$ and $\overline{CT}_{i\neq 0,j}$, we would be asking if the mean of both populations are identical or not. With a finite horizon of 1 year (resulting in a significant difference between each S_i), almost all $\overline{CT}_i$ are different than the mean of $\overline{CT}_{i\neq 0,j}$, thus a significant difference with this test would not bring any particular information on S_0.

As a complementary measure, we check, for each arrival scenario j, whether the mean of the population $\overline{CT}_{i\neq 0,j}$ is statistically different than $\overline{CT}_{0,j}$. For each scenario j, we can answer this question using a t-test and then count the number of times a scenario j gave $\overline{CT}_{0,j}$ respectively statistically smaller, non-statistically different and statistically higher than the mean of the population $\overline{CT}_{i\neq 0,j}$. As the population $\overline{CT}_{i\neq 0,j}$ represents average cycle times from identical simulation sets and parameters, only with randomness coming from different draws of the parameters, we can assume that they follow a normal distribution and can apply the t-test. These numbers are referred to as the "t-test triplets" in the results section III and are computed as a secondary test, keeping the previously explained methodology as the main measure.

F. Handling uncertainties on $\overline{CT}_0$ and I_{95}

As $\overline{CT}_0$ and I_{95} come from a finite population we created through simulation, the values we measure are actually estimates of the true values (that we would get if we had an infinite population) and therefore carry uncertainty. The actual value of $\overline{CT}_0$ is contained in a confidence interval around $\overline{CT}_0$, which we call $[\overline{CT}_0]$. The same applies to I_{95}: as we had to estimate I_{95}, we are not certain of the two limits of this interval, however I_{95} is most likely contained within inner/outer limits which we call I_{95}^- and I_{95}^+.

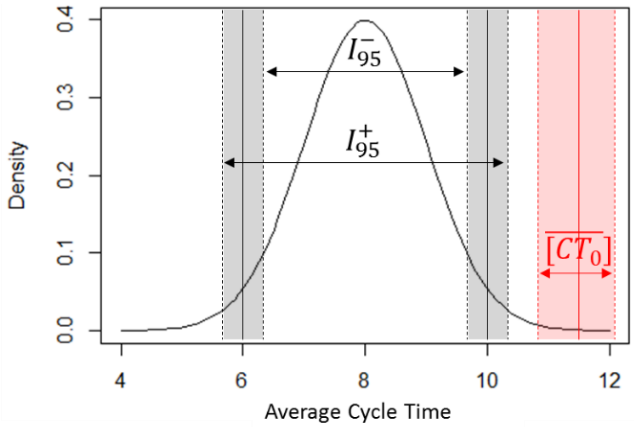


Figure 4: uncertainty areas around $\overline{CT}_0$ and I_{95}

These limits are represented on Figure 4 along with two greyed areas ($I_{95}^+ - I_{95}^-$) and the interval $[\overline{CT}_0]$. If $\overline{CT}_0$ is far away enough from I_{95} as for $[\overline{CT}_0]$ and ($I_{95}^+ - I_{95}^-$) not to intersect, then whatever the value $\overline{CT}_0$ takes within $[\overline{CT}_0]$, and whatever the value I_{95} takes within ($I_{95}^+ - I_{95}^-$), the conclusion is the same (as for Figure 3). The problem arises when the uncertainty areas of $[\overline{CT}_0]$ and I_{95} overlap: depending on what

true value $\overline{CT}_0$ and I_{95} take inside their intervals, the answer can be different, and we can therefore not be sure of the answer. This is a "we do not know" situation as we cannot say with confidence that there is a difference, nor can we say with confidence that there is no difference.

In order to get definitive answers, these uncertainty areas around $\overline{CT}_0$ and I_{95} need to be reduced to a point where they do not overlap anymore: to this extend, we need to increase the number of simulation runs so as to decrease either $[\overline{CT}_0]$ or the gap between I_{95}^+ and I_{95}^-. The appendix gives more details on the computation of these values and the way to minimize the number of simulations in order to get out efficiently of an overlap situation. In any case, there needs to be a minimum amount of simulations in order to start computing the uncertainty intervals. To test one S_0 sequence, we decided to run simulations on at least 20 i.i.d. sequences and 20 arrivals scenarios. Therefore, one simulation set is composed of at least 400 simulations.

G. Ending simulations

Each simulation needs to be run over the period of time defined by the sequence S_0 that is tested. However, at the end of this period of time, some agents may have accumulated in the simulation. Cutting the simulation abruptly can either lead to the loss of the cycle time of these agents or an underestimation of their cycle time depending on how the information is recorded in the simulation.

In order to prevent from such effects, we decided to stop the arrivals after the time duration defined by S_0, and continue the simulation until all agents have exited the simulation. We also chose to leave the tool up after the end of the given up/down sequence. This was first motivated by the fact that any other decision could introduce uncontrolled biases to the results. Secondly, by having a similar end to all simulations, we actually slightly underestimated the difference between the historical sequence and its i.i.d. counterparts: this only strengthens any significant results.

III. Results and implications

A. Down-times of C300 tools

Using the method described in section II, we tested 19 tools from the C300 fab. For each tool, we extracted the empirical sequence of up-times/down-times over a year. We then applied the procedure described in section II. One implicit parameter for all simulations is the utilization rate of the tool: in our simulation this utilization rate is fixed by adjusting the mean inter-arrival time of the scenarios. Since this utilization rate influences the cycle times, we set it as the same value across all simulations: 80%, as this is both a high and common, realistic level.

Table 1 summarizes the results of our experiments. The results are normalized so that the average cycle time for the sample i.i.d. sequences (μ) is equal to 100. Therefore, the values of $\overline{CT}_0$ can be compared straightforwardly between them and relative to the central cycle time for i.i.d. sequences μ. We also provide the confidence interval $[\overline{CT}_0]$ and the I_{95}^* value, which correspond to I_{95}^+ if $\overline{CT}_0$ is outside I_{95}^+ or I_{95}^- otherwise. The column "s" shows the result of our test: YES if there is a significant different, NO if there is no significant different, and NA if we were not able to get a definitive answer. We also add the "t-triplets" described in section II.E. Tools with same letters are from the same tool type.

TOOL	$\overline{CT}_0$	$\overline{CT}_0 - \mu$	I_{95}^*	$[\overline{CT}_0]$	s	t-test triplets
A1	78	-22%	81-119	76-80	YES	(21-0-0)
A2	80	-20%	83-117	77-83	YES	(28-0-0)
B3	175	75%	45-155	165-185	YES	(0-0-20)
C4	115	15%	88-112	113-118	YES	(3-8-89)
D5	668	568%	52-148	645-691	YES	(0-0-20)
E6	128	28%	63-137	119-137	NO	(0-0-21)
F7	182	82%	60-140	176-189	YES	(0-0-20)
F8	301	201%	60-140	279-324	YES	(0-0-20)
G9	131	31%	75-125	126-136	YES	(0-0-29)
H10	111	11%	88-112	109-113	NA	(10-11-79)
I11	144	44%	71-129	136-153	YES	(0-0-20)
J12	156	56%	50-150	151-161	YES	(0-0-20)
K13	208	108%	78-122	191-225	YES	(0-0-23)
K14	1385	1285%	40-160	1348-1421	YES	(0-0-20)
K15	261	161%	79-121	246-276	YES	(0-0-26)
K16	240	140%	62-138	229-250	YES	(0-0-20)
K17	420	320%	66-134	405-436	YES	(0-0-20)
L18	123	23%	66-134	119-128	NO	(0-0-20)
M19	104	4%	92-108	100-107	NO	(18-4-30)

Table 1: Results of the experiment framework tested on 19 tools from STMicroelectronics

As Table 1 shows, out of the 19 historical sequences we tested, 3 did not show any significant difference in the variability potential they carried compared to their i.i.d. counterparts, 13 showed an increase, 2 showed a decrease, and 1 could not draw a definitive answer after 10000 simulation runs. The t-test triplets confirm the results and bring additional information on the impact of the arrival scenarios, but are not enough to assert a significant difference (e.g. see tools E6 and L18).

The differences between $\overline{CT}_0$ and μ in Table 1 first show that the empirical sequences of down-times we tested are statistically different than i.i.d sequences in terms of variability potential. Secondly, it should be noticed that the historical sequences added non negligible cycle time: the average increase in cycle time when the variability potential is positive is +174%, the median increase being +64%. This means that, for the majority of cases we tested, more than one third of the cycle time came from the actual sequence and not from the statistical distribution.

An interesting result is that the sequence of down-times can actually carry negative variability potential. The probability of false-positive is here ruled out by the fact that both tools that showed a sequence of down-times with a negative variability potential are from the same tool type. We explain this as being

the result of "regular" down events amongst "irregular" down events. Indeed, these tools need regular changes of an important resource which create down-time events. However, even though the time between these special events are somewhat constant, there are many other down-time events that happen more or less randomly. Therefore the regularity of these events is lost in the statistics, and when running a simulation with i.i.d. down-time events, this regularity is broken and more variability occurs. The results from table 1 were obtained at a utilization rate of 80%. However, it is also interesting to run distinct experiments for the same sequences, but at different utilization rates. For each empirical sequence, we actually ran experiments on 9 different utilization rates ranging from 10% to 90% for a total number of 180 experiments, each requiring a minimum of 400 simulations. Figures 5 shows operational curves (representing cycle time versus utilization rate) for tools showing respectively an increase in variability potential, no significant change in the variability potential, and a decrease in the variability potential.

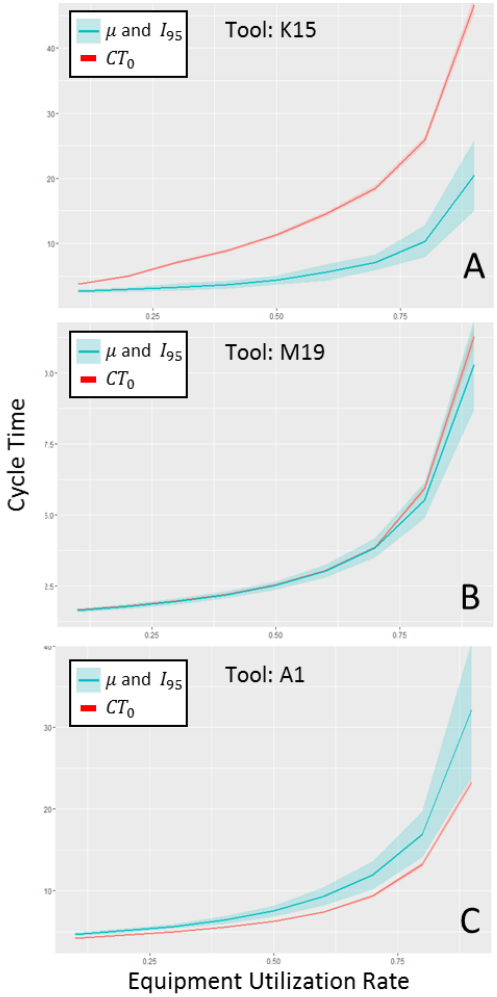


Figure 5: operational curve showing $\overline{CT_0}$ (statistically) higher (A), non-different (B), and lower (C) than μ

The blue curves of Figures 5 are the average cycle times for the sample populations $\overline{CT_{i,i\neq 0}}$ along with its 95% confidence interval. They represent the outputs of the simulations run on the i.i.d down-times sequences. The red curves represent the values $\overline{CT_0}$, i.e. the average cycle times generated by the simulation runs that used the historical down-times sequences. These curves display the expected exponential behavior of cycle time when it comes to utilization rates. The cycle times added (or removed) by the dependencies seem to be proportional to the original cycle times and the utilization rate does not seem to impact whether S_0 is significantly different or not.

B. Results from an operational point-of-view

The results we show indicate that a large part of the cycle time generated by down-time events from historical data in the case of i.i.d. arrivals is unaccounted for in most queuing theories and simulations. However, the difference in cycle times between most current models (using queuing theory or simulation) and reality is not in the order of magnitude that we pointed out. How can it be that we see a significant difference that does not happen in reality?

We believe this difference comes from another dependency phenomenon that is still mostly unaccounted for: the dependency between the arrivals and the tool behaviors. Indeed, in the complex reality of manufacturing, decisions are made dynamically based on the flow of products. For instance, maintenances might be pushed back in the case of temporary over-saturations, more maintenance staff is also affected to tools that are forecasted to be highly used in the near future, etc.

IV. Perspectives and conclusion

A. Quantifying other sources of variability

The results we give in this article show strong evidence that tool down-times actually do not follow i.i.d. distributions and that this difference has a huge impact on the variability potential of down-times. As Fig. 1 shows, there is strong evidence that other sources of variability such as arrivals and processing-times also do not follow i.i.d. sequences. The next step is therefore to apply the same procedure on historical arrivals sequences as well as historical processing-times sequences.

B. Quantifying intra-sources relations

As we previously mentioned, there is strong evidence for a relation between the different sources of variability both from the shop-floor experience and from our quantification of the variability potential of down-time events. Further research should focus on quantifying this relation. Indeed, better modelling this relation could help both making models more accurate and reduce the overall variability fabs experience, as measuring is always the first step to reducing.

C. Conclusion

Accurate modelling of sources of variability (i.e. root causes for the generation of queuing time inefficiencies) is a key element of the manufacturing strategies put forward with industry 4.0 and High Mix Low Volume production. Previous works had questioned the viability of the assumption of independent and identically distributed random variables when it comes to the modelling of sources of variability such as tool down-times, arrivals, process times…

In this article, we proposed an experiment framework based on the repetition of many simulation runs that allows testing on

historical sequences of data if this assumption has any implication in the modelling of variability. The experiment framework was built around the specific scenario of testing down-times, but can be applied to any sequence of data as long as a simulation is able to integrate the source of variability that is tested.

We then tested 19 different industrial down-time sequences of 1 year taken from tools from the Crolles 300 semiconductor fab. The results show that not only do the sequences of down-times carry significant variability potential, they actually contain an important part of the variability potential for most tools. We also showed that, quite interestingly, the sequence of down-times can actually carry negative variability potential and generate less cycle time than if they were sequences of independent and identically distributed variables.

Our results suggest that there is a strong relation between the arrivals and the behaviors of the tools in the reality of complex manufacturing. A next essential step is therefore to measure, quantify and understand the interactions that happen in the real world of complex manufacturing in order to better model these interactions and give industrials the levers they need to follow their transition to industry 4.0.

Appendix

We mentioned in section II.G that the uncertainties on $[\overline{CT}_0]$ and I_{95} might be too high to draw a definitive conclusion. As the uncertainties on $[\overline{CT}_0]$ and I_{95} come from the estimators, they reduce with the number of simulation runs. Both uncertainty areas can be estimated straightforwardly: the standard deviation of the estimator for the mean μ and the standard deviation σ are respectively $\frac{\sigma}{\sqrt{n}}$ and $\frac{\sigma}{\sqrt{2(n-1)}}$ where n is the sample size [8]. Once again, we use these formulas as we make the assumption that the underlying populations follow normal distributions. Thus, we can define $[\overline{CT}_0]$ as $\overline{CT}_0 \mp \frac{2\sigma_0}{\sqrt{m}}$, I_{95}^- as $\mu \mp \left(2\sigma - \frac{\sigma}{\sqrt{n}} - \frac{2\sigma}{\sqrt{2(n-1)}}\right)$, and I_{95}^+ as $\mu \mp \left(2\sigma + \frac{\sigma}{\sqrt{n}} + \frac{2\sigma}{\sqrt{2(n-1)}}\right)$ where σ_0 and σ are respectively the standard deviation of $\overline{CT}_{0,j}$ and $\overline{CT}_{i,i\neq 0}$; n and m are respectively the number of different i.i.d. sequences and of arrival scenarios; And μ is the mean of all $\overline{CT}_{i,i\neq 0}$. To derive the formulas on I_{95}^- and I_{95}^+, we start with the estimation of $I_{95}(\mu \mp 2\sigma)$, and we subtract (I_{95}^-) or add (I_{95}^+) a standard-error for μ $\left(\frac{\sigma}{\sqrt{n}}\right)$ and two standard-errors for σ $\left(\frac{\sigma}{\sqrt{2(n-1)}}\right)$.

After the initial 400 simulations, the condition to either increase the number of sequences S_i or the number of arrivals scenarios is actually the same: an overlap between the red area and either of the grey areas of Figure 5. The decision of which to do first should be taken in order to minimize the computation time required. Drawing another arrival sequence will reduce $[\overline{CT}_0]$ whereas drawing another S_i sequence will reduce the gap between I_{95}^- and I_{95}^+. We therefore need to compute the decrease in the uncertainty area per simulation for either case of making a new draw of arrivals or of down-times, knowing that any new draw of arrivals needs to be run on all existing S_i sequences and vice-versa.

References

[1] D. Delp, J. Si and J. W. Fowler, "The development of the complete X-factor contribution measurement for improving cycle time and cycle time variability", *IEEE Transactions on Semiconductor Manufacturing,* 19(13), pp. 352-362, 2006.

[2] D. J. Kim, L. Wang and R. Havey, "Measuring cycle time through the use of the queuing theory formula (G/G/M)", *Proceedings of the 2014 Winter Simulation Conference,* pp. 2414-2421, 2014.

[3] K. Dequeant, P. Vialletelle, P. Lemaire and M.-L. Espinouse, "A literature review on variability in semiconductor manufacturing: the next forward leap to Industry 4.0", *Proceedings of the 2016 Winter Simulation Conference,* pp. 2598-2609, 2016.

[4] J. G. Shanthikumar, S. Ding and M. T. Zhang, "Queueing theory for semiconductor manufacturing systems: a survey and open problems", *IEEE Transactions on Automation Science and Engineering,* 4(14), pp. 513-522, 2007.

[5] S. Ross, J. G. Shanthikumar and X. Zhang, "Some pitfalls of blackbox queue inference", *Probability in the Engineering and Informational Sciences*, 7(02), pp. 149-157, 1993.

[6] J. A. Buzacott, "The Effect Of Queue Discipline On The Capacity Of Queues With Service Time Dependent On Waiting Times", *INFOR: Information Systems and Operational Research,* 12(12), pp. 174-185, 1974.

[7] W. J. Hopp and M. L. Spearman, Factory Physics: Foundations of Manufacturing Management, Irwin/McGraw-Hill, 2001.

[8] S. Ahn. and J. A. Fessler, "Standard errors of mean, variance, and standard deviation estimators", Lecture notes, EECS Department, The University of Michigan, 2003.

Prioritization of failure modes of fire and rescue vessels with application case study

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mohammad HAIDAR
Marine Operation and Maintenance Supervision
Kuwait Fire Service Directorate, Kuwait
eng.m.haidar@kfsd.gov.kw

El-Awady ATTIA
Mechanical Engineering department, Faculty of engineering (Shoubra), Benha University, Cairo, Egypt
elawady.attia@feng.bu.edu.eg

Ahmed EL-ASSAL
Mechanical Engineering department, Benha faculty of engineering
Benha University, Benha, Egypt
ahmed.el-assal@bhit.bu.edu.eg

Abstract—Marine maintenance department in Kuwait fire service directorate is accountable to perform and organize maintenance activities to guarantee that the fire and rescue vessels are in duty and reliable. To relieve operational failures, critical equipment must be specified and prioritized. In this paper, an application case study is presented to prioritize failure modes of equipment of fire and rescue vessels according to their criticality using MCDM. A set of evaluation criteria is proposed includes equipment function, service experience, obsolescence, and failure-associated risk. Each one of these criteria has the associated set of sub-criteria. Failures with high critical scores are recognized and assigned to preventive maintenance. The information of various vessel equipment is extracted from records of the maintenance department to illustrate the model.

Keywords—Fire and rescue vessels, criticality assessment, maintenance strategies, MCDM, AHP.

I. Introduction

Generally, the magnitude of lives lost at sea has excruciating effect not only for local society, but also for the growth of the national economy of coastal countries. In the early Twentieth century, many countries were endeavoured to set rules, laws and guidelines to build and organize marine search and rescue activities. In 2004, the Indian Ocean Tsunami appears that disasters continue to cause wastage of human life and affect the environment; moreover, it causes infrastructure deficit and economic loss [1]. If we reminisce in the history of marine disasters, we recognise the magnitude of human and economic loss, such as Titanic disaster 1912 that lost more than 2000 passenger [2]. One of the worst passenger-ferry disaster is Scandinavian Star 1990; the ferry was on its way from Norway to Denmark. The fire on board spread extremely and 158 people lost their lives [3]. Other maritime disaster was in the Baltic sea; Estonia ferry 1994 costing over 800 lives [4]. Marine disaster is a type of serious disaster with low chance and high effect. A disaster at sea is an accident that affects the entire vessel, persons on board, cargo and the environment. In 1998, International Maritime Organization (IMO) relied a revised append to the convention, which cleared the responsibility of governments and place major emphasis on the regional path and coordinate between marine and aeronautical Search And Rescue (SAR). Through 2006, SAR convention had been approved by 86 countries represent just over 52% of world trader's shipping tonnage [5]. Accordingly, we recognize the importance to ensure that the fire and rescue vessels are in duty and highly reliable and their effect may touch human, environment and economic loss. This responsibility lies on the maintenance department to assure that rescue vessels are available and effective for any fire, rescue or search operation. This goal requires application of effective maintenance programs. One of the effective maintenance programs is the Reliability Centred Maintenance (RCM). RCM has been developed for aircraft maintenance programs in 1967, it was first applied on the Boing 747. Its significant has been proven in selection of preventive maintenance for critical facilities. It focuses on system function and critical failures, then controlling the failure modes, specify the critical failures, and recognize the suitable and efficient task to apply [6]. After that, it has been applied in many industries. Its objective is to minimize the maintenance cost by focusing on system critical function [7].

In this paper, a multi-criteria decision making model MCDM for ranking failure modes of fire and rescue vessel equipment is presented. The analytical hierarchy process (AHP) is used relying on the experts' evaluation. A set of evaluation criteria is proposed includes equipment function, service experience, obsolescence, and failure-associated risk. Each one of these criteria has the associated set of sub-criteria. The pairwise evaluation has been used to evaluate the importance of each criterion, while the absolute evaluation has been used for weighting failures with respect to the given sub-criterion. The proposed methodology is established for selecting appropriate maintenance strategies for critical vessels' equipment. The next section presents the literature review. Section 3 presents an overview of fire and rescue vessels. Section 4 introduces the research methodology. Section 5 presents the application of the methodology on the fire and rescue vessels.

II. Literature survey

The literature provides many applications on the utilization of MCDM for prioritization of failures modes, e.g. electrical transformers [8], medical equipment [9] [10], health care facilities [11], pavements [12], public buildings [13], electric power substations [14]. Regarding approaches that were used, one can find several: Andrusca *et al.* [14] proposed a weighted sum based real time approach to rank electrical subsystems. They ranked equipment depending on two main indices, the technical conditions and equipment importance. For maritime maintenance [15] proposed three steps based approach to evaluate the reliability and availability of ships equipment. The

first step is the failure mode and criticality analysis (FMECA); the second is the criticality matrix that relies on severity and frequency matrix; and the third is the fault tree analysis. The quality function deployment has been used also to prioritize medical equipment [10]. Bradley and Guerrero [16] used data-elicitation based technique to apply failure modes and effects analysis (FMEA) to relieve the effects of failure modes on products. Moreira *et al.* [8] prioritized the equipment failures using "Preference Ranking Organization Method for Enrichment Evaluations" PROMETHEE. Moreover, they used Fuzzy-PROMETHEE for the same application. Bowles [17] used the methodology of Risk Priority Number (RPN) based on the three factors: failure severity, probability of occurrence and the likelihood of its detectability in design evaluation tests of automobiles. Khanlari *et al.* [18] and [19] implemented fuzzy logic to prioritize equipment relying on RPN. Babashamsi *et al.* [12] integrated fuzzy-AHP and VIKOR method in the prioritization of pavement maintenance activities. They used AHP for ranking the evaluation criteria and VIKOR for ranking alternatives. Sweis *et al.* [11] used a similar approach of [13] for prioritize building maintenance activities. The model is a modified version of AHP that relies on some criteria includes building status, physical conditions, building importance, impact on users and the cost. Some others like [20] used on-line production information to prioritize maintenance work-orders that can enhance productivity within the specified planning horizon. Andersen *et al.* [21] proposed prioritization for maintenance and repair task of monitoring devices of embankment dams. The approach relies on the determination of the importance of each device in terms of dam safety and the measured conditions of dams through field inspection. The device importance is determined based on a set of predefined criteria weighted by experts. Dekker [22] integrated the priority setting in the optimization model of maintenance activities planning. For almost all of these models, the criteria that were considered are general. The obsolescence of the concerned equipment should also be analysed before the prioritization process. This would help in analysing potential and equipment role in developing the performance of the concern organization [23]. Relying on [24], the analysis of equipment age is important because this would help in taking the effective decision by predicting and comparing the return of investment in a particular activity as compared to others. Taghipour *et al.* [9] presented one of effective methodologies that have been used to prioritize medical equipment for maintenance decisions. They proposed to use AHP method. The weight for each criterion was determined relies on the pairwise comparison while the ranking of failures with respect to the specified criterion was proposed to be evaluated in absolute assessment. Moreover, the criteria and the associated sub-criteria have been considered in a detailed manner. In the current study, a similar methodology will be used with some modification.

III. Fire and rescue vessels

The fire and rescue vessels in the Kuwait Fire Service Directorate (KFSD) have the responsibility to provide the protection of human life and properties at local and international sea for the state of Kuwait. The marine fire department has four marine fire stations distributed to cover the coastal area continuously for any raised fire or rescue situation. There are about 400 workers and about 110 of them are working in the maintenance department. Their main responsibilities are the execution of different maintenance activities for 40 vessels of various size contain fire fighter tug and fast response rescue vessels. The main target of this department is putting rescue vessels in high availability, reliability and agility. Figure1 shows one of the fire and rescue vessels during firefighting of a fire disaster in fishery boat. In such restrictions, taking a decision of putting one vessel in dry-docking for maintenance is very difficult, where this vessel will be unavailable for a specific period of time. Figure 2 shows fire tug vessel under maintenance this vessel is designed for firefighting for large ships like container, cargo and oil tankers. Its availability is crucial for commercial and oil ports. In order to develop an effective maintenance program for these essential vessels, one needs to develop a reliable maintenance program at low cost. For each specified vessel, there are highly crucial equipment that affect the overall vessel performance and its functionality. This fact highlights the need of ranking these failure modes of equipment according to the impact of the associated failure(s).

Fig. 1. Fire disaster



Fig. 2. Fire tug vessel



IV. Research methodology

A. Data collection

The data related to different vessels and the associated equipment were collected, in addition to the failure mode(s) of each piece of equipment. The specific data of 23 equipment will be only presented for simplicity. This set of equipment is common in the fire and rescue vessel fleet. Visual inspection has been done to specify the current condition of each equipment. The Failure mode and effect analysis (FMEA) was developed to be used in the study.

B. Evaluation criteria and sub-criteria

1) Equipment function: C_1

Equipment function represents the main purpose to achieve a specified role in fire and rescue missions. This criterion represents the extent by which a specified failure affects the purpose of a given vessel. The failure severity reduces in function of the available equipment with the same role in the firm. Accordingly, here we propose two sup-criteria: the function criticality (C_{11}) and the availability of similar equipment (C_{12}).

2) Service experience: C_2

Service experience reflects the experience of maintenance staff in repairing specified equipment on time at low cost. This criterion can be estimated by the overall labour effectiveness (OLE), and the cost of repair. The OLE can be inferred relying on three main aspects: availability, quality of work done, and performance. The availability and temporal performance can be deduced by the average downtime of equipment. The quality of maintenance staff can be estimated relying on the failure frequency. In this view, four sub-criteria are proposed:

Average downtime C_{21}: Downtime represents the period during which specified vessel is out of service due to specified equipment failure. It can be measured by the mean time to repair for the specific failure.

Operational availability C_{22}: Operational availability is the proportion of time during a mission which particular vessel is available for use satisfactory. It includes all sources of downtime related to technical problems. It can be computed as the ratio between the vessel uptime to the length of the mission period.

Failure frequency C_{23}: Failure frequency is the likelihood of a failure occurrence. It represents the number of failures of certain equipment per specified period. It is often estimated in number of failures per hour. Smaller values of failure rate indicate high quality of maintenance work.

Cost of repair C_{24}: The main objective of service responsible(s) is to provide the required service on time and at an affordable cost. Consequently, the average repair cost of specified failure can be a good indicator for the experience gained overtime.

3) Obsolescence: C_3

Vessels are often designed for long service life, and they often have high construction, operation and maintenance costs. Even though, the technological advancement is rapidly changed. New equipment often realize the same function as the current with different performance levels, e.g. lesser failures, lower energy consumption, higher comfort to persons on board etc. Maintaining vessels in good working conditions requires many maintenance activities relying on [25] about half of the bill of materials used in the vessel's construction are changed over its lifetime. For specified equipment, a degree of obsolescence is presented in case of unavailability of original part manufacturer or non-functional of part(s) due to external reasons, and it may be linked to maintenance activities [26]. The degree of vessel's obsolescence affects the severity of failure and the associated downtime, failure rates, etc. According to [27] the obsolescence of an equipment can be characterized relying on three aspects: Logistical loss, functional loss, and technological obsolescence. Here, these three aspects can be inferred by estimating the residual lifetime of the vessel, and manufacturer support.

Vessel residual lifetime C_{31}: Technical lifetime is known as the total time for which the equipment is technically designed to operate. The technical lifetime is expressed in years or operation hours. The residual lifetime of the equipment is the time-period for which the given equipment can be continued operating before its replacement for technical reasons e.g. safety, or performance deterioration. The residual lifetime of pertinent equipment must be considered prior to the application of any maintenance program.

Manufacturer support C_{32}: it represents an essential factor to keep vessels in efficient operating conditions. This support can be classified into two main features: technical support and availability of original spare parts. Availability of original spare parts cannot be restricted to the original manufacturer. That is to say, many other suppliers are able to deliver original parts. For critical operations like marine fire and rescue, it is crucial to rely on a set of professional suppliers. Here, manufacture support depends on technical Support and availability of spare part.

Technical Support C_{321}: Along with the vessel lifetime, manufacturer technical support is very important especially for uncommon failures. The required knowledge or training can be provided that facilitate solving problems internally. Availability of technical support can reduce the severity of specified failure by reducing repair time and reducing equipment deterioration.

Spare part availability C_{322}: To minimize downtime of critical vessels, a set of specified parts should be available on time. These parts can be classified into two main categories: repairable parts and disposal parts. Both of them should be stocked with specified quantities for emergency failures. After a failure, the required part is picked from inventory. Repairable part will be reconditioned or repaired in the specified workshop or send to a stakeholder. Next, it should be placed again in the inventory for other emergency failures. Losing of such availability of parts leads to high system banality that affects the vessels' reliability and consequences of failure.

4) Failure Risk: C_4

For almost all applications the risk criterion is considered as the most important. In maritime fire and rescue sectors, the consequences associated with unavailability of one or more vessels can generate many catastrophic issues with local and international aspects. The root causes of these consequences are failures. Accordingly, the failure risk criticality can be considered relying on failure aspects, frequency, delectability, and consequences. These three prices can be used to determine the severity degree of failure. RPN is often computed relying on the failure delectability, probability of occurrence and severity. Failure consequences (severity) can be classified according to the associated failure impact on the equipment function, cost, environment, and safety, as discussed by [9].

Failure frequency C_{41}: Failure frequency can contribute in assessing failure risk. It can be estimated as sub-criterion C_{23}.

Failure delectability C_{42}: Failure delectability is the ability to detect a failure before its occurrence. The failure can be detected by inspection or using the sophisticated instruments during P-F interval or the Failure Detection Threshold (FDT). The detection

of a failure mainly depends on the staff experience and the availability of the required instrumentation.

Severity of failure consequence C_{43}: To consider the total consequences of specified failure mode, there are four possible impacts: environment, safety, operational (downtime), non-operational impact (cost) [28].

Environment C_{431}: The vessels influence the atmospheric and maritime environment, there are many aspects of this effect. First, the gas emission: the emission of carbon dioxide from ships was about 2.02% of global human-made in 2012 as presented by IMO. The degradation of vessels' engines increases this percentage. In addition, fires and fire flames contaminate seawater and pollute atmosphere. The availability of fire and rescue vessels reduces or prevents this effect. In case of ships problems, there are some risks of oil spills that is considered also as a serious action against maritime species.

Safety C_{432}: The failure of vessel body by corrosion not only decreases the vessel structural strength and service life, but also can raise the sailing resistance, decreases the speed of sailing, makes the vessel sailing hard and imposes a negative impact on the navigation safety. The bad condition of exterior hull of the vessel can drive to the incidence of maritime sufferer and obtain heavy losses to human lives and property [29]. The degradation of vessel equipment functions negatively affects the safety levels of persons on-board. Moreover, it has a serious effect on the success of the rescue mission. A crucial need for all maritime structure is to have minimum elementary and operational costs, to be rationally safe. Safety is not only concerned with the vessel itself, but also with surrounding damage that may occur because of failure. Safety is not an absolute measurement and must be associated with economic and social consequences [30].

Operational C_{433}: The operational effect can be translated by the vessel downtime. Downtime of an equipment can lead to serious consequences, especially for insufficient number of the same vessel function.

Cost C_{434}: Maintenance task is a dynamic role to realize highest performance goal, each industry establishes a maintenance department to realize their needs effectively. Economic consideration is one of the serious areas of decision-making in the maintenance department. The failure can produce many economic aspects, e.g. cost of repair, cost of failure consequences, e.g. loss of goods for a specified sea ship.

C. Selection of a MCDM method

Multi-Criteria Decision Making (MCDM) is a well-known category of decision making, dividing into multi-objective and Multi-Attribute Decision Making (MADM) [31]. MADM is making decisions such as estimation, prioritization, and select over alternatives, which identify by multiple attributes [32]. It helps decision makers to discover the most effective solution for the problem that comprise multi criteria, and it has been used widely in many areas. In this study, we adopted the Analytic Hierarchy Process (AHP) that enables managers to prioritize failure modes of vessel equipment and find guidance for choosing adequate maintenance strategies. AHP is used to determine the weighting value for each criteria and sub-criteria based on their relative importance for vessel equipment. AHP is a theory of measure over pairwise comparisons that depend on experts' judgments to achieve priority scales [33]. AHP epitomized in a three-step process: it considers a combination problem into a hierarchy, in which the inclusive decision goal lies at the hierarchy head and the criteria, sub-criteria and decision alternatives are at each downward level of the hierarchy composing of specific factors. Decision makers then compare every factor to others at the same level using a pairwise comparison matrix to determine its weight or related importance. The optimum resolution is the alternative with the greater accumulative weight. There are two classes of evaluation can be used in AHP: absolute and relative measurements. Absolute measurement is utilized to rank the alternatives regarding the criteria separated from other alternatives, i.e. alternatives are compared with a criterion in one's mind that has been expanded over experience. However, in relative measurement the priority of an alternative rely as well on other alternatives, i.e. alternatives are compared in pairs to a common attribute. In this study, both methods will be used.

V. Case study

To apply AHP, the hierarchical structure of the problem should constructed. The goal is to prioritize vessels equipment's failure modes. The second level is the identification and placement of all criteria and sub-criteria in the hierarchy. The third level represents alternatives that signify the equipment with the associated failure modes. In the proposed AHP model, there are many ranking procedures: first, is the ranking of criteria according to the associated influence on achieving the desired goal. This step is achieved relying on the pairwise comparison of the set of criteria regarding to the goal to be achieved. The second step is the identification of the contribution of each sub-criteria. It represents the degree of importance of sub-criteria with respect to parental criterion. In this study, this degree is computed relying also on the pairwise comparison between the set of sub-criteria as a traditional AHP process. The third step represents the ranking of equipment failure modes according to each sub-criterion. For this assessment, it is proposed to compute them relying on the absolute ranking, i.e. for each sub-criterion, a set of grades (e.g. high, intermediate, low) is proposed, the experts give each failure one of these grades according to the failure nature. In order to weighting these grades (known here as grade strength): a pairwise comparison is developed to rank them. The last step is the aggregation of these weights in order to compute the absolute rank of each equipment failure mode. By the following, the model is illustrated in details.

A. Weighting criteria and sub-criteria

After identifying all criteria and sub-criteria, their related significance can be calculated with respect to their goal using Saaty's eigenvector [33]. In other words, relative assessment is used to identify weights of criteria and that of the sub-criteria. Table I shows the weights determined for the main criteria and table II presents the pairwise comparison of a set of sub-criteria e.g. that of "Severity of failure consequence" which has four sub-criteria: environment, safety, operational, and cost". The values specified here exemplify experts' opinion; results might vary with the participation of other experts. In our suggested model, the pairwise comparison matrices is structured by employing Saaty's crisp 1–9 scales. The ranking of all sub-

criteria are computed as shown in table III. Moreover, the hierarchy of criteria and sub-criteria is also shown.

TABLE I. Pairwise comparison of criteria (CR = 0.0034)

	C1	C2	C3	C4	Weight
C1	1.00	1.40	1.80	0.56	0.25
C2	0.71	1.00	1.67	0.43	0.19
C3	0.56	0.60	1.00	0.33	0.13
C4	1.80	2.33	3.00	1.00	0.43

TABLE II. Pairwise comparison of sub-criteria of Severity of failure consequence: C_{43} (CR = 0.0024)

	C431	C432	C433	C434	Weight
C431	1.00	1.29	3.00	4.50	0.44
C432	0.78	1.00	1.80	3.00	0.31
C433	0.33	0.56	1.00	1.67	0.16
C434	0.22	0.33	0.60	1.00	0.10

TABLE III. Weights of criteria and Sub-criteria

Criteria (Weight)	*Sub-criteria*	*Sub-criteria*
C1 - Equipment function (0.25): $CS_{Vessel\ hull,\ C1}$ = 1.000	C11-Function importance (0.83): High (1.000) C12- Availability of similar equipment (0.17): High (1.000)	
C2 - Service experience (0.19): $CS_{Vessel\ hull,\ C2}$ = 0.937	C21 - Average downtime (0.47): High (1.000) C22 - Operational availability (0.10): High (1.000) C23 - Failure frequency (0.18): Occasional (0.647) C24 - Cost of repair (0.25): High (1.000)	
C3 – Obsolescence (0.13): $CS_{Vesselhull,\ C3}$ = 0.495	C31 - Vessel residual lifetime (0.25): Old (1.000) C32 - Manufacturer support (0.75): $CS_{Vessel\ hull,\ C32}$ = 0.327	C321-Technical support (0.67): Low (0.242) C322 -Spare part availability (0.33): Medium (0.500)
C4 - Failure risk (0.43): $CS_{Vessel\ hull,\ C4}$ = 0.748	C41 - Failure frequency (0.16) Occasional (0.647) C42 - Failure detectability (0.30) High (1.000) C43-Severity of failure consequence (0.54): $CS_{Vessel\ hull,\ C43}$ = 0.638	C431 – Environment (0.44) Infection (0.522) C432 – Safety (0.31): Damage (0.667) C433 – Operational (0.16): High (1.000) C434 – Cost (0.10): Medium (0.416)

TABLE IV. Grades pairwise comparison of environmental sub-criterion (CR = 0.008)

	G1	G2	G3	G4	Weight	Relative weight
G1	1.00	2.00	4.00	5.00	0.51	1.00
G2	0.50	1.00	2.00	3.00	0.26	0.51
G3	0.25	0.50	1.00	2.00	0.14	0.27
G4	0.20	0.33	0.50	1.00	0.09	0.18

B. Grades preparation of sub-criteria

Absolute measurement is employed for prioritize failure modes in reasons of large numbers of failures and the difficulties associated with pairwise comparison as discussed by [9]. To use absolute measure, specific grades for every sub-criterion should be presented, e.g. for sub-criterion C_{431} (Environment) the corresponding grades: G1 (Pollution), G2 (Infection), G3 (Oil-spill), or G4 (No Consequences). To compute the strength of these grades; the pairwise comparison is conducted as presented in table IV. With the same concept, each one of the other sub-criteria is evaluated with a specific set of grades. With the same concept as [9], the grades' strengths are divided by the most important one in order to magnify the importance of a failure mode with the highest grade.

C. Alternatives evaluation

Here, alternatives are failure modes of vessel equipment. These failure modes will be evaluated with respect to each sub-criterion, using absolute measurement to specify the associated score. After computing the grades' strengths of low-level sub-criteria, each failure mode is assigned a subjective strength for the corresponding sub-criteria. After that, the total scores are synthesized for each failure. The synthetization process starts by aggregating the weighted sum of low-level sub-criteria, then, the second upper level, until the first one for the main criteria. For each criteria that has a set of sub-criteria, the weighted sum is computed as in (1).

$$cs_{ik} = \sum_{j=1}^{n} w_j s_{ij} \quad (1)$$

Where, $i=1,\ldots, m$ and m is the total number of failure modes, $j=1,\ldots,n_k$ and n_k is the number of sub-criteria associated to a criterion $k=1,\ldots,K$ and K is the set of criteria that have sub-criteria, w_j is the weight of the j sub-criterion, and s_{ij} is the strength assigned to failure mode i with respect to sub-criterion j. Considering that $\sum_{j=1}^{n} w_j = 1$. After that, the failure total score (FTS) is computed relying on weighted sum of the main criteria. Table III represents the synthetization of the failure score for the vessel hull. The associated grades and strengths of vessel hull is colored in blue, where CS is colored in brawn (FTS =0.25×1.000+0.19×0.937+0.13×0.495+0.43×0.748=0.814).

After the synthetization process, each failure mode has an associated FTS. To find failure criticality index "FCI", a normalization process can be done by mapping FTSs between the interval of maximum permissible FTS (1.00) and the minimum permissible FTS (0.2942). Such that, the maximum FCI can take one and the minimum permissible value takes zero.

Fuel service tank, vessel hull, transmission gearbox, crankshaft, fuel storage tank, lubricating oil cooler, fuel injectors, sea water pipe and high pressure fuel line have the highest FCI respectively (0.775, 0.737, 0.723, 0.710, 0.685, 0.596, 0.564. 0.559 and 0.548). For these equipment, the main presenting to critical score came from their high score for 'failure risk', 'equipment function' and 'obsolescence' in the prioritization hierarchy structure. Air pipes and valves have the low FCI.

D. Assignment of maintenance strategies

There are three main maintenance strategies that can be considered for vessels equipments: preventive maintenance with replacement, condition based maintenance, and corrective maintenance. Relying on the obtained FCI the equipment failure modes are clustered into three groups: highly critical, moderate critical and non-critical. Failures with high criticality are assigned to preventive maintenance and those with non-critical are assigned to corrective maintenance. Regarding the failures with moderate criticality: If the failure is detectable, it will be

assigned to condition based maintenance otherwise it is assigned to preventive maintenance with replacement.

VI. Conclusions

This paper presents a MCDM model for prioritization of failure modes of fire and rescue vessels equipments for the assignment of appropriate maintenance strategies relying on the criticality assessment. The model is applied on the fire and rescue vessels of the Kuwait Fire Service Directorate. The model utilizes AHP with relative and absolute measurements. The suggested hierarchy structure includes four main criteria to assess criticality: 'Equipment function', 'Obsolescence', 'Service experience' and 'Failure Risk'. For each criterion, the associated set of sub-criteria is presented and discussed. Relying on the criticality factor obtained; failure modes of high criticality are assigned to preventive maintenance, failure mode with moderate criticality are assigned to condition based maintenance if they are detectable other than they assigned to preventive maintenance, and uncritical failure modes are assigned to corrective maintenance. As perspectives the detailed maintenance plan for each equipment will be developed.

References

[1] N. Altay and W. G. Green III, 'OR/MS research in disaster operations management', *Eur. J. Oper. Res.*,175(1),475–493, 2006.

[2] R. D. Peltz, 'The Titanic's Legacy: The History and Legal Developments Following the World's Most Famous Maritime Disaster', *Univ. San Franc. Marit. Law J.*, vol. 12, p. 45, 1999.

[3] T. Solheim, M. Lorentsen, P. K. Sundnes, G. Bang, and L. Bremnes, 'The "Scandinavian Star" ferry disaster 1990 -a challenge to forensic odontology', *Int. J. Legal Med.*, 104(6),339–345,1992.

[4] S.-A. Christianson and E. Engelberg, 'Memory and Emotional Consistency: The MS Estonia Ferry Disaster', *Memory*,7(4), 471–482, 1999.

[5] IMO, SAR Convention, International Maritime Organization (IMO), ISBN 9789280142280', 2006.

[6] T. D.Matteson, 'Airline experience with reliability-centered maintenance', *Nucl. Eng. Des*, 89,385-390,1985.

[7] M. Rausand and J. Vatn, 'Reliability CentredMaintenance', in *Complex System Maintenance Handbook*, Springer London, pp. 79–108, 2008.

[8] M. P. Moreira, C. J. Dupont, and M. M. B. R. Vellasco, 'PROMETHEE and Fuzzy PROMETHEE Multicriteria Methods for Ranking Equipment Failure Modes', in *15thInter.Conference on Intelligent System Applications to Power Systems*, pp. 1–6, 2009.

[9] S. Taghipour, D. Banjevic, and A. K. S. Jardine, 'Prioritization of medical equipment for maintenance decisions', *J. Oper. Res. Soc.*, 62(9), 1666–1687, 2011.

[10] N. Saleh, A. A. Sharawi, M. A. Elwahed, A. Petti, D. Puppato, and G. Balestra, 'Preventive maintenance prioritization index of medical equipment using quality function deployment', *IEEE J. Biomed. Health Inform.*, 19 (3), 1029–1035, 2015.

[11] Sweis G, Sweis R, AbuHusseint R, Hiyassat M, and Suifan T., 'Priority Setting for Healthcare Facilities Maintenance', *Life Sci. J.*,11(2),54–64, 2014.

[12] P. Babashamsi, A. Golzadfar, N. I. M. Yusoff, H. Ceylan, and N. G. M. Nor, 'Integrated fuzzy analytic hierarchy process and VIKOR method in the prioritization of pavement maintenance activities', *Int. J. Pavement Res. Technol.*,9(2),112–120,2016.

[13] Q. Shen, K.-K. Lo, and Q. Wang, 'Priority setting in maintenance management: a modified multi-attribute approach using analytic hierarchy process', *Constr. Manag. Econ.*,16(6),693–702,1998.

[14] M. Andrusca, M. Adam, F. D. Irimia, and A. Baraboi, 'Prioritization of maintenance activities from an electricity company', in*13thInter. Conference on Optimization of Electrical and Electronic Equipment*, pp. 1–7, 2012.

[15] I. Lazakis, O. Turan, S. Alkaner, and A. Olcer, 'Effective ship maintenance strategy using a risk and criticality based approach', in *13thInter. Congress of the International Maritime Association of the Mediterranean*, 2009.

[16] J. R. Bradley and H. H. Guerrero, 'An Alternative FMEA Method for Simple and Accurate Ranking of Failure Modes', *Decis. Sci.*, 42(3),743–771, 2011.

[17] J. B. Bowles, 'An assessment of RPN prioritization in a failure modes effects and criticality analysis', in *Annual Reliability and Maintainability Symposium, 2003.*, pp. 380–386, 2003.

[18] A. Khanlari, K. Mohammadi, and B. Sohrabi, 'Prioritizing equipments for preventive maintenance (PM) activities using fuzzy rules', *Comp. Ind. Eng.*, 54(2), 169–184, 2008.

[19] E. Vijay Kumar and S. k. Chaturvedi, 'Prioritization of maintenance tasks on industrial equipment for reliability: A fuzzy approach', *Int. J. Qual. Reliab. Manag.*, 28(1),109–126, 2011.

[20] Z. Yang, Q. Chang, D. Djurdjanovic, J. Ni, and J. Lee, 'Maintenance priority assignment utilizing on-line production information', *J. Manuf. Sci. Eng.*,129(2), 435–446, 2006.

[21] G. R. Andersen, L. E. Chouinard, C. Bouvier, and W. E. Back, 'Ranking Procedure on Maintenance Tasks for Monitoring of Embankment Dams', *J. Geotech.* Geoenvironmental *Eng.*,125 (4), 247–259, 1999.

[22] R. Dekker, 'Integrating optimisation, priority setting, planning and combining of maintenance activities', *Eur. J. Oper. Res.*, vol. 82, no. 2, pp. 225–240, Apr. 1995.

[23] B. Manganelli, 'The Real Estate Investment', in *Real Estate* Investing, Springer, 2015, pp. 61–70.

[24] J. Wiegmann, A. Sundararajan, and Z. Tao, *Decision Making for Outsourcing and Privatization of Vehicle and* Equipment *Fleet Maintenance*. Transportation Research Board, 2011.

[25] O. Dinu and A. M. Ilie, 'Maritime vessel obsolescence, life cycle cost and design service life', *IOP Conf. Ser.* Mater. *Sci. Eng.*, vol. 95, no. 1, p. 012067, 2015.

[26] G. Ford, C. McMahon, and C. Rowley, 'An Examination of Significant Issues in Naval Maintenance', *Procedia CIRP*, vol. 38, pp. 197–203, Jan. 2015.

[27] B. Bartels, U. Ermel, P. Sandborn, and M. G. Pecht, *Strategies to the Prediction, Mitigation and Management of Product Obsolescence*. John Wiley & Sons, 2012.

[28] J. Moubray, *Reliability Centered Maintenance*. Industrial Press, 1997.

[29] C. Y. Hu, S. W. Kou, and J. J. Guo, 'The Application Research of Ship Metal Corrosion in Marine Environment', *Adv.* Mater. *Res.*,936, 1091–1096, 2014.

[30] M.Shama,'Basic concept of the factor of safety in marine structures', *Ships Offshore Struct.*,4(4),307–314, 2009.

[31] E. Triantaphyllou, *Multi-criteria Decision Making* Methods*: A Comparative Study*. Springer Science & Business Media, 2013.

[32] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision* Making*: Methods and Applications A State-of-the-Art Survey*. Springer Science & Business Media, 2012.

[33] R. W. Saaty, 'The analytic hierarchy process—what it is and how it is used', *Math. Model*, 9(3), 161–176, 1987.

Prediction of complex failures' effects in offshore wind turbines using fuzzy cognitive maps

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Afshin Jamshidi, Daoud Ait-Kadi
Department of Mechanical Engineering
Laval University
Quebec, Canada
afshin.jamshidi.1@ulaval.ca

Amar Ramudhin
Logistics and Humber Development Institute
University of Hull Business School
Hull, England
a.ramudhin@hull.ac.uk

Abstract— Modern offshore wind turbines (OWTs) are subjected to different sort of failures due to the harsh environment and also complex interactions among components. Downtime loss associated with these failures will make huge impacts on the cost of energy. Therefore, the rapid development of wind power industry requires an advanced failure analysis tool which is able to consider all cause-effect interactions among complex failure modes in order to ensure high levels of reliability for OWTs' components. This paper proposes the application of fuzzy cognitive maps (FCMs) as an advanced decision support tool for modeling the behavior of complex failures in OWTs. The proposed tool will allow asset owners to understand how any change in cause of a failure could affect other failures/causes and also on system performance, thereby reducing the cost of unscheduled maintenance interventions by better maintenance decisions makings.

Keywords— Offshore wind turbines; Maintenance; Fuzzy cognitive maps, Failure analysis, Failure mode and effect analysis

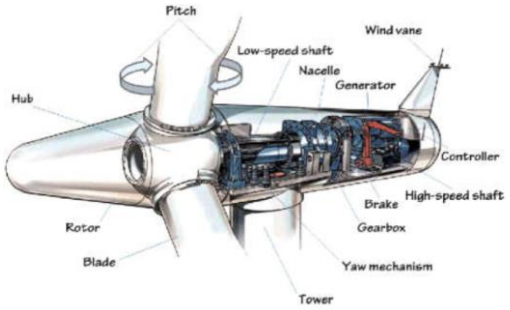
I. Introduction

The expansion of the offshore wind (OW) energy sector has gained increasing attention in recent years due to the surge of renewable energy demand and also high energy potential associated with the vast offshore areas. In these areas, winds are typically stronger and more stable, resulting in significantly higher energy production [1]. OW energy production faces a variety of challenges in design, development, manufacturing, installation, and maintenance and operation. In the case of maintenance, offshore repairs are five to ten times more expensive than onshore repairs because the former are located farther out into the sea where waters are quite deep [2]. Offshore repairs not only need expensive crane vessels, but they are also dependent on suitable weather conditions, which extends the waiting periods for repairs [3]. On average, an OWT shows 2.2 major and several small failures each year [4] but the limited ability to predict the nature of failures and their root causes result in non-optimal maintenances. These non-optimal random maintenances cost 66% of the total maintenance cost incurred during the lifetime for OWT [5]. A technique that can predict failures, identify their nature and root causes and, recognize the effected components, will invariably cut down maintenance costs. However, such a technique is complex to build due to the extent of root causes and failure correlations that exists for OWT.

The offshore wind turbines (OWT) operate in stochastic offshore environments where accessibility is limited, operating costs are higher than onshore locations and prediction of failures/faults in OWT components is difficult to generalize [1]. Therefore, OWT maintenance is largely based on personal judgements/experience. This results in either a Failure Based Maintenance (FBM) and/or Time Based Maintenance (TBM) schemes which is very costly. However, experiences from several industries have shown that FBM and TBM are costlier to plan and implement than Condition Based Maintenance (CBM) [6].

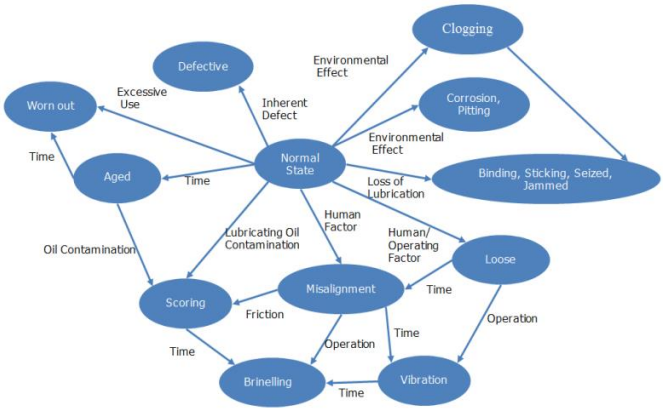
The limited ability to predict the impact of failures on system performance (SP) result in over and/or under maintenance activities that influences inventory management related costs [5]. One possible solution to the above issue is to identify generic components in OWT (Fig.1), determine their failures and causes (Fig. 2) and predict the impact of each failure on the system performance. For Example, from Fig. 2 it is clear that if human factors caused misalignment of the bearing, there will be cases of vibration, brinelling and scoring in time for which maintenance can be pre planned [5]. Such information is critical to plan spares, control occurrence of failure in future and so optimize performance and maintenance. However, current failure analysis tools such as FMEA and even their modified versions are not able to take into account all the direct/indirect interactions among variety of failure modes and causes. Therefore, the aim of this paper is to present the application of Fuzzy Cognitive Maps (FCMs) for failure modeling and predication of OWTs' failures. FCM is an artificial intelligence tool which represents and assess the dynamic behavior of complex systems composed of several cause-effect relationships. FCM represents the behavior of a system in a form that corresponds closely to the way humans perceive it [7]. This tool has been applied widely in variety of domains including business, politics, software engineering, environmental sciences, medical decision system, disease predication, political decision making, decision analysis, process control, and data mining in internet [8, 9, 10, 11, 12, 13, 14]. However, to the best of our knowledge, application of FCM in failure predication of complex OWT systems is not reported in literature. Therefore, in the present paper, we firstly apply FCM in failure predication of OWTs.

Fig. 1 Major wind turbines components.



The reminder of this paper is organized as follows. Section II provides the basic principles of FCM. Sections III and IV present the proposed method with a numerical example related to failure analysis of OWT, respectively. Finally, the conclusions are drawn in Section V.

Fig. 2 Cause-effect interaction of failures in Main Bearing [5].

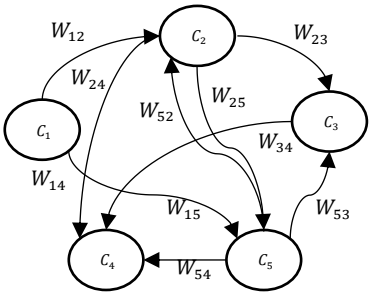


II. The basic theory of FCM

FCM was originally introduced by Kosko [15] as a soft computing technique which is able to take into account the dependencies among the main concepts/nodes and analyse inference patterns [16]. FCMs constitute a modeling methodology that combines fuzzy logic and neural networks and are used to represent both qualitative and quantitative data [17]. FCMs are developed based on the experience and knowledges of experts through an interactive procedure of knowledge acquisition. Various methodologies such as Delphi could be used in order to reach a consensus among the experts in FCM [18].

FCMs are graphs which consist of nodes and weighted arcs between nodes. The following figure illustrates a FCM graph with 5 nodes and 9 arcs. The value of each concept C_i stands in the interval [0, 1], and the weighted arcs among nodes C_i and C_j (W_{ij}) can range in the interval [-1, 1] which represent the influence of each node on the others.

Fig. 3 A simple Fuzzy Cognitive Map.



The values of initial weight matrix (W_{ij}) are suggested by different experts using fuzzy linguistic terms such as Very High (VH), Low (L), etc. in order to determines the dependencies among nodes. Then, the linguistic variables are aggregated and defuzzified to numerical values [19]. When the FCM is initialized, it converges to a steady state through the interaction of equation (1). At each simulation step, the value A_i of the concept C_i is influenced by the values of concepts connected to it and it is updated through the following reasoning process [19]:

$$A_i^{k+1} = f(A_i^{(k)} + \sum_{\substack{j=1 \\ j \neq i}}^{n} W_{ji} A_j^{(k)}) \quad (1)$$

where, W_{ji} shows the initial dependencies weight between concepts C_j and C_i;

A_i^{k+1} is the value of concept C_i at simulation step $k+1$;

$A_j^{(k)}$ is the value of concept C_j at simulation step k;

The initial values of concepts are shown by initial concept vector c as $c = [A_1, \dots, A_j, \dots, A_n]$;

k shows the simulation step;

f is a threshold function, which is used to restrict the concept value into [0,1] range. The most common types of f are: bivalent function (f(x) = 0 or 1), tangent hyperbolic ($f(x) = tanh(x)$), trivalent function ($f(x)$ = -1, 0 or 1), and sigmoid function ($f(x) = 1/(1+e^{-\lambda x})$) [18]. In this study, sigmoid function is adopted.

At each iteration, values of all concepts are recalculated and this process continues until FCM reaches one of the following states [19]:

1) The value of concepts has stabilized at a fixed equilibrium point,
2) A limited state cycle is exhibited, and
3) Chaotic behavior has appeared.

A major deficiency of FCM is its potential convergence to undesired steady states. In order to overcome this drawback, some learning algorithms have been developed such as particle swarm optimization (PSO) [17], Differential Hebbian Learning [11, 14], Simulated Annealing (SA) [20], and etc.

III. The proposed method

In this paper, we propose an advanced approach for dynamic failure analysis of complex OWTs using FCM tool. This tool is able to predict the impact of each cause of failure on other

failures and also SP by developing several what-if analyses and eventually to avoid undesired outcomes. The steps of our proposed model are as follows:

Step 1: Form a group of experts in order to identify potential failure modes and their causes.
Step 2: Depict the FCM for the identified failure modes and their causes and obtain the initial weight matrix ($W^{Initial}$). Experts should first reach consensus on the sign and direction of arcs between nodes. In order to determine the level of influence of each node on the other nodes and vice versa, each expert individually assigns a linguistic term for each arc (W_{ij}) using Table I. In this study, fuzzy triangular numbers parametrized by a triplet (l, m, u) are used in order to consider the uncertainties in experts' opinions. Then, for each arc, the opinions of all expert are aggregated using the average value of assigned linguistic terms in order to obtain the overall linguistic weight. Finally, the overall linguistic weight should be defuzzified in order to find the initial influence weight ($W^{Initial}$). There are different defuzzification methods available in literature [21]. In this paper, we apply the defuzzification method proposed by Çelik & Yamak [22]. According to this method, the defuzzification value t of a triangular fuzzy number (l, m, u) is equal to:

$$t = \frac{l+m+m+u}{4} \tag{2}$$

$$W^{Initial} = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \tag{3}$$

TABLE I. Fuzzy ratings for cause-effect dependencies

Dependencies	Fuzzy rating
Very Low (VL)	(0, 0, 1.5)
Low (L)	(1, 2.5, 4)
Moderate (M)	(3.5, 5, 6.5)
High (H)	(6, 7.5, 9)
Very high (VH)	(8.5, 10, 10)

Step 3: Dynamic analysis of FCM requires the definition of an initial scenario, which represents a proposed initial situation to assess [23]. In this step, several "what-if" scenarios should be defined. In each scenario, a cause of failure/set of causes or a failure mode is activated. The value of activated node/s in the initial concept vector (c) is considered as 1 and this number is 0 for the rest of nodes. In order to achieve precise results, all of the causes and failures should be taken into account and the total effects for each node should be evaluated to determine their influences on the system performance.
Step 4: Calculate the impact of activated nodes by updating the initial concept vector (c). To do so, each initial concept vector is trained through Eq. 1 and using a learning algorithm in order to obtain the steady state vector C^*. The aim of this step is to identify the impact of each node on the system performance. This process is illustrated through a numerical example in the following section.

IV. Numerical Example

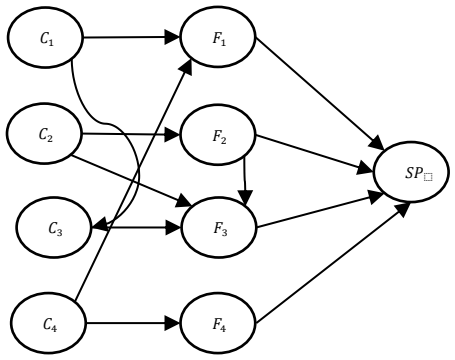
In order to illustrate the proposed tool, consider that 4 failure modes and 4 related cause of failure modes are identified by experts as shown in Table II. In addition, the dependencies among failure modes and their causes are shown in Fig. 3. Besides, one consequence named System Performance is considered as SP in order to show how the proposed tool could consider all the interrelationships among failure modes and their causes on the system performance.

TABLE II. Initial weight matrix

W	F1	F2	F3	F4	C1	C2	C3	C4	SP
F1	0	0	0	0	0	0	0	0	0.6
F2	0	0	1	0	0	0	0	0	0 .84
F3	0	0	0	0	0.8	0	0	0	0.5
F4	0	0	0	0	0	0	0	0	0.2
C1	0.98	0	0	0	0	0	0.77	0	0
C2	0	0.39	0.3	0	0	0	0	0	0
C3	0	0	0.77	0.7	0	0	0	0	0
C4	0.94	0	0	0.85	0	0	0	0	0
SP	0	0	0	0	0	0	0	0	0

The FCM graph in Fig. 2 is depicted based on experts' opinions in order to show the dependencies and feedbacks among factors. In addition, some hypothetical interrelationships among cause of failures/ failures and vice versa are depicted in Fig. 4. In order to make the initial weight matrix (W_{ij}), each expert individually determines the dependencies between concepts (risks), using fuzzy linguistic terms such as Very High (VH), Low (L), etc. Then, the linguistic variables are aggregated and defuzzified to numerical values using Eq. 2 [17]. The defuzzified initial weight matrix is shown in Table 4.

Fig. 4 FCM of the hypothetical numerical example.

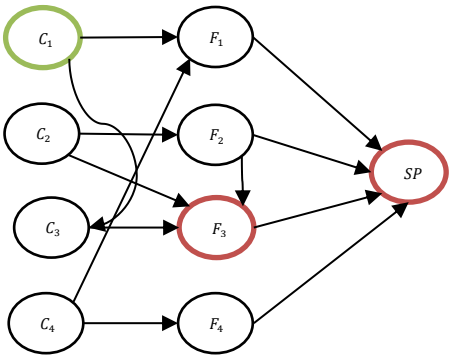


In order to illustrate the failure predication process, in this paper we only assess the impact of activating cause of failure 1 on other failure modes and also SP as following:

F1, F2, F3, F4, C1, C2, C3, C4, SP
c = [0, 0, 0, 0, **1**, 0, 0, 0, 0];

Using W_{ij} matrix, Initial concept vector c, Equation (1) and learning algorithm, the training process starts. In this paper we applied NHL (nonlinear Hebbian learning) algorithm for learning FCM. We imported the data into Matlab code to obtain the updated concept matrix (C^*). The suggested values of learning rate parameter (η) and weight decay learning parameter (γ) have been selected 0.001, 0.97 respectively. The population size is equal to 50. For the algorithm, 100 independent experiments have been performed, to enforce the reliability of the results, and the algorithm was allowed to perform 1000 iterations (generations) per experiment.

Fig.5 Activating C_1 and its impact on F_3 and SP.



C^* = [0.79099, 0.72124, **0.91608**, 0.87501, 0.52362, 0.65905, 0.75142, 0.65905, **0.94277**]

The steady state vector C^* shows that activating cause of failure 1 have a strong influence over the failure F4 and also fairly strong effect over the failure F3. In addition, it has a very strong effect over SP (See Fig. 5). The same procedure should be done for other cause of failures/failures in order to predict their impact on the other failures and system performance. The results reveal that which causes of failure/failures are critical and therefore, experts will be able to manage the failure modes and their causes properly and accurately. In addition to predication and modeling feature of FCM, it has a great potential to be applied for several other problems in complex systems such as risk analysis and management (RAM), big data analytics, data mining, classification, knowledge representation, reasoning, stakeholders' analysis, policy making, decision support, Information Technology, etc. These applications could be interesting future research topics for scholars.

V. Conclusion

In this paper, we proposed an advanced decision support approach for failure predication of modern OWTs systems. Given technology advances, challenging maintenance conditions in OWTs, and significant cost of maintenance activities, this advanced failure predication tool presents a huge potential opportunity for benefit for those organizations such as Siemens who are ready and able to take advantage of this tool. In addition, it could be applied as an advanced failure predication tool for modeling the behavior of failures in other critical industries such as aviation. One of the key features of the proposed tool is that it is able to model all the interactions among failure modes and their causes based on the opinions of experts. In addition, it is able to take into account uncertainties in failure predication process. Moreover, the proposed approach provides valuable information to practitioners for predicting impact of failures on the system performance by developing what-if analyses. In future works, we will evaluate the performance of the proposed tool in a large-scale practical environment and will compare it with other similar tools such as Bayesian Belief Network (BBN). In addition, we will extend the capabilities of the proposed tool for advanced data analytics of OWT databases such as SCADA for failure predication.

Acknowledgment

This research was financed by grants [OPG 0293307 and OPG 0118062] from the Canadian Natural Sciences and Engineering Research Council (NSERC). This support is gratefully acknowledged.

References

[1] V. Baagøe-Engels and J. Stentoft, "Operations and maintenance issues in the offshore wind energy sector; An explorative study," *International Journal of Energy*, vol. 10, no. 2, pp. 245-265, 2016.

[2] M. Shafee, "Maintenance logistics organization for offshore wind energy: current progress and future perspectives," *Renewable Energy*, vol. 77, pp. 182-193, 2015.

[3] S.-P. Breton and G. Moe, "Status, plans and technologies for offshore wind turbines in Europe and North America," *Renewable Energy*, vol. 34, no. 3, pp. 646-654, 2009.

[4] J. Nilsson and L. Bertling, "Maintenance management of wind power systems using condition monitoring systems," *Energy Conversion, IEEE Transactions on*, vol. 22, no. 1, pp. 223-229, 2007.

[5] Y. SINHA, *Optimisation of offshore wind farm maintenance, PhD thesis*, 2016.

[6] M. Ben-Daya, S. O. Duffuaa, A. Raouf, J. Knezevic and D. Ait-Kadi, Handbook of maintenance management and engineering, London: Springer, 2009.

[7] A. Jamshidi, S. Abbasgholizadeh Rahimi, A. Ruiz, D. Ait-Kadi and M. L. Rebaiaia, "Application of FCM for advanced risk assessment of complex and dynamic systems," in *8th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2016*, Troyes, 2016.

[8] J. L. Salmeron and E. I. Papageorgiou, "A Fuzzy Grey Cognitive Maps-based Decision Support System for radiotherapy treatment planning," *Knowledge-Based Systems*, p. 151–160, 2012.

[9] Z. Xiao, W. Chen and L. Li, "An integrated FCM and fuzzy soft set for supplier selection problem based on risk evaluation," *Applied Mathematical Modelling*, vol. 36, p. 1444–1454, 2013.

[10] N. Douali, H. Csaba, J. De Roo, E. I. Papageorgiou and M.-C. Jaulenta, "Diagnosis Support System based on clinical guidelines: comparison between Case-Based Fuzzy Cognitive Maps and Bayesian Networks," *Computer Methods and Programs in Biomedicine*, p. 133–143, 2014.

[11] A. Azadeh, V. Salehi, M. Arvan and M. Dolatkhah, "Assessment of resilience engineering factors in high-risk environments by fuzzy cognitive maps: A petrochemical plant," *Safety Science*, pp. 99-107, 2014.

[12] A. Núnez-Carrera, G. Espinosa-Paredes and H. Cruz-Esteban, "Failure analysis of the standby liquid control system for a boiling water reactor with fuzzy cognitive maps," *Nuclear Engineering and Design*, p. 4004–4012, 2011.

[13] D. K. Lakovidis and E. Papageorgiou, "Intuitionistic Fuzzy Cognitive Maps for Medical Decision Making," *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE*, pp. 100-107, 2011.

[14] A. Jamshidi, S. Abbasgholizadeh Rahimi, D. Ait-kadi and A. Ruiz, "Dynamic risk modeling and assessing in maintenance outsourcing with FCM," in *Industrial Engineering and Systems Management (IESM), 2015 International Conference on*, Seville, 2015.

[15] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man–Machine Studies*, vol. 24, p. 65–75, 1986.

[16] E. Papageorgiou, C. Stylios and P. Groumpos, "Active Hebbian learning algorithm to train fuzzy cognitive maps," *International Journal of Approximate Reasoning*, vol. 37, pp. 219-249, 2004.

[17] E. I. Papageorgiou, K. E. Parsopoulos, C. S. Stylios, P. P. Groumpos and M. N. Vrahatis, "Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization," *Intelligent Information Systems*, vol. 25, no. 1, pp. 95-121, 2005.

[18] M. Glykas, Fuzzy Cognitive Maps: Advances in Theory, Methodologies, Tools and Applications, Springer, 2010.

[19] E. I. Papageorgiou, Fuzzy Cognitive Maps for Applied Sciences and Engineering (From Fundamentals to Extensions and Learning Algorithms), Springer, 2014.

[20] M. G. Somayeh Alizadeh, "Learning FCM by chaotic simulated annealing," *Chaos, Solitons and Fractals*, vol. 41, pp. 1182-1190, 2009.

[21] A. Talon and C. Curt, "Selection of appropriate defuzzification methods: Application to the assessment of dam performance," *Expert Systems With Applications*, p. 160–174, 2017.

[22] Y. Çelik and S. Yamak, "Fuzzy soft set theory applied to medical diagnosis using fuzzy arithmetic operations," *Journal of inequalities and applications*, pp. 1-9, 2013.

[23] C. Lopez and J. L. Salmeron, "Dynamic risks modelling in ERP maintenance projects with FCM," *Information Sciences*, vol. 256, pp. 25-45, 2012.

[24] S. Jimmy Gandhi, A. Gorod and B. Sauser, "Prioritization of outsourcing risks from a systemic perspective," *Strategic Outsourcing*, vol. 5, no. 1, pp. 39-71, 2012.

[25] J. Baitheiemy, "The seven deadly sins of outsourcing," *Academy o/ Management Executive*, vol. 17, no. 2, 2003.

[26] J. V. J. H. Z. Hameed, "Challenges in the reliability and maintainability data collection for offshore wind turbines," *Renewable Energy*, vol. 36, pp. 2154-2165, 2011.

[27] J. Andrawus, *Maintenance optimisation for wind turbines (Doctoral dissertation)*, The Robert Gordon University, 2008.

Optimizing the integrated economic production quantity for a stochastically deteriorating production system under condition-based maintenance

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

A. Khatab*, C. Diallo†, E.-H. Aghezzaf‡and U. Venkatadri§
*LGIPM. Lorraine University, National school of Engineering. Metz, France.
E-mail: abdelhakim.khatab@univ-lorraine.fr
†Department of Industrial Engineering, Dalhousie University, Halifax, Canada.
Email: Claver.Diallo@Dal.ca
‡Department of Industrial Systems Engineering and Product Design.
Faculty of Engineering and Architecture, Ghent University, Belgium.
E-mail: ElHoussaine.Aghezzaf@UGent.be
§Department of Industrial Engineering, Dalhousie University, Halifax, Canada.
Email: Uday.Venkatadri@Dal.ca

***Abstract*—This paper proposes a new integrated economic production quantity (EPQ) and condition-based maintenance (CBM) model for a stochastically deteriorating production system. Inspections are performed periodically to measure the real time degradation. The system fails (out-of-control) whenever its degradation is beyond a critical threshold level. In the out-of-control state, a proportion of nonconforming items are produced. To assess the degradation of the system and to increase the production of conforming items, preventive maintenance (PM) actions are carried out. An integrated EPQ and CBM optimization model that minimizes the total expected cost rate over an infinite time horizon is developed. The objective is to determine a joint optimal EPQ and PM strategy minimizing the sum of inspection/maintenance and setup costs, cost of nonconforming items in addition to inventory holding cost. Numerical experiments are provided to illustrate the proposed approach.**

Production lot-sizing; EPQ; Condition-based maintenance; Stochastic process; Optimization.

I. Introduction

In the production and inventory management setting, the economic production quantity (EPQ) model has been extensively investigated and extended under relaxation of various assumptions initially made in the basic EPQ model [1]. Many research investigations have been made to integrate production quantity, maintenance and quality issues in a single model where their interrelations are explicitly accounted for.

Ben-Daya and Makhdoum [2] investigated the effects of various PM policies on the joint optimization of the EPQ and the economic design control charts. In [3], the author developed an integrated model for the joint determination of EPQ, quality and PM level for a process with a general deterioration distribution and increasing failure rate. The durations of inspection intervals are chosen to make the integrated hazard rate function equal over each inspection interval [3]. The reader is referred to [4] for further details on the EPQ approaches published before 2001. Wang [5] proposed an integrated EPQ model with rework activity in addition to imperfect preventive maintenance including minimal repair. Chen [6] investigated an integrated EPQ model with inspection, rework and preventive maintenance with error. The correct execution of preventive maintenance reduces the system failure rate, whereas a preventive maintenance error shifts the system to an out-of-control state with a certain probability. Chen and Lo [7], and Wang [8], and more recently, Lia [9] studied an integrated EPQ model for an imperfect production system producing items that are sold with a warranty. An age based preventive maintenance has also been used along with the EPQ model to jointly determine the production lot size and PM schedule for randomly deteriorating production system producing both conforming and nonconforming items (see [10], [11] and the references therein).

From this literature review, one can conclude that the maintenance policies integrated into the EPQ model are based on the age of the production system and the statistical information from the system's lifetimes. As a result, the changes in the system reliability caused by how it is being used is not accounted for. As pointed out in [12], the main drawback of the lifetimes distributions is that only the aging process is accounted for when evaluating whether a system is functioning or not. However, many real life production systems suffer damage and deteriorate with both age and usage. There is therefore a need to develop new integrated EPQ models where the production system failures are explicitly related to both age and usage. In the literature, deterioration process of a system is generally modeled as a time-dependent stochastic process. For example, the random deterioration rate, Markov, Wiener, Gamma and Inverse Gaussian processes

are commonly used approaches that are particularly good for their mathematical properties and clear physical interpretations [13], [14], [15]. When the deterioration of the production system reaches a failure threshold, maintenance actions should be performed on the system. If the system's degradation is appropriately modeled and measurable, maintenance actions can be carried out on the basis of the observed degradation data before the system enters the out-of-control state. This paper aims to develop a cost-effective joint EPQ and degradation-based model.

The remainder of the article is organized as follows. Section II describes the production system and presents the scope of the problem under consideration. Section III presents the development of the mathematical formulation of the problem. A solution procedure is proposed in Section IV along with several numerical experiments and the discussion of their results. Conclusions are drawn and future extensions discussed in the last section.

II. System description and problem definition

Our paper develops a novel joint EPQ and degradation-based maintenance model and investigates its contribution to the reduction of non-conforming items produced. The production system produces a single product type at any given time to meet a constant and continuous customer demand rate d. The system is subjected to stochastic degradation which is defined by a measurable scalar time-dependent random variable $X(t)$ which can take linear or nonlinear forms. A dormant failure occurs whenever the accumulated degradation reaches a critical threshold X_f, which can be specified according to either economical or safety reasons. The dormant failure of the system is revealed only through periodic inspection. The system is preventively maintained (PM) when the degradation level reaches a predetermined threshold X_p. The dormant failure causes the production system to go out-of control with a fraction non-conforming rate α. These non-conforming items add non-quality costs to the producer due to additional expenses for rework or economic losses for poor service. An inventory of good products is built up when the production system is in-control. When the degradation level is found to have exceeded X_p or X_f the production is stopped for the preventive and corrective maintenance operations to be carried out. Demand is then satisfied from the built-up inventory. Production resumes as soon as the inventory is completely depleted (Figure 1).

Our objective is to determine an optimal joint EPQ and PM strategy to minimize the sum of inspection, maintenance, setup, inventory, non-quality costs. An evaluation method is provided to optimize the expected cost rate function.

The following assumptions are considered in this paper:

1) The system deteriorates with a monotonically increasing stochastic degradation process. The damages at each time increment are cumulative. Hence, the degradation is increasing meaning that the system worsens with time due to aging and accumulated wear or damage.
2) The system is assumed to sojourn either in *out-of-control* or in *in-control* states.
3) In the in-control state, the system is producing only conforming items, while in the out-of-control state, the production system has a fraction non-conforming rate α.
4) Shortages are not allowed.
5) Failures are revealed only through error-free inspections (non-self announcing failures).
6) Inspections are carried out periodically until one or both of the thresholds are crossed. The inspection periodicity is a decision variable. Inspection duration is negligible.
7) PM is performed at a cost C_p when the degradation is beyond the threshold X_p, which is a decision variable.
8) Corrective maintenance (CM) is performed at a cost C_c when the degradation is beyond the threshold X_f. Furthermore, $C_c >> C_p$.
9) After either a PM or CM, the system becomes "*as good as new*" and is restored to the *in-control* state. Maintenance durations are negligible.
10) Ample inspection and maintenance resources are always available.

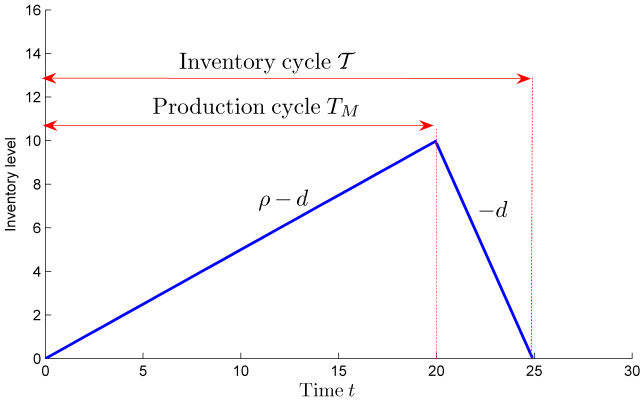


Fig. 1. Inventory cycle of the EPQ model with inventory build-up and depletion.

The continuous degradation of the production system is assumed to follow a stationary Gamma process. The Gamma process is appropriate for characterizing monotonically accumulating gradual damage over time [16], [14]. It has therefore been extensively used in condition-based maintenance optimization problems [17], [18]. An excellent survey dealing with the application of the Gamma process in maintenance modeling and optimization can be found in the seminal paper by van Noortwijk [14]. The proposed approach is general enough to accommodate any other degradation process.

The Gamma degradation process of the production sys-

tem is a time-dependent stochastic process $\{X(t) : t \geq 0\}$ with the following characteristics:

1) $Pr\{X(0) = 0\} = 1$,

2) $X(t)$ has independent increments,

3) For all $0 \leq s < t$, the random variable $\Delta X(s,t) = X(t) - X(s)$ follows a Gamma distribution whose pdf $f(s,t,x)$ and cdf $F(s,t,x)$ are defined for all $x \geq 0$ as:

$$f(s,t,x) = \frac{x^{[(t-s)\gamma-1]}}{\Gamma\left[(t-s)\gamma\right]\eta^{(t-s)\gamma}} \exp\left(\frac{-x}{\eta}\right), \quad (1)$$

and

$$\begin{aligned} F(s,t,x) &= \Pr\left(\Delta X(s,t)\right) < x) \quad (2) \\ &= \int_0^x f(s,t,y)\mathrm{d}y, \\ F(s,t,x) &= \frac{\Gamma\left[(t-s)\gamma, \left(\frac{x}{\eta}\right)\right]}{\Gamma\left[(t-s)\gamma\right]}, \end{aligned}$$

where $(t-s)\gamma$ and η are the shape and scale parameters, respectively. The function $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\mathrm{d}u$ is the Gamma and $\Gamma(\alpha, x) = \int_0^x u^{\alpha-1}\exp(-u)\mathrm{d}u$ is the lower incomplete Gamma function defined for $\alpha > 0$ and $x \geq 0$. From time 0 up to time t, the expected degradation is $\mathbb{E}[X(t)] = \eta\gamma t$ and its variance is $\mathbb{V}ar[X(t)] = \eta^2\gamma t$.

The lifetimes of the system are represented by the random variable $T_f = \inf\{t : X(t) \geq X_f\}$ given as the first passage time when the degradation exceeds the failure threshold X_f. Its cdf $G_f(t) = \Pr\{T_f \leq t\}$ is computed as:

$$\begin{aligned} G(t) &= \Pr\{X(t) > X_f\} \\ &= 1 - F(0,t,X_f) \\ G(t) &= \overline{F}(0,t,X_f), \end{aligned} \quad (3)$$

where $\overline{F}(0,\tau,X_p - x) = 1 - F(0,\tau,X_p - x)$. The pdf $g(t) = \frac{\partial G(t)}{\partial t}$ corresponding to the system' lifetimes T_f is given as:

$$g(t) = \frac{\gamma}{\Gamma(\gamma t)} \int_{\frac{X_f}{\eta}}^{\infty} \left[\ln(u) - \Psi(\gamma t)\right] u^{\gamma t - 1} \exp(-u)\mathrm{d}u, \quad (4)$$

where $\Psi(u)$ is the digamma function defined as the logarithmic derivative of the Gamma function:

$$\Psi(u) = \frac{\mathrm{d}\ln\left(\Gamma(u)\right)}{\mathrm{d}u}.$$

The production process starts initially with a new production system. The system degrades while producing items. When the system degradation is lower that the failure threshold X_f, the system is said to be in an *in-control* state, it is in an out-of-control state whenever its corresponding degradation exceeds the failure threshold X_f. The sojourn time in the in-control state is modeled by the random variable T_f whose cdf is given by Equation (3). While in its in-control state, items produced satisfy the quality requirements. However, in the out-of-control state, the system continues to produce but only a percentage

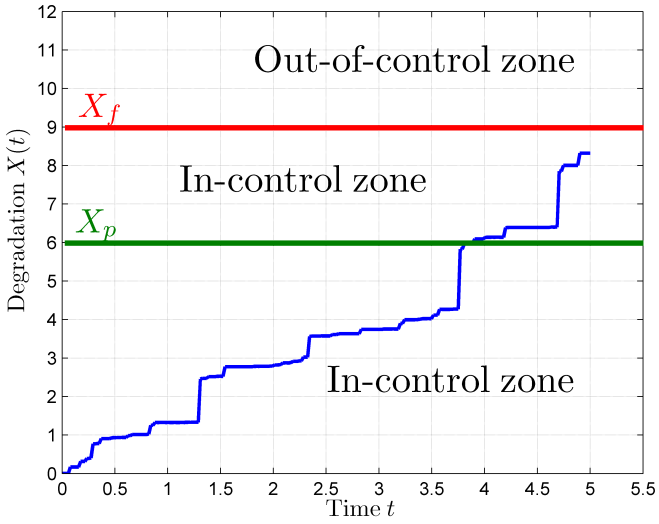


Fig. 2. Production system's degradation path and its operating zones.

$(1-\alpha)$ of the products are acceptable. An example of a possible degradation path of the production system is shown in Figure (2) where the degradation is governed by a Gamma process with a shape parameter $\gamma = 2.5$ and a scale parameter $\eta = 0.5$. The figure shows also the system's failure threshold X_f separating the two possible zones where the system is either in the in-control or in the out-of-control state.

To assess the production system's degradation and to improve production quality, the production system is subjected to an inspection and maintenance strategy described as follows. The system is periodically inspected at times $k\tau$ $(k = 1, 2, \ldots)$ where τ is the inspection interval which is a decision variable. Failures of the system are revealed only through inspections (i.e. failures are not self-announcing). During an inspection, the magnitude of the system's degradation is measured at cost C_I. If the degradation level measured is found to be less than X_p, no maintenance action is performed and the system's condition remains as it was just before the inspection. If the degradation value exceeds the failure threshold X_f, a corrective maintenance (CM) is carried out at cost C_c. If the degradation level is higher than a given degradation level X_p (see Figure 2), then a preventive maintenance (PM) action is performed. The degradation level X_p is a decision variable and $X_p \leq X_f$. It should be noted that the choice of the PM threshold X_p greatly impacts the performance of the production system. Indeed, if X_p is chosen to be close to the failure threshold X_f, then the probability of the system shifting into the out -of-control state increases significantly. Conversely, for low values of X_p, the probability of failure is reduced and the residual life of the production system is increased. Low values of X_p may reduce the risk of producing bad items but they increase the maintenance cost by generating more PM actions. There is therefore a need to find the optimal trade-

offs between the production, inspection and maintenance costs through the optimization of the economic production quantity Q, the threshold X_p and the inspection interval τ.

III. The integrated EPQ optimization model

The objective of the optimization model is to find the optimal values of the decision variables, namely the inspection interval τ and the preventive maintenance threshold X_p, which minimize the expected total cost per unit of time $C(\tau, X_p)$ over an infinite time horizon. We have a regenerative process starting and ending at the instants of complete depletion of the inventory, which follows either a preventive or a corrective replacement (see Figure 1). It follows from the theory of renewal reward processes that the long-run expected total cost per unit of time $C(\tau, X_p)$ is the average total cost $\mathbb{E}[\mathcal{C}]$ in a cycle divided by the average length $\mathbb{E}[\mathcal{T}]$ of that cycle:

$$C(\tau, X_p) = \frac{\mathbb{E}[\mathcal{C}]}{\mathbb{E}[\mathcal{T}]}. \tag{5}$$

The average total cost $\mathbb{E}(\mathcal{C})$ in a cycle is defined as the sum of the setup cost S_c, the expected inventory holding cost H_c, the expected maintenance cost M_c, and the cost of producing nonconforming items NC_c. In what follows, these costs are fully defined and discussed as well as the average length of the inventory cycle. The expected inventory cycle length $\mathbb{E}[\mathcal{T}]$ is the sum of the expected maintenance (production) cycle and the expected time required for inventory depletion (Figure 1). To compute these costs, let us first evaluate the expected inventory cycle length $\mathbb{E}[\mathcal{T}]$.

A. The expected inventory cycle length

According to Figure (1), the expected inventory cycle length $\mathbb{E}(\mathcal{T})$ is computed as:

$$\mathbb{E}[\mathcal{T}] = \frac{\rho}{d}\mathbb{E}[T_M], \tag{6}$$

where ρ is the production rate, $\mathbb{E}[T_M]$ is the expected maintenance (production run) cycle which ends either by a preventive or corrective replacement. The following lemma computes the value of $\mathbb{E}[T_M]$.

Lemma 1: The expected production cycle $\mathbb{E}[T_M]$ is computed as:

$$\mathbb{E}[T_M] = \sum_{i=1}^{\infty} i\tau \int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_p-x)\mathrm{d}x, \tag{7}$$

Proof: Comes straightforward from the fact that the probability of a maintenance action being performed after the i^{th} inspection is equal to the probability that the accumulated system's degradation in the interval $[0,(i-1)\tau]$ is lower than the threshold X_p and the degradation $X(i\tau)$ measured at time $i\tau$ is greater than or equal to the PM threshold X_p. This probability is computed as:

$$\begin{aligned}&\Pr\{X((i-1)\tau) < X_p; X(i\tau) > X_p\}\\ &= \int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_p-x)\mathrm{d}x.\end{aligned}$$

■

B. The expected inventory holding cost

According to Figure (1) together with the result of Lemma (1), the expected inventory holding cost $\mathbb{E}[H_c]$ is computed as:

$$\mathbb{E}[H_c] = C_h\frac{\rho(\rho-d)(\mathbb{E}[T_M])^2}{2d} \tag{8}$$

where C_h is the unit holding cost per unit of time.

C. The expected inspection and maintenance cost

According to the maintenance policy adopted, the production system is periodically inspected. Whenever either a PM or a CM is performed, the production systems becomes as good as new. The following lemma gives the expected maintenance cost of a maintenance cycle.

Lemma 2: The expected maintenance cost $\mathbb{E}[M_c]$ during the maintenance cycle T_M is:

$$\begin{aligned}\mathbb{E}[M_c] = &\sum_{i=1}^{\infty} iC_I \int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_p-x)\mathrm{d}x+\\ &\sum_{i=1}^{\infty}\int_0^{X_p} f(0,(i-1)\tau,x)\left(C_p\overline{F}(0,\tau,X_p-x)+\right.\\ &\left.(C_c-C_p)\overline{F}(0,\tau,X_f-x)\right)\mathrm{d}x.\end{aligned} \tag{9}$$

Proof: Let us consider two stochastic events E_i^p and E_i^c. The event E_i^p occurs when a PM is to be performed after the i^{th} inspection, while the event E_i^c occurs when the CM is to be carried out following the the i^{th} inspection. The occurrence probability of the event E_i^p is computed as:

$$\Pr\{E_i^p\} = \Pr\{X((i-1)\tau) < X_p; X_p \le X(i\tau) < X_f\}, \tag{10}$$

It follows that the probability to perform a PM after the i^{th} inspection is:

$$\Pr\{E_i^p\} = \int_0^{X_p} f(0,(i-1)\tau,x)\left(\int_{X_p-x}^{X_f-x} f(0,\tau,y)\mathrm{d}y\right)\mathrm{d}x \tag{11}$$

Similarly, the occurrence probability of the event E_i^c is computed as:

$$\Pr\{E_i^c\} = \Pr\{X((i-1)\tau) < X_p; X(i\tau) > X_f\} \tag{12}$$

We then have that the probability to perform a CM after the i^{th} inspection is:

$$\Pr\{E_i^c\} = \int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_f-x)\mathrm{d}x. \tag{13}$$

If the maintenance cycle is equal to $i\tau$, it follows that from Equations (11) and (13), the resulting expected inspection and maintenance cost is:

$$(C_p + iC_I)\Pr\{E_i^p\} + (C_c + iC_I)\Pr\{E_i^c\}.$$

Therefore the expected total maintenance cost can be written as

$$\sum_{i=1}^{\infty}(C_p + iC_I)\Pr\{E_i^p\} + (C_c + iC_I)\Pr\{E_i^c\}.$$

After basic algebraic operations, the result of the lemma is directly obtained. ∎

D. The expected cost of producing nonconforming items

It is assumed that a percentage α of nonconforming items are produced during the period where the production system sojourns in its out-of-control state, i.e. during the time period where the system's degradation is greater than the failure threshold X_f. The expected total cost $\mathbb{E}[NC_c]$ induced by producing such nonconforming items is given by the following lemma.

Lemma 3: The expected total cost $\mathbb{E}[NC_c]$ corresponding to nonconforming items is:

$$\mathbb{E}[NC_c] = C_{nc}\alpha\rho\sum_{i=1}^{\infty}\left(\int_{(i-1)\tau}^{i\tau}(i\tau - t)g(t)\mathrm{d}t\right)\times \left(\int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_f - x)\mathrm{d}x\right), \tag{14}$$

where $g(t)$ is the pdf of the system lifetimes in Eq. (4).

Proof: The proof is obtained by computing the conditional expectation by conditioning on the event E_i^c which represents the case where the production run cycle is ended by a CM.

$$\Pr\{E_i^c\} = \int_0^{X_p} f(0,(i-1)\tau,x)\overline{F}(0,\tau,X_f - x)\mathrm{d}x.$$

We then have:

$$\begin{aligned}\mathbb{E}[NC_c] &= C_{nc}\alpha\rho\sum_{i=1}^{\infty}\mathbb{E}[NC_c|_{E_i^c}]\Pr\{E_i^c\}\\ &= C_{nc}\alpha\rho\sum_{i=1}^{\infty}\left(\int_{(i-1)\tau}^{i\tau}\mathbb{E}[NC_c|_{T_f}]\mathrm{d}G(t)\right)\Pr\{E_i^c\}\\ &= C_{nc}\alpha\rho\sum_{i=1}^{\infty}\left(\int_{(i-1)\tau}^{i\tau}\alpha\rho(i\tau - t)\mathrm{d}G(t)\right)\Pr\{E_i^c\}.\end{aligned} \tag{15}$$

∎

From the above results, the optimization problem considered is to find the decision variables defining the optimal joint values of the inspection time period τ and the PM threshold X_p, which minimize the total expected cost rate:

$$C(\tau, X_p) = \frac{Sc + \mathbb{E}[H_c] + \mathbb{E}[M_c] + \mathbb{E}[NC_c]}{\mathbb{E}[\mathcal{T}]}, \tag{16}$$

where we recall here that Sc, $\mathbb{E}[H_c]$, $\mathbb{E}[M_c]$, $\mathbb{E}[NC_c]$ and $\mathbb{E}[\mathcal{T}]$ are, respectively, the setup cost, the expected holding cost, the expected inspection and maintenance costs, the expected cost of producing nonconforming items, and the expected inventory cycle length. Unfortunately, the optimal solutions that minimize Equation (16) are in general difficult to obtain analytically and proof of global convexity would be intractable. Therefore, a numerical method is needed to solve this optimization problem. A numerical procedure is developed based on the fix-and-optimize method to minimize Equation (16).

IV. Numerical examples

In this section, we investigate the problem of solving the integrated EPQ and maintenance for a production system whose random degradation is governed by a stationary gamma stochastic process $\{X(t) : t \geq 0\}$. The latter is characterized by its scale and shape parameters set, respectively, to $\eta = 0.8$ and $\gamma = 1.15$. The production system fails whenever its degradation reaches the failure threshold $X_f = 4$. The demand and production rates are $d = 50$ and $\rho = 100$. Costs corresponding to setup, unit non-conforming product, PM, CM and inspection are set, respectively, to $S = 150$, $C_{nc} = 400$, $C_p = 60$, $C_c = 100$, $C_I = 0.5$, and the holding cost is set to $C_h = 0.5$. Data used within this experiment are arbitrary and considered for illustration purposes, and they are assumed to be given, if any, in appropriate time and monetary units.

Let us recall that the objective is to determine simultaneously the optimal values of the the decisions variables, namely the inspection period τ and the preventive degradation threshold X_p.

To show how important the value of the preventive threshold X_p is, let us first assume that X_p is an input parameter set by the decision-maker rather than being a decision variable. The objective of the manufacturer is then reduced to finding the optimal value of the inspection period τ. In the case where $X_p = 2.5$, the optimal solution suggests to perform inspections at a period $\tau = 0.5$. This solution induces an expected total cost rate $C(\tau, X_p) = C(1.1, 2.5) = 71.94$. If we consider the extreme case where the PM is ignored ($X_p = X_f$), in this case the optimal inspection period is $\tau = 0.6$ which induces an expected total cost rate evaluated to $C(0.6, 4) = 82.50$. This simple example clearly illustrates the balancing role played by the PM threshold in the joint EPQ optimization problem.

Now, if the integrated EPQ model is solved jointly for both decision variables τ and X_p, the optimal solution suggests to perform inspection at period $\tau = 1.4$ while PM are carried out whenever the degradation threshold reaches the level $X_p = 1.55$ (see Figure 3). This inspection/maintenance policy induces an expected total cost rate of $C(1.4, 1.55) = 70.89$.

V. Conclusion

In this paper, we developed a new integrated model for the joint optimization of EPQ and preventive maintenance for a stocastically deteriorating production system. Unlike the existing approaches, our model uses condition-based

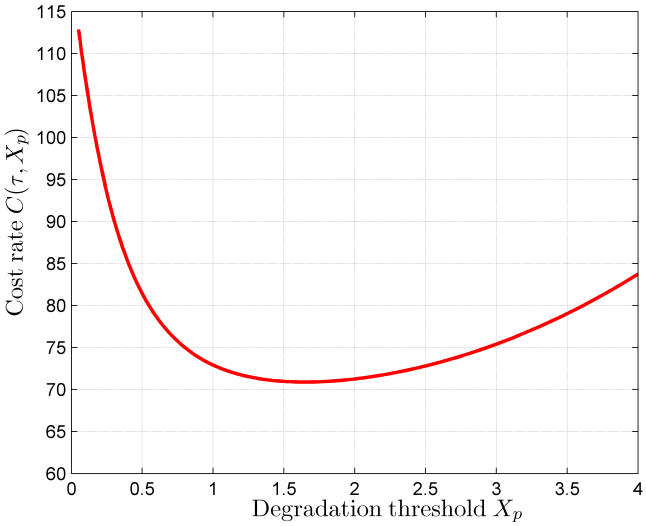


Fig. 3. Total expected cost rate versus the degradation threshold X_p: case of $\tau^* = 1.4$.

maintenance decision making. The degradation of the system is modeled as a stationary Gamma process. However the proposed approach is still general enough to encompass other kind of degradation processes. Inspections are carried out periodically to monitor the system degradation. A PM is performed whenever the degradation exceeds a threshold which is a decision variable. A CM is performed whenever a specified critical threshold is reached. A numerical example was provided to illustrate the proposed approach. Effects of both inspection period and PM degradation threshold is shown to play an important role on the joint EPQ and maintenance decision making.

References

[1] F. Harris, "How many parts to make at once?" *Factory Magazine Management*, vol. 10, no. 2, pp. 135–136, 1913.

[2] M. Ben-Daya and M. Makhdoum, "Integrated production and quality model under various preventive maintenance policies," *Journal of the Operational Research Society*, vol. 49, pp. 840–853, 1998.

[3] M. Ben-Daya, "The economic production lot-sizing problem with imperfect production process and imperfect maintenance," *International Journal of Production Economics*, vol. 76, pp. 257–264, 2002.

[4] M. Ben-Daya and A. Rahim, *Integrated production, quality & maintenance models: An overview.* in: M.A. Rahim, M. Ben-Daya (Eds.), Integrated Modeling in Production Planning, Inventory, Quality and Maintenance, Kluwer Academic Publishers, 2001.

[5] W.-Y. Wang, "Optimum production and inspection modeling with minimal repair and rework considerations," *Applied Mathematical Modelling*, vol. 37, pp. 1618–1626, 2013.

[6] Y.-C. Chen, "An optimal production and inspection strategy with preventive maintenance error and rework," *Journal of Manufacturing Systems*, vol. 32, pp. 99–106, 2013.

[7] C.-K. Chen and C.-C. Lo, "Optimal production run length for products sold with warranty in an imperfect production system with allowable shortages," *Mathematical and Computer Modelling*, vol. 44, no. 3-4, pp. 319–331, 2006.

[8] C.-H. Wang, "The impact of a free-repair warranty policy on emq model for imperfect production systems," *Computers & Operations Research*, vol. 31, no. 12, pp. 2021–2035, 2004.

[9] G. L. Liao, "Production and maintenance policies for an epq model with perfect repair, rework, free-repair warranty, and preventive maintenance," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 8, pp. 1129–1139, 2016.

[10] A. Chelbi, N. Rezg, and M. Radhoui, "Simultaneous determination of production lot size and preventive maintenance schedule for unreliable production system," *Journal of Quality in Maintenance Engineering*, vol. 14, no. 2, pp. 161–176, 2008.

[11] K. Dhouib, A. Gharbi, and M. B. Aziza, "Joint optimal production control/preventive maintenance policy for imperfect process manufacturing cell," *International Journal of Production Economics*, vol. 137, no. 1, pp. 126 – 136, 2012.

[12] N. D. Singpurwalla, "Survival in dynamic environments," *Statistical Science*, vol. 10, no. 1, pp. 86–103, 1995.

[13] Z. Ye and M. Xie, "Stochastic modelling and analysis of degradation for highly reliable products," *Applied Stochastic Models in Business and Industry*, vol. 31, no. 1, pp. 16–32, 2015.

[14] J. M. van Noortwijk, "A survey of the application of gamma processes in maintenance," *Reliability Engineering and System Safety*, vol. 94, pp. 2–21, 2009.

[15] M. Zhang, O. Gaudoin, and M. Xie, "Degradation-based maintenance decision using stochastic filtering for systems under imperfect maintenance," *European Journal of Operational Research*, vol. 245, pp. 531–541, 2015.

[16] M. Abdel-Hameed, "A gamma wear process," *IEEE Transactions on Reliability*, vol. 24, no. 2, pp. 152–153, 1975.

[17] T. Cheng, M. D. Pandey, and J. A. M. van der Weide, "The probability distribution of maintenance cost of a system affected by the gamma process of degradation: Finite time solution," *Reliability Engineering and System Safety*, vol. 108, pp. 65–76, 2012.

[18] P. Do, A. Voisin, E. Levrat, and B. Iung, "A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions," *Reliability Engineering and System Safety*, vol. 133, pp. 22–32, 2015.

Integrated maintenance/production planning plan for a multiple-product manufacturing system with variable maintenance and changeover costs considerations

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mifdal Lahcen/Majdouline Ilias
Laboratory of sustainable innovation and applied research
International University of Agadir - Universiapolis
Bab Elmadina, quartier tillila, 80000 Agadir, Morocco
mifdal@e-polytechnique.ma

Dellagi Sofiene / Hajej Zied
Laboratory of Industrial Engineering and Mechanical
UFR MIM, Ile du Saulcy
57000 Metz, France

Abstract—This paper deals with the problem of integrated maintenance/production strategy for multiple-product manufacturing system. We consider a manufacturing system consisting of one machine which ensures the production of several types of products in order to satisfy randomly demands corresponding to every type of product. At any given time, the machine is able to produce one type of product, and then switches to another. From the point of view of reliability, the manufacturing system is subject to random failures. The goal of this study is to establish an economical production planning followed by an optimal maintenance strategy, taking into account the influence of production rate on the system degradation. Compared to previous works dealing with the same subject, the first contribution of this study consists of the considering a variable setup costs according to the type of product to switch. More than that, we adopt an increased preventive maintenance cost according the system degradation. Analytical models are developed in order to minimize sequentially the production/holding cost and the total maintenance cost. Finally, a numerical example is presented to illustrate the usefulness of the proposed approach.

Keywords—Maintenance/production policy; multiple-product; optimization; setup cost.

I. Introduction and literature review

The Manufacturing companies have to cope successfully with several functional capacities, such as production, and maintenance. One of the keys to accomplishment consists of treating simultaneously all services. We note that the satisfaction of the customer is one of the important objectives of a business. Therefore, it becomes compulsory to develop a new integrated maintenance policies relating to production, which reduce the total costs integrating maintenance and production. This objective is needed in the case of multiple product manufacturing systems.

Establishing an optimal production planning and maintenance strategy has always been the great challenge for industrial companies. Moreover, during the last few decades, the integration of production and maintenance policies problem has attracted much research attention. In fact, integrated maintenance-production strategies which take into consideration subcontracting have been studied by [1]. They developed and optimized a maintenance policy incorporating subcontractor constraints. They demonstrated through a case study, the influence of the subcontractor constraints on the optimal integrated maintenance-production strategy.

An analytical model and a numerical procedure which allow to determine a joint optimal inventory control and an age based on preventive maintenance policy for a randomly failing production system was presented by [2].

The present study examined both of the problems of the optimal production planning formulation and the optimal maintenance strategy of a manufacturing system. The system considered is composed of a single machine which produces several products in order to meet corresponding several random demands. The problem is presented in (Fig. 1).

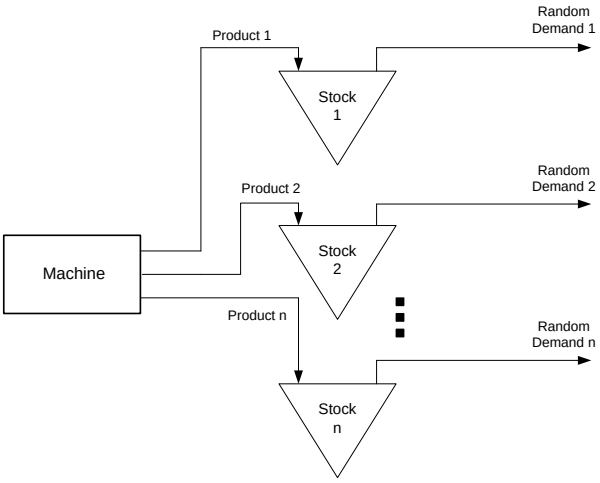


Fig. 1. Problem description

This type of problem, concerning multiple products or multiple machines, was treated by [3]. They presented an analysis of production control and corrective maintenance problem in a multiple-machine, multiple-product manufacturing system. They realized a near optimal control policy of the system through numerical techniques by controlling both production and repair rates. Sloan *el al.* [4] treated a Markov decision process model that simultaneously established maintenance and production schedules for a multiple-product, single-machine production system. They considered that the equipment condition can affect the yield of different product types differently.

In other hand, the literature on integrated maintenance policies, we noticed that the influence of the production rate on the system degradation over a finite planning horizon was rarely addressed in depth. Hajej et al. [5] took into account the influence of production plan on the equipment degradation, but in the case of a system composed by single machine producing one type of product under randomly failing and satisfying a random demand over a finite horizon. In the same context [6] presented a model, where the failure rate of a machine depends on its age; hence, maintenance policies are machine-age dependent.

Motivated by the work of [5], we treated the production and maintenance problem in another context that we consider a more complex and real industrial system, composed by one machine that produces several products during a finite horizon divided into periods. This study shows that it has a novelty and originality relative to this type of problem which reflects the influence of several products on the degradation state of the considered manufacturing system. More than keys study consists of, firstly, considering a variable setup cost between products; secondly we adopt an increased preventive maintenance cost according to the system degradation.

We note that, our problem can easily adopted to case of multiple machine with multiple product, and the same methodology will adopted considering every machine

This paper is organized as follows: In the next section we specify the targeted contributions of this work. Section 3 presents the notations and assumptions. The production and maintenance models are developed respectively in section 4 and 5. A numerical example is presented in section 6. Finally, the conclusion is included in Section 7.

II. Targeted contribution

The present study examined a problem of the optimal production planning formulation of a manufacturing system consisting of one machine producing several products in order to meet several random demands. Then we find the optimal maintenance strategy according to the optimal production plan established for every product. In fact we take into account the impact of the production rate on the system degradation. We note that we considered a variable setup cost between products; and we adopt an increased preventive maintenance cost according to the system degradation. In fact, the variable setup and maintenance costs, which present the industrial reality, are rarely treated in literature.

III. Notations and assumptions

In this paper, we use the following notations and assumptions:

A. Notations

$Cp(i)$: The unit production cost of product i

$Cs(i)$: The unit holding cost of one unit of product i during Δt

$St(i,l)$: Changeover cost from product i to product l

Mc : Corrective maintenance action cost

Mp : Preventive maintenance action cost

H : Total number of periods

n : Total number of products

p : Total number of sub-periods during each period

Δt : Production period duration

U_{inom} : Nominal production rate of product i during Δt

θ_i : Probabilistic index (related to customer satisfaction) of product i

$\varphi^{-1}(.)$: Inverse distribution function

$\hat{d}_i(k)$: The average demand of product i during period k

$\sigma(.)$: The standard deviation

$\hat{S}_{i,(k\times p)-(p-j)}$: The average stock level of product i at the end of sub-period j of period k

$Z\ (U)$: The total expected cost of production and holding over the finite horizon

$I(N)$: The total cost of maintenance

$\lambda_{(k\times p)-(p-j)}(.)$: Failure rate function at sub-period j of the period k

$\lambda_n(.)$: Nominal failure rate

$\phi(.)$: The average number of failures

N : The number of actions of preventive maintenance

T : The intervention period for preventive maintenance actions

Decision variables:

$U_{i,j,k}$: The production rate of product i during sub-period j of period k

$\delta_{(k\times p)-(p-j)}$: The duration of sub-period j at period k

$y_{i,j,k}$: A binary variable, which is equal to 1 if product i is produced in sub-period j of the period k, and 0 otherwise

N : The number of preventive maintenance actions during the finite horizon

B. Assumptions

To develop the production model, the following assumptions are specifically adopted:

- setup costs, holding and production costs of each product are assumed constant;
- Only a single product can be produced in each sub-period;
- As described in (Fig. 2), the period k is divided in exactly p sub-periods, in order to simplify the mathematical model developed;
- The standard deviation of demand $\sigma(di)$ and the average demand $\hat{d}_i$ for each product and each period k are known and constant.

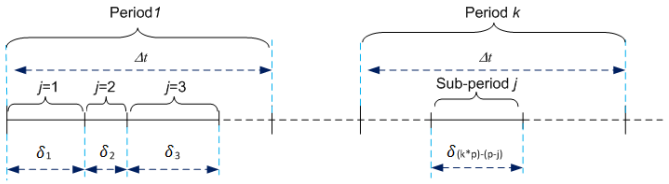


Fig. 2. Split of production horizon

To develop the maintenance model, it was assumed that

- Durations of maintenance actions are negligible;
- Mp and Mc costs incurred by the preventive and corrective maintenance actions are known and constant, with $Mc >> Mp$.
- Preventive maintenance actions are always performed at the end of the sub-periods of production.

IV. Production Policy

The aim of this section is to develop an analytical model that allows us to determine the economical production plan for a finite time horizon $H \times \Delta t$. The decision variables are the production rates $U_{i,j,k}$, the duration of sub-periods $\delta_{(k\times p)-(p-j)}$ and the binary variable $y_{i,j,k}$.

The mathematical formulation of the proposed problem is based on the extension of the model described by [7]. The difference is that we considered that the setup costs are variable.

$$Min\ (\mathrm{Z}(U))$$
$$U = U_{i,j,k}\ \forall \{i = 1 \dots n\}, \{j = 1 \dots p\}, \{k = 1 \dots H\} \quad (1)$$

With:

$$\mathrm{Z}(U) = \sum_{k=1}^{H}\sum_{j=1}^{P}\sum_{i=1}^{n}\left[\begin{array}{c} y_{i,j,k}\times\left(\begin{array}{c} St\left(i, q\left((k\times p)-(p-j)-1\right)\right) \\ +\left(Cp(i)\times\left(U_{i,j,k}\right)^2\right)\end{array}\right) \\ +Cs(i)\times\frac{\delta t_{(k\times p)-(p-j)}}{\Delta t}\times\left[\begin{array}{c}\sigma^2(S_{i,0}) \\ +\left(\sum_{Q=1}^{k-1}\sum_{l=1}^{p} Ent\left(\frac{l}{p}\right)\times\sigma^2(d_i(Q))\right) \\ +\left(\sum_{l=1}^{j} Ent\left(\frac{l}{p}\right)\times\sigma^2(d_i(k))\right) \\ +\left(\hat{S}_{i,(k\times p)-(p-j)}\right)^2\end{array}\right]\end{array}\right] \quad (2)$$

Under the following constraints:

$$\hat{S}_{i,(k\times p)-(p-j)} = \hat{S}_{i,(k\times p)-(p-j)-1} + \left(y_{i,j,k}\times U_{i,j,k}\right) - Int\left[\frac{j}{p}\right]\times\hat{d}_i(k) \quad (3)$$
$$\forall \{i = 1 \dots n\}, \{j = 1 \dots p\}, \{k = 1 \dots H\}$$

This first constraint represents the stock balance equation for each product i, during each sub-period j, of period k.

$$0 \le U_{i,j,k} \le \frac{\delta_{(k\times p)-(p-j)}}{\Delta t}\times U_{inom} \quad (4)$$
$$\forall \{i = 1 \dots n\}, \{j = 1 \dots p\}, \{k = 1 \dots H\}$$

The second constraint is defined by (4). This constraint expresses the upper production rate of the machine for each product i.

$$\sum_{j=1}^{p}\left(y_{i,j,k}\times U_{i,j,k}\right) \ge \sigma^2\left(d_i(k)\right)\times\varphi^{-1}(\theta_i) + \hat{d}_i(k) - S_{i,(k-1)\times p} \quad (5)$$
$$\forall \{i = 1 \dots n\}, \{k = 1 \dots H\}$$

The above constraint refers to the satisfaction level of demand of product i in each period k. This equation is obtained on the basis of the stochastic formula below:

$$Prob\left(S_{i,(k\times p)} \ge 0\right) \ge \theta_i \quad (6)$$
$$\forall\ \{i = 1 \dots n\}, \{k = 1 \dots H\}$$

Where $S_{i,(k\times p)}$ represents the stock level of product i at the end of sub-period j of period k.

$$\sum_{j=1}^{p}\delta_{(k\times p)-(p-j)} = \Delta t \qquad \forall\ \{k = 1 \dots H\} \quad (7)$$

$$0 < \delta t_{(k\times p)-(p-j)} \le \Delta t\ \ \forall\ \{j = 1 \dots p\}, \{k = 1 \dots H\} \quad (8)$$

The aim of (7) and (8) is to divide each period k into p different sub-periods and the sub-periods duration must be between 0 and Δt.

$$q\left((k\times p)-(p-j)\right) = \sum_{i=1}^{n} y_{i,j,k}\times i \quad (9)$$
$$\forall \{j = 1 \dots p\}\ \{k = 1 \dots H\}$$

The purpose of (9) is to memorize the treated product at each sub-period $j$of period k.

The constraints below should also be taken into account.

$$\sum_{i=1}^{n} y_{i,j,k} = 1 \qquad \forall\ \{j = 1 \dots p\}\ For\ \{k = 1 \dots H\} \quad (10)$$

$$\sum_{j=1}^{p} y_{i,j,k} = 1 \qquad \forall \{i = 1 \dots n\}\ For\ \{k = 1 \dots H\} \quad (11)$$

$$y_{i,j,k} \epsilon\ \{0,1\} \quad (12)$$
$$\forall \{i = 1 \dots n\}, \{j = 1 \dots p\}, \{k = 1 \dots H\}$$

The equations (10) and (11) indicate that only one type of product will be produced in each sub-period j of period k.

The constraint (12) states that $y_{i,j,k}$ is a binary variable. Maintenance Strategy.

V. Maintenance Policy

A. Description

In this study we adopted a maintenance strategy with minimal repair. As illustrated in the figure below, the preventive maintenance actions are carried out at each instant $q \times T$ $(q = 1, 2 \dots)$ in order to replace the system by a new one (as good as new). If the system fails between preventive maintenance actions, only minimal repair is implemented.

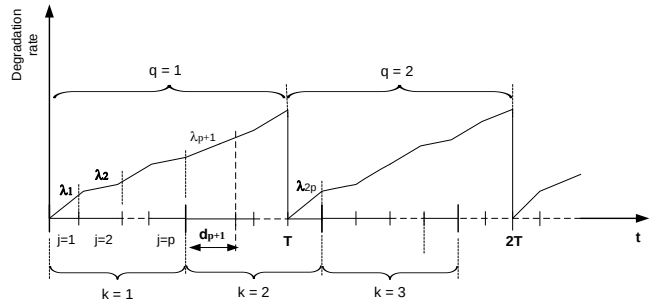


Fig. 3. The degradation rate

If we note Mc the cost of corrective maintenance actions, Mp the initial cost of preventive maintenance actions. It's clear that's difficult to make a perfect preventive maintenance action with restoring the system as good as new with adopting the usually the initial cost of preventive maintenance action Mp. That's why we will consider an increased preventive maintenance cost.

Formally we assume that the preventive maintenance action increased between successive actions according to the geometric sequence with constant rate α. Since that the total cost of maintenance is expressed as follows:

$$\Gamma(N) = Mc \times \phi_{(N,U)} + \sum_{i=1}^{N} Mp_i \qquad (13)$$

The aim of this maintenance strategy is to determine the optimal number of preventive maintenance actions $N^*(N = 1, 2 \dots)$ minimizing the total cost of maintenance over a given horizon $H \times \Delta t$.

Nakagawa *et al.* [8] has proven that there is an optimal number of partitions N^* and therefore, the optimal preventive maintenance period T^*, in the case of increased failure rate.

B. The expression of the average number of failures

As illustrated in "Fig. 3", we assume that the actions of preventive maintenance are made at the end of sub-periods, in order to reduce the complexity of the generation of the optimal number of preventive maintenance.

Hence, the function of the intervention period for preventive maintenance actions is presented as follows:

$$T = Round\left[\frac{H \times p}{N}\right] \qquad (14)$$

$Round[x]$: A round number of x

Lemma1:

$$\phi_{(U,N)} = \sum_{q=1}^{N+1} \left[\begin{array}{l} \sum_{j=\left((q-1)\times T+1\right)-\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]\times p\right)}^{p} \int_{0}^{\delta_{\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]+1\right)\times p-(p-j)}} \lambda_{\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]+1\right)\times p-(p-j)}(t)dt \\ + \sum_{k=In_{sup}\left[\frac{(q-1)\times T+1}{\Delta t}\right]+1}^{In\left[\frac{q\times T}{\Delta t}\right]} \sum_{j=1}^{p} \int_{0}^{\delta_{(k\times p)-(p-j)}} \lambda_{(k\times p)-(p-j)}(t)dt \\ + \sum_{j=1}^{q\times T-In\left[\frac{q\times T}{\Delta t}\right]\times p} \int_{0}^{\delta_{\left(In\left[\frac{q\times T}{\Delta t}\right]+1\right)\times p-(p-j)}} \lambda_{\left(In\left[\frac{q\times T}{\Delta t}\right]+1\right)\times p-(p-j)}(t)dt \end{array} \right]$$

$In[x]$: Integer part of number x
$In_{sup}[x]$: Superior integer part of number x

Proof: See [7]

C. The expression of failure rate

Recall, that the failure rate is influenced by the variation of the production rates. As illustrated in "Fig. 3", the failure rate is reset after each $q \times T$.

Thus, the expression of the failure rate can be written as follows:

Lemma2:

$$\lambda_{(k\times p)-(p-j)}(t) = \left[\begin{array}{l} \left(\begin{array}{c} \lambda_0 \\ + \sum_{Q=1}^{k-1} \sum_{l=1}^{p} \sum_{i=1}^{n} \frac{U_{i,l,Q} \times \Delta t}{U_{imax} \times \delta_{(Q\times p)-(p-l)}} \times \lambda_n\left(\delta_{(Q\times p)-(p-l)}\right) \\ + \sum_{l=1}^{j-1} \sum_{i=1}^{n} \frac{U_{i,l,k} \times \Delta t}{U_{imax} \times \delta_{(k\times p)-(p-l)}} \times \lambda_n\left(\delta_{(k\times p)-(p-l)}\right) \end{array} \right) \\ \times \left(1 - In\left[\frac{(k \times p) - (p - j + 1)}{q \times T}\right]\right) \\ + \sum_{i=1}^{n} \frac{U_{i,j,k} \times \Delta t}{U_{imax} \times \delta_{(k\times p)-(p-j)}} \times \lambda_n(t) \end{array} \right]$$

$$t \in \left[0, \delta_{(k\times p)-(p-j)}\right] \quad \forall \{k = 1 \dots H\}, \{j = 1 \dots p\}$$

Proof: See [7]

D. The expression of the total cost of maintenance

We recall that we considered we will consider an increased preventive maintenance cost to determine the total cost of maintenance.

Therefore:

$$Mp_{i+1} = Mp_i \times \alpha \quad (with\ \alpha > 1) \qquad (15)$$

Supposed that the initial preventive maintenance cost is:

$$Mp_1 = Mp \qquad (16)$$

The total cost of maintenance expressed in (13) can be represented as follows:

$$\Gamma(N) = Mc \times \phi_{(N,U)} + Mp_1 \times (1 + \alpha + \alpha^2 + \cdots + \alpha^N) \quad (17)$$

Thus:

$$\Gamma(N) = Mc \times \phi_{(N,U)} + Mp \times \frac{1-\alpha^N}{1-\alpha} \quad (18)$$

Using the average number of failure $\phi_{(U,N)}$ established in lemma 1, we can deduce that the analytical expression of the total maintenance cost is expressed as follows:

$$\Gamma(N) = \left[Mc \times \sum_{q=1}^{N+1} \left[\sum_{j=((q-1)\times T+1)-\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]\times p\right)}^{p} \int_0^{\delta_{\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]+1\right)\times p-(p-j)}} \lambda_{\left(In\left[\frac{(q-1)\times T}{\Delta t}\right]+1\right)\times p-(p-j)}(t)dt + \sum_{k=In_{sup}\left[\frac{(q-1)\times T+1}{\Delta t}\right]+1}^{In\left[\frac{q\times T}{\Delta t}\right]} \sum_{j=1}^{p} \int_0^{\delta_{(k\times p)-(p-j)}} \lambda_{(k\times p)-(p-j)}(t)dt + \sum_{j=1}^{q\times T-In\left[\frac{q\times T}{\Delta t}\right]\times p} \int_0^{\delta_{\left(In\left[\frac{q\times T}{\Delta t}\right]+1\right)\times p-(p-j)}} \lambda_{\left(In\left[\frac{q\times T}{\Delta t}\right]+1\right)\times p-(p-j)}(t)dt \right] + Mp \times \frac{1-\alpha^N}{1-\alpha} \right]$$

VI. Numerical Exemple

A simple example of a system that produces three different types of products is considered in order to satisfy three random demands according to every type of product.

Using the models described in previous sections, we will determine first the economical production plan followed by the optimal strategy of maintenance over a finite planning horizon: $H = 8\ trimesters$ (two years). We consider that the duration of periods $\Delta t = 3\ months$.

A. Numerical data

- The data relating to production:

We supposed that the standard deviation of demand of each product i, is the same for all periods $\left(\sigma(d_i(k)) = \sigma(d_i(k+1)) = \sigma(d_i)\right)$.

TABLE I. The average demands

	Average demand	Standard deviation
Product 1	200	1.5
Product 2	110	0.9
Product 3	320	1.2

The demands for each product by trimester are presented in the following table:

TABLE II. Demands

	Tr1	Tr2	Tr3	Tr4	Tr5	Tr6	Tr7	Tr8
Prod1	201	199	198	199	201	202	200	199
Prod2	111	119	108	201	111	112	110	119
Prod3	321	322	323	319	321	317	320	319

The setup costs are represented as follows:

TABLE III. Setup costs

		To		
		Product 1	Product 2	Product 3
From	Product 1	0	60	55
	Product 2	65	0	45
	Product 3	40	35	0

The other data are represented as following:

TABLE IV. Other data relating to production

	$S_{i,0}$ (pu)	U_{inom} $(pu/\Delta t)$	$Cp(i)$ (mu)	$Cs(i)$ (um/ut)	Θ_i (%)
Prod1	110	750	13	3	87
Prod2	85	530	17	5	95
Prod3	145	1150	9	2	90

- The data relating to maintenance:

System reliability law (Weibull) and preventive/corrective maintenance costs are defined by the following data:

TABLE V. Parameters of weibull function

Scale parameter (η)	Shape parameter (β)	Position parameter (γ)
12 months	2	0

The other data are given in the table below:

TABLE VI. Other data relating to maintenance

$Mp\ (mu)$	$Mc\ (mu)$	λ_0	α
800	1500	0	0.05

B. The economical production plan obtained

The production rates and the duration of sub-periods are given in the table below.

TABLE VII. The economical production plan

	Trimester 1			Trimester 2		
	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
Durations	0.85	0.71	1.44	0.65	1.21	1.14
Prod1	0	129	0	0	298	0
Prod2	120	0	0	0	0	185
Prod3	0	0	507	230	0	0
	Trimester 3			Trimester 4		
	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
Durations	0.18	1.18	1.01	0.43	0.74	1.83
Prod1	0	295	0	0	151	0
Prod2	134	0	0	0	0	312
Prod3	0	0	387	158	0	0

	Trimester 5			Trimester 6		
	δ_{13}	δ_{14}	δ_{15}	δ_{16}	δ_{17}	δ_{18}
Durations	1.82	0.87	0.31	0.56	0.55	1.89
Prod1	0	212	0	0	138	0
Prod2	0	0	52	58	0	0
Prod3	354	0	0	0	0	542
	Trimester 7			Trimester 8		
	δ_{19}	δ_{20}	δ_{21}	δ_{22}	δ_{23}	δ_{24}
Durations	0.76	1.11	1.13	1.05	0.77	1.18
Prod1	0	172	0	0	81	0
Prod2	0	0	92	130	0	0
Prod3	187	0	0	0	0	235

C. The optimal strategy of maintenance obtained

In the figure below, we represent the evolution of the total cost depending to the number of preventive maintenance actions.

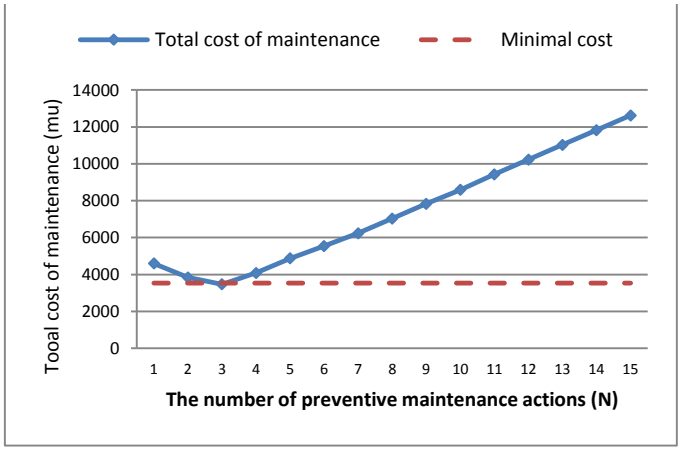


Fig. 4. The evolution of the total cost of maintenance

We deduce that the optimal number of preventive maintenance actions is $N^* = 3\ times$. Hence, the optimal period to apply the preventive maintenance is $T^* = 8\ months$, ensuring a minimal total cost of maintenance $\Gamma^*(N) = 3\ 469\ mu$.

VII. Conclusion

In this paper we have discussed the problem of integrated maintenance/production strategy for a manufacturing system consisting of a single machine which ensures the production of several types of products in order to satisfy random demands according to every product. The machine is subject to random failures, with increased failure rate according both time and production rate. Consequently, preventive maintenance actions, bloc type with minimal repair, are adopted in order to improve the system reliability. At failure, a minimal repair is carried out to restore the system into the operating state without changing its failure rate. In contrast to several studies in the literature, the originality of our study is the consideration of a variable setup cost and an increased preventive maintenance cost between successive actions. A variable setup costs used in our study and according to different product play an import role in the decision making in order to choose the order of product to produce at every period. More than that, the majority of the studies in maintenance frame in literature use a fixed cost of preventive maintenance action and assume that the system is restored to the state "as good as new" after every preventive maintenance action. It's clear that this assumption is unreal. That's why in this study, we considered an increased preventive maintenance cost between successive actions. A mathematical model is presented in order to formulate this assumption.

According to the proposed problem we start by developing analytically a stochastic production problem. Solving the analytical model expressing production, storage and setup cost and respecting proposed constraints, we obtained the economical production plan, presenting the quantity to produce for every product in each sub-period over a finite horizon. In the second phase, taking into account the economical production plan obtained, we have studied and optimized the maintenance policy. An analytical model, expressing the total maintenance cost is developed and optimized in order to determine the optimal number of preventive maintenance to do over a finite horizon, ensuring a minimal maintenance cost. We note that the maintenance policy adopted is characterized by considering the influence of production rate variation on the system degradation in the case of multiple products and an increased preventive maintenance cost between successive actions.

References

[1] S. Dellagi, N. Rezg, X. Xie, "Preventive Maintenance of Manufacturing Systems under Environmental Constraints", International Journal of Production Research. 45, Issue 5, 1233-1254, 2007.

[2] N. Rezg, S. Dellagi, and A. Chelbi, " Optimal Strategy of Inventory Control and Preventive Maintenance", International Journal of Production Research.Vol. 46 No. 19: 5349–5365, 2008.

[3] J.P. Kenne, E. Boukas, A. Gharbi, "Control of production and corrective maintenance rates in a multiple machine multiple-product manufacturing system", Mathematical and computer Modelling. 38, 351-365, 2003.

[4] T. W. Sloan and J. G. Shanthikumar, "Combined production and maintenance scheduling for a multiple-product, single- machine production system", Production and Operations Management. 9: 379–399. doi: 10.1111/j.1937-5956.2000.tb00465, 2000.

[5] Z. Hajej, S. Dellagi, N. Rezg, "Optimal integrated maintenance/production policy for randomly failing systems with variable failure rate", International Journal of Production Research, vol. 49, Issue 19, p5695- 5712, 2011.

[6] J.P. Kenne, L.J. Nkeungoue, "Simultaneous control of production, preventive and corrective maintenance rates of failure prone manufacturing system", Applied numerical mathematics, 58, 180-194, 2008.

[7] L. Mifdal, Z. Hajej, and S. Dellagi, "Joint Optimization Approach of Maintenance and Production Planning for a Multiple-Product Manufacturing System," Mathematical Problems in Engineering, vol. 2015, Article ID 769723, 17 pages, 2015. doi:10.1155/2015/769723.

[8] T. Nakagawa and S. Mizutani, "A summary of maintenance policies for a finite interval", Reliability Engineering and System Safety, 94(1), pp. 89–96, 2009.

Integrated Preventive Maintenance and Inspection Planning in a Deteriorating Serial Multi-Stage Manufacturing System

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

M. Rezaei-Malek [1,2], R. Tavakkoli-Moghaddam [1,2], A. Siadat [1,2], and J-Y. Dantan [2]
[1] School of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran
m.rezaeimalek@ut.ac.ir, tavakoli@ut.ac.ir

[2] LCFC, Arts et Métier ParisTech, 57078 Metz, France

mohammad.rezaeimalek@ensam.eu, reza.tavakkolimoghaddam@ensam.eu, ali.siadat@ensam.eu, jean-yves.dantan@ensam.eu

Abstract—This paper presents an integrated mathematical model for preventive maintenance and inspection planning problem to determine the most effective strategy for employing inspection and preventive maintenance activities regarding the different cost components: repair, scrap, production, inspection, maintenance, and penalty of shipped defective items. The results show that the determination of inspection locations along the manufacturing line in the different periods of time regarding the impact of preventive maintenance activities on the defective production probability causes a huge economic benefit. Therefore, it can be concluded that the integrated planning of preventive maintenance and inspection activities seems an appropriate technology that can be adopted by manufacturing companies to remain profitable in today's highly competitive market.

Keywords—Preventive maintenance; inspection; mathematical optimization

I. Introduction

Manufacturing high-quality products is a significant factor in a company aiming to maintain its global competitiveness. In Manufacturing Systems (MSs), almost all the production stages are technologically incapable for processing items in a perfect quality. Hence, to increase the guarantee of delivering non-defective products to customer, it is required to adopt an effective quality-assurance program by planning and managing resources applied to the inspection and testing of important product quality attributes [1]. Noteworthy, unnecessarily processing of undetected defective items during different downstream production stages results in wasting of material, equipment, and time. Such kind of waste, in the present economic environment, has become a major concern for many companies because of its negative effect on the environment and the final production cost [2].

When, where, and how an inspection activity should be implemented in a manufacturing process, is a significant and challenging decision in quality control research area [3]. On the other side, after detection of a defect, determining proper defect management strategies, including scrap, replacement, rework and repair, allows smoothing the defect propagation throughout process and to end-consumer(s) [4]. The researches dealt with these questions are classified under the sub-research area of *Inspection-oriented quality-assurance strategies* or in a short form *Inspection Strategies (ISs)*. An IS tries to allocate an economically appropriate level of inspection by approaching a trade-off among the various cost components such as inspection, repair, rework, scrap, replacement, and warranty (takes place when a defective product has been shipped to customer) because of quality failure [1]. The research into the IS problem has been started by Lindsay and Bishop [5]. There are some important review papers in the literature such as [1]–[3], and [6]. The literature of the IS problem can be classified into two main categories according to the proposition of Mandroli et al. [1]: Allocation Strategy (AS) and Parametric Strategy (PS). The research studies dealing with the question about the location of inspection activities in MS (e.g., see [7]) are classified as the AS category. The papers attempting to answer questions such as the fraction of items to be inspected and the number of repetitions, are classified as the PS category (e.g., see [8]). There are some articles which are concurrently placed in the two categories (i.e., AS-PS) (e.g., see [9]). The present literature review concentrates on the AS and AS-PS categories.

II. Literature Review

There are five important features for the papers dealing with the AS and AS-PS problems: System feature, Constraint(s), Inspection capability, Cost components (i.e., Internal failure, External failure, Inspection cost, and Manufacturing cost), and Solution technique. Based on the systems studied in the existing literature, there are three major process configurations regarding the flow of conforming items [1]: serial (e.g. considered at [10]), assembly (e.g. considered at [11]), and non-serial systems (e.g. considered at [12]). In a serial MS, input material passes sequentially the successive processing production stages [2]. There are some papers considered constraint(s) for their optimization problems. Indeed, in the real MSs, there are usually practical constraints for implementing an inspection strategy. For example, Korytkowski [13] studied a production system with the Average Outgoing Quality Limit (AOQL) (please refer to Ercan [14] for the AOQL definition). A limit on the budget/number of inspection stations was considered in many studies, e.g. [11] and [15]–[26].

There are two types of error that may happen during an inspection activity: Type-I and Type-II errors. Type-I error means the rejection of good items (e.g. considered at [27]). Type-II error means the acceptance of non-conformance items

(e.g. considered at [28]). For the optimization purpose, different cost components can be minimized. Internal cost refers to the costs incur inside the company like reworking, scrapping and replacing the detected non-conformance items (e.g. considered at [29]). The external failure costs incur after the defective products are delivered to customer like replacement, repair, and cost of loss of goodwill (e.g. considered at [30]).

The multi-stage structure of a MS leaded most of the papers to apply Dynamic Programming (DP) as a solution approach (e.g. see [5], [11], and [31]). However, when the number of production stages increases, because of the formulation complexity, the DP method is not applicable. Therefore, the non-exact solution procedures like heuristics and meta-heuristics are more appropriate solution approaches in this situation [9]. Furthermore, there are some studies that employed the simulation technique for performance assessment of the inspection plan while focusing on simple processes to simplify their models (e.g. see [32]).

Comparing the existing literature to the present work, this paper considers a serial multi-stage MS. Type-I and Type-II error can happen during each inspection activity. The internal and external cost components are considered for the optimization purpose and it tries to present an exact optimal solution for the system.

We have already mentioned the general features of the present work. Now, we aim at justifying the main contribution of this work which is the proposition of an integrated optimization approach for the simultaneous planning problem of preventive maintenance and inspection strategy. For this regard, first we clarify the relation of inspection and maintenance activities. Manufacturing equipment is designed to successfully do operation during the anticipated service life; however, deterioration starts as soon as it is commissioned [33]. Hence, equipment downtime, quality problems, and slower production are the obvious outcomes of this deterioration process. As expressed by Madu [34], equipment maintenance is an important factor affecting the MS's ability to provide quality and timely services to customer. In this regard, British Standard Institution defines Maintenance as "a combination of all technical and associated administrative activities needed to maintain equipment, installations and other physical assets in the appropriate operating condition or bring them back to this condition" [33]. In the real MSs, when a machine runs, it starts deteriorating after a while and finally it will fail. To be exact, the deterioration process starts when a failure mode is arrived. In the deterioration period, the defective production probability of machine is increasing continuously. Regarding this situation, we should determine, on one side, the appropriate time for doing a preventive maintenance activity and, on the other side, where and when is the right place and appropriate time for doing an inspection. When a maintenance activity is done, the defective production probability of machine is improved. And, when you decide to do an inspection activity, you can detect the defect items and scrap, rework, or repair them. The second question (i.e., right inspection time and place) was answered in the current IS literature. But, when you face to a changing defective production probability (deteriorating condition) and there is a possibility of doing a maintenance activity, it is necessary to simultaneously respond to the first (right maintenance time and place) and second question. However, to the best of our knowledge, save the present paper, there is no study for dealing with these two decision problems concurrently.

The rest of this paper is organized as follows. Section III describes the problem and provides a mathematical model for that. Section IV is dedicated to the computational experiment. Section V draws the conclusions and future research direction.

III. Problem Description and Mathematical Formulation

Consider a Serial Multi-Stage Production Line (SMSPL). Material enters the SMSPL and passes all the stages and after the last one, it is transformed to a final product ready for selling to customer. Each production stage is responsible for a certain quality characteristic. The production stages are technologically incapable for processing items in a perfect quality, so there is a possibility to do an inspection activity (inspection station establishment) after each stage. Each inspection station is capable to detect the defective items associated to the preceding stages and then they can be scrapped or be repaired in a known cost which depends on the processing stage number. There are two types of error for each inspection activity (i.e. Error type-I and Error type-II).

In the considered SMSPL, each stage is deteriorating in time. To be exact, when a stage starts operating, it starts deteriorating after a while and finally it fails. For instance, consider a drill machine, the drill bit becomes blunt after a while and does not have the precise and capability of its beginning. During the deterioration period, probability of a conforming item acquires a defect at a stage, ε, is increasing in time and it is equal to 1 at the end (i.e., the stage fails). Also, it is possible to do a PM activity in a known cost which depends on the extent of remained time to the complete failure. Fig. 1 shows an example for the ε behavior of a stage and the impact of the PM activities on it. Fig. 2 shows the schematic plan of the systems. Regarding the above-described SMSPL, the problem is to concurrently plan the inspection and PM activities during a planning horizon while minimizing the total costs (i.e. repair, production, inspection, maintenance, and penalty of shipped defective items). The planning horizon consists of equal time periods and it is needed to determine in which time periods, inspection, maintenance activities require to be done. There are n stages and in each period t, there is a possibility of doing an inspection and/or PM activity. Below, we first define the notations, then develop the equivalent model of our non-linear mixed-integer program.

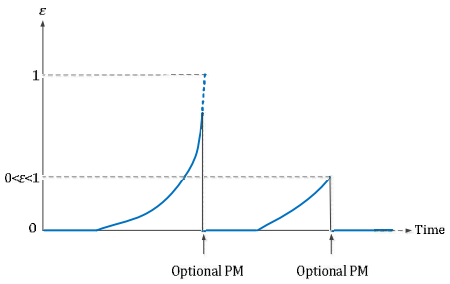


Fig. 1. The impact of the PM on the value of ε

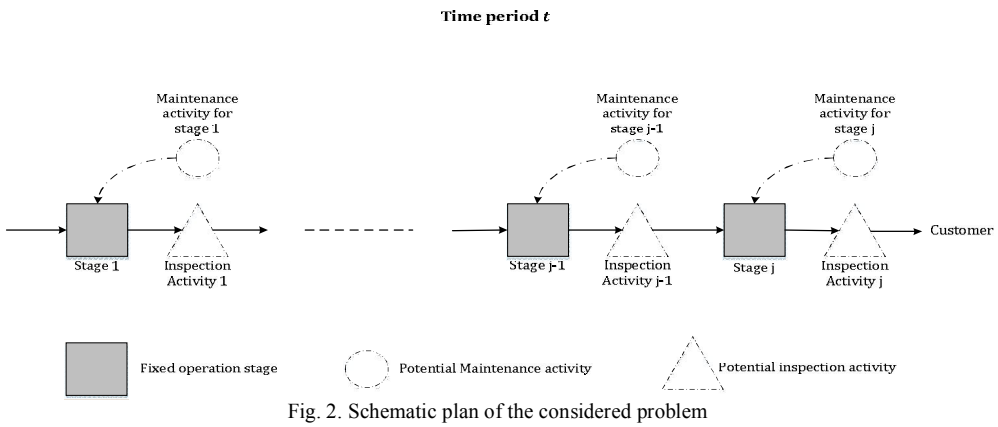


Fig. 2. Schematic plan of the considered problem

A. Assumptions

- A PM activity restores a machine as good as a new one (i.e., $\varepsilon = 0$),
- The only reason that a production stage transforms a conforming item to a defective one, is the machine/machine tool decay. It is notable that this paper does not consider the other causes of quality issues such as misadjustment, misalignment or human errors.
- The values of Error type-I and Error type-II are certain constant quantities during the planning horizon,
- The repair function is done perfect and transforms the rejected items to the conforming ones.
- ε_{tj} behavior function are assumed the same for all the production stages.

B. Indices

j index of production stages ($j = 1, \ldots, n$),
t,t' index of time periods ($t, t' = 1, \ldots, T$).

C. Parameters

n number of production stages in the SMSPL,
w_{t01} number of unit material that enters the SMSPL in the period t,
α_j probability of Type-I error at the inspection station j,
β_j probability of Type-II error at the inspection station j,
ε_{t0} non-conforming fraction of material that enters the SMSPL in the period t,
I_j unit inspection cost at the inspection station j,
R_{j1} unit replacement/repair cost of a conforming item rejected by the inspection station j,
R_{j2} unit replacement/repair cost of a non-conforming item rejected by the inspection station j,
p_j unit production cost in the stage j,
p_0 unit material cost entering the SMSPL,
S_j scrap cost per unit at the stage j,
C penalty cost of delivering a non-conforming item to customer,
RE revenue generated by delivering an item to customer,
f_{j1} fraction of the rejected items repaired at the inspection station j,
f_{j2} fraction of the rejected items scraped at the inspection station j,
$g(.)$ ε_{tj} behavior function,
$v(.)$ mc_{tj} behavior function,
K number of unit material entering the SMSPL in each period.

D. Variables

d_{tj} 1; if inspection station associated to the stage j is established in the period t, 0; otherwise,
m_{tj} 1; if PM activity associated to the stage j is performed in the beginning of the period t, 0; otherwise,
mc_{tj} cost of performing a PM activity for the stage j in the period t; $mc_{t,j} = v(\omega_{tj})$,
w_{tj1} expected conforming items entering the production stage j in the period t,
w_{tj2} expected non-conforming items entering the production stage j in the period t,
u_{tj1} $= \alpha_j(1-\varepsilon_{tj})w_{tj1}d_{tj}$; expected conforming items rejected at the j^{th} inspection station in the period t,
u_{tj2} $= (1-\beta_j)\ \varepsilon_{tj}\ (w_{tj1}+w_{tj2})\ d_{tj}$; expected non-conforming items rejected at the j^{th} inspection station opportunity in the period t,
ω_{tj} number of periods have been passed since the last PM activity on the operation stage j in the period t,
ε_{tj} probability of a conforming item acquires a defect during processing in the stage j in the period t; $\varepsilon_{tj} = g(\omega_{tj})$,
IC total inspection cost of the SMSPL,
RC total replacement/repair cost of the SMSPL,
PC total production cost of the SMSPL,
MC total PM cost of the SMSPL,
SC total scrap cost of the SMSPL,
$Profit$ what is left over at the very end after the company has paid for the penalty cost of non-conforming goods sold.

E. Mathematical Model

Based on the above-mentioned definitions, the mixed-integer non-linear mathematical formulation of the problem is provided as follows.

$$\text{Min } z = RC + SC + PC + IC + MC - Profit \quad (1)$$

s.t.

$$RC = \sum_{t=1}^{T}\sum_{j=1}^{n} f_{j1}R_{j1}u_{tj1} + \sum_{t=1}^{T}\sum_{j=1}^{n} f_{j1}R_{j2}u_{tj2} \quad (2)$$

$$SC = \sum_{t=1}^{T}\sum_{j=1}^{n} f_{j2}S_{j}u_{tj1} + \sum_{t=1}^{T}\sum_{j=1}^{n} f_{j2}S_{j}u_{tj2} \quad (3)$$

$$PC = \sum_{t=1}^{T}\sum_{j=1}^{n} p_j(w_{tj1} + w_{tj2}) \quad (4)$$

$$IC = \sum_{t=1}^{T}\sum_{j=1}^{n} I_j(w_{tj1} + w_{tj2})d_{tj} \quad (5)$$

$$MC = \sum_{t=1}^{T}\sum_{j=1}^{n} mc_{tj}m_{tj} \quad (6)$$

$$Profit = \sum_{t=1}^{T} RE \times \left(w_{t,n+1,1} + w_{t,n+1,2}\right) - C \times w_{t,n+1,2} \quad (7)$$

$$w_{t,1,1} = (1-\varepsilon_{t0})w_{t01} \quad \forall t \quad (8)$$

$$w_{t,1,2} = \varepsilon_{t0}w_{t01} \quad \forall t \quad (9)$$

$$\begin{aligned} w_{t,j+1,1} = {} & (1-\varepsilon_{tj})w_{tj1} \\ & + \left(f_{j1}\left(\alpha_j(1-\varepsilon_{tj}) + (1-\beta_j)\varepsilon_{tj}\right)\right. \\ & \left. - \alpha_j(1-\varepsilon_{tj})\right)w_{t,j,1}d_{tj} \\ & + f_{j1}(1-\beta_j)w_{tj2}d_{tj} \quad \forall t, j \neq n \end{aligned} \quad (10)$$

$$\begin{aligned} w_{t,n+1,1} = {} & (1-\varepsilon_{tn})w_{tn1} \\ & + \left(f_{n1}(\alpha_n(1-\varepsilon_{tn}) + (1-\beta_n)\varepsilon_{tn})\right. \\ & \left. - \alpha_n(1-\varepsilon_{tn})\right)w_{t,n,1}d_{tn} \\ & + f_{n1}(1-\beta_n)w_{tn2}d_{tn} \quad \forall t \end{aligned} \quad (11)$$

$$\begin{aligned} w_{t,j+1,2} = {} & w_{tj2} + \varepsilon_{tj}w_{tj1} - (1-\beta_j)\varepsilon_{tj}w_{tj1}d_{tj} \\ & - (1-\beta_j)w_{tj2}d_{tj} \quad \forall t, j \end{aligned} \quad (12)$$

$$\omega_{tj} = (\omega_{t-1,j} + 1)(1-m_{tj}) \quad \forall j, t \neq 1 \quad (13)$$

$$\omega_{1,j} = 0 \quad \forall j \quad (14)$$

$$md_{1j} = 1 \quad \forall j \quad (15)$$

$$\varepsilon_{tj} = g(\omega_{tj}) \quad \forall t, j \quad (16)$$

$$mc_{t,j} = v(\omega_{tj}) \quad \forall t \quad (17)$$

$$w_{t01} = K \quad \forall t \quad (18)$$

$$w_{tj1}, w_{tj2}, mc_{tj}, u_{tj1}, u_{tj2}, \omega_{tj}, \varepsilon_{tj}, RC, IC, PC, MC, SC, Profit \geq 0 \quad \forall t, j \quad (19)$$

$$m_{tj}, d_{tj}: \text{binary} \quad \forall t, j \quad (20)$$

Objective function (1) minimizes the total cost including repair, scrap, production, inspection, preventive maintenance, and negative form of profit. Equations (2) – (7) calculate these different cost components. Equations (8) – (12) calculate the expected number of conforming and non-conforming items entering the different production stages in each period. Equation (13) calculates the number of passed time periods after the last performed PM activity. It is assumed that a PM activity is done for all the production stages in the first period (please see Equations (14) and (15)). Equation (14) indicates that ε is a function of the number of passed time periods after the last performed PM activity. Equation (16) calculates the value of ε_{tj} which is a function of ω_{tj}. Equation (17) calculates the maintenance cost of the machine *j* in the period *t*. Equation (18) shows the number of unit material entering the SMSPL in the period *t*. Equation (19) is a non-negativity limitation, and Equation (20) indicates that opening an inspection station and doing a PM activity are binary decisions.

IV. Example

To show the presented model works well and provides reasonable and useful results, we solve a similar example to which was investigated in [35]. The data of that 3-stage system is used for all the 12 periods. The deterioration processes of all the three stages are considered the same. To be exact, we assume that the ε_{ij} behavior of each stage is a function of ω_{tj} as $\varepsilon_{tj} = 0.1 \times \omega_{tj}$. Similarly, the maintenance cost is a function of ω_{tj} as we have $mc_{tj} = 50 + 10 \times \omega_{tj}$. The number of unit material entering the SMSPL is 100 units per period and %5 of them are non-conforming, i.e. $\varepsilon_{t0} = 0.05$. The revenue generated by delivering a conforming item to customer is 10 US$ and the penalty cost of a non-conforming item shipment is estimated 20 US$. The rest of required data is provided in TABLE I. The model is coded in GAMS23.5 and solved with a zero absolute gap by the BONMIN solver on the data of example, using a laptop with Core i7 CPU, 1.73 GHz and 4 GB of RAM.

The most important achievement of this research is that the proposed technology of integrated planning of PM and inspection activities results in a decrease in the total manufacturing cost of the SMSPL as this was predictable because of the inherent interconnectivity of these two types of operation. In the investigated example this saving is %4.67. The value of objective function is 28624 and by adding the implementation possibility of PM activities, it decreases to 27287 where this is a huge saving for a manufacturing company.

TABLE II and TABLE III shows the place and period that the PM and inspection activities are done. As it can be concluded by TABLE II and TABLE III, the model decided to establish inspection stations when and where the PM activities were not taken to be done. The reason for the PM activities which were done mostly for the last stage is that the scrap cost for the last stage is too high in comparison with the other stages and the model avoided a considerable waste of money.

The obtained solution of the presented model and accordingly the associated saving cost highly depend on the fixed cost of PM. As it is shown in Fig. 3, by increasing the fixed cost of PM, the cost saving is being decreased as there is not any monetary benefit of applying the proposed integrated approach when the fixed cost of a PM is higher than 150. The impact of changing the variable cost of a PM is not investigated because this kind of expense is related to the required material and it is not being changed during the planning horizon. Fig. 4 depicts the values of the objective function when the unit cost of inspection is changing. As can be seen, the "With PM Possibility" approach is a cost-effective approach even when there is no cost for inspection. In addition, the cost saving of the "With PM Possibility" approach is increasing when the unit inspection cost is being raised. Hence, the "With PM Possibility" is more cost-effective when the unit cost of inspection is high.

TABLE I. Production parameters of the 3-stage MS

Parameter	Stage 1	Stage 2	Stage 3
α	0.02	0.05	0.01
β	0.01	0.1	0.01
I	0.5	0.5	0.5
R_1	5	0	0
R_2	20	0	0
p	10	20	5
S	0	-2	40
f_1	0.2	0	0.5
f_2	0.8	1	0.5

TABLE II. The PM activities which are done in the SMSPL

Period	Stage 1	2	3
1	1	1	1
2	1	0	1
3	0	1	1
4	0	0	1
5	0	1	1
6	0	1	1
7	0	0	1
8	0	1	1
9	0	0	1
10	0	0	1
11	0	0	1
12	0	0	0

TABLE III. The inspection activities which are done in the SMSPL

Period	Stage 1	2	3
1	1	0	0
2	1	1	0
3	1	0	0
4	1	1	0
5	1	0	0
6	1	0	0
7	1	1	0
8	1	0	0
9	1	1	0
10	1	1	0
11	1	1	0
12	1	1	0

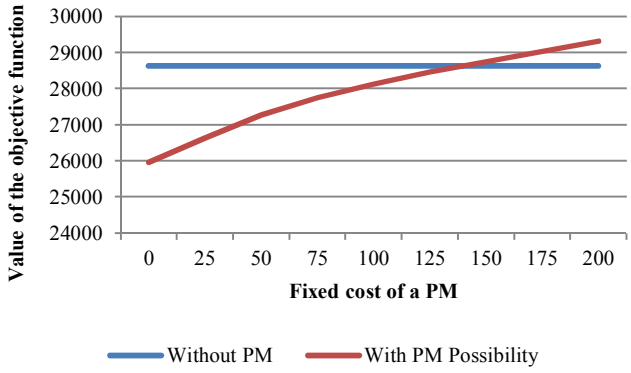


Fig. 3. Value of the objective function with considering different values for the fixed cost of a PM

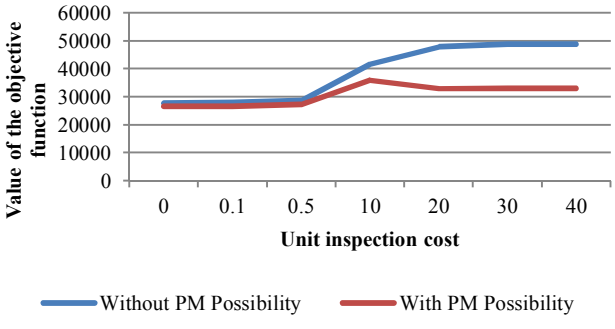


Fig. 4. Value of the objective function with considering different values for the unit inspection cost

V. Conclusion and Future Research Direction

This paper has developed a mixed-integer mathematical model for the integrated problem of preventive maintenance and inspection planning in a serial multi-stage manufacturing system. It has been assumed that each production stage is deteriorating in time and consequently the probability of a conforming item acquires a defect in each stage is increasing. The proposed model determines the optimum time and place for the preventive maintenance and inspection activities while minimizing the total manufacturing cost including repair, scrap, production, inspection, and preventive maintenance cost. A numerical example was solved for a serial 3-stage manufacturing system and the obtained results showed that the concurrent decision-making on preventive maintenance and inspection activities decreased the total manufacturing cost around %5. In this paper, it has been considered that the life time of each production stage is known and thus the downtime is certain. However, it does not seem a real assumption and the downtime has a stochastic behavior. Hence, for future researches, it is highly recommended to incorporate this stochastic behavior to the mathematical model.

Acknowledgment

The first author gratefully acknowledges the support given by French Ministry of Foreign Affairs and the embassy of France in Iran for their financial support under Scholarship No. [889461B].

References

[1] Mandroli, S.S., Shrivastava, A.K., and Ding, Y., 2006, "A Survey of Inspection Strategy and Sensor Distribution Studies in Discrete-Part Manufacturing Processes," IIE Transactions, 38(4), 309–328.

[2] Shetwan, A.G., Vitanov, V.I., and Tjahjono, B., 2011, "Allocation of Quality Control Stations in Multistage Manufacturing Systems," Computers and Industrial Engineering, 60(4), 473–484.

[3] Raz, T., 1986, "A Survey of Models for Allocating Inspection Effort in Multistage Production Systems," Journal of Quality Technology, 18(4), 239–247.

[4] Colledani, M., Tolio, T., Fischer, A., Iung, B., Lanza, G., Schmitt, R., and Vancza, J., 2014, "Design and Management of Manufacturing Systems for Production Quality," CIRP Annals - Manufacturing Technology, 63(2), 773–796.

[5] Lindsay, G.N., and Bishop, A.B., 1964, "Allocation of Screening Inspection Effort-A Dynamic-Programming Approach," Management Science, 10(2), 342–352.

[6] Shi, J. and Zhou, S., 2009, "Quality Control and Improvement for Multistage Systems: A Survey," IIE Transactions, 41(9), 744–753.

[7] Emmons, H., Rabinowitz, G., 2002, "Inspection Allocation for Multistage Deteriorating Production Systems," IIE Transactions, 34(12), 1031–1041.

[8] Zheng, S., 2000, "Dynamic Quality Control in Assembly Systems," IIE Transactions, 32(9), 797–806.

[9] Azadeh, A., Sangari, M.S., Sangari, E., and Fatehi, S., 2015, "A Particle Swarm Algorithm for Optimising

Inspection Policies in Serial Multistage Production Processes with Uncertain Inspection Costs," International Journal of Computer Integrated Manufacturing, 28(7), 766–780.
[10] Mohammadi, M., Siadat, A., Dantan, J.-Y., and Tavakkoli-Moghaddam, R., 2015, "Mathematical Modelling of a Robust Inspection Process Plan: Taguchi And Monte Carlo Methods," International Journal of Production Research, 53(7), 2202–2224.
[11] Penn, M., and Raviv, T., 2008, "A Polynomial Time Algorithm for Solving a Quality Control Station Configuration Problem," Discrete Applied Mathematics, 156(4), 412–419.
[12] Rau, H., Chu, Y.H., and Cho, K.H., 2005, "Layer Modelling for the Inspection Allocation Problem in Re-Entrant Production Systems," International Journal of Production Research, 43(17), 3633–3655.
[13] Korytkowski, P., 2011, "A Genetic Algorithm with Tournament Selection for Optimising Inspection Allocation in Multiproduct Multistage Production Systems," International Journal of Simulation and Process Modelling, 6(3), 238.
[14] Ercan, S.S., Hassan, M.Z., and Taulananda, A., 1974, "Cost Minimizing Single Sampling Plans with AIQL and AOQL Constraints," Management Science, 20(7), 1112–1121.
[15] White, L.S., 1969, "Shortest Route Models for the Allocation of Inspection Effort on a Production Line," Management Science, 15(5), 249–259.
[16] Ballou, D.P., and Pazer, H.L., 1985, "Process Improvement versus Enhanced Inspection in Optimized Systems," International Journal of Production Research, 23(6), 1233–1245.
[17] Taneja, M., and Viswanadham, N., 1994, "Inspection Allocation in Manufacturing Systems: A Genetic Algorithm Approach," Proc. of the IEEE International Conference on Robotics and Automation, May 8-13, San Diego, California, 3537–3542.
[18] Rebello, R., Agnetis, A., and Mirchandani, P.B., 1995, "Specialized Inspection Problems in Serial Production Systems," European Journal of Operational Research, 80(2), 277–296.
[19] Bai, D.S., and Yun, H.J., 1996, "Optimal Allocation of Inspection Effort in a Serial Multi-Stage Production System," Computers and Industrial Engineering, 30(3), 387–396.
[20] Lee, J., and Unnikrishnan, S., 1998, "Planning Quality Inspection Operations in Multistage Manufacturing Systems with Inspection Errors," International Journal of Production Research, 36(1), 141–155.
[21] Shiau, Y.R., 2002, "Inspection Resource Assignment in a Multistage Manufacturing System with an Inspection Error Model," International Journal of Production Research, 40(8), 1787–1806.
[22] Jang, W., and Shanthikumar, J.G., 2002, "Stochastic Allocation of Inspection Capacity to Competitive Processes," Naval Research Logistics, 49(1), 78–94.
[23] Shiau, Y.R., 2003, "Quick Decision-Making Support for Inspection Allocation Planning with Rapidly Changing Customer Requirements," The International Journal of Advanced Manufacturing Technology, 22(9–10), 633–640.
[24] Shiau, Y.R., 2003, "Inspection Allocation Planning for a Multiple Quality Characteristic Advanced Manufacturing System," The International Journal of Advanced Manufacturing Technology, 21(7), 494–500.
[25] Penn, M., and Raviv, T., 2007, "Optimizing the Quality Control Station Configuration," Naval Research Logistics, 54(3), 301–314.
[26] Shiau, Y.R., Lin, M.H., and Chuang, W.C., 2007, "Concurrent Process/Inspection Planning for a Customized Manufacturing System Based on Genetic Algorithm," The International Journal of Advanced Manufacturing Technology, 33(7–8), 746–755.
[27] Vaghefi, A., and Sarhangian, V., 2009, "Contribution of Simulation to the Optimization of Inspection Plans for Multi-Stage Manufacturing Systems," Computers and Industrial Engineering, 57(4), 1226–1234.
[28] Deliman, N.C., and Feldman, R.M., 1996, "Optimization of Process Improvement and Inspection Location for Serial Manufacturing," International Journal of Production Research, 34(2), 395–405.
[29] Kim, J., and Gershwin, S.B., 2008, "Analysis of Long Flow Lines with Quality and Operational Failures," IIE Transactions, 40(3), 284–296.
[30] Tayi, G., and Ballou, D., 1988, "An Integrated Production-Inventory Model with Reprocessing and Inspection," International Journal of Production Research, 26(8), 1299–1315.
[31] Finkelshtein, A., Herer, Y.T., Raz, T., and Ben-Gal, I., 2005, "Economic Optimization of Off-Line Inspection In a Process Subject to Failure And Recovery," IIE Transactions, 37(11), 995–1009.
[32] Tannock, J., and Saelem, S., 2007, "Manufacturing Disruption Costs due to Quality Loss," Int. J. Qual. Reliab. Manag., vol. 24, no. 3, pp. 263–278.
[33] Muchiri, P.N., Pintelon, L., Martin, H., and Chemweno, P., 2014, "Modelling Maintenance Effects on Manufacturing Equipment Performance : Results from Simulation Analysis," International Journal of Production Research, vol. 52, no. 11, pp. 3287–3302.
[34] Madu, C.N., 2000, "Competing through Maintenance Strategies," International Journal of Quality and Reliability Management, 17(9), 937–949.
[35] Yum, B.J., and McDowell, E.D., 1987, "Optimal Inspection Policies in a Serial Production System Including Scrap Rework and Repair: An MILP Approach," International Journal of Production Research, 25(10), 1451–64.

Manufacturing System Design

Dual strategy for the evaluation of the structures of series production

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Matthias Heinicke
Institute of Ergonomics, Manufacturing Systems and Automation (IAF)
Otto von Guericke University
Magdeburg, Germany
Matthias.Heinicke@ovgu.de

Ulf Bergmann
Institute of Ergonomics, Manufacturing Systems and Automation (IAF)
Otto von Guericke University
Magdeburg, Germany

Abstract—Turbulent environment challenges series production and induces an intrinsic complexity of production systems. Stepwise optimization of production processes may lead to pseudo-optimums during factory operation. Even small changes of the production program cause a suboptimal configuration of the production system due to path dependencies in the long run. Production systems are forced to cope with these changes by adjusting their internal structure. In this context, temporal and / or spatial adjustments should be considered. The evaluation of an adequate adaption requires proper figures that assess the suitability of the production structure regarding the pending tasks and demands. Thus, this paper presents an approach that enables a deduction of appropriate measures. It aims at matching the temporal and spatial structure with the production program by revealing suitable options for adaption.

Keywords—Production Structure; Suitability of System´s Configuration; Structural Adaption; static / dynamic Key Figures

I. Introduction

Most planning approaches pre-determine a flexible system structure [1], whereas few studies present analytical approaches assessing the appropriateness of the production structure regarding pending tasks and demands [2]. For this reason, we have developed Structural Quality (SQ) as a vector of the spatial and temporal structure, an integrated assessment standard for efficient evaluation of production structures [3]. In this paper, we illustrate an SQ-based approach and its application potential.

Structuring is the arrangement of the elements of a production system (spatial reference) and the planned configuration of the relations a product undergoes to its production (time reference) [4]. In order to evaluate the spatial structural component, an assessment of the relevant spatial layout is required. Focusing on the complexity of the material flow, the minimum value therefore is found within the purest form of a production line. Precisely one upstream and one downstream workstation without return flow characterize such a production line. However, the maximum is found in a fully meshed job-shop structure [4]. The temporal structure takes account of the workflow, i.e. the logical and temporal linking of the partial tasks of production. The sequence of relations between the system elements (workstations) expresses the temporal component of the processes running in the production system. Based on the intensities of the relations as dynamic variables, different temporal courses can be characterized [5].

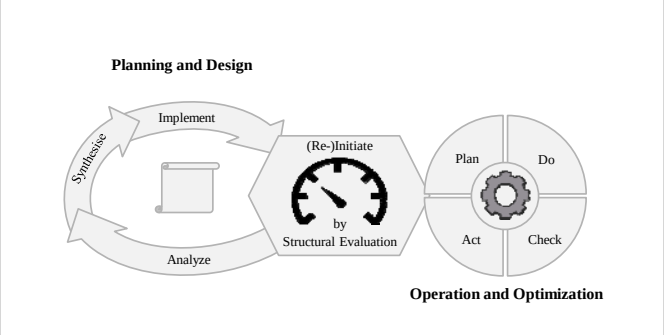
Moreover, structuring determines the implementation costs of the production processes. Based on the specific type and number of workstations (system´s composition) and the relationship between them, a production structure of this kind defines the basic structure of a production system (system's setup) as well as its function (system's behavior) [6]. The vector SQ enables measuring and verifying the appropriateness of installed system structures by concerning a specific production program and a proper function of the production system. According to Chandler's insights into the interrelations between structure and strategy, the production task (expressed by the production program) is one of the main inputs to structural design within the framework of production structure planning [7]. Theoretically, any change in the production program leads to a change in structure in terms of a modified spatial-temporal organization of the production system. On the contrary, industrial production processes require a long-term stable cost function. Therefore, it is aimed at long-term valid structures for production systems [8]. In practice, any established production structure is a transition between the function-oriented and the object-oriented type of structure [9]. Consequently, the definition of a structure is commonly a long-term decision of the general principle of structuring.

However, any defined production structure is counteracted by the increasing individualization of customer requirements. Due to significant shifts in production program, there is latent pressure to change an installed production structure. Beyond the limits of the equipment's flexibility [8], adaptations to the inherent system-related properties are necessary, i.e. a re-configuration of the system's composition, setup, and behavior. Companies do not use every shift in the production program as an opportunity to change the installed production structure. Typically, more often rationalization measures are introduced, such as process-specific modification or integration of production steps. Finally, there are adapted structures in the form of case-specific operation solutions. This results in huge amount of different transitional structures.

Structural quality as key figure reveals the potential of an adaptation of the production structure. At this point, structural quality facilitates an assessment of changing the spatial and temporal structure. Both in the planning phase as well as in the

operation and optimization phase, an evaluation of the suitability of production structures for the present production task is appropriate (Fig. 1). If the existing production structure is a planning construct, an anticipatory structure evaluation takes place, whereas active structure monitoring is possible during operation.

Fig. 1. Planning and Optimization cycles initiated by Evaluation of Production Structures.



II. Key Figures for Evaluation of Production Structures

A. Spatial Structure

In order to evaluate the spatial structural component, an assessment of the relevant spatial layout is required. Basically, we focus on the complexity of material flow. Minimal "complexity of the material flow" is found within the purest form of a production line. Such a production line is characterized by precisely one upstream and one downstream work station without return flow. However, maximum "complexity of the material flow" is found in a fully meshed job-shop structure. In order to assess the "complexity of the material flow", we use two indicators: The degree of linearity, and the standardized degree of cooperation. The degree of linearity takes into account the number of return flows in contrast to the prevalent sequence of operations, while the standardized degree of cooperation determines the number of material flow relations existing between the workstations [6].

For the standardized degree of cooperation, the value 1 is defined as a threshold value: In the range $0 \geq \kappa \leq 1$, the working stations are directly connected to one another in a sequence; this arrangement tends to be a linear structure. For values in the range $\kappa > 1$, more than one workstation is connected to more than two other workstations; this arrangement tends to be a job-shop structure. A change in intensity and basic orientation of the relations imply the trend of a potential structural change. Minimum "complexity of the material flow" exists when the degree of linearity assumes the value of 1 and the standardized degree of cooperation has a value of 0.

B. Temporal Structure

The temporal structure takes account of the workflow, i.e. the logical and temporal linking of the partial tasks of production [6]. The sequence of relations between the system elements (workstations) expresses the temporal component of the processes running in the production system. On the basis of the intensities of the relations as dynamic variables, different temporal courses can be characterized.

Parameters of the system-related temporal structure are mean value and variance of material flow´s intensity. These figures refer specifically to the (variable or constant) amplitude of the system load. The regularity of the intensity is determined by the load-oriented coefficient of variation. It provides reliable information on the variability of the load resulting from a changed product and / or quantity mix.

If the production system is fundamentally capable of dealing with the upcoming production program, the order-induced average system load as estimated value of processing times characterizes the dynamics of the load situation. It combines the mix of the production task and the intervals of the specific customer demands. If both aspects are differentiated, on the one hand the workload for each order as well as the average time span between the respective customer orders result. However, the effect on the production system is only apparent when considered aggregated.

C. Functional Structure

An orientation of the functional production structure on the basis of a different product characteristic (demand profiles) leads to a segmentation of the production in bestseller (quantity focus) and exotics (diversity focus) [6]. The distinction with regard to a product specification requires a certain diversity of the product range with divergent continuity and contrasting order volumes, i.e. the primary structuring takes place according to the type of production demand. Both the assessment of quantities as well as product changes are thus of central importance for evaluating the suitability of an existing or planned production structure.

The quantity index indicates the average number of products for each order. The average order volume combined with the corresponding coefficient of variation expresses the quantity mix of a production program. In contrast, the variance index represents the number of product variants of an organizational unit (product mix). In this case, nominally different types of orders can be combined in the form of identical product families due to their process-related system load.

An increasing quantity index and a decreasing variation index indicate a concentration of the production program on a decreasing number of rather homogeneous products. Similarity is expressed particularly by a reduction of the coefficient of variation. On the contrary, the organizational unit tends to a broader and more heterogeneous product spectrum if the trends are reverse. By using these indicators, a decisive characterization of the production program is carried out which takes account of both the quantity and the product mix. Thus, a decision regarding the basis segmentation of the production system into organizational units is enabled.

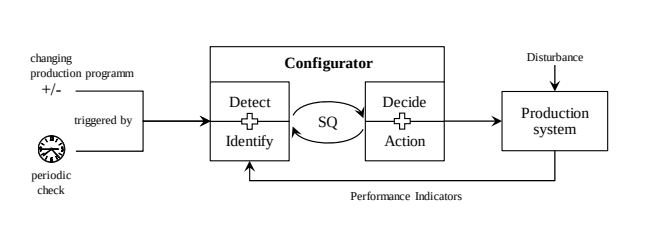
III. Assessment of the Structural Quality

A. Basic Process Model

Derived from control theory, we illustrate our re-configuration model as a block diagram of a closed-loop control system (Fig. 2). The object of observation is a production system with a planned or already installed production structure as described above. Typically, relevant performance indicators are recorded during system operation. Such performance indicators

along with sufficient input variables are necessary for an overall re-configuration or adjustment of production structures. Starting point for such a re-configuration is a change to the production program. Such changes affect variations in the type and quantity of the products to be manufactured in relation to the original design specifications.

Fig. 2. Configuration Procedure as closed-loop system.



SQ-based structure evaluation is performed in the configurator, from which necessary adaptation measures can be derived. But unlike the actual control theory, such measures refer to the change of the control system, i.e. the present structure of the production system. Thus, the following integrated configuration procedure is based on human decisions. However, proper indicators and detailed analysis of the Structural Quality of the production system support the detection of changes and the adopting of appropriate measures.

B. Integrated Configuration Procedure

The configurator carries out a multi-stage evaluation and decision-making process. At first, changes of the production program are detected by a preliminary evaluation. Moreover, the amount of value alternation facilitates decisions about further steps. After this rough assessment, a detailed analysis is based on Structural Quality [3]. The more in-depth evaluation procedure enables an identification of significant shifts within the production program. Thereby planning and design activities are taken into account simultaneously. As a result, proper measurements in terms of structural adaptions are revealed. This allows the planner both, to evaluate the suitability of the current production structure and to identify adequate structural solutions in case the present production system is not sufficient regarding the pending tasks anymore. Thus, the integrated configuration procedure is an iterative evaluation and decision-making process.

Initially, a preliminary assessment of the effect of a changed production program on the already installed production structure is conducted. Thus, the planner can detect if there are certain alterations within the production program. If the limits defined by the planner in advance are adhered to, then no action is necessary. Otherwise, further evaluation steps are appropriate. However, if the indicator values go beyond the defined thresholds, then a detailed structure evaluation is carried out using SQ-determination. Based on this SQ-determination, final decision about specific changes is made. This refers to both, temporal and spatial aspects of the production structure.

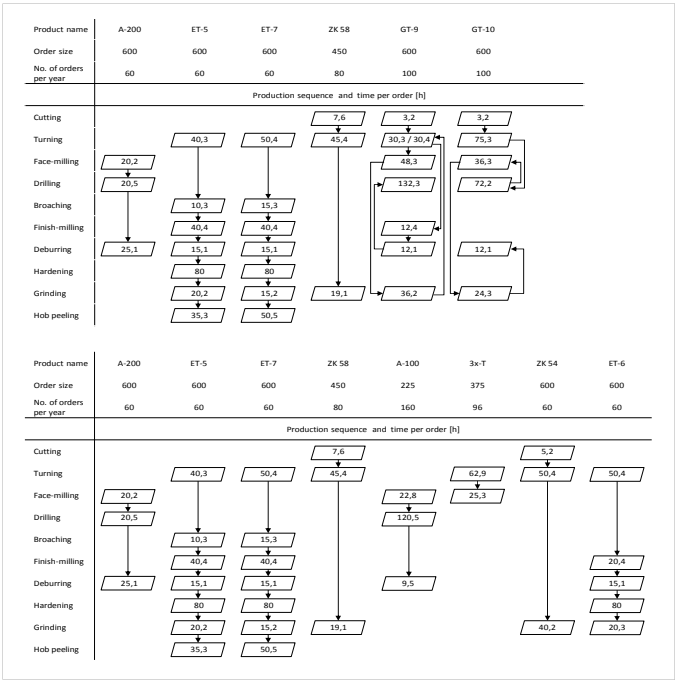
The determination of the structural quality (SQ) within the integrated configuration procedure is an event-driven process, triggered by a changing production program. In this case, the vector SQ serves as an important criterion for allowing the planner to weigh the need for changes in the production structure. Structural decisions are consequently long-term decisions regarding the general principle. For this reason, a periodic check of the functional structure also is recommended. Then, the configurator runs as a time-discrete application.

IV. Case Study

A. Indicator-based Evaluation of Production Structures

This section presents the use of SQ by means of case study data, whereby the Basic Process Model constitutes the framework. In addition to the integrated SQ assessment for the evaluation of the spatial and temporal aspects of production structures (event-driven process), the periodic check of the functional structure of a production system (time-discrete process) is also shown using the Integrated Configuration Procedure. According to the presented approach, the mode of operation of the configurator is described; both for the evaluation of the spatial and temporal structure components (event-driven mode) as well as the assessment of the functional structure of the given system (time-discrete mode).

Fig. 3. Former production program (top) and actual production program (bottom).



Based on a certain production program, production structure was installed. Currently, a change in this production program is apparent, as shown in Fig. 3. The relevance of this change is analyzed regarding the need for structural adaptation.

Consequently, at first the preliminary assessment of relevant indicators is executed. In order to assess the complexity of the material flow, we use two indicators: The degree of linearity takes into account the number of return flows in contrast to the prevalent sequence of operations, while the standardized degree of cooperation determines the number of material flow relations existing between the workstations [6]. Minimum complexity of the material flow exists when the degree of linearity assumes the

value of 1 and the standardized degree of cooperation has a value of 0. For values in the range $\kappa > 1$, more than one workstation is connected to more than two other workstations; this arrangement tends to be a job-shop structure.

In our example, the standardized degree of cooperation amounts 0.92 (previous: 1.29). The degree of linearity amounts 1 (previous: 0.85). Thus, after the change of the production program, a consistently forward-directed sequence of operations is present. With a view to the standardized degree of cooperation, a production line appears to be suitable. Thus, the complexity of material flow decreases. According to the change of the indicator values, the production structure gradually migrates from a job shop production towards a production line.

The temporal structure takes account of the workflow, i.e. the logical and temporal linking of the partial tasks of production [6]. The sequence of relations between the system elements (workstations) expresses the temporal component of the processes running in the production system. On the basis of the intensities of the relations as dynamic variables, different temporal courses can be characterized.

Parameters of the system-related temporal structure are mean value and variance of material flow´s intensity. These figures refer specifically to the (variable or constant) amplitude of the system load. The regularity of the intensity is determined by the load-oriented coefficient of variation. It provides reliable information on the variability of the load resulting from a changed product and / or quantity mix.

If the production system is fundamentally capable of dealing with the upcoming production program, the order-induced average system load as estimated value of processing times characterizes the dynamics of the load situation. It combines the mix of the production task and the intervals of the specific customer demands. If both aspects are differentiated, on the one hand the workload for each order as well as the average time span between the respective customer orders result.

In our example indicators for the temporal aspect of the production structure show an ambiguous development. The mean workload remains approximately constant. Accordingly, the capacity of the production system and thus the system composition is still sufficient. In contrast, the coefficient of variation alters from 98% to 132%. Diversity of process times increases, which is indicative for a shift towards a job shop production. This contradicts the development of the other indicators.

B. Detailed analysis of Production structures

Finally, we present the results of the detailed structure evaluation based on the SQ-determining. For a detailed SQ-assessment, we refer back to known (ideal-typical) arrangement algorithms for minimizing transport effort. For the purpose of a comparison of future and actual structure types a visualization in form of scaled block-layouts is helpful. Regarding the material flow, the transport effort decreases from 59,023 to 51,809 transport units per year. Based on the previous production program which indicates a job shop production, spatial component of Structural Quality decreases from 1.9 to 1.7. In this context, the ideal spatial arrangement was used as normalization value of 100% transport effort. The shift from a job-shop structure to a production line reveals a potential improvement of about 13.9%.

The quality of Temporal Structure refers to the operation point of the production systems which is derived from three logistic performance indicators: throughput time, output rate and work in progress (WIP). To sum up, all figures decrease when retaining the current control strategy: CONWIP (Constant Work-in-Progress). WIP decreases from an average work load of 5719 h to 3438 h, output rate declines from 338.3 h per SCD to 313.2 h per SCD, and mean throughput time lowers from 17.3 days to 11.3 days. While a reduction of WIP and lead time represents an improvement, the decrease of the output rate reveals a less effective system behavior. However, the simulation-based assessment of operating points shows that the existing system composition is still capable to fulfill respective production tasks. Thus, the Temporal Quality improves from 2.7 to 1.4.

C. Findings and Discussion

The case study shows that the presented approach facilitates the detection and decision making regarding structural adaptions of production systems. Due to the ambiguous conclusion derived from the indicator´s development, a detailed analysis based on the SQ-determining was inevitable. Further evaluation steps revealed a shift from a job-shop structure to a production line. In this context, the Spatial Quality improves from 1.9 to 1.7. Regarding the temporal structure, different control strategies may be investigated. However, WIP level and throughput time already improve, when retaining the current CONWIP system. Output rate also is sufficient. Thus, the Temporal Quality improves from 2.7 to 1.4.

However, possible segmentation of the organizational unit may require other structural solutions respectively. In this case, the integrated configuration procedure and in particular the SQ-determining is conducted iteratively until an adequate segmentation including suitable production structures were found.

V. Conclusion and Outlook

The integrated configuration procedure enables a detection of structural changes. Therefor a multi-stage approach is developed. If defined thresholds of indicators are exceeded, an in-depth analysis is inevitable. This detailed analysis comprises an evaluation of the suitability of current production structures by determining the Structural Quality vectors. During this procedure, planning and design alternatives are considered systematically. This enables the iterative improvement of structural suitability.

The case study highlights the application of the stepwise procedure. If indicator´s development shows an ambiguous trend or the thresholds of indicators are exceeded, an in-depth analysis of Structural Quality is recommended. Simulation studies facilitate the assessment of different structural variants. This is the only way to consider the dynamic behavior and performance of an altered production system.

References

[1] Tompkins, JA. Facilities Planning. Hoboken, N.J.: Wiley; 2010.

[2] Hildebrand T, Mäding K, Günther U. Plug + Produce – Gestaltungsstrategien für die wandlungsfähige Fabrik. IBF. TU Chemnitz; 2005.

[3] Bergmann U., Heinicke, M. Resilience of productions systems by adapting temporal or spatial organization. Procedia CIRP. Amsterdam: Elsevier, 2016;57:183-188.

[4] Costanzo F, Kanda Y, Kimura T., Kühnle H, Lisanti B, Srai JS, Thoben KD, Wilhelm B, Williams PM. Enterprise Organization and Operation. In: Grote KH, Antonsson EK, editors. Springer Handbook of Mechanical Engineering. New York, N.Y.: Springer; 2009. p. 1267–1359.

[5] Nyhuis P, Wiendahl HP. Fundamentals of Production Logistics Theory, Tools and Applications. Berlin: Springer; 2014.

[6] Kühnle H. Lecture Notes in Manufacturing Systems Design and Manufacturing Process Organisation. Göttingen: Cuvillier; 2017

[7] Chandler AD Jr. Strategy and Structure: Chapters in the History of the American Industrial Enterprise. Cambridge, MA: MIT Press; 1962

[8] Wiendahl HP, ElMaraghy HA, Nyhuis P, et al. Changeable manufacturing - classification, design and operation. CIRP Annals. Manufacturing Technology 56 (2007), No. 2, pp.783-809

[9] Chase RB, Jacobs FR, Aquilano NJ. Operations management for competitive advantage. 10. ed., internat. ed. Boston [etc.]: McGraw-Hill-Irwin; 2004.

Toward a generic method to design reconfigurable manufacturing systems

(Presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Khaled LAMECHE
Institut de recherche technologique Jules Verne
IRT-VJ
Nantes, France
khaled.lameche@irt-jules-verne.fr

Khalid KOUISS
Institut Français de la Recherche Mécanique Avancée
IFAM
kkouiss@ifma.fr
Clermont-Ferrand, France

Najib Najid, Pierre CASTAGNA
Laboratoire des sciences du numérique de Nantes
LS2N
Nantes, France
najib.najid@univ-nantes.fr
pierre.castagna@univ-nantes.fr

Abstract— Nowadays, the market is characterized by a high level of competitiveness and very frequent and sudden variations in the production context. The critical review of the existed manufacturing paradigms which are the dedicated manufacturing lines DMLs and the flexible manufacturing systems FMSs reveals that these systems are not capable of fulfilling the requirements imposed by the actual market; these requirements are mainly resumed in cost, quality and reactivity. Therefore, the need for a new manufacturing paradigm that could fulfill these requirements has arisen. Reconfigurable Manufacturing System or RMSs is this new paradigm; it is supposed to be reactive enough to cope with the sudden changes of the market while keeping the products quality high with a low cost. The main challenge in RMS is their design. Most of the proposed methods in the literature do not address the RMS design issue as a whole; they treat just part of the problem. Hence, as a contribution, we propose in this paper, a generic RMS design methodology based on systems engineering SE.

Keywords—Reconfigurable Manufacturing Systems; Systems Engineering; ISO/IEC/IEEE15288; Modularity;

I. Introduction

Nowadays, the market is undergoing increasingly, unpredictable, and frequent changes caused mainly by global competition. In order to survive and to be efficient in this new constantly changing environment, a common set of requirements has been established. These requirements have to be fulfilled by the manufacturing system; they can be briefly summarized as follows: short lead-time, more variants, low and fluctuating volumes, low price. Quality, safety, sustainability and other requirements have to be considered too [1]. Hence, to remain competitive, companies must design manufacturing systems that not only produce low-cost and high-quality products, but also enable a rapid response to market changes. The conducting critical review of the literature over the recent years shows that dedicated manufacturing lines (DMLs) and flexible manufacturing systems (FMSs) cannot satisfy all these requirements. Therefore, in the actual market, it is crucially needed more than ever a manufacturing system that can react cost-electively to production context changes. Reconfigurable manufacturing systems RMSs, whose components are reconfigurable machines and reconfigurable controllers, as well as methodologies for their systematic design and rapid ramp-up, are new manufacturing paradigm proposed as a solution to the actual needs.

During the recent years, a lot of works have been done to achieve the ultimate goal which is to create a 'dynamic factory' that can quickly adapt its production capacity and variety while maintaining a high level of quality and affordable cost. The review of these works shows that several RMSs design issues were addressed and many of the proposed solutions address solely part of the problem, for more details, a general overview of methods and supportive tools to design RMS was conducted in [2]. Taking into consideration the large size and the complexity of manufacturing systems, it is quite obvious that is difficult to establish a global, standard and a detailed method that addresses all the RMS design issues. To adrees this problem, Systems Engineering (SE) which is a descipline to develop complex systems could be applied to design RMS. It brings together the best practices of industrials in developing successful complex systems, optimizing costs and respecting deadlines. These good practices are generally: processes, methods or methodologies which are the outcomes of long experience in the development of complex systems. Therefore, in this paper we propose, as a contribution, a generic and a global methodology based on systems engineering principals to design RMS.

The rest of this paper is organized as follows: In section two, we will give a brief introduction about RMSs. In section three, we will present the SE discipline and the standard ISO/IEC/IEEE15288 that was used as a framework or a baseline to develop a general methodology to develop RMS. This methodology will be exposed and discussed in section four. This paper will be concluded by discussing the lacks and

the weaknesses of the proposed methodology, and as a response, some perspectives for future work will be proposed.

II. Reconfigurable Manufacturing Systems

According to [3], [4], [5], [6], an RMS is a manufacturing system whose physical and logical structures at all levels of system composition can be changed rapidly and cost effectively to adjust the production capacity and functionality around a family of products in response to sudden changes in the market. The RMS composition is based on standard physical and logical modules that can be added, modified, rearranged, interchanged or replaced quickly and reliably as needed to respond to changing requirements. The location of these modules can be changed as well at the shop-floor level to achieve the configuration that fills the requirements of the production context. RMS provides a customized flexibility for a particular products family and it will be open architecture so that it can be improved, updated and reconfigured rather than replaced.

During its operational life, an RMS changes its configuration following the occurrence of reconfiguration trigger events or RTEs. The system is in its initial configuration when deployed for the first time. This configuration has to be changed when it is necessary; the passing process from one configuration to another is called the reconfiguration process. This process requires hard/physical and soft/logical modifications. The soft/logical reconfiguration is usually driven by the hard/physical reconfiguration. The physical reconfiguration includes system structure modification as well as modifications concerning cells, machines, handling equipment, storage equipment, tools and fixture and all the physical modules that compose the system. Soft/logical reconfiguration include controllers programming, production re-planning, and re-scheduling, operator training, etc. [6]

Manufacturing systems are complex systems and reconfigurability brings an additional complexity to these systems. So, addressing the design of RMSs is a large issue, and it cannot be addressed entirely in this limited study. In this paper, we will not address the design of RMSs at all composition levels, our work will be limited to the shop-floor level and all the composition levels that fall under this level.

III. System Engineering

Complex industrial systems are difficult to design and are composed of several subsystems and components of different technologies. In order to manage this complexity, a new discipline known by the name of Systems Engineering SE was created and popularized in the industrial field. This discipline brings together the best practices of industrials in developing successful complex systems, optimizing costs and respecting deadlines; These good practices are generally: processes, methods or methodologies which are the outcomes of long experience in the development of complex systems. Many definitions have been given to this discipline. In the standard IEEE1220 [7], the SE is defined as a methodological, cooperative, and interdisciplinary approach that encompasses all appropriate activities to design, develop, and verify a set of products, processes and human expertise providing an economical and efficient solution to the needs of stakeholders and acceptable by all. The International Council on System Engineering INCOSE [8] defines the SE as an interdisciplinary approach and means that enable the realization of successful systems. SE is not a new method that would replace the already existed methods, but it is a comprehensive and systematic approach capable of bringing together the existing methods and approaches such as functional analysis, value analysis, operational safety, methods of specific engineering, etc. while benefiting from their experience feedback.

SE associations such as the INCOSE and major standardization organizations such as the American National Standards Institute ANSI and the Institute of Electrical and Electronics Engineers IEEE have worked together to establish SE standards that describe the entire course of projects regarding processes by listing the necessary activities for the smooth running of the project. The expected result of these activities and the requirements of their application complete these descriptions. The most used SE standards are: IEEE 1220: Standard for Application and Management of the Systems Engineering Process [7], EIA/ANSI 632: Processes for Engineering a System [9] and ISO/IEC/IEEE 15288: Systems and Software Engineering - System Life Cycle Processes [10].

The ISO/IEC/IEEE 15288 is the most mature and recent standard in SE discipline. This standard has evolved from many other older standards. It covers the entire system life cycle starting from the concept stage and ending by the retirement stage going throughout the development, production, utilization and support stages. The standard consists of four categories of processes (Technical processes, Project processes, Agreement processes, and Enterprise processes); each process contains a set of activities to do and a set of outcomes to deliver. These processes can be invoked, as required, at any time throughout the system lifecycle and there is no definitive order in their use. ISO/IEC/IEEE 15288 is a common SE process framework. It forms an abstract framework to develop any technical or socio-technical system. In the proposed methodology, this standard will be used as a framework to which different methods and techniques will be added to get a generic and global methodology to develop RMSs.

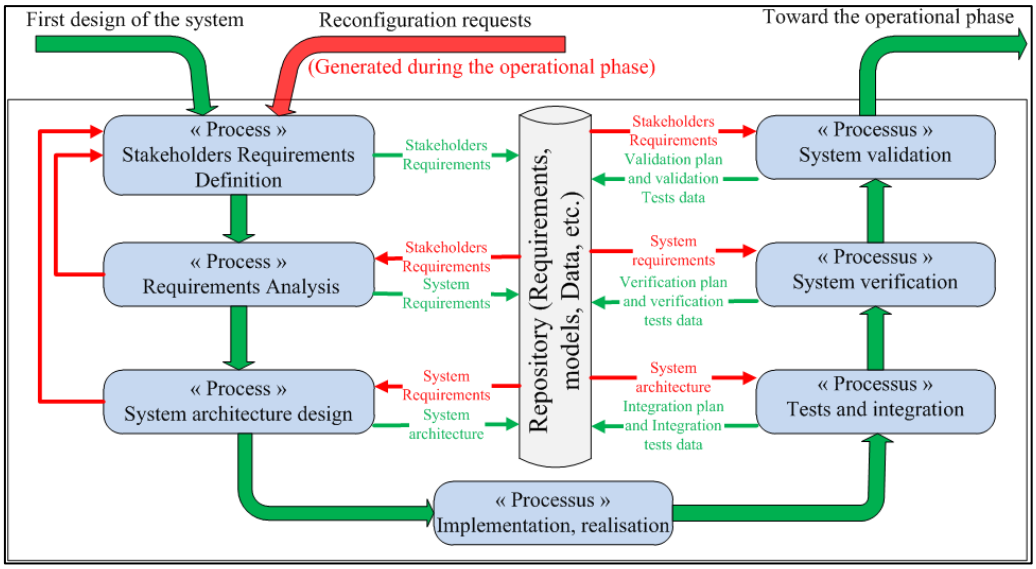
IV. RMS Design methodology

The Fig. 1 provides a general overview of the methodology processes, their execution sequence and different links between them. In this methodology, to design an RMS, we start with defining the stakeholder needs, and we end with the system validation in its operational environment. The methodology does not implement all the technical processes

proposed in the standard ISO/IEC/IEEE15288. The adopted processes are those presented in Fig. 1 except the implementation and realization processes. They are executed in the order shown by the green path (Fig. 1), but the methodology is agile and returns to previous or already completed processes to make the necessary corrections are always possible.

Development of large and complex systems like manufacturing systems, including RMS, involves people and teams from different fields. Communication between all these actors is very important. Thus, sharing project data (requirements, specification, models, verification tests, validation tests, test reports, etc.) is an effective solution to ensure the communication between those actors. In Fig. 1, data to be exchanged between different processes is stored in a central repository. The execution of each process will require a set of data and will consequently generate other data that will be sent back to the repository to be used by other processes.

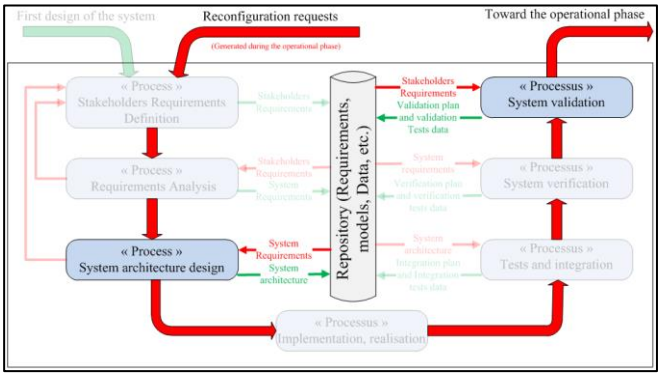
Fig. 1 : SE-based methodology to design reconfigurable manufacturing systems.



The proposed methodology follows a modified version of the 'V' Cycle. When the RMS is deployed for the first time, it will be in its initial configuration and during its operational phase, the system should be reconfigured to adapt to its production context. Thus, each time the system reconfiguration is necessary, the methodology is resumed to review the system design.

In the case where the requirements of the new production context could be fullfilled solely by the system reconfiguration, with an already designed solution (physical modules and logical modules), only the configuration selection will be necessary. Therefore the methodology will be resumed starting from the system architecture design process, not all the detailed process will be considered, but only the system configuration which includes system layout design as well as the configuration of the system modules. More than one configuration that can cope with the new production context could be found, in this case, an evaluation and comparison phase is needed to select the best configuration according to some criterion as the production cost, the delivery time, etc. The simulation could be used to assess, compare, and select the future configurations of the system, for more details see [11]. Once the future configuration is selected; it should be deployed and validated. In conclusion, only the processes: system architecture design and system validation will be considered in this case (Fig. 2).

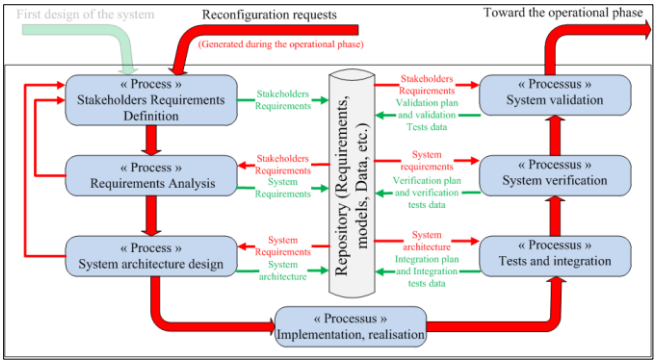
Fig. 2 : System reconfiguration without adding capabilities to the system.



It is important to consider the case when all the possible configurations of the system cannot satisfy the requirements of the new production context. In this case, the system has to evolve. Therefore other capabilities should be added to the system. Thus, the methodology should be resumed from the beginning where additional requirements will be added,

analyzed, and transformed to additional modules or options that will improve the system capabilities. All the methodology processes will be considered in this case (Fig. 3).

Fig. 3 : System reconfiguration with adding capabilities to the system.



In the case of developing complex and large systems, decomposition is a very helpful technique to manage the complexity of these systems. Manufacturing systems at the shop floor level could be decomposed to four sub-systems: 1) the added-value system, which is the part responsible for adding value to the raw material (machining system, assembly system, etc., 2) the storage system, 3) the material handling system, 4) and the control system. When developing manufacturing system, we start by developing its added-value subsystem. The storage subsystem and material-handling subsystem are developed to support the added-value system, they are considered as support functions. The control system is developed to drive and supervise all these sub-systems. This decomposition in subsystems and design precedence constraints could be applied to RMSs. Therefore, an RMS can be developed by developing its subsystems which should be reconfigurable too. The RMS features and design principals should be applied to develop its subsystems.

A. Stakeholder requirements definition

The first process to execute in the proposed RMS design methodology is the 'stakeholders requirements definition process.' According to the standard ISO/IEC/IEE15288, the purpose of this process is to define the stakeholders' requirements for a system that can provide the capabilities needed by users and other stakeholders in a defined environment. It identifies stakeholders involved with the system throughout its life cycle and their needs. It analyzes and transforms these needs into a common set of stakeholders requirements that express the intended interaction the system will have with its operational environment and that are the reference against which each resulting operational capability is validated.

In the case of manufacturing system, the main stakeholder is the end user who will use this system to manufacture the desired products. Once the stakeholders are defined, all the system functions should be defined, the required functions are mainly the added-value functions (machining functions; milling, cutting, drilling, assembly functions, or any functions that add value to raw materials along the manufacturing process). These functions could be identified from the process plan of each product that will be made by the manufacturing system. In the case of RMS, the system is designed around a family of products with enough flexibility to make all these products using reconfiguration capabilities. The design of RMS will be easy if these products are grouped into families according to their common features like similarities in process plan operations.

The added-value subsystem should be designed first, therefore, after grouping products family, the process plan operations should be allocated to the future cells, machines or station of the system. Storage subsystem should be defined to support the different needs of the added-value subsystem. The functions of the handling material or the logistic system will be defined based on the material and information flows between the different parts of the added value subsystem and the storage subsystem. Interactions between the system and its users should be defined. In the case of manufacturing systems, operators are in direct interactions with the system. They will interact with different resources of the system like operating machines, filling buffers, driving transportation means, setting the control system, etc. all these interactions should be defined during this process.

B. Requirements analysis process

Requirements analysis process aims to transform the stakeholder requirements into a technical view of a solution that meets the operational needs of the user. The process creates a set of measurable system requirements that specify, what characteristics, attributes, and functional and performance requirements the system is to possess, to satisfy stakeholders requirements. As far as constraints permit, the requirements should not imply any specific implementation. [10].

C. System Architecture Design

Once the system is well-defined and specified, we start to design its architecture; often we start with the system functional architecture then the physical architecture. The defined requirements must be allocated to the various functional elements of the system, and these will be allocated to the physical elements of the system. Traceability links must be defined between the requirements and elements of the system architecture. Data and models of the system architecture will be stored in the repository. The integration plan (assembly and testing) of the system will be based on these models and Data.

1) Modularity

During the system architecture design process, modularity, a very important feature of RMS, that should be

taken into consideration. The objective is to realize a modular architecture where basics modules, functional and physical, are clearly defined as well as the interfaces between them. In [12] three basic types of modularity are mentioned: modularity in design, modularity-in-production, and modularity-in-use. The one that should be considered during the system design process is the modularity in use because the objective is to design a set of standard modules that could be used to form different configurations of the system to adapt easily and cost-effectively to the system's changing context. In [13] modularity in use was defined as the use of standard units or modules to create products/systems variants. It aims at the identification of independent, standardized or interchangeable units/modules to satisfy a variety of functions/changing context. Therefore, a system can be modular in use if the end-user can mix and match elements/modules to come up with a final configuration that suits his needs. The benefits of modularity are discussed in [14].

Modularity is a very important feature of RMSs [15] and it should be considered and integrated into the system design, at all composition levels, during the earlier phases of the development process. For this reason, developing methods and techniques to develop a modular system is very important to succeed in designing RMSs. Regardless the method or the technique used to attain modular architecture, there are some principals that have to be considered when designing a modular system, for more details see [14].

2) Simulation

Simulation is a very important step in the proposed methodology and it could be even unavoidable in the RMS design. It could be used during the development phase of the system to evaluate and compare different solutions for a particular problem. And it could also be used during the operational phase of the system to support the selection of its future configurations. When the system reconfiguration is needed, simulation could be used to evaluate and compare the possible future configurations of the system to select the best one. In [11], in parallel with the system development, virtual modules were developped. Each virtual module represents the future corresponding real module of the system. The development of these virtual modules has helped enormously to design the real modules. Furthermore, the developed virtual modules were grouped together to form a virtual library that will be used later during the system operational phase to build virtual models for different system configurations. This will help to compare these configurations and select the best one for each production context.

D. System Verification and Validation

1) Verification process

In the standard ISO/IEC/IEE15288, the verification process is described as the process that provides objective evidence that a system or system elements/modules fulfill its specified requirements and characteristics. The aim of this process is to identify the anomalies (errors, defects, or faults) in any information item (system requirements or architecture description), implemented systems elements, or life cycle processes using appropriate methods, techniques, standards or rules. The verification process provides the necessary information to determine the resolution of identified anomalies. It is used to check whether the system has been correctly implemented, i.e. the realized system respects the architecture designed and the specifications imposed.

In addition to the standard verification tests (safety and security tests, operational tests, robustness tests, etc.), reconfigurability tests should be performed on the system, as well as on its sub-systems and its modules. Each part of the system should be verified in its different configurations. In each configuration of the system, sub-system or module, standard verification tests should be performed. Also, the reconfigurations (transition from one configuration to another) should be verified for each part of the system as well as for the whole system. Based on the system requirements, a verification plan is drawn up with an already defined strategy. The results of the verification tests must be recorded in the data repository, and these results will be analyzed to identify the system's non-conformities with the system requirements.

2) Validation process

If all verification tests are passed, the system will be deployed, and the validation process can begin. The purpose of this process is to provide objective evidence that the system, when in use, fulfills its business or mission objectives and stakeholders requirements, achieving its intended use in its intended operational environment. [10]. Based on the needs of stakeholders, a validation plan is drawn up with an already defined strategy. The results of the validation tests must be recorded in the data repository; these results will be analyzed to identify the non-conformities of the system regarding the needs of the stakeholders. The validation process is used to check if the system meets the expectations of the stakeholders, i.e. it provides the services requested by the stakeholders in the defined operational environment.

V. Conclusion and perspectives

In this paper, we have briefly discussed the importance of RMSs, and highlighted that these systems are the most efficient for manufacturing companies to stay competitive in today's market which is characterized by globalization. Due to its changing structure (physical and logical) at different levels of system composition, the RMS is a complex system and its design is not an easy task. Most of the methods and techniques used to design complex systems could be used to design the RMS. Some modifications, adaptations, and improvements need to be done to apply these methods and

techniques on the RMS design. Thus, based on this idea, a global methodology based on the standard ISO/IEC/IEEE15288 to design RMS was proposed in this paper. This methodology assists the developers in their work from the stakeholders' needs definition until the system deployment and validation.

This work is part of the project (Système Transitique Agile et Robotisé) or STAR, which aims to develop agile and automated logistics system based on Automated Guided Vehicles AGVs. In this project, two industrial partners are involved, Faurecia which is an automotive supplier, and Daher which is a logistics subcontractor. The objective is to design a reconfigurable logistics system that should work for the two case studies. Obviously, for each case study, a different configuration of the logistics system will be used. This project is underway, after defining the stakeholders' needs, specifying the system, and designing its architecture, we are now at the stage of realizing the system (other parts of the system are still at the designing stage like the control system).Therefore, the results of this project cannot reported in this work yet. This will be done in future publications.

Currently, we are testing and improving the proposed methodology by integrating other activities and tasks specific to RMSs domain. Also, we are looking for techniques, methods, and tools to support the different activities of this methodology. The ultimate objective is to develop a Model Based System Engineering MBSE methodology, which is a methodology based on system engineering principals where the different tasks and activities are realized using mainly models. We are currently looking into these problems and hope to report the findings as well as the project STAR results in future publications.

Acknowledgment

This study is part of the STAR project managed by IRT Jules Verne (French Institute for Research and Technology in Advanced Manufacturing Technologies for Composite, Metallic and Hybrid Structures). The authors wish to associate the industrial and academic partners of this project; respectively DAHER, FAURECIA, BA SYSTEMES, and IRCCyN.

References

[1] Z. M. Bi, S. Y. . Lang, W. Shen, and L. Wang, "Reconfigurable manufacturing systems: the state of the art," *Int. J. Prod. Res.*, vol. 46, no. 4, pp. 967–992, Feb. 2008.

[2] A.-L. Andersen, T. D. Brunoe, K. Nielsen, and C. Rösiö, "Towards a generic design method for reconfigurable manufacturing systems," *J. Manuf. Syst.*, vol. 42, pp. 179–195, 2017.

[3] M. G. Mehrabi, A. G. Ulsoy, Y. Koren, and P. Heytler, "Trends and perspectives in flexible and reconfigurable manufacturing systems," *J. Intell. Manuf.*, vol. 13, no. 2, pp. 135–146, 2002.

[4] M. G. Mehrabi, A. G. Ulsoy, and Y. Koren, "Reconfigurable manufacturing systems: Key to future manufacturing," *J. Intell. Manuf.*, vol. 11, no. 4, pp. 403–419, 2000.

[5] Y. Koren, U. Heisel, F. Jovane, T. Moriwaki, G. Pritschow, G. Ulsoy, and H. Van Brussel, "Reconfigurable manufacturing systems," *CIRP Ann. - Manuf. Technol.*, vol. 48, no. 2, pp. 527–540, 1999.

[6] H. A. ElMaraghy, "Reconfigurable Process Plans For Responsive Manufacturing Systems," in *Digital Enterprise Technology: Perspectives and Future Challenges*, P. F. Cunha and P. G. Maropoulos, Eds. Boston, MA: Springer US, 2007, pp. 35–44.

[7] IEEE, "IEEE 1220 Standard for Application and Management of the Systems Engineering Process," vol. 1220. IEEE, New York, p. 101, 2005.

[8] C. Haskins, K. Forsberg, and M. Krueger, *Systems Engineering Handbook : A guide for system life cycle processes and activities*, no. August. INCOSE, 2007.

[9] EIA/ANSI, *EIA-632: Processes for Engineering a System*. Arlington: ELECTRONIC INDUSTRIES ALLIANCE, 1999.

[10] ISO/IEC/IEEE15288, *ISO/IEC/IEEE15288: Systems and software engineering – System life cycle processes*, 1st ed. Switzerland: ISO/IEC/IEEE, 2015.

[11] K. Lameche, M. N. Najid, P. Castagna, and K. Kouiss, "Utilisation de la simulation dans la conception des systèmes logistiques reconfigurables," in *11e Conférence Francophone d'Optimisation et Simulation- MOSIM'16*, 2016.

[12] C. Y. Baldwin and K. B. Clark, "Modularity in the Design of Complex Engineering Systems," in *Complex Engineered Systems: Science Meets Technology*, D. Braha, A. A. Minai, and Y. Bar-Yam, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 175–205.

[13] C. C. Huang and A. Kusiak, "Modularity in design of products and systems," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans.*, vol. 28, no. 1, pp. 66–77, 1998.

[14] K. Lameche, M. N. Najid, P. Castagna, and K. Kouiss, "Modularity in the design of reconfigurable manufacturing systems," in *The 20th World Congress of the International Federation of Automatic Control*, 2016.

[15] Y. Koren and M. Shpitalni, "Design of reconfigurable manufacturing systems," *J. Manuf. Syst.*, vol. 29, no. 4, pp. 130–141, 2011.

Modelling Flexible Manufacturing Systems with Power Constraints and Machine Switch on/off

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Sylverin Kemmoe-Tchomte
CRCGM EA 38 49
Université Clermont Auvergne
Clermont Ferrand, France
sylverin.kemmoe_tchomte@uca.fr

Pierre Fenies
Université Paris Ouest Nanterre La Défense
Nanterre, France
pierre.fenies@u-paris10.fr

Damien Lamy, Nikolay Tchernev
LIMOS UMR 6158
Université Clermont Auvergne
Aubière, France
nikolay.tchernev@uca.fr, lamy@isima.fr

Abstract - This paper deals with Flexible Manufacturing System in the context of the future's industry. The problem under study is the Flexible Job-shop which models various production systems. This study aims at scheduling operations efficiently by considering a power limitation. To this purpose, each operation has a power profile depending on the machine it is assigned to. Furthermore, in industry, machines are often left idle. This practice could lead to loss in the available power for the manufacturing system, especially when a machine requires a lot of power when it is not processing any operation. Hence, it is proposed to address the benefits of switch *on*/*off* in the Flexible Job-shop problem with power limitations. A mathematical formulation for this problem is presented in this paper. The results are promising and show that it is possible to schedule efficiently operations with power requirements in a production system.

Keywords - Flexible Job-shop; Power Limitation; Energy Efficiency; Machine Switch *on*/*off*.

I. Introduction

It is known that the industrial sector worldwide is responsible for consuming more than 50% of the global delivered energy per year [1]. Furthermore, these energy resources represent important expenses for companies hence, better practices should be applied to reduce their consumption. However, these problems are still unsolved, partly because of the gap between research and industrial needs [2]. To provide solutions to these problems, two decisions are possible: technological and/or organisational [3]. Technological measures could lead to strong reductions but are really expensive since they concentrate on new machines or production processes, whereas organisational measures are dedicated to the current system. These organisational measures include scheduling methods which can drastically improve the energy-efficiency of the production system. In the literature, the studies on scheduling with energy-saving objectives mainly concern total energy consumption [4][5][6]. However, another way to reduce costs related to power billings in production systems is to manage the production without exceeding a contracted power load (threshold) [7]. In this last paper, the authors explore a method to schedule efficiently tasks in a steel plant according to a given energy load profile. In this approach, the electrical consumption of the plant is sent to the electricity provider after what the given load profile is accepted or a new one is proposed to the plant owner. The load profile is then sequentially modified until a definitive one is accepted by both stakeholders. Once a load profile is chosen, penalties are applied for every over- or under-consumption. Respecting this contracted power load has a direct impact in power generation for suppliers as stressed in [8]. Such approaches can be seen as Demand-side Management [9], which consists to manage the load and consumption patterns of end users in order to reduce the production of energy on suppliers' side. With such an approach it is possible for a company to reduce the extra expenses related to the excess of instant power by strictly respecting its contract with the supplier. To this purpose, it is proposed to address the impact of power consumption (PC) on schedules in the context of a Flexible Job-shop manufacturing environment. In this paper, a general model of a Flexible Job-shop scheduling problem (FJSP) is given; its objective is the minimisation of total completion time (makespan) by taking into account the power consumption of operations and a power threshold. Furthermore, machines can be switched *on*/*off* to assess the performance of a production system including this kind of rules.

The remaining of this paper is as follows: in the next section a brief review of papers related to the field of energy efficient manufacturing is presented. In section 3, the problem and its mathematical model are presented, followed in section 4 by the results on generated instances. Finally, the last section is devoted to the conclusion and research directions.

II. Background and motivations

A lot of industrial problems consisting in minimising the makespan, total tardiness, and other time-oriented objective functions can be found in the literature. Only a few works deal with energy optimisation as an important objective in scheduling of production systems. However, researches concerning "Green Manufacturing" issues are increasing since 2011 [2]. A non-exhaustive review on the reduction of energy and/or power consumption in manufacturing systems through manufacturing scheduling is presented below.

In [10], an optimisation approach for a single machine where the objective is to minimise its energy consumption is explored by the authors. A mathematical formulation in order to minimise both the makespan and the total energy consumption (TEC) is proposed. The model takes into account different states of a given machine: idle, running, *on*/*off*. In [11], Time Of Use (TOU) pricings are taken into account with consideration of transitions between machine states in a manufacturing system involving one single machine. The authors developed a genetic

algorithm that could be used in extension of a Manufacturing Resource Planning System.

The authors of [12] proposed a mathematical model consisting in minimising the makespan, the carbon footprint and the peak power load in a Flow-shop (FS) with machines having variable speeds. In [13] an approach consisting in minimising power consumption peaks (PC) in a Flexible Flow-shop (FFS) is proposed. This approach uses an Energy Aware Scheduling (EAS) module to adapt a given schedule obtained with an Advanced Planning and Scheduling system. The EAS does not modify the schedule, but optimises it from the energy consumption viewpoint by definition of new starting dates for the operations. The authors of [14] used a Genetic Algorithm (GA) combined with a Simulated Annealing in order to provide solutions to a bi-objective FS problem. Their objectives are the minimisation of the makespan and the TEC. In [15], a time indexed mathematical formulation is proposed in order to minimise the energy expenses under a TOU pricing and the carbon footprint in an eight level FS. An Ant Colony Optimisation algorithm to solve a bi-objective Hybrid FS problem including TOU rates where the objective is to optimise the makespan and the cost of the energy is given in [16].

A Job-Shop Problem (JSP) with machines able to process operations with different speeds and energy consumptions is studied in [6]. The authors study the correlation between makespan, energy and robustness of such a production system. According to them, an energy-efficient schedule is more robust and less sensitive to machine breakdowns. Authors of [4] handled minimisation of both the total tardiness and the TEC in a JSP by reducing the non-processing times of machines. A Non-dominated Sorting Genetic Algorithm is used to obtain a set of solutions to the problem. A Flexible Job-shop where the TEC should be minimised is studied in [17]. Their work is motivated by the heterogeneity of machines in flexible manufacturing systems, which could lead to strong improvements in energy efficiency. In [5] a JSP with different states of machines is considered. Several Key Performance Indicators (KPI) for energy-oriented manufacturing are proposed. A multi-objective GA optimises three of these KPI: makespan, TEC, and WEC (Worthless Energy Consumption). Different machine behaviour policies are explored to find the most promising way to improve the criteria.

Actually, few works are dealing with power threshold as an important constraint in scheduling and most of researches try to optimise the TEC or TOU. The objective of the present paper is to propose operational methods to provide schedules respecting a power threshold constraint while considering operations with specific power profiles. In order to address this problem, an improved mathematical model based on [18] is introduced. This model takes into account a power threshold and operations with peak power requirements. It is completely formalised in the following section.

III. Problem definition and mathematical formulation

The purpose of this study is to provide an effective schedule to a Flexible Job-shop problem considering power limitations.

The FJSP problem has been first introduced in [19]. It is formally formulated as follows: a set J of r jobs, $J = \{J_1, \dots, J_r\}$, must be scheduled on a set M of v machines, $M = \{m_1, \dots, m_v\}$. Each job $\{J_i\}_{i\in[1,r]}$consists in a number of n_i operations. Hence, n operations must be scheduled with $n = \sum_{i=1}^{r} n_i$. Each operation is noted $\{O_i\}_{i\in[1,n]}$. An operation O_i is allowed to be executed in any machine of a given set $M_i \subseteq M$. The processing time of an operation O_i on a machine $m \in M_i$ is noted $P_{i,m}$. Furthermore, each operation can be assigned to only one machine at a time. Note that several operations of a same job can be assigned to the same machine. Considering that there is an assignment and a schedule for all the operations, the starting date of an operation O_i is noted s_i. The objective is to minimise the total completion time of all the operations, also called makespan and noted c_{max}. Several assumptions are made for the classical problem, including availability of all machines, no preemption, and release dates of all jobs at time 0.

To better understand the following mathematical formulation, the FJSP can be modelled with a conjunctive-disjunctive graph G as stressed in [20], while power limitations in scheduling problems may be modelled as flows in the graph [18]. An example of an evaluated graph G for the classical Flexible Job-shop problem with operations already assigned to machines is given in Fig. 1.

For ease of understanding, in Fig. 1, operations are already assigned to machines. The objective of this research work is to schedule operations considering their power requirements. Since machines are heterogeneous, the power profile may vary according to the selected machine for processing an operation. In this paper, this power profile is varying along the manufacturing process: (i) at first, a peak power consumption is considered, followed by (ii) a nominal power consumption. Thus, an operation can be viewed as the succession of sub-operations corresponding to each specific power requirement during the production process [21]. The distinction between these two different power consumptions is made by splitting operations into sub-operations. Each operation O_i consists in two sub-operations $O_{i,1}$ and $O_{i,2}$ with durations $P_{i,1,m}$ and $P_{i,2,m}$ and consumptions $W_{i,1,m}$ and $W_{i,2,m}$ respectively. An example of splitting an operation (O_1 from Fig. 1) into two sub operations, and the power profile related to the machine assignment is given in Fig. 2.

In Fig. 2, power profiles are in exponent of the operations as follows: ${O_{i,k}}^{W_{i,k,m}}$. Also, it can be seen that the power profile and processing time of the operation O_1 are not the same whether it is assigned to machine M_1 or machine M_2. Furthermore, since operations must be executed without pre-emption, no-wait arcs are added between sub-operations to ensure these sub-operations to be executed without delay (the negative arcs).

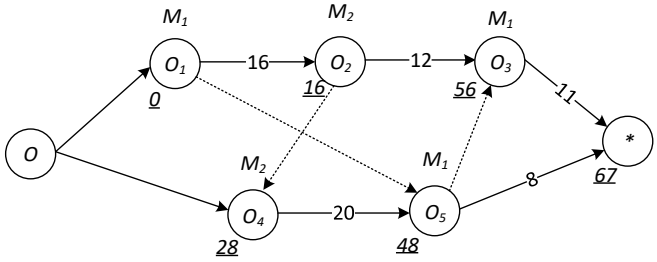


Fig. 1. An evaluated graph for the classical FJSP.

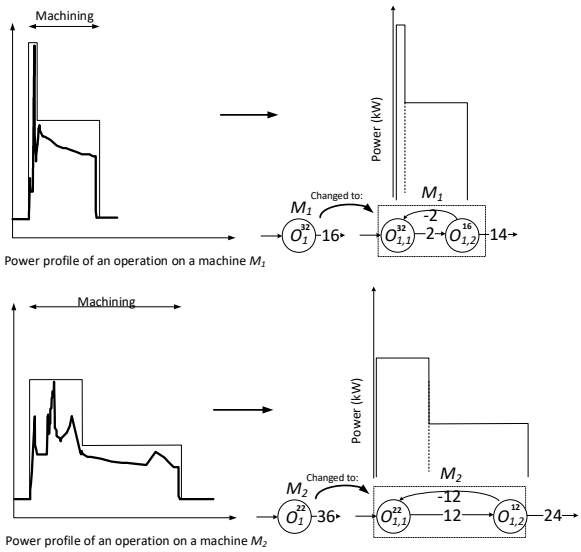


Fig. 2. Example of power profiles depending on the machine assigned to an operation.

For instance, in the top right of Fig. 2, if the starting date of operation $O_{1,1}^{22}$ is set to 0, then $O_{1,2}^{16}$ may start after 2 time units. However because of the negative arc, if $O_{1,2}^{16}$ starts later, let say at 20, then $O_{1,1}^{22}$ will have to start at 18.

Another assumption of the study is the possibility to switch *on*/*off* (or *off*/*on*) machines to allow operations to be scheduled earlier by benefitting from the power unused when the machine is shut down. These switch *on*/*off* operations depend on the order of operations on a machine. As it is not possible to know ahead of the optimisation process the assignment of the operations, and hence, if it is possible to switch *off*/*on* a machine between two consecutive operations, two additional sub-operations modelling switch *on*/*off* are added to the already split operation as shown in Fig. 3.

In Fig. 3(B), $O_{1,1}^{32}$ becomes $O_{1,2}^{32}$. The new $O_{1,1}^{16}$ is an operation modelling a possible switch *on*, and the operation $O_{1,4}^{4}$ models a switch *off*. The switch *on*/*off* operations require also power in order to be executed. After a machine is turned *on*, it will consume continuingly the power required to keep it active. For instance, suppose that machine M_1 needs 14 power units at the moment when it is turned *on*, only 2 power units will be returned to the system after $O_{1,1}^{16}$ is ended since the switch *on* procedure requires 16 power units. If the operation $O_{1,4}^{4}$ is executed it will require 4 power units. However, it will return 18 power units to the system (4 units plus 14 for the power used to keep the machine idle). With these assumptions, it is possible to generate a new graph G' of a problem as in Fig. 4, where the allowed power threshold is fixed at 46. In G' it will be impossible, considering the given assignment, to schedule operations $O_{1,2}^{32}$ and $O_{4,2}^{22}$ at the same moment, as they need 54 power units to be processed which is higher than the power threshold. As can be seen, operations $(O_{1,1}, O_{1,4})\ldots(O_{5,1}, O_{5,4})$ corresponding to switch *on*/*off* do not have power requirements nor durations. This is because it is not known yet if these operations (*on*/*off*) are executed or not, even though the machining operations are assigned to a machine. Finally, a solution is obtained by managing the power requirements with flows in the graph.

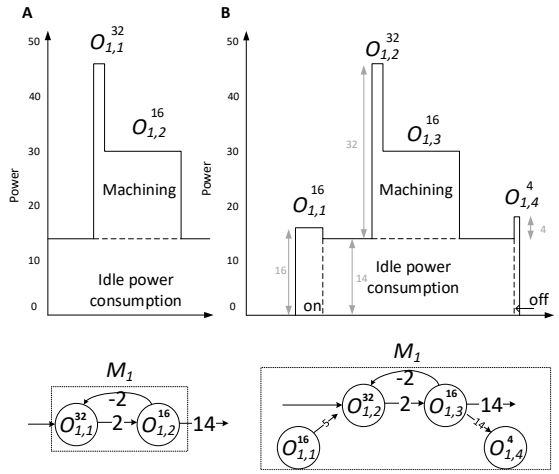


Fig. 3. Example of new added sub-operations for switch *on*/*off* policies ($O_{1,1}$ and $O_{1,4}$).

It is expected that operations can be scheduled earlier by using the power required to keep machines *on* while they are not in a machining phase. By switching a machine *off* during an idle phase between two operations, more power will be available for processing the other operations. The objective is then to schedule all operations considering a power threshold. This power threshold is a constraint given by the power supplier and it cannot be changed. Hence, when new orders arrive, the enterprise must schedule its production according to the contracted power load. The mathematical formulation of this problem is given as follows, starting with parameters:

Parameters

V	:	set of operations;
i, j	:	indices for operations;
J_i	:	job of operation *i*;
N_i	:	set of sub-operations for operation *i*;
M_i	:	set of available machine for operation *i*;
m	:	possible machine for an operation;
$P_{i,k,m}$	:	duration of sub-operation *k* on machine *m;*
W_{max}	:	maximum available power (Power Threshold) ;
$W_{i,k,m}$	:	power requirement of sub-operation *k* on machine *m;*
$W^b{}_m$	:	power requirement during idle phases of machine *m*;
H	:	a large positive number;

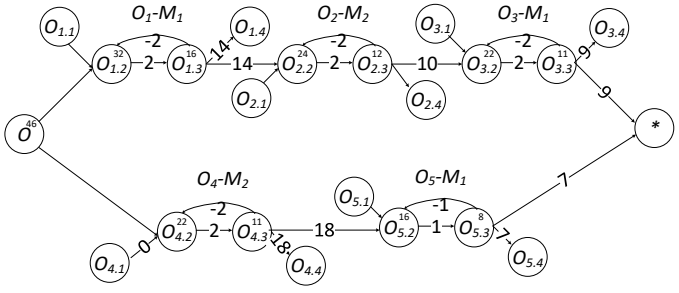


Fig. 4. An example of a non-evaluated graph representing a problem with power profiles for operations and switch *on*/*off* operations (Former graph in Fig. 1).

Variables

c_{max} : makespan of the schedule;

$s_{i,k}$: starting date of sub-operation *k* of operation *i*;

$p_{i,\mathrm{k}}$: processing time of sub-operation *k* of operation *i*;

$w_{i,k}$: power requirements of sub-operation *k* of operation *i*;

$w^b{}_{i,k}$: basis power 'lost', or returned to the system after a switch *on*/*off*;

$\varphi_{i,k,j,l}$: number of power units transferred from sub-operation *k* of operation *i* to sub-operation *l* of operation *j* ($\varphi_{0,0,j,l}$ is for the power units sent from the source node);

$y_{i,m}$: binary variable equal to 1 if machine *m* is assigned to operation *i*, 0 otherwise;

$x_{i,j,m}$: binary variable equal to 1 if operation *i* is scheduled before operation *j* on machine *m*, 0 otherwise;

$z_{i,k,j,l}$: binary variable equal to 1 if there is a flow from sub-operation *k* of operation *i* to sub-operation l of operation *j*, 0 otherwise;

$l_{i,j,m}$: binary variable equal to 1 if there is a switch *off*/*on* between the end of operation *i* and the start of operation *j*, 0 otherwise;

Mathematical Formulation

$$M\text{in } c_{max} \quad (1)$$

Subject to:

Flexible Job-shop related constraints:

$$c_{max} - s_{i,|\mathrm{N_i}|-1} - p_{i,\mathrm{N_i}-1} \geq 0, \forall i \in V \quad (2)$$

$$\sum_{m\in M_i} y_{i,m} = 1, \forall i \in V \quad (3)$$

$$p_{i,k} - \sum_{m\in M_i} y_{i,m} P_{i,k,m} = 0, \forall i \in V, \forall k \in \{2,3\} \quad (4)$$

$$2x_{i,j,m} + 2x_{j,i,m} - y_{i,m} - y_{j,m} \leq 0, \forall (i,j) \in V, \quad \forall (i,j) \in V, \forall m \in M_i \cap M_j, J_i \neq J_j \vee i < j \quad (5)$$

$$\sum_{v\in M_i, v\neq m} x_{0,i,v} + y_{i,m} \leq 1, \forall i \in V, \forall m \in M_i \quad (6)$$

$$\sum_{j\in V^*, J_j \neq J_i \cup i<j} \sum_{m\in M_i\cap M_j} x_{i,j,m} = 1, \forall i \in V$$
$$\sum_{i\in V^0, J_j \neq J_i \cup i<j} \sum_{m\in M_i\cap M_j} x_{i,j,m} = 1, \forall j \in V \quad (7)$$

$$\sum_{i\in V, m\in M_i} x_{0,i,m} \leq 1, \forall m \in M \quad (8)$$

$$s_{j,2} - s_{i,|\mathrm{N_i}|-1} - p_{i,|\mathrm{N_i}|-1} \geq 0, \forall (i,j) \in V, J_i = J_j, i < j \quad (9)$$

$$s_{i,\mathrm{l}} - s_{i,\mathrm{k}} - p_{i,\mathrm{k}} = 0, \forall i \in V, \forall (k,l) \in N_i \backslash \{1, |N_i|\}, k < l \quad (10)$$

$$s_{i,k} - s_{i,l} - p_{i,l} \geq 0, \forall i \in V, \forall (k,l) \in N_i, k > l \quad (11)$$

$$s_{j,1} - s_{i,|\mathrm{N_i}|} - p_{i,|\mathrm{N_i}|} + H(1 - x_{i,j,m}) \geq 0, \forall (i,j) \in V, \quad J_i \neq J_j \vee i < j, \forall\, m \in M_i \cap M_j \quad (12)$$

Power related constraints:

$$w_{i,k} - \sum_{m\in M_i} y_{i,m} W_{i,k,m} = 0, \forall i \in V, \forall k \in \mathrm{N_i} \backslash \{1, |\mathrm{N_i}|\} \quad (13)$$

$$\sum_{j\in V, k\in N_j} \varphi_{0,0,j,k} \leq W_{max} \quad (14)$$

$$\varphi_{0,0,j,l} + \sum_{i\in V} \sum_{k\in N_i} \varphi_{i,k,j,l} - w_{j,l} = 0, \forall j \in V, \forall l \in N_j \quad (15)$$

$$\sum_{j\in V} \sum_{l\in N_j} \varphi_{i,k,j,l} - w_{i,k} \leq 0, \forall i \in V, \forall k \in N_i \backslash \{1, |N_i|\}$$
$$\sum_{j\in V} \sum_{l\in N_j} \varphi_{i,1,j,l} - w_{i,1} + w^b{}_{i,1} \leq 0, \forall i \in V$$
$$\sum_{j\in V} \sum_{l\in N_j} \varphi_{i,|N_i|,j,l} - w_{i,|N_i|} - w^b{}_{i,|N_i|} \leq 0, \forall i \in V \quad (16)$$

$$\varphi_{i,k,j,l} \leq H.z_{i,k,j,l}, \forall (i,j) \in V, \forall k \in N_i, \forall l \in N_j \quad (17)$$

$$z_{i,k,j,l} \leq \varphi_{i,k,j,l}, \forall (i,j) \in V, \forall k \in N_i, \forall l \in N_j \quad (18)$$

$$s_{j,l} - s_{i,k} - H.z_{i,k,j,l} - p_{i,k} \geq -H, \forall (i,j) \in V, \quad \forall k \in N_i, \forall l \in N_j \quad (19)$$

$$z_{i,k,j,l} = 0, \forall (i,j) \in V, \forall k \in N_i, \forall l \in N_j, \quad J_i = J_j \wedge (j < i \vee l < k) \quad (20)$$

$$l_{i,j,m} - x_{i,j,m} \leq 0, \forall i \in V^0, \forall j \in V^*, \forall m \in M_i \cap M_j \quad (21)$$

$$x_{i,j,m} - l_{i,j,m} \leq 0, \forall (i,j) \in (0,V) \vee \forall (i,j) \in (V,*), \quad \forall m \in M_i \cap M_j \quad (22)$$

$$p_{j,1} - \sum_{i\in V^0\backslash j} \sum_{m\in M_i\cap M_j} l_{i,j,m} P_{j,1,m} = 0, \forall j \in V$$
$$p_{i,|N_i|} - \sum_{j\in V^*\backslash i} \sum_{m\in M_i\cap M_j} l_{i,j,m} P_{i,|N_i|,m} = 0, \forall i \in V \quad (23)$$

$$w_{j,1} - \sum_{i\in V^0\backslash j} \sum_{m\in M_i\cap M_j} l_{i,j,m} W_{j,1,m} = 0, \forall j \in V$$
$$w_{i,|N_i|} - \sum_{j\in V^*\backslash i} \sum_{m\in M_i\cap M_j} l_{i,j,m} W_{i,|N_i|,m} = 0, \forall i \in V \quad (24)$$

$$w^b{}_{i,k} - \sum_{j\in V^*\backslash i} \sum_{m\in M_i\cap M_j} l_{i,j,m} W^b{}_m = 0, \quad \forall i \in V^0, \forall k \in N_i, k = \{1, |N_i|\} \quad (25)$$

The first line (eq. 1) refers to the objective of the problem: minimising the completion time of all operations (the makespan). Constraints (2) give the expression of this makespan. Only the ending dates of processing jobs (i.e.: operations $o_{i,|\mathrm{N_i}|-1}$) are considered and not switch *offs* (sub-operations $o_{i,|\mathrm{N_i}|}$). Because of constraints (3), each operation must be assigned to one and only one machine. Constraints (4) set the duration of processing sub-operations once a machine is selected. Constraints (5-8) are adapted from the FJSP to have only one output/input arc for each set of sub-operations related to each processing operations. Constraints (5) define precedence constraints of operations, if they are occurring on the same machine. If one of the operation is not assigned to a machine, necessarily, $x_{i,j,m}$ and $x_{j,i,m}$ are valued 0. If both operation are assigned to the same machine ($y_{i,m} = y_{j,m} = 1$), hence, $x_{i,j,m}$ or $x_{j,i,m}$ may be equal to 1. Constraints (6) ensure that, if a machine is assigned to an operation then necessarily the ingoing arcs from the source node 0 to the operation, or the outgoing arcs from the operation to the sink node * and relevant to other machines are set to 0. Constraints (7) ensure that, for all processing operations, only one arc exists between two operations assigned to the same machine, which means that only one arc is entering an operation block (i.e.: all sub-operations of an operation), and only one arc is leaving such a block. Operations *0* can have several outgoing arcs, and no entering arcs; operation * may have several entering arcs but no leaving arcs. Constraints (8) ensure that, there is only one arc per machine going from the source node to all operations. If a machine processes no operation then there is no arcs relevant to this machine at all. Constraints (9-12) are adapted from the FJSP to consider sub-operations. Constraints (9) update starting dates of operations in order to respect the processing order of operations of a job. Constraints (10) ensure no preemption of sub-operations corresponding to a manufacturing process. Constraints (11) stipulate that process order of sub-operations of an operation *i* must remain the same, which also means that eventual switch *on* (respectively: *off*) are always done before (after) the beginning (end) of a manufacturing process on a machine. Constraints (12) adjust the

starting dates of operations that belong to different jobs but need the same machine operations.

All following constraints (13-25) are specific to the problem under study. Constraints (13) define the power actually needed to process sub-operations of operation *i* on machine *m*. The cases k=1 or k=$|N_i|$, relevant to switch *on*/*off* are handled separately in constraints 24. Constraint (14) avoid to exceed the power threshold when processing the operations as it cannot be allocated more energy to the operations than W_{max}. Constraints (15) ensure that the sum of energy flowing from all sub-operations and initial energy node (W_{max} kept in operation *0*) is equal to the energy needed to process the l^{th} sub-operation of operation *j*. Constraints (16) are divided into three different sets. The first one ensures that the sum of energy flowing from the considered sub-operation k to the other ones never exceeds the energy that was used for its processing. Here, all sub operations are considered however, operations $O_{i,1}$ are returning to the system less than what they consume because after a switch *on*, a machine stays idle and needs power anytime (second set); meanwhile, operations $O_{i,4}$ are returning to the system more than what they consume (third set). Constraints (17) ensure that if there is a power flow from the k^{th} part of operation *i* to the l^{th} part of operation *j*, then $z_{i,k,j,l}$ =1. If $z_{i,k,j,l}$=0, then no flow is possible from the k^{th} part of operation *i* to the l^{th} part of operation *j*. Constraints (18) stipulate that if there is no need of a flow from $O_{i,k}$ to $O_{j,l}$ ($\varphi_{i,k,j,l}$=0), then necessarily $z_{i,k,j,l}$ = 0. If $z_{i,k,j,l}$ = 1 then at least one unit of power goes from $O_{i,k}$ to $O_{j,l}$. Constraints (19) specify that no reverse flow is possible from the l^{th} sub-operation of operation *j* to the k^{th} sub-operation of operation *i*, if *i* and *j* belong to the same job and the l^{th} sub-operation of operation *j* is processed after the k^{th} sub-operation of operation *i*. Constraints (20) adjust the starting dates of sub-operations which need to wait because of a power flow. Constraints (21) detect if a switch *off*/*on* is possible. Furthermore, if *i* and *j* are not in disjunction, then necessarily, no switch *on*/*off* is possible between these operations. Constraint (22) detect if an operation is the first scheduled on the machine. If so, the machine must be started before processing the operation *j*. Constraints (23-24) set the processing times and power requirements of the switch *off*/*on* operations. If $l_{i,j,m}$ is null for any *m*, hence *p* and *w* of these operations are necessarily equal to 0, thus no power flows go *in* or *out* of these operations and they are not deteriorating the makespan either since the processing time is null. Finally, constraints (25) set the basis consumption of machines. This variable is then used in constraints (16) to model the constant consumption of machines if they are switched *on*. Hence, if no switch *off* occur, the $w^b_{i,k}$ variable is like a sink node, and the absorbed power cannot be allocated to other operations unless a switch *off* occur.

All these constraints applied together allow to find schedules with switch *on*/*off* policies in order to reduce the power consumption of a manufacturing system. Results obtained with this mathematical formulation are given in the next section.

IV. Results and discussion

Results have been computed using CPLEX 12.4 on a set of 10 generated instances. These instances have been generated randomly by considering jobs varying from 2 to 8, number of operations per job between 2 and 6, machines varying from 3 to 5. One or two machines are available to process an operation. Each instance is tested with three different power thresholds. As the impact of switch *on*/*off* policies is the main purpose of this work, these instances are also tested with and without switch *on*/*off* policies, which brings to 60 the number of tests performed. For each test, the solver is left running for 5400 seconds. Results are provided in Table I. In this table, **Ins.** refers to the instance tested. **n** is for the number of machining operations in the instance. **W_{max}** is for the different power thresholds tested. **SUP** and **INF** refer to the upper and lower bounds. Solutions with an asterisk(*) are proven optima. **CPU** is the time requested to prove the optimality, whereas **CPU*** is the approximate time for reaching the best found solution. Dashes represent computation times exceeding the 5400s limit, or unavailable solutions.

As can be seen in Table I, switch *on*/*off* policies have an impact on the quality of solutions. In all instances proven optima, the solutions obtained with consideration of the switch *on*/*off* policy are always equal or better than the results without considering extinctions of machines (see for example instance Inst_3). This is the expected behaviour however, because the possibility to switch *on*/*off* machines results in additional binary variables in the mathematical program, computation times are longer in average. Note please that some instances really show the importance of such approaches. For example Inst_4 with a 143 power threshold and switch *on*/*off* has a makespan potentially better than 251 (the upper bound) whereas the same instance without the switch *on*/*off* will not reach a solution with a makespan better than 284 (the lower bound).

TABLE I. Results of switch *on*/*off* policies in a FMS with power considerations

			Without *on-off*				With *on-off*			
Ins.	n	W_{max}	SUP	INF	CPU	CPU*	SUP	INF	CPU	CPU*
Inst_1	7	50	**49***	49	4	4	**49***	49	7	7
		40	**52***	52	4	4	**52***	52	4	4
		30	**83***	83	750	8	**83***	83	2666	145
Inst_2	9	54	**40***	40	13	13	**40***	40	1490	20
		44	**41***	41	20	20	**41***	41	13	13
		34	77	61	-	150	-	38	-	-
Inst_3	10	77	**160***	160	5	5	**160***	160	13	13
		67	**168***	168	5	5	**168***	168	15	15
		57	**218***	218	20	20	**215***	215	490	60
Inst_4	12	163	270	233	-	80	229	209	-	60
		153	319	221	-	385	230	224	-	1265
		143	313	284	-	2875	251	216	-	3730
Inst_5	12	51	**51***	51	15	15	**51***	51	20	20
		41	**51***	51	21	21	**51***	51	30	30
		31	62	61	-	100	62	54	-	375
Inst_6	14	136	**224***	224	40	40	**221***	221	930	50
		126	**240***	240	80	80	237	222	-	4900
		116	**248***	248	115	115	249	224	-	630
Inst_7	14	141	**233***	233	1740	1640	260	209	-	75
		131	268	241	-	2000	298	212	-	950
		121	-	220	-	-	312	206	-	4400
Inst_8	15	162	**286***	286	-	100	**285***	285	-	340
		152	**316***	316	300	100	**296***	296	-	375
		142	**341***	341	2000	820	305	300	-	1750
Inst_9	17	146	293	249	-	2825	284	245		3450
		136	399	246	-	5250	358	245	-	3500
		126	470	245	-	4750	-	245	-	-
Inst_10	26	59	-	42	-	-	-	43	-	-
		49	-	42	-	-	-	42	-	-
		39	-	42	-	-	-	41	-	-

Also, it has been observed in some cases that the switch *on*/*off* policies allows to obtain schedules with lower power thresholds, whereas no solutions could be found without the switch *on*/*off* policies. Furthermore, it appears from these first generated instances that switch *on*/*off* are very useful for situations where the idle power requirements of machine are high, which is the case of steel and plastic industries for instance. However, this might not be always profitable because of machines that may suffer from lot of switch *on/off* [2]. A constraint to limit the number of switch *on/off* can easily be added to the previous model for such cases.

V. Conclusion

In this study the problem of the Flexible Job-shop with a power limitation and machines with switch *on*/*off* policies is addressed. Operations have a variable power profile and they must be scheduled without exceeding the power threshold. In this problem, machines can be switched *on* or *off* to benefit from the power loss when they are idle.

Test instances have been generated, where power profiles of operations are limited to two different power consumptions. Exact solutions for these small scale instances can be obtained, but the model is quickly unable to return results for instances with more operations or low power thresholds. In the future, large scale instances based on the instances from the Flexible Job-shop literature will be addressed. These instances may consider wider range of power profiles for operations to better model real world machining processes, and a variable power threshold could be considered. The use of CP optimizer is also a perspective as it is supposed to be better for scheduling problems [22]. Furthermore, it could be interesting to add financial objectives in order to consider penalties for exceeding the power threshold, which could better represent real case situations.

Acknowledgment

This work was financially supported by the French Public Investment Bank (BPI) and granted under the ECOTHER project.

References

[1] Z. Zhang, R. Tang, T. Peng, L. Tao, and S. Jia, ‘A method for minimizing the energy consumption of machining system: integration of process planning and scheduling’, J. Clean. Prod., vol. 137, pp. 1647–1662, Nov. 2016.

[2] A. Giret, D. Trentesaux, and V. Prabhu, ‘Sustainability in manufacturing operations scheduling: A state of the art review’, J. Manuf. Syst., vol. 37, pp. 126–140, Oct. 2015.

[3] J. R. Duflou et al., ‘Towards energy and resource efficient manufacturing: A processes and systems approach’, CIRP Ann. - Manuf. Technol., vol. 61, no. 2, pp. 587–609, Jan. 2012.

[4] Y. Liu, H. Dong, N. Lohse, S. Petrovic, and N. Gindy, ‘An investigation into minimising total energy consumption and total weighted tardiness in job shops’, J. Clean. Prod., vol. 65, pp. 87–96, Feb. 2014.

[5] G. May, B. Stahl, M. Taisch, and V. Prabhu, ‘Multi-objective genetic algorithm for energy-efficient job shop scheduling’, Int. J. Prod. Res., vol. 53, no. 23, pp. 7071–7089, Dec. 2015.

[6] M. A. Salido, J. Escamilla, A. Giret, and F. Barber, ‘A genetic algorithm for energy-efficiency in job-shop scheduling’, Int. J. Adv. Manuf. Technol., vol. 85, no. 5–8, pp. 1303–1314, Jul. 2016.

[7] K. Nolde and M. Morari, ‘Electrical load tracking scheduling of a steel plant’, Comput. Chem. Eng., vol. 34, no. 11, pp. 1899–1903, Nov. 2010.

[8] L. Merkert, I. Harjunkoski, A. Isaksson, S. Säynevirta, A. Saarela, and G. Sand, ‘Scheduling and energy – Industrial challenges and opportunities’, Comput. Chem. Eng., vol. 72, pp. 183–198, Jan. 2015.

[9] A. Gupta and S. Venkataraman, ‘Reducing Price Volatility of Electricity Consumption for a Firm’s Energy Risk Management’, Electr. J., vol. 26, no. 3, pp. 89–105, Apr. 2013.

[10] G. Mouzon, M. B. Yildirim, and J. Twomey, ‘Operational methods for minimization of energy consumption of manufacturing equipment’, Int. J. Prod. Res., vol. 45, no. 18–19, pp. 4247–4271, Sep. 2007.

[11] F. Shrouf, J. Ordieres-Meré, A. García-Sánchez, and M. Ortega-Mier, ‘Optimizing the production scheduling of a single machine to minimize total energy consumption costs’, J. Clean. Prod., vol. 67, pp. 197–207, Mar. 2014.

[12] K. Fang, N. Uhan, F. Zhao, and J. W. Sutherland, ‘A new approach to scheduling in manufacturing for power consumption and carbon footprint reduction’, J. Manuf. Syst., vol. 30, no. 4, pp. 234–240, Oct. 2011.

[13] A. A. G. Bruzzone, D. Anghinolfi, M. Paolucci, and F. Tonelli, ‘Energy-aware scheduling for improving manufacturing process sustainability: A mathematical model for flexible flow shops’, CIRP Ann. - Manuf. Technol., vol. 61, no. 1, pp. 459–462, 2012.

[14] M. Dai, D. Tang, A. Giret, M. A. Salido, and W. D. Li, ‘Energy-efficient scheduling for a flexible flow shop using an improved genetic-simulated annealing algorithm’, Robot. Comput.-Integr. Manuf., vol. 29, no. 5, pp. 418–429, Oct. 2013.

[15] Y. Zhang, G. H. Huang, and L. He, ‘A multi-echelon supply chain model for municipal solid waste management system’, Waste Manag., vol. 34, no. 2, pp. 553–561, Feb. 2014.

[16] H. Luo, B. Du, G. Q. Huang, H. Chen, and X. Li, ‘Hybrid flow shop scheduling considering machine electricity consumption cost’, Int. J. Prod. Econ., vol. 146, no. 2, pp. 423–439, Dec. 2013.

[17] Y. He, Y. Li, T. Wu, and J. W. Sutherland, ‘An energy-responsive optimization method for machine tool selection and operation sequence in flexible machining job shops’, J. Clean. Prod., vol. 87, pp. 245–254, Jan. 2015.

[18] S. Kemmoe, D. Lamy, and N. Tchernev, ‘A job-shop with an energy threshold issue considering operations with consumption peaks’, IFAC-Pap., vol. 48, no. 3, pp. 788–793, 2015.

[19] P. Brucker and R. Schlie, ‘Job-shop scheduling with multi-purpose machines’, Computing, vol. 45, no. 4, pp. 369–375, 1990.

[20] S. Dauzère-Pérès and J. Paulli, ‘An integrated approach for modeling and solving the general multiprocessor job-shop scheduling problem using tabu search’, Ann. Oper. Res., vol. 70, pp. 281–306, 1997.

[21] N. Weinert, S. Chiotellis, and G. Seliger, ‘Methodology for planning and operating energy-efficient production systems’, CIRP Ann. - Manuf. Technol., vol. 60, no. 1, pp. 41–44, 2011.

[22] M. A. Salido, J. Escamilla, F. Barber, A. Giret, D. Tang, and M. Dai, ‘Energy efficiency, robustness, and makespan optimality in job-shop scheduling problems’, Artif. Intell. Eng. Des. Anal. Manuf., vol. 30, no. 03, pp. 300–312, Aug. 2016.

In search of an optimal tooling capacity for the vulcanising planning decision problem

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Federico Trigos
Tecnológico de Monterrey
EGADE Business School
Ave. Eugenio Garza Lagüera y Rufino Tamayo
66269 MX, San Pedro Garza García
Email: ftrigos@itesm.mx

Eduardo M. López
NCA Europa S.C.
Mississippi 231 A Pte.
66220 MX, San Pedro Garza García
Email: direccion@ncaeuropa.com.mx

***Abstract*—The vulcanising planning decision problem (VPDP) is a tactical periodically solved problem that models a number of real world applications assuming tooling constrained capacity as input data. VPDP importance lies on its effect on planning manufacturing cycles periodically. However, previous strategic decisions need to be determined prior to solving VPDP. In particular decisions on tooling capacity are significant for financial performance. This paper states the tooling capacity problem (TCP) for VPDP making its contribution twofold: the development of part-number wise lower bounds on tooling capacity; and the development of a heuristic to find tooling capacity values. Numerical illustration shows the efficiency of the heuristic.**

I. Introduction

The Vulcanising Planning Decision Problem (VPDP) [1] is a NP-hard one consisting on a number of orders (part numbers), each one containing a number of units (demand) to be processed on a single machine with fixed capacity under a single manufacturing period. Each unit from every order has to use a tool which is part number dependent in a one to one basis. Therefore a number of tools per part number have to be available. The machine has the ability to operate a number of products from the same family at the same time, subject to tool availability. Each product has to pass through the machine once and tools are reusable for as many cycles as needed. Each time the machine is operated over a number of products is called a manufacturing cycle. Every manufacturing cycle uses the same amount of time. Thus, VPDP consists on finding the sets of products that have to be manufactured together per manufacturing cycle satisfying tool availability and minimising the number of total manufacturing cycles to meet demand. This problem is a special version of the Cutting Stock Problem (CSP) as shown in [1]. There are many industrial applications that are modeled by VPDP [2], [3], [4], [5], [6], [7], [8]. For illustration purposes, the case of manufacturing Rubber Curved Hoses (RCHs) will be used as illustration for the VPDP in this article.

The raw material for RCHs is a rubber hose (designed specifically for its part number) that has to be molded throughout a vulcanising process; the rubber hose is inserted on a mandrel (a metallic tool that has the geometrical specifications of the part number) and then set in a Vulcanising Unit (VU) for a fixed amount of time called Vulcanising Cycle (VC). It is assumed that all part numbers in the catalogue belong to the same family of products, thus the VC time is the same for all part numbers in the family. A number of RCHs are set in the VU for a VC restricted by mandrel availability. A Vulcanising Cart Design (VCD) consists of a set of part numbers and their corresponding number of units related; each VCD may be run more than once so the frequency of each VCD (number of cycles to run the same VCD) has also to be determined. VPDP consists on finding the sets of VCDs that minimise the number of VCs to meet total demand at equality.

VPDP is a tactical problem that is solved every manufacturing period (daily, weekly, monthly, etc.) and mandrel availability is part of the input data. When a new family of products have to be quoted by a manufacturer for the first time, it is not clear how to compute the appropriate number of mandrels per part number. In [1] it was suggested to study this problem prior to solve VPDP. Mandrel availability has a significant influence on the optimal solution of VPDP. The Tool Capacity Problem (TCP) for VPDP consists on finding the appropriate number of mandrels per part number that minimises total VCs for VPDP.

The contribution of this paper is two fold: To obtain a lower bound on the number of available mandrels that minimises the objective function of VPDP; and to provide a heuristic to solve TCP. Given the scant treatment of vulcanising optimisation decision making in general literature, TCP becomes a strategic decision making tool for the overall manufacturing process of RCHs.

The rest of the paper is as follows: Section II sets the mathematical model and notation to identify VPDP; section III identifies the mathematical properties of VPDP as well as some figures of mandrel availability required to influence VPDP; section IV develops a heuristic to solve TCP; section V shows a small numerical example to illustrate the heuristic; and finally, section VI sets the final conclusions and further research to be performed.

II. The vulcanising planning decision problem

In reference [1] VPDP is originally formalised. This is a decision problem that is solved in every manufacturing planning period (daily, weekly, monthly, etc.). Table I summarizes the notation used in this paper. Let c be the size of the RCH catalogue, s be the total possible number of different

VCDs, and $i \in I = \{1, ..., c\}$ be the index set of RCHs; d_i, m_i and l_i will be defined as the total demand, maximum mandrel availability and number of stars required in the VU for each RCH i respectively. Define $R \in \mathbb{Z}_+$ as theVU capacity, i.e. the total number of stars available in the VU for every VCD to use. The problem consists of defining a set of VCDs $a^j = (a_1^j, \ldots, a_c^j)$, where a_i^j is the number of mandrels of RCH i to be included in the $j-th$ VCD. Each a^j is run x^j times in the vulcanizer.

Literature [1], [4] identifies at least four different Vulcanising Cart Geometries (VCGs); this paper focuses on linear VCGs only.

The VPDP model to minimise the number of VCs to be studied in the remainder of this paper becomes (1) through (6), where $A = (a^1, \ldots, a^s) \in Z_+^{c \times s}$, $D = (d_1, \ldots, d_c)^T \in Z_+^c, M = (m_1, \ldots, m_c)^T \in Z_+^c, L = (l_1, \ldots, l_c)^T \in Z_+^c$ and $X = (x^1, \ldots, xs)^T \in Z_+^s$; in addition, $\mathbf{1} \in R^s$ is the vector of ones and $\mathbf{0} \in R^c$ is the vector of zeros at each position. D, M, and L are strictly positive vectors. A represents the matrix of all feasible VCDs, vector D represents RCH demands, vector M represents the number of available mandrels per RCH type, vector L represents the number of stars out of the VU required per RCH type, and X represents the vector of VCs to be used for each VCD. The objective function (1) seeks to minimise the total number of VCs to be used to satisfy complete demand. Constraint (2) ensures that the total demand is met exactly for each type of RCH. Let $J = \{1, \ldots, s\}$ be the index set of all feasible VCDs. Thus, each feasible VCD a^j must satisfy (4) through (6).

$$\min \mathbf{1}^T X \tag{1}$$

subject to

$$AX = D \tag{2}$$

$$X \in \mathbb{Z}_+^c, \tag{3}$$

$$\mathbf{0} \leq a^j \leq M, \quad \forall j \in J \tag{4}$$

$$L^T a^j \leq R, \quad \forall j \in J \tag{5}$$

$$a^j \in \mathbb{Z}_+^c, \quad \forall j \in J. \tag{6}$$

Because of the combinatorial nature of the problem, the cardinality of J makes it unpractical to state matrix A explicitly. Problem (1) through (6) is a special extended case of the 1D-CSP because there is a single vulcanizer (single roll) type with R stars (5), integer frequencies X (3), part-dependent tooling constraints with upper bound M (4), and integer demands D that have to be met at equality (2) as stated in [1].

III. The Tooling Capacity Problem for the VPDP

In the VPDP defined in the previous section the number of available mandrels (m_i) is known in advance. In the case of planning for a new family of product under a given contract demand, the assumption on mandrel availability is no longer valid. The family of products is going to be quoted to be manufactured over a single period (typically several years in this new case). In TCP mandrel demand (d_i) is related to the duration of the contract (which may address years). In order to develop a RCH quote for the contract, mandrel availability (tooling capacity m_i) has to be defined for every RCH in the family to be quoted. Let us call that problem Tooling Capacity Problem (TCP) for VPDP. The problem becomes to find the optimal number of mandrels per RCH in order to minimise the number of VC to meet demand for all products in the family for the duration of the contract. Hence, TCP is a serious strategic problem.

Reference [1], developed a number of bounds on the the number of total cycles. The first bound can be computed using the total number of stars needed divided by the stars availability (R) in each VCD as:

$$TC = \left\lceil \frac{\sum_{i \in I} l_i d_i}{R} \right\rceil, \tag{7}$$

where $\lceil\ \rceil$ is the ceiling function. Notice that this bound does not takes into account mandrel capacity (m_i).

On the other hand the minimum number of cycles to meet demand per part number depends on mandrel availability, i.e. if we use all available mandrels in a VCD for the $i-th$ RCH then total demand d_i divided by mandrel availability also sets a lower bound on the optimal number of cycles required to meet demand for that RCH, as:

$$\alpha_i = \left\lceil \frac{d_i}{m_i} \right\rceil, \forall i \in I. \tag{8}$$

If there is no budget restriction for tooling, the optimal solution for VPDP can become independent of mandrel availability. In this case, in order to choose properly the number of mandrels per RCH, m_i can be set to all mandrels fitting the VU or the total demand in the contract if the latter is less than the former. In this case mandrel availability is no longer an active constraint for VPDP making it a classic cutting stock problem [9], [10], [11], [12], [13], [14]. Thus, m_i can be represented as:

$$m_i = \min \left\{ \left\lceil \frac{R}{l_i} \right\rceil, d_i \right\}, \forall i \in I. \tag{9}$$

On the other hand, since mandrels are expensive and have a significant influence on the optimal solution of VPDP, then TCP can be stated as finding m_i, $\forall i \in I$ in such a way that the minimum VCs to meet total demand is set as close as (7) as possible.

In order to do so, let us set m_i in such a way that no part number i in (8) sets a lower bound larger than (7) $\left(\frac{d_i}{m_i} \leq \frac{\sum_{i \in I} l_i d_i}{R}\right)$ thus:

$$m_i \geq \lceil f d_i \rceil, \tag{10}$$

where

$$f = \frac{R}{\sum_{i \in I} l_i d_i}. \tag{11}$$

Hence, (10) sets a lower bound on mandrel capacity. Notice that this lower bound becomes a demand mandrel factor (f) of the $i-th$ RCH. At this point it is not clear how to develop better lower bounds for m_i, so one option is to take (10) at equality and multiply f by a mandrel-multiple (λ) making mandrel assignment as follows:

$$m(\lambda)_i = \lambda\lceil f d_i \rceil. \quad (12)$$

The following section shows a heuristic based on (12) which searches for several values of λ.

IV. TCP FORWARD HEURISTIC

The heuristic is based on a search for λ from (12) to get total number of VCs as close as possible to (7).

TCPFH: A forward heuristic

1: $\lambda = 1, VC(0) = \infty, f = \frac{R}{\sum_{i\in I} l_i d_i^k}, TC = \lceil f^{-1} \rceil$.
2: $m(\lambda)_i = \lambda\lceil f d_i \rceil, \forall i \in I$.
3: Run forward heuristic from [1] as follows
(with slight modifications for $m(\lambda)_i$):
. 3.1: Make $k = 1$, $d_i^k = d_i$, $\psi_i^k = d_i^k / m(\lambda)_i$, $\forall i \in I$.
. 3.2: While $\psi_{[1]}^k > 0$
. 3.3: Compute order $\psi_{[1]}^k \geq \cdots \geq \psi_{[c]}^k$, break ties by
. non-increasing $l_{[i]}$.
. 3.4: $C^k = R$.
. 3.5: For $i = [1]$ to $[s]$
. 3.6: $a_i^k = \min(m(\lambda)_i, d_i^k, \lfloor C^k / l_i \rfloor)$.
. 3.7: $C^k = C^k - l_i a_i^k$.
. 3.8: End For
. 3.9: $x^k = \lfloor d_{i^*}^k / a_{i^*}^k \rfloor = \lfloor \min_{\{i|a_i^k>0\}} (d_i^k / a_i^k) \rfloor$.
. 3.10: $k = k + 1$.
. 3.11: $d_i^k = d_i^{k-1} - a_i^{k-1} x^{k-1}$, $\psi_i^k = d_i^k / m(\lambda)_i$, $\forall i \in I$.
. 3.12: End While.
4: $VC(\lambda) = \sum_{i=1}^{c} x_i^k$.
5: If $VC(\lambda) = TC$ then set $\lambda^* = \lambda$ and goto 7.
6: If $VC(\lambda) < VC(\lambda - 1)$, then $\lambda = \lambda + 1$ and goto 2, otherwise set $\lambda^* = \lambda - 1$.
7: Stop.

Step 1 initializes the mandrel-multiple λ, the vulcanising cycles $VC(0)$, computes the demand mandrel factor f, and the lower bound on total vulcanising cycles TC. Step 2 computes current mandrel availability as in (12). Step 3 implements the Forward Heuristic as in [1] with the modification of using $m(\lambda)_i$ instead of m_i as originally proposed. At the end of step 3 a solution of VPDP is reached under the current tooling capacity $m(\lambda)_i$, for $i \in I$. Step 4 computes current VC to meet demand. In Step 5, if current number of $VC(\lambda)$ has reached the lower bound TC, then the current solution of VPDP associated to $\lambda^* = \lambda$ has reached its final solution; then stop, otherwise continue to step 6. If the current VPDP solution is better than the previous iteration, step 6 increases λ by one unit and goes back to step 3 once more; otherwise the previous solution associated to $\lambda^* = \lambda - 1$ is the final solution and stop.

At the end of step 7, mandrel capacity $m(\lambda^*)_i$ might differ from the maximum use of mandrels per RCH defined by:

$$m_i^{'} = \max_{\forall k}(a_i^k). \quad (13)$$

Hence, step 3 may be run only (replacing $m(\lambda)_i$ with m_i^*) until the maximum use of mandrels reaches mandrel capacity or the current solution increases the number of cycles $VC(\lambda^*)$ already achieved, making the solution that keeps $VC(\lambda^*)$ with the least number of mandrels explored the final solution of the TCP for the VPDP.

Notice that in every iteration of TCPFH, the feasible region of VPDP is expanded by increasing mandrel availability. Therefore, the optimal objective function value is either reduced or remains constant. Since the objective function is bounded below by TC, TCPFH terminates in a finite number of steps.

Once the solution of TCP is obtained from TCPFH, it is used to quote mandrel cost in the contract related to the family of products demanded.

V. SMALL NUMERICAL EXAMPLE

Table II shows the data of a small TCP example with $c = 6$ RCHs, a single VU with a capacity of $R = 25$ stars, and all $l_i = 1$ for $i = 1, \ldots, 6$.

The total number of stars needed to meet demand is $\sum_{i\in I} l_i d_i^k = 456$. From (10) the lower bound on mandrel availability is $\{1, 1, 7, 5, 1, 12\}$ and the lower bound (7) on the optimal number of cycles is $TC = 19$.

Table III shows the results of TCPFH for $\lambda = 1$ where six VCDs $(a^1, \ldots, a^6)$ are developed as shown in columns 2 to 7; the 8-th column shows the maximum number of mandrels used per RCH as in (13). The 9-th column shows the number of available mandrels in the iteration. By adding the numbers in the last row one finds $VC(1) = \sum_j x^j = 30$. The first VCD $a^1 = (1, 1, 7, 5, 0, 12)$, contains one mandrel of RCH one, one mandrel of RCH 2, seven mandrels of RCH 3, five mandrels of RCH 4, and twelve mandrels of RCH 6, and it is run $x^1 = 15$ times on the VU. Since our lower bound $TC = 19$ is smaller than our currently $VC(1) = 30$ cycles, one more iteration of TCPFH is in order. Table IV shows the results for $\lambda = 2$ making $VC(2) = \sum_j x^j = 24$ still above our lower bound of $TC = 19$. Table V shows the results for $\lambda = 3$ making $VC(3) = \sum_j x^j = 19$ which makes it equal to TC, hence the optimal number of VC is achieved. By inspection of columns 8 and 9, RCHs 3 and 6 are not using all their available mandrels. Table VI shows the results for M=(3,3,19,15,3,25) as they were used in table V. Notice that the total VCs are still 19 (optimal for VPDP), but the third RCH uses 16 mandrels instead of the 19 available. Table VII is obtained by using $M = (3, 3, 16, 15, 3, 25)$ as they were used in table VI. This times the solution makes the total number of $VCs = 20$ making this solution no longer optimal for the VPDP.

Hence, our final solution for the TCP is found at $M = (3, 3, 16, 15, 3, 25)$ (as used in table VI), with the final VPDP solution shown in table VI.

Notice that our TCPFH seeks for the number of mandrels that keep the final solution of VPDP as close as possible to the lower bound defined in (7). It does not consider differences in: mandrel manufacturing cost, maintenance cost, or financial rescue value at the end of the contract among RCHs.

VI. CONCLUSIONS AND FURTHER RESEARCH

TCP is an important strategic problem faced by companies (mainly automotive, aerospace among others) that quote for

the first time (or update) a family of products for a long term contract. It is related to the periodical NP-hard VPDP that is run in short term manufacturing cycles (shifts, days, months, etc.). This paper offers a set of lower bounds on the solution of TCP and an efficient heuristic that can be used by practitioners. Solving TCP is a preliminary process that supports the cyclical production planning derived from VPDP. Economical consequences of TCP affect financial performance of companies in the long run.

Further research pends ahead. Deeper analysis on bounds regarding mandrel availability have to be studied. The cost of mandrel manufacturing, maintenance and salvage value per RCH have to be incorporated in TCP. The analysis of multiple identical or non identical VUs has to be carried on.

REFERENCES

[1] Trigos, F. and López, M. 2017. A vulcanizing decision planning as a particular one-dimensional cutting stock problem with limited part-related tooling in make-to-order industrial environments, *International Journal of Production Research* **55** (10) 2881-2896.

[2] Fowler, J.W., Hogg, G.L., Phillips, D.T. 1992. Control of multi product bulk service diffusion/oxidation process, *IIE Trans.* **24** 84-96.

[3] Lakshminarayanan, B. 2008. High Reliability Miniature RF-MEMS. Switched Capacitors , *Microwave Theory and Techniques IEEE Transactions* **April 56** (4) 971-981.

[4] López, H. 2001. *Algoritmo para el diseño óptimo de carros de vulcanizado utilizando programación lineal con un número infinito de variables.*, Master's thesis, ITESM - Campus Toluca, México.

[5] Mathirajan, M., Chandru, V. and Sivakumar, A.I. 2007. Heuristic algorithms for scheduling heat-treatment furnaces of steel casting industries, *Sādhanā* **October 32** (5) 479–500.

[6] Uzsoy, R., Lee C.H., Martin-Vega, L.A. 1992. Scheduling semiconductor test operations: Minimizing maximum lateness and number tardy jobs on a single machine, *Naval Research Logistics* **39** 369-388.

[7] Wayne, B.N. 2004. *Accelerated Testing: Statistical models, Test Plans and Data Analysis*. John Wiley and Sons, Hoboken NJ.

[8] Zee, D.J. ,Van Harten, A., Schuur, P.C. 1997. Dynamic job assignment Heuristics for multi-server batch operations-A cost based approach, *International Journal of Production Research* **35** (11). 3063-3093.

[9] Belov, G., Letchford, A.N., and Uchoa, E. 2005. A Net-Flow model for 1D Stock Cutting: Robust Branch-Cut-and-Price, *Technical report, Institute of Numerical Mathematics, Technische Universitat Dresden, April 10* 1-14.

[10] Belov, G., and Scheithauer G. 2007. Setup and Open-Stacks Minimization in One-Dimensional Stock Cutting, *Informs Journal on Computing* **19** (1) 27-35.

[11] Gilmore, P. C., R. E. Gomory. 1961. A linear programming approach to the cutting-stock problem, *Operations Research* **9** (6) 849–859.

[12] Gilmore, P. C., R. E. Gomory. 1963. A linear programming approach to the cutting-stock problem Part II, *Operations Research* **11** 863–888.

[13] Valério de Carvalho, J.M. 2002. LP models for bin packing and cutting stock problems, *European Journal of Operational Research* **141** (2) 253-273.

[14] Varela, R., Vela, C. R., Puente, J., Sierra, M., González-Rodríguez, I. 2009. An effective solution for a real cutting-stock problem in manufacturing plastic rolls, *Annals of Operational Research* **166** 125-146.

TABLE I. PROBLEM PARAMETERS AND VARIABLES.

Parameters	Description
c	Size of RCH catalogue.
i	Index of RCHs, $i \in I = \{1, \ldots, c\}$.
d_i	Demand for the $i-th$ RCH with $D = (d_1, \ldots, d_c)^T$, where $d_i \in \mathbb{Z}_+$ (the set of non-negative integers). Single period demand for VPDP and contract demand for TCP.
l_i	Length of the $i-th$ RCH (number of VU stars used per mandrel) with $L = (l_1, \ldots, l_c)^T \in \mathbb{Z}_+^c$.
R	VU capacity (in number of stars available per vulcanising cycle), $R \in \mathbb{Z}_+$.
0	The vector of all zeroes in R^c
1	The vector of all ones in R^s
a^j	The $j-th$ VCD, where $a^j = (a_1^j, \ldots, a_c^j)^T$, with $a^j \in \mathbb{Z}_+^c$.
a_i^j	Number of mandrels of the $i-th$ RCH to be included in the $j-th$ VCD. **VPDP and TCP decision variable**.
A	Matrix of all possible VCDs, where $A = (a^1, \ldots, a^s)$.
C^k	Current unused stars in the $k-th$ VCD (iteration) at step 3 of TCPFH.
d_i^k	Remaining demand of the $i-th$ RCH after running the first $k-1$ VCDs (iterations) at step 3 of TCPFH.
f	Demand mandrel factor.
j	Index for VCDs, where $j \in J = \{1, \ldots, s\}$.
m_i	Number of available mandrels for the $i-th$ RCH, with $M = (m_1, \ldots, m_c)^T \in \mathbb{Z}_+^c$. **TCP decision variable, data for VPDP.**
$m(\lambda)_i$	Number of available mandrels for the $i-th$ RCH at the $\lambda-th$ iteration of TCPFH.
$m_i^{'}$	Maximum number of mandrels used in current VPDP/TCP solution for the $i-th$ RCH.
s	Number of feasible VCDs.
TC	Lower bound on the number of VCs to meet total demand based on the capacity of the VU and RCH total demand.
$VC(\lambda)$	Number of total VCs obtained (to meet total demand) when $m(\lambda)_i, \forall i \in I$ mandrels are available.
x^j	Frequency of the $j-th$ VCD, with $X = (x^1, \ldots, x^s)^T \in \mathbb{Z}_+^s$. **VPDP and TCP decision variable**.
α_i	Lower bound on total VCs based on the $i-th$ RCH mandrel availability.
λ	Mandrel-multiple.
λ^*	Output value of λ found by TCPFH.
ψ_i^k	Revenue assigned to each mandrel used by the $i-th$ RCH in the $k-th$ iteration at step 3 of TCPFH, where $\psi_i^k = d_i^k / m_i$.

TABLE II. SMALL EXAMPLE DATA ($c = 6, R = 25,$ AND $L = 1$).

i	1	2	3	4	5	6
d_i	15	17	110	88	11	215

TABLE III. SMALL EXAMPLE ($c = 6, R = 25,$ AND $L = 1$), TCPFH($\lambda = 1$), $VC(\lambda) = 30$.

i	a^1	a^2	a^3	a^4	a^5	a^6	$m_i^{'}$	$m(\lambda)_i$
1	0	1	1	1	1	1	1	1
2	1	1	1	0	0	0	1	1
3	7	5	0	0	0	0	7	7
4	5	5	5	3	0	0	5	5
5	0	1	1	1	1	0	1	1
6	12	12	12	11	0	0	12	12
x^j	15	1	1	1	8	4		

TABLE IV. SMALL EXAMPLE ($c = 6, R = 25$, AND $L = 1$), TCPFH($\lambda = 2$), $VC(\lambda) = 24$.

i	a^1	a^2	a^3	a^4	a^5	a^6	a^7	a^8	$m_i^{'}$	$m(\lambda)_i$
1	0	0	2	2	2	2	2	1	2	2
2	0	2	0	1	0	0	0	0	2	2
3	0	13	0	6	0	0	0	0	13	14
4	1	10	0	0	0	0	0	0	10	10
5	0	0	2	2	2	1	0	0	2	2
6	24	0	21	2	0	0	0	0	24	24
x^j	8	8	1	1	3	1	1	1		

TABLE V. SMALL EXAMPLE ($c = 6, R = 25$, AND $L = 1$), TCPFH($\lambda = 3$), $VC(\lambda) = 19$, OPTIMAL.

i	a^1	a^2	a^3	a^4	a^5	a^6	$m_i^{'}$	$m(\lambda)_i$
1	0	0	3	3	3	0	3	3
2	0	3	0	0	2	0	3	3
3	0	7	19	9	9	0	**19**	**21**
4	0	15	0	13	0	0	15	15
5	0	0	3	0	2	0	3	3
6	25	0	0	0	9	6	**25**	**26**
x^j	8	5	3	1	1	1		

TABLE VI. SMALL EXAMPLE ($c = 6, R = 25, L = 1$), FIRST ADJUSTMENT OF MANDRELS, VC=19 STILL OPTIMAL.

i	a^1	a^2	a^3	a^4	a^5	a^6	$m_i^{'}$	m_i
1	0	0	3	3	3	0	3	3
2	0	0	3	3	3	2	3	3
3	0	10	16	4	4	4	**16**	**19**
4	0	15	0	13	0	0	15	15
5	0	0	3	2	0	0	3	3
6	25	0	0	0	15	0	25	25
x^j	8	5	3	1	1	1		

TABLE VII. SMALL EXAMPLE ($c = 6, R = 25, L = 1$), SECOND ADJUSTMENT OF MANDRELS, VC=20, NO LONGER OPTIMAL.

i	a^1	a^2	a^3	a^4	a^5	a^6	a^7	$m_i^{'}$	m_i
1	0	0	3	3	3	3	0	3	3
2	0	0	3	3	3	3	2	3	3
3	0	16	1	12	0	0	0	16	16
4	0	9	15	0	4	0	0	15	15
5	0	0	3	3	2	0	0	3	3
6	25	0	0	4	11	0	0	25	25
xj	8	6	2	1	1	1	1		

Operating Room Scheduling and Health Care Services

A heuristic for an operating room scheduling problem with durations' uncertainty

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mohamed Amine Abdeljaouad
Université de Tunis El Manar, Faculté des Sciences de Tunis, LIP2-LR99ES18, 2092 Tunis, Tunisia
amine.abdeljaoued@fst.utm.tn

Zied Bahroun
Industrial Engineering, College of Engineering American University of Sharjah, Sharjah, UAE
zbahroun@aus.edu

Nour El Houda Saadani
Université de Tunis El Manar, Faculté des Sciences de Tunis, LIP2- LR99ES18, 2092 Tunis, Tunisia
Houda.Saadani@ensi.rnu.tn

Belhassen Zouari
Mediatron Lab, Higher School of Communications of Tunis, University of Carthage, Tunisia
belhassen.zouari@supcom.tn

Abstract— The problem studied in this paper consists of a daily operating room scheduling issue where surgeons have to perform sets of surgical operations on identical operating rooms. The durations of the surgical operations are uncertain and the objective is to minimize the ending date of the last completed operation (makespan).

A stochastic mathematical formulation of this NP-hard problem is provided, as well as a new heuristic that can reach a satisfying solution in a reasonable time. In both methods, the uncertainty aspect is modeled by a set of scenarios The experimental study shows that the proposed heuristic is high performing and can handle efficiently the different scenarios' cases.

Keywords—scheduling; operating rooms; uncertainty; heuristic;

I. Introduction

Optimizing the resources' administration has nowadays become a main issue for hospitals in their attempt to reduce the costs linked to their use. These resources can be of different kinds: ranging from the medical staff, such as the surgeons and the nurses [11], to the medical equipments as the stretchers, the operating rooms or the recovery beds [8]. In this context, researchers have had a growing interest in improving the use of the operating rooms over the last decades. Indeed, the operating theatre is one of the most important parts of a hospital, both in terms of costs related to it and of patient number requiring it.

The operating rooms management is an attempt to find the most efficient assignment and sequencing of the surgical operations in order to optimize a given objective, such as the rooms' opening duration, which is one of the most considered objectives. This process has to deal with many challenging aspects. One of them is that the durations of the surgical interventions are not known in advance; they can only be estimated on the basis of the surgeons' experience or the hospitals historical data. But scheduling operations relying on simple estimations can lead to ineffective results. Indeed, operations that are longer than expected can disrupt the rest of the schedule, or cause the operating rooms to remain open after the regular hours, which leads to additional costs.

In this paper, we study a daily operating room scheduling problem where a set of operations have to be scheduled on identical operating rooms. The operations are already assigned to surgeons and thus two operations that are assigned to the same surgeon cannot be performed at the same time. The durations of these surgical operations are uncertain and are represented by a set of scenarios generated according to a probability distribution. For each operation, a scenario corresponds to one of its possible durations. The objective is to minimize the ending date of the last completed operation (i.e. makespan). This problem is part of the parallel machine scheduling under resource constraints issue, and even its deterministic version (i.e. without uncertainty over the durations of the operations) is NP-hard according to the complexity classification provided by Blazewicz et al. [3].

The rest of the paper is structured in this way: In section II, we present a brief literature review on the operating room scheduling problems under uncertainty constraints. In section III, we provide a stochastic mathematical formulation of the problem and its constraints. We then present, in section IV, a new heuristic method that can find a satisfying solution. Then, we assess its performances in section V by carrying out an experimental study. We give our conclusion and remarks in Section VI.

II. Literature review

Researches related to the operating room planning and scheduling problems under uncertainty constraints have become more and more important over the last decade. Numerous articles have thus been published on the subject. In [4], a survey is provided on the operating rooms planning and scheduling problem. We can see that the uncertainty treated by the works usually concern two possible aspects: the uncertainty over the patients' arrival (i.e. the occurrence of urgent cases) and the uncertainty over the durations of their operations.

Many papers dealt with the operating rooms scheduling problems with uncertain durations [7] [17]. Among them, Mancilla and Storer [12] worked on a scheduling problem where the operations of a surgeon have to be scheduled on two identical operating rooms. They provide a stochastic mathematical model to minimize the weighted sum of the surgeon's waiting time, the rooms' idle time and the overtime work (i.e. the working time after the rooms' regular hours). The duration of an operation is made of three parts: the patient's preparation, the surgical act which is the only part that requires the surgeon's presence and the cleaning of the room. The durations of these three parts are all uncertain and are represented by a set of scenarios and the objective function is calculated by a sample average approximation (i.e. the average objective function calculated over the set of scenarios). As the solving times of this model quickly become unreasonable, the authors propose a decomposition based approach to solve to optimality real-sized instances.

The sample average approximation is one of the most popular techniques to deal with uncertainty. Denton et al. [6] use it when they provide their two-level mathematical model to minimize the patients and surgeons' waiting times as well as the operations' tardiness in a daily operating room scheduling problem, which consists in sequencing the set of surgeons' operations in one operating room. The scheduling is determined by the first level of the model while the second computes the objective function of each scenario. The model returns the scheduling that minimizes the average objective function over the set of scenarios. Denton et al. [5] present two mathematical models for an operating room planning problem where the objective is to assign the operations to the rooms by minimizing the costs related to their opening duration. The first proposed model is inspired from the bin-packing problems where uncertainty is represented by a set of scenarios. The objective of the second proposed model is to minimize the worst scenario's objective function.

Batun et al. [2] also represent the stochastic aspect of the operating room scheduling problem they study by a set of scenarios. They considered identical operating rooms and pre-defined sequences of surgeons' operations. The authors presented a two-level mathematical model where the decisions taken at the first level concern the number of operating rooms to open and the scheduling sequence inside each room. The second level computes the objective values of each scenario which are the ending date of the rooms, the rooms' idle time and the overtime work.

Saadouli et al. [17] provide two MILP models respectively for the planning and scheduling phases of a Tunisian orthopedic surgery department, where the uncertainty over the operations durations is represented by a percentile distribution. The first proposed model maximizes the rooms' utilization rate while selecting the operations to be performed over the planning horizon, and the second schedules the assigned operations for each day by minimizing the makespan and the waiting times. They used a simulation to compare their models to the actual hospitals' policies. Indeed, simulation methods are frequently used in solving problems with stochastic aspects as they allow evaluating the quality of the proposed solutions.

While the uncertainty of the operations' durations can be treated when dealing with the scheduling aspect, the uncertainty over the patients' arrivals is handled when dealing with the planning phase. Indeed, most of the papers on the uncertainty about the emergencies' arrivals attempt to reserve a certain capacity from the operating rooms while allocating the operations to them, in order to handle these emergencies if they occur. Thus, the aim is to find the best rooms' capacity reservation, which avoids the underutilization, in the cases where there are less emergencies cases than expected, and the overtime work when the emergencies require more time than the duration allocated to them.

Researchers dealing with this kind of problem usually represent this uncertainty by a set of scenarios of the actual capacity required by the emergency cases. Lamiri et al. [10] are among those. They study an operating room planning problem which consists in assigning the operations to different periods over a planning horizon, taking into account that the emergency cases have to be scheduled in the same day of their arrivals. The authors present a stochastic mathematical model and a Monte Carlo simulation to minimize the average costs of the patients waiting time and the overtime work over the sets of scenarios. The scenarios techniques were also used by Razmi et al. [16] to deal with a similar problem that presents the specificity that some of the operations to plan may also require the use of a unique equipment. The authors present a stochastic model and a heuristic to minimize the planning's total costs. Persson and Persson [15] deal with both the uncertainties over the patients' arrivals and the operations' durations while dealing with the operations planning of an orthopedic department in Sweden composed of two operating rooms. They provide a simulation model inspired from the bin-packing problems to minimize the costs of the overtime, waiting time and operations cancelling. They use a Poisson distribution, whose parameters are based on the historical data of the hospital, to determine the arrival of the operations. The estimated durations of the operations, which are considered by the model, are also based on the historical data while their actual ones are generated using a log-normal distribution. The authors propose a new strategy of capacity reservation and compare it to the actual policies used by the hospital.

In our paper, we consider a daily operating room scheduling problem, where a set of surgeons' operations have to be scheduled on identical operating rooms. The objective is to minimize the makespan. In literature, few works deal with groups of operations on identical operating rooms. The only papers we found concern the deterministic version of the problem. Abdeljaouad et al. [1] present a mathematical formulation and a dichotomic algorithm to solve the problem while Guinet and Saadani [9] provide a comparative study between a set of two-step heuristics. These heuristics start by arranging the operations according to a specific order and then assigning them to the operating rooms on the basis of different assignation rules. The authors propose different operations orders, such as SPT (i.e. shortest processing time), LPT (i.e.

longest processing time), or an adaptation of the NEH heuristic [14], and also several assignation rules, such as FAM (i.e. first available machine) or LBM (last busy machine). They test several combinations of these procedures and the obtained results show that the heuristics NEH-LBM and LPT-LBM are those that perform the best.

In our problem, the uncertainty over the durations of the operations is an important aspect. Although the stochastic problems treated in the literature are numerous, no paper dealt with our same issue. In the next section, we propose a stochastic mathematical model, then we discuss its limits (i.e. which instances sizes can be solved in a reasonable time) and finally, we present a new heuristic that can quickly reach satisfying solutions.

III. Mathematical Model

Let's provide a stochastic mathematical formulation of the studied problem. We first recall it by formally presenting its constraints and characteristics: n surgical operations have to be scheduled on a set of identical operating rooms. Each operation has to be performed by a given surgeon, thus two operations that have to be performed by the same surgeon cannot be processed at the same time. Every operating room can process at most one operation at a time and the surgeons are able to move from one room to another to pursue their operations. The preemption of the operations are however not allowed (i.e. once an operation has started, it cannot be interrupted).

This problem is part of the parallel machine scheduling under resource constraints issue. Indeed, in our problem the operating rooms can be assimilated to identical parallel machines while the surgeons can be considered as resources that the operations require to be performed. This problem can be formally noted $P|res.11|Cmax$ using the three-field notation of Blazewicz et al. [3], with P meaning that the machines are identical, $res.11$ meaning that every operation requires one specific type of resources and that there are only one resource from each type and $Cmax$ referring to the objective to minimize (i.e. the makespan).

The mathematical model we provide is inspired from the strip-packing problems and is based on the one provided by Abdeljaouad et al. [1] to solve the deterministic version of the problem. To consider the stochastic aspect of the operations' durations, we opted for the sample average approximation technique. The duration of every operation is represented by a set of possible scenarios. The objective is therefore to find the scheduling that minimizes the average makespan over the set of scenarios. Below we introduce the notations used and formally define the model:

Notations :

- K : The number of scenarios.
- n : The number of operations.
- W : The number of operating rooms.
- S : The number of surgeons.
- nb_s: Number of operations of surgeon s.
- H_h : A large number.
- HH_k : The makespan of scenario k.
- h_i^k : The duration of operation i in scenario k.
- $x_{i,}$: Ranging from 0 to $W-1$, $x_{i,}$ indicates to which operating room is assigned the operation i.
- y_i^k : The starting time of the operation i in scenario k.

Binary variables of the model:

- l_{ij} = 1 if $x_{i,}$ is lower to $x_{j,}$.
  = 0 otherwise.
- b_{ij} = 1 if the ending date of the operation i is lower than the starting date of operation j.
  = 0 otherwise. In other words, the variable b_{ij} indicates if the operation i precedes the operation j, no matter if it is not a direct precedence or if the two operations are not processed on the same operating room.

Model:

$$Min \sum_{k=1}^{K} \frac{HH_k}{K} \quad (1)$$

Subject to:

$$l_{ij} + l_{ji} + b_{ij} + b_{ji} \geq 1 \quad \forall i,j = 1,\dots n \quad with\ i \neq j \quad (2)$$

$$x_i - x_j + W l_{ij} \leq W - 1 \quad \forall i,j = 1,\dots n \quad with\ i \neq j \quad (3)$$

$$0 \leq x_i \leq W - 1 \quad \forall i = 1,\dots n \quad (4)$$

$$b_{ij} + b_{ji} = 1 \quad \forall i,j = 1,\dots nb_s \quad with\ i \neq j \quad \forall s = 1,\dots S \quad (5)$$

$$l_{ij}\ ,\ b_{ij} \in [0,1] \quad \forall i,j = 1,\dots n \quad (6)$$

$$x_i \in IN \quad \forall i = 1,\dots n \quad (7)$$

$$y_i^k - y_j^k + H_h b_{ij} \leq H_h - h_i^k \quad \forall i,j = 1,\dots n \quad with\ i \neq j\ , \forall k = 1,\dots K \quad (8)$$

$$HH_k \geq y_i^k + h_i^k \quad \forall i = 1,\dots n\ ,\ \forall k = 1,\dots K \quad (9)$$

$$HH_k\ , y_i^k \in IN \quad \forall i = 1,\dots n\ ,\ \forall k = 1,\dots K \quad (10)$$

The objective function (1) is to minimize the average makespan over the set of scenarios. Inequality (2) ensures that the operations that are performed at the same time cannot be processed in the same operating room. Inequalities (3) and (4) ensure that every operation is assigned to one operating room. Equation (5) reflects the disjunctive constraints between the operations of a same surgeon. Inequality (8) determines the the starting times of the operations for each scenario and finally, inequality (9) computes the makespan of each scenario. Equations (6), (7) and (10) indicate the nature of the variables used in the model.

We ran some tests to assess the quality of this model. The scenario durations are generated as follows: We start by generating the estimated duration of every operation by a log normal distribution, with a mean equal to 120 and a standard deviation of 60. Then, the different scenarios of every operation are generated with a uniform distribution between 90% and 120% of its estimated duration. For each of the considered instance's size, we ran 5 tests. The model was coded with the IBM Cplex solver and a 4 GB computer with an I3 processor was used for the tests.. As expected, the mathematical model is not efficient since the problem is NP-hard. The (2 rooms, 3 surgeons, 9 operations) instances are solved in 7 minutes when considering 5 scenarios, but the same instances take nearly one hour for 10 scenarios, and more than 3 hours for 20 scenarios. These high solving times are due to the nature of the problem and the combinatory explosion of solutions. Indeed, the absence of order between the operations of a same surgeon leads to a considerably large solution space.

IV. The proposed heuristic

Since the exact solving is not efficient for the NP-hard problems and takes too long to return optimal solutions, we have thus to use heuristic methods that can reach solutions of a good quality in a reasonable time. In this section, we provide a new heuristic for our problem. This heuristic starts by arranging the operations in a specific priority order and then assigning at each iteration the first ordered operation to an operating room that we select according to a pre-defined rule. The assignation order of the operations is updated at the end of each iteration.

The priority order of the operations is obtained as follows: We start by calculating the average duration of each operation over the set of scenarios. Then, at each iteration, we select the operation having the longest average duration from the surgeon with the longest remaining workload (i.e. the sum of the average durations of its remaining operations). This operation is then assigned to the first operating room to be available, in every scenario, at a time t that is greater or equal to the ready time of the operation on that scenario (i.e. the time at which the operation can be performed at the earliest, which is equivalent to the time at which the surgeon that will perform it has ended his/her previous operation). Below we give the main steps of our heuristic more in details:

Heuristic's steps:

- Step 1: Calculate the average duration of every operation over the set of scenarios.
- Step 2: Compute the average remaining workload of every surgeon by summing the average durations of his unscheduled operations.
- Step 3: Select the operation i with the longest duration from the surgeon with the longest remaining workload.
- Step 4: Let t_i^k be the time at which i can be performed at the earliest in scenario k (i.e. the time at which the surgeon ends his previous operation). Assign i to the first operating room m to be available at a time tr_m^k that verifies $tr_m^k \geq t_i^k$ for every scenario k.
- Step 5: If, there are still unscheduled operations, go back to Step 2. Else, go to Step 6.
- Step 6: The solution HH_k of every scenario k is equal to the maximal value among the operating rooms' ending dates. The algorithm returns the average solution $\frac{HH_k}{K}$ (i.e. the sum of the solutions of the scenarios, divided by the number of scenarios).

Given that n is the operations' number, the complexity of this heuristic is $O(nlog(n))$ since it includes a sorting procedure to arrange the operations.

V. Experimental study

In this section, we evaluate the quality of the presented heuristic by carrying out a battery of tests. We use two performances' measures through our tests: the relative gap between the average solution and the average lower bound, which are calculated over the set of scenarios and also the worst relative gap given by the scenarios. This latter value allows us to assess the quality of our heuristic in the worst cases.

The lower bound that we use is inspired from the one provided in [9] for the deterministic version of the problem. This deterministic lower bound is equal to the maximum value between two values: the duration of work of the surgeon with the longest workload and the solution of the scheduling with preemption of all the operations on the operating rooms (regardless of the disjunctive constraints between the operations of a same surgeon). This lower bound is applied for each scenario k and is calculated as follows:

$$LB_k = Max[\sum_{i=1}^{n} \frac{h_i^k}{\mathrm{W}}, \max(\sum_{i=1}^{nb_s} h_i^k, \forall s = 1, \dots S)] \quad (11)$$

We assessed the quality of this deterministic lower bound by leading comparative tests on the deterministic model of Abdeljaouad et al. [1]. Over 15 tests, the average gap between the bound and the solution is about 0.22%.

The lower bound for the stochastic problem corresponds to the average value of the deterministic bounds over the set of scenarios. It can be written as follows:

$$LB = \sum_{k=1}^{K} \frac{LB_k}{K} \quad (12)$$

A. Instances of tests

In our tests, we attempted to work on realistic sizes of operating theatres. For the number of operating rooms, we set three values: 2, 3 and 5. We calibrated the numbers of surgeons according to the number of operating rooms, since in real life, an operating room rarely welcomes more than 10 operations a day. Thus, the considered values are the following:

- 5 and 7 surgeons for the 2 operating rooms' instances.
- 7 and 10 surgeons for the 3 operating rooms' instances.
- 15, 20, 25 and 30 surgeons for the 5 operating rooms' instances.

The number of operations for each surgeon is randomly generated between the interval [2, 4] using a uniform distribution. For every operation, we considered 100 scenarios of durations. These scenarios are obtained as follows: We start by generating the estimated duration of the considered operation. This estimation is obtained by a log-normal distribution whose parameters are based on real historical data provided by a Norwegian hospital and available online [13]. These data include the durations over three years of six different surgical specialties: gynecology, urology, cardiology, gastroenterology, orthopedics and general medicine. We used different parameters of the log-normal distribution, all based on the historical data, for each of these specialties. Once the estimation of the considered operation is obtained with this distribution, we generated the 100 scenarios of its durations using a uniform distribution in the interval between the 90% and 120% of this estimation.

Ten tests are performed for each instance size, which means that 80 tests are performed in total.

B. Results of the tests

The heuristics were coded with C++ and the tests were held on a 4 GB computer with an I3 processor. The heuristic returned its solution for all the performed tests within 1 second. Table I displays the average relative gap for each instance size between the returned solutions and the lower bound that we presented above. We can see that the obtained results given are rather good, with a relative gap rarely exceeding 4% and an overall average value close to 3%. We also notice that for a given number of operating rooms, the more the number of surgeons (as well as the number of operations) is bigger, the more our heuristic behaves better, giving solutions closer to the lower bound. Indeed, from Table I we can see from the instances with 2 operating rooms, the results obtained with 7 surgeons are better than those obtained with 5 surgeons. The same remark can be made about the instances with 3 and 5 operating rooms. This can be explained by the following: when the number of surgeons is big compared to the number of operating rooms, only few surgeons can work at the same time. Therefore, the cases where the operations have to wait for their surgeons to be free do not happen frequently. Most of the time, the selected operation is simply assigned to the first operating room to be available. The assignation rule used by our heuristic is thus more effective.

Table II gives for each instance size the average relative gap of the worst scenarios (i.e. the scenario whose solution gives the worst relative gap compared to its lower bound). This is useful to analyze how the provided scheduling would behave in the worst case. Knowing that, the rates obtained by our methods in the worst cases are respectable. As we can see from the Table II, the relative gap of the worst scenario is about 7.95% in average. As for the average relative gaps calculated over the set of scenarios, the relative gaps of the worst scenarios get also better when the number of surgeons get bigger in comparison with the number of operating rooms.

TABLE I. Heuristic's average relative gap to the lower bound

Instance's size	Average gap to the lower bound
2 rooms 5 surgeons 100 scenarios	4.45%
2 rooms 7 surgeons 100 scenarios	1.79%
3 rooms 7 surgeons 100 scenarios	4.02%
3 rooms 10 surgeons 100 scenarios	3.18%
5 rooms 15 surgeons 100 scenarios	4.17%
5 rooms 20 surgeons 100 scenarios	3.53%
5 rooms 25 surgeons 100 scenarios	2.96%
5 rooms 30 surgeons 100 scenarios	2.91%
Average	3.37%

TABLE II. The average relative gap to the lower bound of the worst scenario

Instance's size	Average gap of the worst scenario
2 rooms 5 surgeons 100 scenarios	9.34%
2 rooms 7 surgeons 100 scenarios	6.39%
3 rooms 7 surgeons 100 scenarios	9.24%
3 rooms 10 surgeons 100 scenarios	8.26%
5 rooms 15 surgeons 100 scenarios	8.77%
5 rooms 20 surgeons 100 scenarios	8.21%
5 rooms 25 surgeons 100 scenarios	6.98%
5 rooms 30 surgeons 100 scenarios	6.47%
Average	7.95%

VI. Conclusion

In this paper, we dealt with an operating room scheduling problem with uncertainty constraints. The problem consists of a set of operations that have to be scheduled on identical operating rooms. The operations are divided into groups and each group has to be performed by a single surgeon. The objective is to minimize the ending date of the last completed operation (i.e. makespan). The problem is NP-hard and is part of the parallel machine scheduling under resource constraints issue.

We represented this uncertainty by a set of scenarios, whose values are generated between 90% and 120% of the estimated durations of the operations, which are based on the historical data provided by a Norwegian hospital and available online. We start by providing a stochastic mathematical model inspired by the 2-dimensional strip packing problems. Then, as the model cannot solve large instances since the problem is NP-hard, we present a new heuristic that can reach satisfying solutions in a reasonable time. At each iteration, this heuristic uses a priority rule to select one from the unscheduled operations and then assigns it to one of the operating rooms according to an assignation rule. The experimental study that we carried out shows that this heuristic gives very satisfying results, with an average gap to the lower bound around 3.34%. We also analyzed the behavior of our heuristic in the worst cases and the relative gap of the worst scenario is about 7.95% in average.

The next step of this work is to introduce a metaheuristic method, such as a tabu search or a simulated annealing, to improve the quality of the solution returned by the heuristic and get near optimal solutions. As future work, we can also include more realistic aspects from the healthcare field, such as separating the durations of the operations into three phases: the patient's preparation, the surgical act and the cleaning of the operating rooms.

References

[1] Abdeljaouad M.A., Bahroun Z., Saadani N.E.H., "A dichotomic algorithm for an operating room scheduling problem", in proceeding of international conference on control, decision and information technologies: *CoDIT*, pp 134-139, 2014.

[2] Batun S, Denton BT, Huschka TR. Schaefer AJ, "Operating Room Pooling and Parallel Surgery Processing Under Uncertainty" INFORMS Journal on Computing Vol. 23, No. 2, pp. 220–237, 2011.

[3] Blazewicz, J., Lenstra, J. K., & Kan, A. H. G. R.. "Scheduling subject to resource constraints: classification and complexity", Discrete Applied Mathematics, 5(1), pp 11-24, 1983.

[4] Cardoen B., Demeulemeester E., Belien J., "Operating room planning and scheduling: A literature review", European Journal of Operational Research, vol. 201(3), pp. 921-932, 2010.

[5] Denton BT., Miller AJ., Balasubramanian HJ., Huschka TR, "Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty". OPERATIONS RESEARCH Vol. 58, No. 4, Part 1 of 2, July–August, pp. 802–816, 2010.

[6] Denton BT., Viapiano J., Vogl A., "Optimization of surgery sequencing and scheduling decisions under uncertainty". Health Care Management Science 10, pp 13–24, 2007.

[7] Dexter, F., Ledolter J. "Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data". Anesthesiology 103(6), pp 1259–1267, 2005.

[8] Guinet, A., & Chaabane, S. "Operating theatre planning". International Journal of Production Economics, 85(1), pp 69–81, 2003.

[9] Guinet, A., & Saadani, N. E. H. "Nomadisme chirurgical et ordonnancement de blocs opératoires". In Proceeding of Conférence francophone en Gestion et Ingénierie des Systèmes Hospitaliers. Clermont-Ferrand, France: *GISEH*, 2010.

[10] Lamiri M, Xie X, Dolgui A, Grimaud F, "A stochastic model for operating room planning with elective and emergency demand for surgery". European Journal of Operational Research 185, pp 1026–1037, 2008.

[11] Maenhout, B., & Vanhoucke, M. "Branching strategies in a branch-and-price approach for a multiple objective nurse scheduling problem". Journal of Scheduling, 13(1), pp 77-93, 2010.

[12] Mancilla C, Storer RH, "Stochastic Sequencing of Surgeries for a Single Surgeon Operating in Parallel Operating Rooms". IIE Transactions on Healthcare Systems Engineering Volume 3, Issue 2, pp 127-138, 2013.

[13] Mannino, C., Nilssen, E.J., Nordlander, T.E., SINTEF ICT: MSS-Adjusts Surgery data [WWW] Available from: http://www.comihc.org/index.php/Test-Bed/3-years-of-surgery-data-at-sab.html [Accessed date: February 2017].

[14] Nawaz, M., Enscore, E. E., & Ham, I. "A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem". *Omega*, 11(1), pp 91–95, 1983.

[15] Persson MJ, Persson JA, "Analysing management policies for operating room planning using simulation" Health Care Management Science 13: pp 182–191, 2010.

[16] Razmi J,Yousefi M.S, and Barati M. "A stochastic model for operating room unique equipment planning under uncertainty". IFAC-PapersOnLine 48-3, pp 1796–1801, 2015.

[17] Saadouli H, Jerbi B, Dammak A, Masmoudi L, Bouaziz A. "A stochastic optimization and simulation approach for scheduling operating rooms and recovery beds in an orthopedic surgery department", Computers & Industrial Engineering 80, pp 72–79, 2015.

A mathematical model and a heuristic for an operating room scheduling problem with setup times

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Mohamed Amine Abdeljaouad
Université de Tunis El Manar, Faculté des Sciences de Tunis, LIP2-LR99ES18, 2092 Tunis, Tunisia
amine.abdeljaoued@fst.utm.tn

Nour El Houda Saadani
Université de Tunis El Manar, Faculté des Sciences de Tunis, LIP2- LR99ES18, 2092 Tunis, Tunisia
Houda.Saadani@ensi.rnu.tn

Zied Bahroun
Industrial Engineering, College of Engineering American University of Sharjah, Sharjah, UAE
zbahroun@aus.edu

Belhassen Zouari
Mediatron Lab, Higher School of Communications of Tunis, University of Carthage, Tunisia
belhassen.zouari@supcom.tn

Abstract – This paper deals with an operating room scheduling issue modeled as a parallel machine problem with setup times and resource constraints. This problem is characterized by sets of operations to be scheduled on identical operating rooms. Each set of operation has to be performed by a specific surgeon. Moreover, a patient's preparation and a room's cleaning are required respectively before and after the process of each operation. The objective is to minimize the makespan.

A mathematical formulation, a lower bound and a heuristic approach to solve the problem are presented in this paper. Tests are carried out to assess the performances of the methods and the quality of the lower bound.

Keywords— scheduling; operating rooms; setups; parallel machines;

I. Introduction

The operating room scheduling problems are among the most studied issues in the healthcare optimization field. Indeed, operating rooms are one of hospitals' most used resources and optimizing their functioning can thus lead to an important cost reduction [7].

The problem concerned by our study consists of a daily operating room scheduling issue where n operations have to be scheduled on m parallel and identical operating rooms. These operations are distributed over r surgeons. This implies that two operations that have to be performed by the same surgeon cannot be processed at the same time (i.e. disjunctive constraints). We assume that all the hospital's upper level decisions preceding the daily scheduling phase are already set, including the assignment of the operations to the surgeons which is not a part of our problem. The surgeons are assumed to be available to perform their operations in the considered day.

Each one of the n operations is made up of the three following phases: The first phase consists of the patient and room preparation, the second phase consists of the surgical act and the third phase consists of the cleaning of the operating room. Only the second phase requires the presence of the surgeon. The two other phases can be performed while the surgeon is occupied with a surgical act in another operating room. No restriction on the operations sequence inside each room is considered. The preemption of the operations is however banned and no waiting time is allowed between the phases of a same operation. The considered objective is to minimize the makespan (i.e. the ending date of the schedule).

This problem is NP-hard for $m \geq 2$. This can be easily proven since the scheduling of n operations on m parallel machines (noted $Pm||Cmax$ using the notation of, Blazewicz et al. [4]) is already NP-hard for a number of machines equal or greater than 2. Since our problem is more general than the $Pm||Cmax$ (i.e. the two problems coincide when every surgeon has only one operation to perform), it is also NP-hard for $m \geq 2$.

The remainder of the paper is structured as follows: In the next section, we present a brief literature review on the problems that are close to ours. Then, in section III, we mathematically define the problem by providing a linear model. In section IV, we introduce three possible lower bounds for the problem and in section V we present a new heuristic that can reach a satisfying solution on a reasonable time. We lead an experimental study in section VI and we conclude the paper with our remarks in section VII.

II. Literature review

The literature on the operating rooms optimization field is very large. Cardoen et al. [6] present a well-structured review on the operating room planning and scheduling problems. Several constraints were dealt with and several performance measures were accounted for (makespan [11], costs' minimization [5], the maximization of the room's utilization rate [8] and the minimization of the overtime work [10]).

Among the articles that include the patients' preparation and rooms' cleaning times, we can mention Mancilla and Storer [14], which study the scheduling of a single surgeon's operations on two operating rooms, taking into account the uncertainty over the durations of the operations' three phases. As for our problem, only the second phase needs the surgeon's presence. The objective is to minimize the weighted sum of the rooms' idle time and overtime work (i.e. the working time

after the rooms' regular hours) as well as the surgeon's waiting time. The authors represent this uncertainty by a set of scenarios and provide a stochastic MILP model that uses a sample average approximation, which calculates the average value of the objective function over the set of scenarios. They also present a decomposition based approach to solve instances of bigger sizes. Batun et al. [3] also take into account the three phases of the operations as well as the uncertainty over their durations, which is represented by a set of scenarios. The considered operating rooms here are identical, but the sequence of operations of every surgeon is already defined. The objective function is to minimize the weighted sum of the costs linked to the number of opened rooms, the overtime work and the surgeons' idle time. The authors present a two-steps stochastic mathematical model, a set of valid inequalities and an L-shaped method to solve the problem.

Guinet and Saadani [12] consider the scheduling of surgeons' operations on identical operating rooms, but without taking into account the three phases of the operations. The authors considered the makespan minimization and compared a set of two-steps heuristics, which start by arranging the operations in a specific order and then assigning them to the operating rooms. Among the procedures used to arrange the operations we can mention SPT (shortest processing time), LPT (longest processing time) and an adapted version of the NEH heuristic [16]. The assignment of the operations to the operating rooms is done by procedure like FAM (first available machine), LBM (last busy machine) or the heuristic of Herrmann and al. [13]. The results they obtained show that the NEH-LBM and the LPT-LBM heuristics are the best performers. Abdeljaouad et al. [1] present a strip-packing formulation for the same problem and a dichotomic algorithm to solve real-sized instances.

We extended our literature review to the parallel machine scheduling with setup times issue, since it presents several similar characteristics to our problem. Allahverdi et al. [2] present a review on the scheduling problems involving setup considerations. The main difference between the parallel machines with setup times problems and ours is that they do not consider groups of operations (i.e. each job is equivalent to one operation and is independent from the other jobs), and that the setups occur only before the operations. In these problems, two kinds of setups are generally considered: the setups that depend on the scheduling sequence and those that do not. Among the problems that considered the makespan minimization, we can mention França et al. [11] which provide a three-step heuristic to solve the parallel machine scheduling problem with sequence setup times. This method builds an initial solution by grouping the operations that share the lowest setup times on the same machine, then, the second step consists in looking for a better solution using a tabu search method. The third step looks for a possible improvement of the final solution by rearranging the operations assigned on the busiest machine. Tang [17] also consider the sequence dependent setup times and families of operations: when two operations from the same family are scheduled one after another on the same machine, a minor setup time is needed, and when the successive operations do not belong to the same family, a major setup time is needed. The authors provide two lower bounds and a heuristic that consists in assigning the operations to the machines by batches.

The problem that we study differs from those treated in the literature in many ways. First, the groups of operations and the disjunctive constraints make its structure very distinct from the usual parallel machine scheduling with setups issue, where the operations considered are independent from one another. Another special feature is that the setup times that we consider come before and after each operation. Our problem is thus an extended version of the scheduling with setup considerations, combined with the parallel machine scheduling under resources constraints issue [4] (i.e. the surgeons can be considered as the resources used by the operations). We propose in this paper a mathematical model and a heuristic method to solve it, as well as three new lower bounds to which we compare our solutions.

III. Mathematical model

In this section, we mathematically define the problem. Let us first introduce the notations that will be used by the model:

- $n,\ i$: Number and index of operations.
- $m,\ j$: Number and index of operating rooms.
- $r,\ k$: Number and index of surgeons.
- h_k : Number of operations of surgeon k.
- p_{ik}: The processing time of surgical operation i performed by surgeon k.
- c_{ik}: The ending date of surgical operation i performed by surgeon k.
- pre_{ik}, pos_{ik} : Respectively the patient's preparation duration and the operating room's cleaning duration before and after the surgical operation i performed by surgeon k. These variables are indexed by i and k because we consider that they only depend of the surgical operation they are related to: the patient's preparation is only specific to the surgical act it precedes. Same thing for the room's cleaning.
- M: A large number.
- C_{max}: The closing date of the operating theatre.

The decision variables are the following:

- Z_{ik}^{j} = 1 if the surgical operation i performed by surgeon k is assigned to operating room j. And 0 otherwise.
- X_{ik}^{s}= 1 if i is the s^{th} operation performed by surgeon k. And 0 otherwise.
- Y_{ikj}^{t} = 1 if the surgical operation i performed by surgeon k is assigned to rank t on the operating room j. And 0 otherwise.

Below we give the mathematical model:

$$Min\ C_{max} \qquad (1)$$

Subject to

$$\sum_{j=1}^{m} Z_{ik}^{j} = 1 \qquad \forall k = 1, \dots r, \quad \forall i = 1, \dots h_k \tag{2}$$

$$\sum_{i=1}^{h_k} X_{ik}^{s} = 1 \quad \forall k = 1, \dots r, \quad \forall s = 1, \dots h_k \tag{3}$$

$$\sum_{s=1}^{h_k} X_{ik}^{s} = 1 \quad \forall k = 1, \dots r, \quad \forall i = 1, \dots h_k \tag{4}$$

$$c_{ik} + M(1 - X_{ik}^{s}) \geq c_{lk} + p_{ik} - M(1 - X_{lk}^{s-1}) \quad \forall k = 1, \dots r, \ \forall i, l = 1, \dots h_k, \ \forall s = 2, \dots h_k \tag{5}$$

$$c_{ik} \geq pre_{ik} + p_{ik} \quad \forall k = 1, \dots r, \ \forall i = 1, \dots h_k \tag{6}$$

$$C_{max} \geq c_{ik} + pos_{ik} \quad \forall k = 1, \dots r, \ \forall i = 1, \dots h_k \tag{7}$$

$$c_{ik} + M\left(1 - Y_{ikj}^{t}\right) \geq c_{lq} + pos_{lq} + pre_{ik} + p_{ik} - M\left(1 - Y_{lqj}^{t-1}\right) \quad \forall k, q = 1, \dots r, \ \forall i = 1, \dots h_k, \ \forall l = 1, \dots h_q, \ \forall t = 2, \dots n \tag{8}$$

$$\sum_{t=1}^{n} Y_{ikj}^{t} = Z_{ik}^{j} \quad \forall k = 1, \dots r, \ \forall j = 1, \dots m, \ \forall i = 1, \dots h_k \tag{9}$$

$$\sum_{k=1}^{r} \sum_{i=1}^{h_k} Y_{ikj}^{t} \leq 1 \quad \forall j = 1, \dots m, \ \forall t = 1, \dots n \tag{10}$$

$$\sum_{k=1}^{r} \sum_{i=1}^{h_k} Y_{ikj}^{t} \leq \sum_{k=1}^{r} \sum_{i=1}^{h_k} Y_{ikj}^{t-1} \quad \forall j = 1, \dots m, \ \forall t = 2, \dots n \tag{11}$$

$$Z_{ik}^{j} \in \{0, 1\} \quad \forall k = 1, \dots r, \ \forall i = 1, \dots h_k, \ \forall j = 1, \dots m \tag{12}$$

$$Y_{ikj}^{t} \in \{0, 1\} \quad \forall k, q = 1, \dots r, \ \forall i = 1, \dots h_k, \ \forall t = 1, \dots n \tag{13}$$

$$X_{ik}^{s} \in \{0, 1\} \quad \forall k = 1, \dots r, \forall i, s = 1, \dots h_k, \tag{14}$$

The objective function (1) is the minimization of the closing date of the operating theatre. This closing date is given in constraints (7) and is greater or equal to the ending date of every operation plus the cleaning time that comes after. This objective is equivalent to a makespan minimization (i.e. the ending date of the schedule). Equations (2) ensure that every operation is assigned to only one operating room. Constraints (3), (4) and (5) reflect the disjunctive constraints between the operations of a same surgeon (i.e. the operations of a same surgeon cannot be performed at the same time). Inequalities (6) ensure that the ending date of an operation is greater or equal to the preparation's duration it requires plus the duration of its surgical phase. Constraints (8) ensure that each operating room handles one operation at a time; a surgical operation cannot start unless its preparation and the room's cleaning following the previous operation (if it exists) are finished. Equations (9) assign to each operation a rank on an operating room. Inequalities (10) reflect that each rank on an operating is assigned to at most one operation; while inequalities (11) ensure that the ranks assigned to the operations on the operating rooms are successive (this constraint is needed since we do not know how many operations are assigned to each operating room. Constraints (12), (13) and (14) indicate the binary nature of variables Z_{ik}^{j}, X_{ik}^{s} and Y_{ikj}^{t}.

We performed a battery of tests to assess the quality of this model, using the 12.5 IBM Cplex solver. The durations of the operations (in minutes) are generated by a log-normal distribution. The mean value and the standard deviation that we used for this distribution function are respectively 120 and 60. The preparation's duration that precedes every operation is equal to 5% of the surgical duration, and the cleaning following each operation is equal to 10 minutes. The tests were performed on a 4 GB RAM computer with an I3 processor. Their results are displayed in Table I. In the remaining of the paper, OR, SG and OP respectively refer to the operating rooms, surgeons and operations when defining the test instances. 10 tests are performed for each instance size. Although the (2 OR/ 3 SG/ 9 OP) and the (3 OP/ 4SG/ 8 OP) instances are solved in a quite quick time, the solving times of the bigger instances quickly become unreasonable. The (3 OR/ 4 SG / 12 OP) instances require more than 2 hours to be solved. The disadvantage of this model is that we do not know in advance how many operations would be scheduled in every room. By considering, as we did in constraints (9), (10) and (11), that the number of operations per room ranges from 1 to n, which is the total number of operations, the model may have to perform unnecessary computations. Restricting this number can improve the model's solving times.

TABLE I. Mathematical model's solving times

Instance's size	**Solving time**
2 OR / 3 SG / 6 OP	1 second
2 OR / 3 SG / 9 OP	2 minutes
2 OR / 3 SG / 12 OP	1 hour 30 minutes
2 OR/ 4 SG / 12 OP	More than 3 hours
3 OR / 4 SG / 8 OP	10 seconds
3 OR / 4 SG / 12 OP	More than 3 hours

IV. Lower bound

Since the exact solving is not efficient to solve big instances from an NP-hard problem and takes a very long time, the alternative is to use heuristic methods that can reach a satisfying solution in a reasonable time. To assess the quality of such methods, and since we cannot know the value of optimal solutions for the big-sized instances, we have to

compare the heuristic's solutions to lower bounds. In this section, we present three lower bounds for our problem.

The first lower bound ($LB1$) consists of measuring the workload of each surgeon (i.e. the sum of durations of his/her surgical operations) and selecting the surgeon with the longest workload. In the best case, this surgeon performs his operations without any waiting time between them, when the required preparation and cleaning before and after his operations are done while he is performing another operation in another operating room. In this case, only the preparation time prior to his first operation as well as the cleaning time following his last one are accounted for, since they cannot be hidden by the "operating time". This lower bound can thus be written as follows:

$$LB1 = min_k(pre) + \sum_{i=1}^{h_k} p_{ik} + min_k(pos) \quad (15)$$

With k referring to the surgeon with the longest workload, $min_k(pre)$ referring to the preparation with the lowest duration among the preparations preceding the surgeon's operations, and $min_k(pos)$referring to the cleaning with the lowest duration among the cleanings following the surgeon's operations. Let us note that these preparation and cleaning with the lowest durations should not belong to the same operation, since the preparation precedes the first operation of the surgeon, and the cleaning follows his last operation.

In the second lower bound ($LB2$), we consider the dividing of the sum of all the surgical durations on the number of operating rooms. If we have m operating rooms, each one of them requires at least one preparation preceding its first operation and one cleaning following its last one, in the case where all the other preparations and cleanings on this room are "hidden" by the surgical operations that are being performed on the other operating rooms. So, in the best case, these two setups that are accounted for are the preparation with the lowest duration and the cleaning with the m^{th} lowest duration – or the cleaning with the lowest duration and the preparation with the lowest duration. We thus select the couple of operations with the lowest sum of the setup durations. This lower bound can therefore be written as follows:

$$LB2 = \max\left(SPT_{[1]}(pre) + SPT_{[m]}(pos),\ SPT_{[m]}(pre) + SPT_{[1]}(pos)\right) + \sum_{k=1}^{r}\sum_{i=1}^{h_k}\frac{p_{ik}}{m} \quad (16)$$

With $SPT_{[m]}(pre)$(respectively $SPT_{[m]}(pos)$) referring to the preparation (respectively cleaning) with the m^{th} lowest duration.

The third lower bound also considers the dividing of the sum of all the surgical durations on the number of operating rooms. Then, we deduce the minimal number of operations by operating room (by summing the operations with the longest durations until reaching this division value). Let nb be this number of operations. We can thus deduce the minimal number of preparations and cleaning on each operating room (i.e. each one from these nb operations requires a preparation and a cleaning before and after it). In the best case, these setups are those having the minimal durations and this bound can therefore be written as follows:

$$LB3 = \sum_{j=1}^{n} SPT_{[j]}(pre) + \sum_{k=1}^{r}\sum_{i=1}^{h_k}\frac{p_{ik}}{m} + \sum_{j=1}^{n} SPT_{[j]}(pos) \quad (17)$$

The lower bound that we keep is LB, which is equal to the maximum value between $LB1$, $LB2$ and $LB3$. To assess the quality of this bound, we led 30 comparative tests to the optimal solutions provided by our mathematical model. The average relative gap over these 30 tests between LB and the optimal solution is about 6.04%. Details of these tests are given in Table II, 10 tests are performed for each instance size.

TABLE II. Average relative gap between the lower bound and the optimal solutions

Instance's size	Average relative gap
2 OR / 3 SG / 6 OP	9.72%
2 OR / 3 SG / 9 OP	6.24%
3 OR / 4 SG / 8 OP	2.17%
Average	6.04%

V. New heuristic to solve the problem

In this section, we present the heuristic that we propose to solve the problem. The idea of this iterative algorithm is to look for the scheduling that can hide the setup times by the operating phases as much as possible. At each iteration, the algorithm selects one operation and assigns it to the operating room that brings the starting time of its operating phase as close as possible to the ready time of the surgeon (i.e. the date at which the surgeon finishes his previous operation if it exists). The priority of selection is given to the longest operation of the surgeon that has the longest remaining workload, and this order is recalculated at the end of each iteration. More details about the steps of this algorithm are given below:

Heuristic's steps:

- Step 1: Initialize the ready time t_k of each surgeon k (i.e. the date at which the surgeon is available) and the ending date tm_j of each operating room j to 0.
- Step 2: Arrange the surgeons in the decreasing order of their remaining workload (i.e. the sum of durations of their unscheduled operations).
- Step 3: Select the surgeon k with the longest remaining workload and pick his/her longest operation i.

- Step 4: Let pre_{ik}, $p_{ik}$and pos_{ik}respectively be the preparation's duration, the processing time and the cleaning's duration linked to operation i. Assign i to the first operating room j available at a time $tm_j \geq t_k - pre_{ik}$. In other words, the operation is assigned to the operating room that can process the most part of the preparation pre_{ik} while the surgeon is performing his current operation.
- Step 5: Update the values of t_k and tm_j as follows:

$$t_k := t_k + p_{ik};$$
$$tm_j := tm_j + pre_{ik} + p_{ik} + pos_{ik};$$

- Step 6: If, there are still unscheduled operations, go back to Step 2. Else, go to Step 7.
- Step 7: The solution of the scheduling is equal to the ending date of the most busy machine (i.e. the tm_j having the greater value among all the operating rooms).

The heuristic's complexity is $O(nlog(n))$ due to the sorting procedure it uses to arrange the operations.

VI. Experimental study

In this section, we lead experimental tests to assess the quality of our heuristic, by comparing it to the optimal solutions provided by our model for the small instances, and then to the lower bound that we introduced in the previous section for the instances of bigger size.

A. Instances of tests

We now present the instances on which we will experiment our heuristic. We start by testing it on the instances that we solved by the mathematical model. This allows us to compare the solutions returned by our heuristic to optimal ones. These instances are given below:

- 2 OR / 3 SG / 6 OP
- 2 OR / 3 SG / 9 OP
- 3 OR / 4 SG / 8 OP

Then, we test our heuristic on instances of bigger size. We tried to generate real-sized instances of operating theatres. We tested three numbers of operating rooms: 2, 3 and 5. The numbers of surgeons tested for each instance are the following:

- 5 and 7 surgeons for the 2 operating rooms' instances.
- 7 and 10 surgeons for the 3 operating rooms' instances.
- 15, 20, 25 and 30 surgeons for the 5 operating rooms' instances.

The number of operations by surgeon is generated by a uniform distribution in the in the interval [2, 4]. The durations of the operations (in minutes) are generated by a log-normal distribution whose parameters are based on the historical data provided by a Norwegian hospital and available online [15]. The duration of every preparation is equal to 5% of the duration of the operation it precedes. We also set that the cleaning duration following each operation is equal to 10 minutes. For each instance size, we performed 10 tests, so 110 tests were performed in total.

B. Tests' results

We coded our heuristic using the C++ language. The tests were held on a 4 GB RAM computer with an I3 processor. For all of the 110 tests performed, the heuristic returned its solution within one second. Table III displays the average relative gap between the solutions returned by the heuristic and those returned by the mathematical model for each instance size. From this table, we can see that the heuristic is high performing. Indeed, the average relative gap returned over all the instances tested is about 4%.

TABLE III. Average relative gap between the solutions of the heuristic and the solutions of the model

Instance's size	**Average relative gap to optimal solutions**
2 OR / 3 SG / 6 OP	4.64%
2 OR / 3 SG / 9 OP	2.76%
3 OR / 4 SG / 8 OP	4.65%
Average	4.01%

TABLE IV. Heuristic's average relative gap to the lower bound

Instance's size	**Average gap to the lower bound**
2 rooms 5 surgeons	9.33%
2 rooms 7 surgeons	8.87%
3 rooms 7 surgeons	10.65%
3 rooms 10 surgeons	8.05%
5 rooms 15 surgeons	11.95%
5 rooms 20 surgeons	9.54%
5 rooms 25 surgeons	10.09%
5 rooms 30 surgeons	9.27%
Average	9.71%

Table IV displays the average relative gap between the solutions returned by the heuristic and the lower bound. We can notice that the returned average relative gap substantially rises for all the instances tested. The average gap is about 9.71%. The raise of this value can be explained by the fact that in many cases the lower bound does not correspond to an optimal solution. Indeed, in most of the tests, there may not be a scheduling that can hide the setup times by the operating durations.

Let us note that we compared the three bounds that we presented in section IV. For the 30 tests that we performed on the small instances (Table III) that are characterized by close numbers of surgeons and operating rooms, the lower bound $LB1$ performs the best; it gives the greatest value among the bounds tested in 14 tests, followed by the bound $LB3$ which has the best value in 11 tests, and the bound $LB2$, which is the best in 5 tests. Over the 80 tests realized on the bigger instances (Table IV) however, where the number of surgeons is much bigger than the number of operating rooms, the lower bound $LB3$ is always the best one, giving the greatest value in each of the 80 tests.

VII. Conclusion

In this paper, an operating room scheduling problem is studied. This problem is characterized by a set of surgeons that have to perform their operations on identical operating rooms. Each operation is made of three phases: the patient's preparation, the surgical act and the room's cleaning. Only the second phase requires the presence of the surgeon and the operations that have to be performed by a same surgeon cannot be processed at the same time. This problem is NP-hard and has been compared to well known problems such as the parallel machine scheduling problems, the scheduling under resource constraints problems and the scheduling involving setups considerations problems.

We provide a linear mathematical formulation that reflects the problem and its constraints, and then we present a heuristic method that can reach a satisfying solution in a reasonable time. This iterative heuristic selects the operations one by one on the basis of a priority order and then assign them to the operating rooms in such a way that the most part of the preparation and cleaning times by the operating phases. To assess the quality of this heuristic, we introduce three lower bounds of the optimal solution. The comparative tests between the greatest value of the lower bounds and the optimal solutions returned by the mathematical model gave an average relative gap close to 6%. In the experimental study that we led, we compared the solutions returned by our heuristic to the optimal ones that we obtained with the mathematical model for the small instances, and then to the lower bound for the instances of bigger size. The average relative gap to the optimal solutions and to the lower bound is respectively about 4.01% and 9.71%.

The next step of this work is to look for a possible improvement of the solutions returned by the heuristic. This can be done using a metaheuristic method that may require a longer running time but visits a large number of solutions. One of the metaheuristics that we can adapt is the simulated annealing that will exchange the operations' positions in the assignment order.

References

[1] Abdeljaouad M.A., Bahroun Z., Saadani N.E.H., "A dichotomic algorithm for an operating room scheduling problem", in proceeding of international conference on control, decision and information technologies: *CoDIT*, pp 134-139, 2014.

[2] Allahverdi, A., Gupta, J.N.D., & Aldowaisan, T. "A review of scheduling research involving setup considerations", Omega, the International Journal of Management Science, 27, pp 219-239, 1999.

[3] Batun S., Denton B.T., Huschka T.R and Schaefer A.J, "Operating room pooling and parallel surgery processing under uncertainty", INFORMS Journal on Computing, Volume 23, pp 220-237, 2011.

[4] Blazewicz, J., Lenstra, J. K., & Kan, A. H. G. R.. "Scheduling subject to resource constraints: classification and complexity", Discrete Applied Mathematics, 5(1), pp 11-24, 1983.

[5] Calichman M.V., "Creating an optimal operating room schedule", AORN Journal 81, pp 580-588, 2005.

[6] Cardoen B., Demeulemeester E., Belien J., "Operating room planning and scheduling: A literature review", European Journal of Operational Research, vol. 201(3), pp. 921-932, 2010.

[7] Denton, B., Viapiano, J., & Vogl, A. "Optimization of surgery sequencing and scheduling decisions under uncertainty". Health Care Management Science, 10(1), 13–24, 2007.

[8] Dexter F., "Operating room utilization: Information management systems", Current Opinion in Anaesthesiology 16, 2003, pp 619-622.

[9] Fei, H., Combes, C., Meskens, N., & Chu, C. "Endoscopies scheduling problem: a case study". Information Control Problems in Manufacturing,12(1), 635-640, 2006.

[10] Fei H., Meskens N., Chu C., "An operating theatre planning and scheduling problem in the case of a block scheduling strategy", in: Proceedings of the International Conference on Service Systems and Service Management, 2006.

[11] França, P.M., Gendreau, M., Laporte, G., & Müller, F.M. "A tabu search heuristic for the multiprocessor scheduling problem with sequence dependent setup times", International Journal of Production Economics, 43, pp 79-89, 1996.

[12] Guinet, A., & Saadani, N. E. H. "Nomadisme chirurgical et ordonnancement de blocs opératoires". In Proceeding of Conférence francophone en Gestion et Ingénierie des Systèmes Hospitaliers. Clermont-Ferrand, France: *GISEH*, 2010.

[13] Herrmann, J., Proth, J., & Sauer, N. "Heuristics for unrelated machine scheduling with precedence constraints". European Journal of Operational Research, 102(3), pp 528-537, 1997.

[14] Mancilla C, Storer RH, "Stochastic Sequencing of Surgeries for a Single Surgeon Operating in Parallel Operating Rooms". IIE Transactions on Healthcare Systems Engineering Volume 3, Issue 2, pp 127-138, 2013.

[15] Mannino, C., Nilssen, E.J., Nordlander, T.E., SINTEF ICT: MSS-Adjusts Surgery data [WWW] Available from: http://www.comihc.org/index.php/Test-Bed/3-years-of-surgery-data-at-sab.html [Accessed date: February 2017].

[16] Nawaz, M., Enscore, E. E., & Ham, I. "A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem". *Omega*, 11(1), pp 91–95, 1983.

[17] Tang, C.S. "Scheduling batches on parallel machines with major and minor set-ups", European Journal of Operational Research, 46, pp 28-37, 1990.

Planning surgical activities considering post-surgical capacity

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Valérie Dorval[*†], Abdessamad Ait El Cadi[†], Daoud Aït-Kadi[*], Abdelhakim Artiba[†]
[*]Department of Mechanical Engineering, Laval University, Canada
[†]LAMIH, University of Valenciennes, France

Abstract— Surgical activities constitute an important cost center for health care establishments and receive great attention from the stakeholders. It is found that hospitalization beds on the wards can be a bottleneck in the service and are commonly not considered in the planning stage of the activities. To help dealing with this situation, a mathematical model has been developed, allowing to include the occupancy on the ward in the scheduling process. Since the length of stay of patients after a surgery is highly variable, a simulation model is also used to measure the robustness of the solution and compare it to the actual situation.

Keywords—Operating rooms; planning; post-surgical capacity

I. Introduction

Access to specialty and subspecialty services is a fundamental concern for the Health and Social Services Department of the province of Québec, in Canada, in order to improve the services dispensed to the population. Facing a constant increase in the demand for care, the health care organizations need to: manage effectively their resources (operating theaters, human resources, beds, etc.) and analyse and anticipate the needs.

Surgery services are facing difficulties to meet demand and patients face long waiting times before being treated. In order to improve services, maximum waiting time limits have been established for certain types of surgery, however, adding a constraint to the already overloaded system. Finally, the cancelation of surgeries due to the lack of available beds in intensive care units and wards is considered fairly common, causing a bottleneck in the flow of patients.

In this context, the aim of this paper is to propose a procedure for planning surgical activities that takes into account the capacity of post-surgical care units in order to improve bed utilization and thus increase the flow of patients through the system.

Fig. 1 shows the surgical activity planning from the patient's perspective and his path through the system.

Initially, the patient will consult a specialist who will decide if surgery is needed. If it is the case, the patient will be placed on a waiting list depending on the type of intervention to be performed.

The pre-booking service will determine the approximate date of surgery. This date is based on the availability of the surgeon and priority access to operating rooms granted to each surgical specialty. A pre-operative appointment will also be fixed to perform some prep tests and provide the patient with necessary information enabling him to prepare for his intervention. Obtaining the results of the pre-operative tests will ensure that everything is in order for the intervention and the surgery date will be officially validated depending on the availability of operating rooms.

Finally, the surgery will take place and the patient will be admitted to a ward for post-surgical care until the surgeon decides he has recovered enough to leave the hospital.

Although the patient's path seems relatively straightforward, since the availability of a bed is not considered when setting a surgery date, the schedule can be disrupted by the shortage of beds on the wards.

This problem can manifest itself in different ways. Sometimes patients have to stay longer in the recovery room, waiting for a bed to be available downstream, delaying the post-operative care for rehabilitation. It may also happen that patients will be placed on another care unit, thus complicating the tours for various professionals involved. In the worst case, some interventions will be canceled or postponed because of a lack of bed availability.

On a decisional perspective, we situate this work in the gray zone, straddling between the tactical and operational decision levels. It is indeed during this planning stage where a preliminary, and then final, date is assigned to each surgery that it would be relevant to include the expected duration of stay of the patient, avoiding overflow and allowing to level the occupancy.

The aim is to develop a decision support tool integrating, in the planning phase, the whole admission and treatment process of each patient to reduce the number of canceled surgeries or postponements, reduce the number of patients placed on other wards, improve timeliness and reduce waiting lists.

The remainder of the paper is organized as follows. The second section presents various works that have addressed similar problems. The third section exposes our method to create the surgical schedule; the linear programming model and the simulation model used to validate the model. Finally, the forth section provides the summary and conclusion.

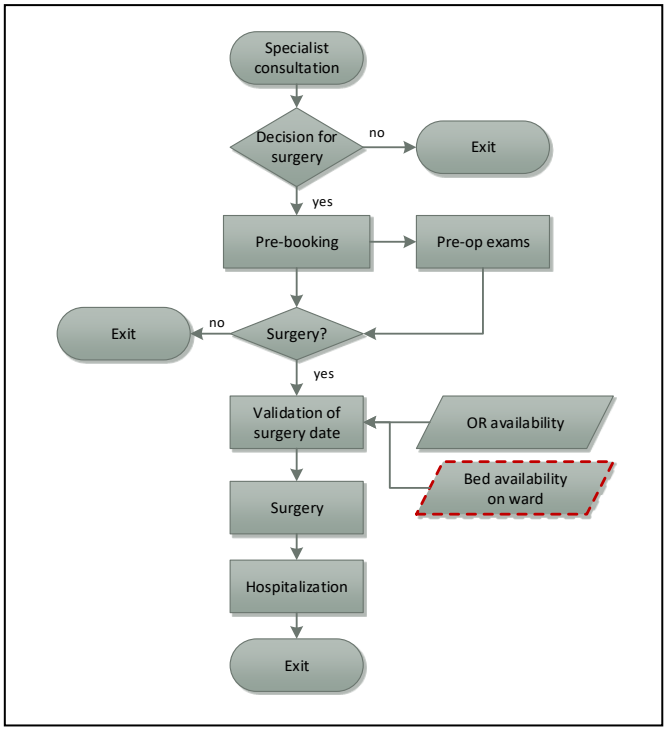


Fig. 1. Patient's path

II. Literature review

From one country to another (and even from one province to another) the funding of hospitals, and more specifically in this case, of operation theaters varies a lot. Therefore, difficulties are not necessarily found at the same level and stakes are different. Due to those differences, management priorities will not be the same [1]. In several countries, such as France for example, patients are hospitalized the day before their surgery. That way, when an intervention is delayed, the patient occupies a bed longer, before intervention. In that case, bed availability does not have the same importance. The aim is usually to optimize the operating room utilization time and resources. The vision of the problem will not be quite the same as in the case of Québec province, where the number of beds licensed for each hospital is fixed and allocated to different specialties (general medicine, urology, orthopedics, obstetric, gastroenterology, ophthalmology, etc.).

In general, planning surgical activities is a well-covered topic in the literature [2], [3], [4], [5]. At the operational level, planning is usually done in two stages. Initially, patients are selected to undergo surgery within a period ranging from one day to a few weeks. Thereafter, the surgeries for each period are scheduled according to various objectives.

In [6], a model is proposed where patient satisfaction (in terms of length of stay before the intervention) and performance of resources (in terms of overtime) are the criteria to optimize. Several issues are taken into account such as capacity (regular hours, maximum overtime allowed by labor laws, availability of surgeons and type of equipment available in each operating room) and the time windows (patients' dates of admission and maximum deadlines) when selecting interventions. Once the choices are made, scheduling of interventions is made for each day and each operating room considering synchronization of needs in human and material resources.

A four-step approach is proposed by [7]. First, the surgical cases are assigned to a particular day of the planning horizon, based on a value assigned to each case corresponding to the emergency level. Second, for each day, cases are scheduled according to different strategies, evaluated by simulation in step three. Finally, when necessary, rescheduling is made based on feedbacks.

On their side, [8] define a score computed with the difference between the maximum time limit and the due date, in addition to the duration of intervention. This score is used to select interventions that will be carried out over a period of one week. This way of organizing the objective function allows a tradeoff between surgeries that are longer (and therefore more difficult to plan) and those more urgent.

The aim of [9] is to plan as many cases as possible considering the availability of operating rooms, human resources and equipment while minimizing the number of staffed operating rooms and the violation of soft constraints. A two-phase heuristic is proposed to achieve this goal.

The papers listed above have in common that they do not address the availability of downstream resources. Indeed, from the point of view of post-operative care units and the consideration of their occupation in the planning stage, studies are scarcer [2].

However, some works have approached this problem. In particular, [10] proceed substantially in the same way as [6], but their planning horizon is rather a day. In that manner, they can take into account disruptions that would occur in the current day to schedule the next one. The scheduling part of the problem aims at minimizing the overtime in addition to consider bed availability in the recovery room.

In [11], at the stage of the case mix problem, the expected length of stay is taken into account to adjust the release dates and minimize the occupation on Saturdays and Sundays, limiting necessary resources during weekends.

Some works treat the problem of downstream capacity and its variability, but at the tactical decision level, the stage where time windows are allocated to each surgical group. Thus, the finer planning, where cases are selected and assigned to specific time slots is left untreated.

For example, [12] build the operating schedule by trying to level the bed occupancy. In order to do so, the "expected bed shortage" is minimized. This allows to predict performance measures such as the expected occupancy, the variance of this occupancy, the bed shortage and the probability of bed shortage for every day.

The impact of a given master surgical schedule (MSS) on the downstream capacity (ICU and wards) is measured in [13] and different methods (exact and heuristic) are tested to minimize the cost for the downstream resources.

In the same spirit, [14] seek to reduce the number of required beds by rearranging the OR blocks schedules in the planning

horizon, taking into account the bed occupancy probability distribution.

The aim [15] is to build a MSS that minimizes the number of patients waiting, the underutilization of operation theater and reduces inpatients bed occupancy peaks in mid-week based on the number of available hours for each surgical specialty. Nonetheless, those models do not take into account the priority level of patients and time past in the waiting list.

Finally, closer to our problematic, [16] propose a stochastic optimisation approach for scheduling operating rooms and recovery beds in an orthopedic surgery department, but only consider the stochasticity in the surgery duration and the beds' availability in the recovery room.

III. Proposed approach

The aim of the project is to level the bed occupancy by taking length of stay into consideration in the planning process, limiting the placement of patients on other wards or postponement of their surgery. Since every specialty has a ward with a certain number of beds reserved, it is reasonable to consider each specialty individually. Orthopedic surgery is chosen for this study due to the growth of demand caused by the aging population.

As presented in the Fig. 2, our methodology works in two steps: first, a mathematical model uses a data set as input to find the operating room schedule that will allow the maximum number of surgeries considering the limited bed capacity. With the obtained schedule (the patient-date pairing) a simulation model is used to measure the quality of the solution using real data.

A. Mathematical model

First, a mathematical model is used to build the surgical schedule. In this step, the length of stay of patients is usually not known. Since it is not an easily predictable information and is highly variable, we used historical data to separate the patients into three different categories. Firstly, we have the ambulatory surgeries, in which patients are not staying overnight at the hospital. Secondly, the overnight patients are split in two other groups, one with a short length of stay and low variability, and the other one with longer length of stay and higher variability.

Each surgeon has some booked time in an operating room and a pool of patients awaiting surgery. The goal is to build a surgical schedule at the operational level (which patient?, which day?), respecting the surgeons' schedule and ensuring the availability of a bed on the post-operative care unit for each patient undergoing intervention. In order to do so, we first used a linear programming model defined as follow.

TABLE I. Nomenclature

J	Set of available surgeons
K	Set of days in the planning horizon
I_j	Set of cases that can be performed by surgeon j
d_i	Estimated operating time of case i
c_{ik}	Value of case i if operated on day k
s_i	Length of stay of case i
L_{jk}	Available time of OR for day k and surgeon j
P	Number of beds in the ward
p	Number of available beds in the ward

Maximize

$$\sum_{i\in I}\sum_{k\in K} c_{ik}x_{ik} \quad (1)$$

Subject to:

$$\sum_{k\in K} x_{ik} \leq 1,\ \forall\, i \in I \quad (2)$$

$$\sum_{i\in I_j} d_i x_{ik} \leq L_{jk},\ \forall\, j \in J, \forall\ \mathrm{k} \in K \quad (3)$$

$$y_{ik} = \sum_{k'=max\{1,\ \mathrm{k}-s_i+1\}}^{k} x_{ik'}, \forall i \in I_j, \forall j \in J, \forall k \in K \quad (4)$$

$$\sum_{j\in J}\sum_{i\in I_j} y_{ik} \leq P,\ \ \forall k \in K \quad (5)$$

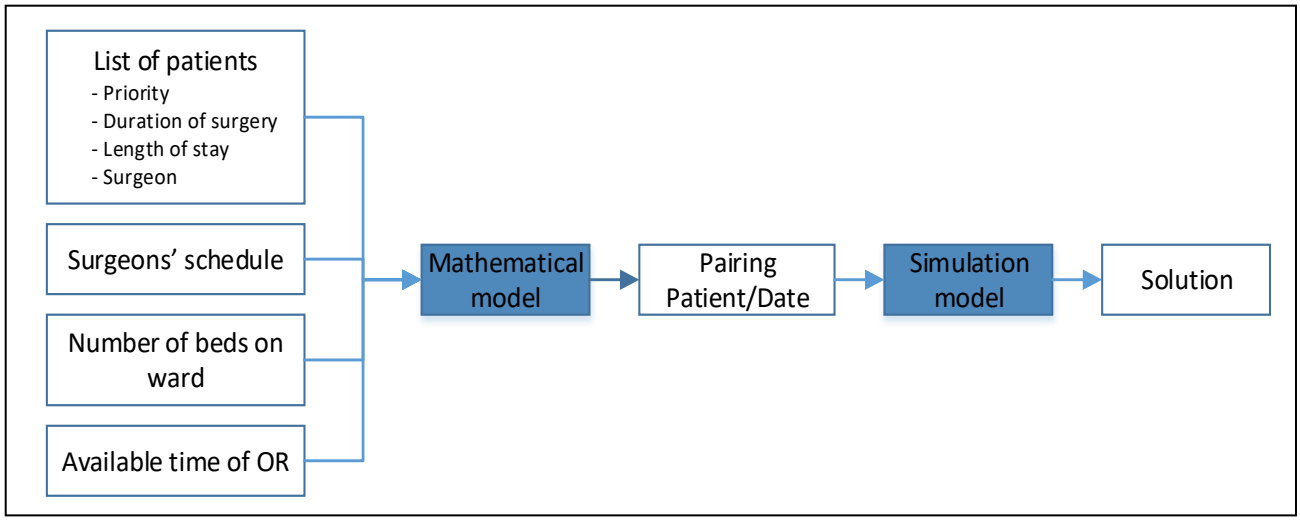


Fig. 2. Methodology

The objective function (1) is to maximize the total value of the chosen cases over the planning horizon. The value of a case c_{ik} is decided by the surgeon according to the urgency level of the case and the patient's waiting time before surgery The decision variable x_{ik} is equal to 1 if the case i is chosen for surgery on the day k, 0 if not

Constraints (2) ensure that each surgery i is performed at most once while constraints (3) make sure the available time L in OR is not exceeded. The equalities (4) define another decision variable, y_{ik}, equal to 1 if patient i is still in hospital on day k, 0 if not. This new variable allows (5) to let no more patient than available beds P on the ward for each day k.

The input data for the model are the planning horizon in days, the number of opening hours each day for the operating rooms, the number of available beds on the care unit and the surgeons. We must also provide the model with the waiting list of cases, based on which the schedule will be built. Thus, for each case, we will have an identification number, a specific surgeon, the duration of surgery in hours, the length of stay and a priority level. The working schedule for each surgeon is also supplied for the planning horizon (1 if the surgeon has an operating room available on day k, 0 if not).

As output, the model produces a pairing patient/date, indicating which day of the planning horizon will be operated each chosen patient, respecting the surgeons' working schedule, the number of available hours for each operating room and the number of beds on the care unit.

Concretely, as input to the model solved with Cplex, a list of 50 patients is chosen based on historical data, as well as the schedule of 8 surgeons. Based on the variability of the length of stay, a planning horizon of 21 days is used and the model is set for a ward of 13 beds. With those settings, the pairing patient/date is obtained and 38 cases are planned over the time horizon.

Some cases will not be selected, they will therefore stay in the waiting list for the next planning horizon. Since they are waiting longer, it is possible to increase their level of priority so they have a higher chance of being selected in the next planning exercise.

B. Simulation model

To validate the obtained schedule, a simulation tool has been developed using Arena software. Each patient is associated with its respective surgeon, operation time and length of stay. For the simulation, the length of stay is considered stochastic which means that for each patient, depending on its category (no overnight stay, short stay or long stay), a random length of stay is generated from a distribution based on historical data. The model retrieves those information from an Excel file.

As shown in Fig. 3, the arrival of a patient takes place as scheduled by the mathematical model. The patient then passes through an operating room where the surgeon takes care of him during the preset time of the intervention and goes through the recovery room. Finally, the patient is sent to the care unit if there is a bed available. If a bed is going to be freed later in the same day, the patient will wait in the recovery room. On the other hand,

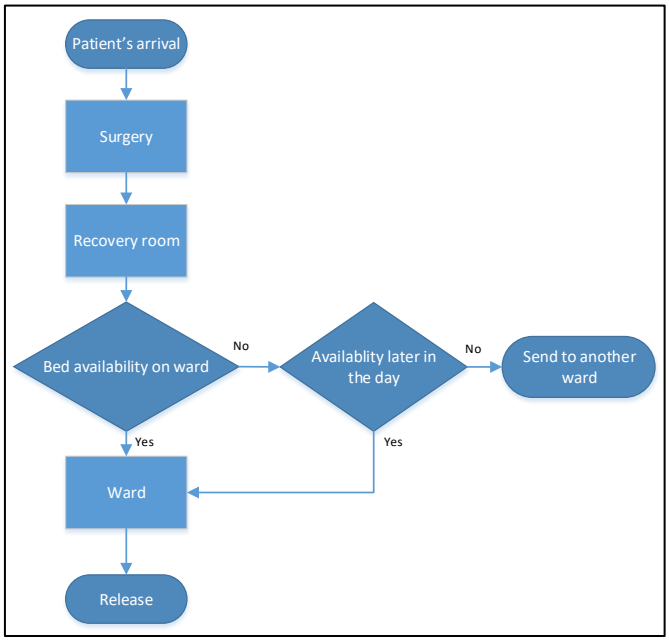


Fig. 3. Flow chart of the simulation

if no bed is released during the day, the patient will be sent to another care unit.

To decide the number of replications for the simulation, the equation (6) is used, where p represents the average number of non-treated patients and N, the total number of patients through the system. With a maximal error of 2% and a 95% confidence interval, it is found that 60 replications is satisfying.

$$\varepsilon = Z\sqrt{\frac{p(1-p)}{N}} \leq 2\% \qquad (6)$$

C. Results

On the 38 cases planned with the mathematical model to be hosted on the ward, the simulation shows that an average of 32.93 patients will actually be hosted, which means the remaining 5.07 patients will have to be placed elsewhere or the intervention postponed at the last minute. The high variability of the length of stay explains this difference between the number of patients planned with the mathematical model and the number actually hosted on the ward according to the simulation.

To compare the result of our model with the historical data of different wards' sizes, a productivity ratio is computed: the number of handled cases over the planning horizon divided by the total number of beds on the ward. Our model is designed for a 13 beds ward but the historical data comes from a 33 beds ward. This ratio calculated for our model with the simulation gives a 2.53 productivity ratio while in reality, over a year, the productivity ratio has an average of 1.35. Even if we look only at the productivity ratio for the best period, 2.06 is obtained, which is also inferior to the solution proposed.

IV. Conclusion

In a context where the availability of beds on the wards is a problem and patients see their surgery cancelled or postponed because of this lack of resource, we propose a model where the length of stay and its variability is taken into account in the planning stage of the surgical activities.

First, a mathematical model is used to maximize the number of surgeries while taking into account the bed availability on the ward. In this stage, three different kind of patients are planned, with a different average length of stay and standard deviation for each group. A plan specifying the date of surgery to be performed on the given planning horizon for each chosen patient is then obtained, respecting the surgeons' available time in the OR. This schedule is then used as entry to a simulation model, allowing to measure the robustness of the solution by using random length of stay drawn from the distribution for each category of patient.

For this stage of the study, we consider that the availability of human resources and equipment is not a problem. We also consider the problem solely from the elective point of view, since emergency cases are treated in a dedicated operating room. Finally, the variability in the duration of surgery is neglected. All those considerations are part of our concern to extend this work and bring it to reality and its implementation.

Acknowledgment

The present research work has been carried out in the context of the LIA (International Associate Laboratory) ROI-TML (Operational Research and Computer Science in Transportation, Mobility and Logistics) between the LAMIH UMR 8201 (France) and the CIRRELT (Canada). The authors gratefully acknowledge the support of these laboratories and relating institutions.

References

[1] E. Marcon, A. Guinet, and C. Tahon, *Gestion et performance des systèmes hospitaliers*, Lavoisier. Paris: Hermes Science, 2008.

[2] E. Demeulemeester, J. Beliën, B. Cardoen, and M. Samudra, "Operating Room Planning and Scheduling," in *Handbook of Healthcare Operations Management*, vol. 184, B. T. Denton, Ed. 2013, pp. 1–18.

[3] E. W. Hans and P. T. Vanberkel, "Operating Theatre Planning and Scheduling," in *Handbook of Healthcare System Scheduling*, R. Hall, Ed. Boston, MA: Springer US, 2012, pp. 105–130.

[4] F. Guerriero and R. Guido, "Operational research in the management of the operating theatre : a survey," pp. 89–114, 2011.

[5] B. Cardoen, E. Demeulemeester, and J. Beliën, "Operating room planning and scheduling : A literature review," *Eur. J. Oper. Res.*, vol. 201, no. 3, pp. 921–932, 2010.

[6] A. Guinet and S. Chaabane, "Operating theatre planning," *Int. J. Prod. Econ.*, vol. 85, no. 1, pp. 69–81, 2003.

[7] I. Beaulieu, M. Gendreau, and P. Soriano, "Operating rooms scheduling under uncertainty," in *Advanced Decision Making Methods Applied to Health Care*, E. Tànfani and A. Testi, Eds. Milano: Springer Milan, 2012, pp. 13–32.

[8] A. Agnetis, A. Coppi, M. Corsini, G. Dellino, C. Meloni, and M. Pranzo, "A decomposition approach for the combined master surgical schedule and surgical case assignment problems," *Health Care Manag. Sci.*, vol. 17, no. 1, pp. 49–59, 2014.

[9] W. Vancroonenburg, P. Smet, and G. Vanden Berghe, "A two-phase heuristic approach to multi-day surgical case scheduling considering generalized resource constraints," *Oper. Res. Heal. Care*, vol. 7, pp. 27–39, 2015.

[10] A. Jebali, A. B. Hadj Alouane, and P. Ladet, "Operating rooms scheduling," *Int. J. Prod. Econ.*, vol. 99, no. 1–2, pp. 52–62, 2006.

[11] E. Tànfani and A. Testi, "A pre-assignment heuristic algorithm for the Master Surgical Schedule Problem (MSSP)," pp. 105–119, 2010.

[12] J. Beliën and E. Demeulemeester, "Building cyclic master surgery schedules with leveled resulting bed occupancy," *Eur. J. Oper. Res.*, vol. 176, no. 2, pp. 1185–1204, 2007.

[13] A. Fügener, E. W. Hans, R. Kolisch, N. Kortbeek, and P. T. Vanberkel, "Master surgery scheduling with consideration of multiple downstream units," *Eur. J. Oper. Res.*, vol. 239, no. 1, pp. 227–236, 2014.

[14] J. T. van Essen, J. M. Bosch, E. W. Hans, M. van Houdenhoven, and J. L. Hurink, "Reducing the number of required beds by rearranging the OR-schedule," *OR Spectr.*, vol. 36, no. 3, pp. 585–605, 2014.

[15] X. Li, N. Rafaliya, M. F. Baki, and B. A. Chaouch, "Scheduling elective surgeries: the tradeoff among bed capacity, waiting patients and operating room utilization using goal programming," *Health Care Manag. Sci.*, vol. 20, no. 1, pp. 33–54, 2017.

[16] H. Saadouli, B. Jerbi, A. Dammak, L. Masmoudi, and A. Bouaziz, "A stochastic optimization and simulation approach for scheduling operating rooms and recovery beds in an orthopedic surgery department," *Comput. Ind. Eng.*, vol. 80, pp. 72–79, 2015.

A ConWIP approach to manage an integrated surgical suite: the case of the ophthalmology surgical suite in a Tunisian hospital

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Chabouh Safa
Université Tunis El Manar, ENIT, LR11ES20, Tunis, Tunisia.

Hammami Sondes
Université de Carthage, ENICAR, Université Tunis El Manar, LR11ES20 Tunis, Tunisia.

Hanen Bouchriha
Université Tunis El Manar, ENIT, LR11ES20, Tunis, Tunisia.

Eric Marcon
Université de Lyon INSA-LYON, Université Jean Monnet Saint-Etienne, DISP, EA 4570, 69621 Villeurbanne, France

Abstract— This paper evaluates the performance of Constant Work-In-Process (ConWIP) control strategy in an Integrated Surgical Suite (ISS) while accounting for uncertainty in service times. ConWIP is largely implemented in industrial environments but not yet in surgery departments. In this paper, we show the benefits of the implementation of the ConWIP system in an ISS. We first determine the theoretical number of ConWIP cards. This latter is further adjusted via a discrete event simulation model to take into account the variability in process durations. The computational experiments are performed based on real data of a hospital-integrated facility within an ophthalmology surgery department in a Tunisian hospital.

Keywords—ConWIP; simulation; Integrated surgical suite, resources'utilization.

I. Introduction

ConWIP (Constant Work-In-Process) is a production control system used mainly in industrial environments and introduced by [1] as a pull alternative to the Kanban system. It refers to any system maintaining constant the maximum amount of work in a production line or workshop in order to avoid the phenomena of congestion and to maintain a constant lead time. ConWIP is described in detail by [2]. Like Kanban, ConWIP is based on the use of cards or tickets [1]. However, the cards are allocated to a system of workstations instead of just one. In fact, a card is attached to a job at the beginning of the system and released when a job is processed in the last station. The card is then sent again to the beginning of the system, where it is attached to the next job to be processed. Therefore, the production system for ConWIP is perceived as a "black box" which justifies its ease-of- implementation and its applicability to complex environments.

Despite its ease of implementation in a complex environment and its benefits, the application of ConWIP is limited to industrial environments [3, 4]. To our knowledge, only [5] studied ConWIP in the health care management field to re-organize and optimize the workflow for robotic stereotactic radiotherapy treatments. In our study, we apply ConWIP to an Integrated Surgical suite ISS. The ISS is a complex environment: on one hand, it is composed of three stages (pre-operation, surgery and recovery); on the other hand, it treats two types of patients (inpatient and outpatient) having different care paths and sharing critical resources.

The surgical suite presents a hospital's greatest revenue source as well as its largest cost center including the costs of personnel (surgeons, anesthetists, nurses, etc.) and facilities (operating rooms, intensive care beds, etc.) [6]. That is why the managers aim to reduce continually the operational costs by a better utilization of their resources. We aim to propose to the hospital manager an effective and easy-to implement management tool (based on ConWIP) in order to optimize the utilization of the ISS resources. Many studies focused on optimizing the use of the surgical suite resources by applying operational research techniques such as optimization algorithms, stochastic programming and simulation-optimization approaches [7]. Their main drawback is related to the hardness of implementation in reality. This is due not only to reluctance of people but also to small unexpected issues not taken into account beforehand [2, 5]. That is why, many studies performed industrial engineering methods especially lean methods. For instance, process Improvement Workshops and 5S techniques [8] were performed to improve workflow in an ambulatory paediatric surgery centre by eliminating non-value added activities mainly patient waiting time. The implementation of Lean six sigma in the operating room (OR) allowed [9] to point out that patients experience long waiting times before entering the OR. In the same context, [10] applied value stream mapping method in order to improve the OR process. The author proposed to switch from a Push system to a Pull system [i.e. the patients are only transferred inward when the OR gives the go ahead]. The main limitation of these studies consists of focusing on lean's traditional implementation strategies without taking into account the variability in process durations inherent to surgical suite processes.

In this paper, we apply "ConWIP" approach for managing operations in an ISS. We first determine the theoretical number of ConWIP cards needed to implement the pull system. The WIP level is then adjusted via a discrete event simulation model to take into account the variability in process durations.

The remainder of this paper is organized as follows: section II describes the case study. Section III comprises the methodology for ConWIP's implementation in the ISS. The Section IV is devoted to the presentation and discussion of the experimentation results. Finally, concluding remarks and directions for future research are given in section V.

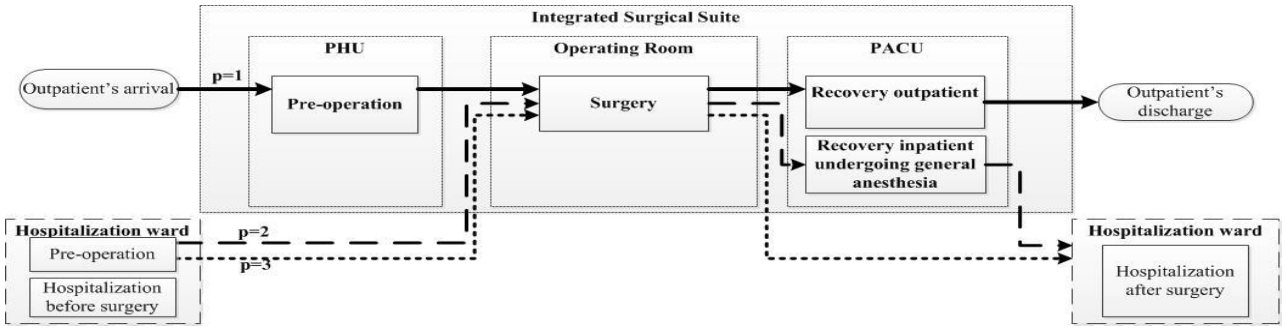


Fig. 1 Patients trajctories in the Tunisian ISS

II. The case study

The ISS deals mainly with two types of elective patients: outpatient and inpatient. During a surgery, in- and outpatients flows pass through three stages: the preoperative, operative, and postoperative stage. We studied patient flow in an ophthalmology ISS of a public Tunisian hospital surgery department illustrated in Fig. 1. It is composed of an inpatient ward and an ISS. Our field of study is limited to the surgical suite which involves three stages and does not consider the hospitalization ward. Hence, preparation and recovery resources in the hospitalization ward are supposed unlimited. In the ISS, the preoperative phase is done in the Preoperative Holding Unit (PHU) (respectively hospitalization ward) for outpatients (respectively for inpatients). The perioperative phase is performed in operating rooms for both patient types. The postoperative phase is completed whether in Post Anesthesia Care Unit (PACU) or hospitalization ward depending on patient's type and its anesthesia type. Only, outpatient and inpatient undergoing general anesthesia are routed to the recovery room. An inpatient undergoing local anesthesia is directly transferred to his bed in the hospitalization ward. As depicted in Fig. 1, there are three paths of patients: outpatient (which we refer to as p=1), inpatient undergoing general anesthesia (which we refer to as p=2) and inpatient undergoing local anesthesia (which we refer to as p=3). We distinguish three servers/stages with limited capacity as presented in TABLE I.

We note that some resources (PHU/PACU beds) are dedicated per patient type; others (ORs and related resources) are shared. For each path of patients, different types of acts are considered. Each type of acts is performed by specific surgeon types.

ConWIP can be applied to the ISS environment for several reasons. First of all, the ISS is working in a make to order environment given that surgeries are performed based on the patient's need/ demand. Second, the number of patients in the system (WIP) is conditioned, (or limited) by the limited capacity of resources during a certain period (ORs, recovery beds ...). Finally, the ISS is a complex environment that treats different patient types having different care paths.

III. Methodology

A. ConWIP Implementation

Under a basic ConWIP system, patient flow can be illustrated in Fig. 2. Upon his arrival, an outpatient takes his place in the waiting room for outpatients. Inpatients are already in the hospitalization ward. Once a ConWIP card is available (or empty), it is assigned to an in or outpatient according to a certain defined priority (e.g. outpatient first, according to a type of surgery,…). Once the surgical process is completed, the outpatient (respectively inpatient) is discharged (respectively is transferred to the hospitalization service) and the ConWIP card becomes available again and joins the queue of empty cards.

TABLE I. Limited resources per stage in the ISS

Server/ stage	Limited resources	Specification
Preoperative	Preparation beds	Dedicated to p=1
Surgery	ORs and Surgeons	Shared by all patient types
Recovery	Recovery beds	A predefined quota on the number of beds dedicated to patients of type p=1 and type p=2.

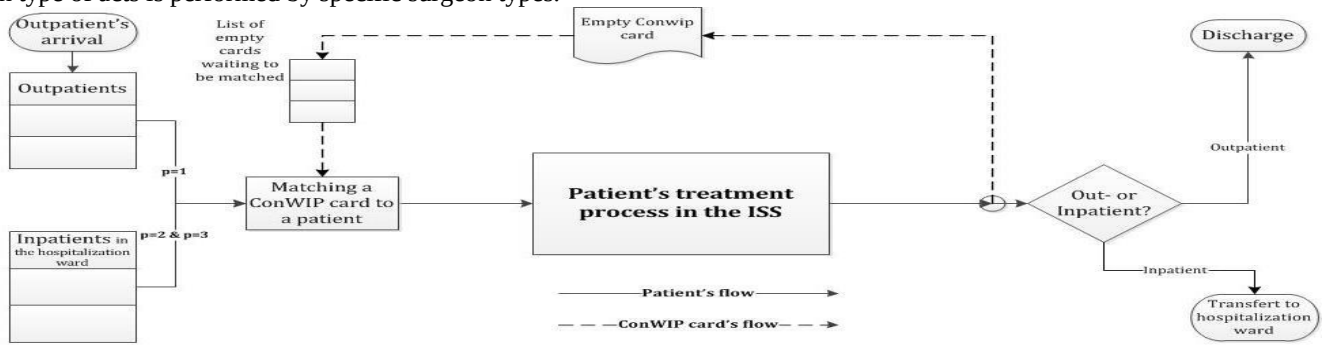


Fig.2 Workflow in the ISS under a basic ConWIP system

The implementation of a ConWIP system requires the determination and control of these parameters: (1) The production quota i.e. the desired production quantity (2) The Maximum work ahead; (3) Capacity shortage trigger; and (4) The card count "*m*" which determines the WIP level. The first three parameters should be set at the tactical level before implementing ConWIP to ensure the production of planned work at each planning period [3]. The card count can be determined at this level, but must be adjusted and monitored at the operational level. Although this parameter is the most important factor influencing the performance of the system, no clear rules have been formulated to quantify it [2, 3]. Finding the optimal value of the number of cards *m* will clearly involve economic tradeoffs. Increasing m will tend to increase service levels at the expense of higher inventories (WIP). Decreasing *m* will tend to decrease WIP while decreasing service levels. The number of cards m is routed in the LITTLE's LAW [11]:

$$m = Throughput\ TH * Lead\ Time\ FT \quad (1)$$

The throughput rate "*TH*" is determined by the bottleneck resource (number of jobs per time unit). The average flow time or lead time "*FT*" presents the time a job spends in the system in time units.

Firstly, in subsection B, we determine a "theoretical" number of cards "m" taking into account the patient types partition and their resource consumptions. In order to validate the cards number, a simulation model is developed in subsection C to take into consideration the stochastic nature of the process in evaluating the performance of CONWIP system.

B. ConWIP cards' number determination

The aim here is to determine the average number of ConWIP cards to be used in order to reach the maximum production load rate of the patient treatment process according to the system bottleneck. To set the number of cards, we propose an approach based on five steps [2].

The first step consists of determining patient types and their probability of presence, In this case study, we distinguish six patient types. A patient type is determined based on patient path p (presented in section II) and surgery type n. In the considered ISS, three major types of ophthalmologic surgeries are performed. The first two types are related to cataract surgery: Phacoemulsification (which we refer to as n=1) and Extracapsular (which we refer to as n=2) and are performed respectively by surgeon type s=1 and s=2. The third type encompasses Strabismus surgery and others (which we refer to as n=3) and is performed by surgeon type s=3. There are 2 surgeons available for the first and second types and one surgeon for the last type. The cataract surgery is the most frequent (45% for n=1 and 40% for n=2), while the relative frequency of the remaining types is 15%. Local (respectively General) anesthesia rate is about 30% (respectively 70%). In summary, for the studied ISS we distinguish 6 patient types which we note j = (p,n) (i.e. the patient type p having the surgery type n). For instance, outpatients (p=1) can have surgery types n=1and n=2 (which we refer to j= (1,1) and j= (1,2) respectively), inpatient undergoing general anesthesia (p=2) can have all surgery types and inpatient undergoing local anesthesia (p=3) can only have surgery types n=3. Each type j of patients has a probability of presence which we refer to as P_j. Patient type's partition is presented in TABLE II.

The second step comprises the identification of the different resources (machines) for each patient type. Based on patient trajectory presented in Fig.1, we identify 4 major resources corresponding to each stage, preparation beds (which we refer to as M_1), surgeons and operating rooms (OR) and recovery beds dedicated to outpatients (which we refer to as $M_{3.1}$) and to inpatients under general anesthesia (which we refer to as $M_{3.2}$). For surgery stage, both surgeons and ORs are used simultaneously so, we should consider one of them at once as the representative resource. Two possible representations of the machines are to consider. In the first case (which we refer to System 1), we use the OR as a representative machine of the surgery stage (which we refer to M2). Hence we obtain patient flow presented in Fig.3. In the second case (which we refer to System 2), since we have different surgeon types we obtain patient trajectory as depicted in Fig.4. In summary, for each system, each patient type j goes through machines as presented in TABLE III.

Third, to determine the throughput $\boldsymbol{TH}$ of the global treatment process, we determine firstly, for each resource M_i the mean flow time ($\boldsymbol{FT_i}$). The equation (2) stipulates that it is equal to the sum of the mean processing time ($\boldsymbol{D_{i,j}}$) of each patient type j using M_i multiplied by their probability of presence $\boldsymbol{P_j}$. We determine then the mean throughput ($\mathbf{TH_i}$) via (3) which stipulates that it is equal to the capacity of M_i ($\boldsymbol{Cap_i}$) divided by its mean flow time. Once calculated, we identify the bottleneck resource which has the minimal throughput (4).

$$FT_i = \sum_j P_j * D_{i,j} \quad (2)$$

TABLE II. Patient types partitions

Global patient types		**Patient types considering surgery types**	
Type (p)	**Probability of attendance**	**Type j=(p,n)**	**Probability of attendance Pj**
p=1	0.5	*j=(1,1)*	0.28=0.56*0.5
		j=(1,2)	0.22=0.44*0.5
p=2	0.4775	*j=(2,1)*	0.225=
		j=(2,2)	0.2=0.45*0.5
		j=(2,3)	0.0525=0.15*(1-0.5)*0.7
p=3	0.0225	*j=(3,3)*	0.0225=0.15*(1-0.5)*0.3

TABLE III. Resources used by each patient type in both systems

Type j=(p,n)	**Machines in the first case**	**Machines in the second case**
j=(1,1)	$M_1 + M_2 + M_{3.1}$	$M_1 + M_{2,1} + M_{3.1}$
j=(1,2)		$M_1 + M_{2,2} + M_{3.1}$
j=(2,1)	$M_1 + M_2 + M_{3.2}$	$M_1 + M_{2,1} + M_{3.2}$
j=(2,2)		$M_1 + M_{2,2} + M_{3.2}$
j=(2,3)		$M_1 + M_{2,3} + M_{3.2}$
j=(3,3)	M_2	$M_{2,3}$

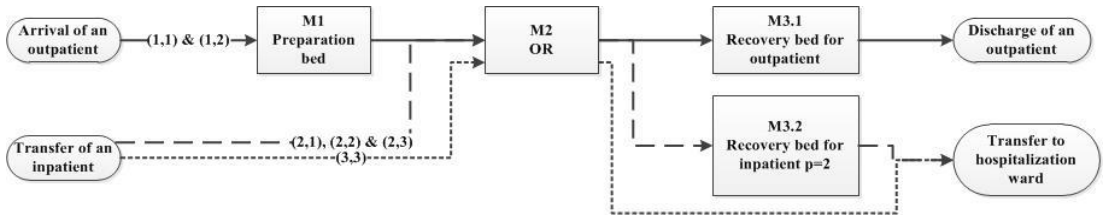


Fig. 2 Resources identification (considering ORs)

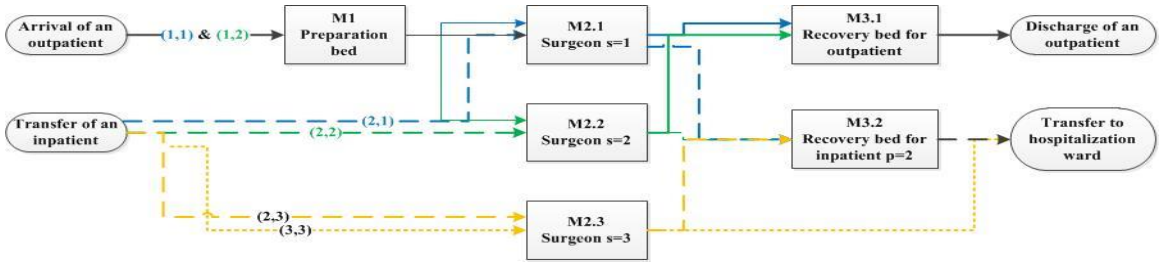


Fig. 3 Resources identification (considering surgeons)

$$TH_i = \frac{Cap_i}{FT_i} \quad (3)$$

$$TH = \min_i TH_i \quad (4)$$

Fourth, we determine the weighted flow time. To do so, we calculate for each patient type j, the flow time $\boldsymbol{FT_j}$ corresponding to the sum of the mean processing time $\boldsymbol{D_{i,j}}$ on all the machines used by patient type j as illustrated by (5). We determine then the flow time FT which is equal to the weighted sum of $\boldsymbol{FT_j}$ as illustrated by (6).

$$FT_j = \sum_{i/j \text{ uses } M_i} D_{i,j} \quad (5)$$

$$FT = \sum_j P_j * FT_j \quad (6)$$

Finally, we calculate *m* using the formulae (1). The obtained value is obviously the lower bound because we take into account deterministic process durations. To take into account the variability in process durations, we develop a simulation model of the ISS under a ConWIP system which allows us to control and refine the number of cards.

C. Card number adjustment

The objective of the simulations is to adjust in a stochastic context, the number of ConWIP cards necessary to manage the operations of the ISS. For this purpose, we build a discrete event simulation (DES) model of the ISS operating under a ConWIP system as illustrated in Fig.2. The model development is discussed in subsection.1). The procedure of card number adjustment using the DES model is discussed in subsection 2).

1) The simulation model

The conceptual model was computationally implemented using the ARENA software (Version 14, Rockwell Automation). Its first part consists of patients' arrival and ConWIP cards creation. In the studied case, all outpatients arrive at the beginning of the day. Hence, at the beginning of the simulation, all patients and m cards are created. Extensions such as tardiness and no-shows are straightforward with our model; however, they are uncommon in the ISS we studied, and for simplicity we do not include them in our analysis. Hence, we assume arrivals are on time and all patients show up for their scheduled procedure. Once created, these entities (patients and cards) enter the ConWIP loop. To realize this latter in the simulation model, we used a "match module" with two queues. The first one (respectively the second one) is dedicated for cards (for patients of all types). Once queues are not empty, the module matches an available card to a patient waiting in the queue. The matched entities are then synchronized to depart from the module. We note that there is no priority defined for the patient type to be the first matched to a card. The patient begins the surgical process and the card is hold in a "Hold module". The card waits for a signal to be released and re-routed to the Match module. Once the surgical process is complete, the signal for a card release is given. The card becomes available again and joins the queue of empty cards (in the Match module). The outpatient (respectively inpatient) is discharged (respectively is transferred to the hospitalization service). The empty card is matched to another patient waiting and so on until all patients are treated.

We supply the DES model with different data such as service times per patient types, surgery types and per process. These data were collected over a two-month period. A total of 70 patients were observed and tracked. Probability density functions were fit for all stages of a patient's movement through the ISS using the Chi-square and Kolmogorov Smirnov tests. The results were obtained by the Arena Input analyser tool and are shown in TABLE IV. For each operation within each stage, we provide service time distributions. The surgery stage is composed of 2-3 operations depending on patient's type: anesthesia, surgery (for all types) and postoperative treatment. This latter operation involves the wake up of patients under general anesthesia (patient type p=2). For

outpatient recovery stage, we distinguish three operations; wake-up, doctor's prescription and discharge.

At the beginning of each simulation run, the simulation model accepts, as inputs, the patient appointment schedule and the parameters of the problem (which include the number of patients per type, service times at each stage, number of resources per stage etc. We assume that the surgeries are considered independent from one another; ORs are identical and parallel meaning any operation can be scheduled on any OR at any time, and all the patients are scheduled before the day begins.

Before beginning the procedure for card controlling, the simulation model parameters (i.e. the warm up period, the number of replications and the replication length) should be adjusted. We notice that in our case we have a steady state simulations and no a terminating ones. In fact, since the ConWIP system allows to maintain a constant amount of work in process (WIP) regardless the number of products to be produced (in our case number of patients to be treated), the system should deliver a constant value of WIP for a big amount of patients. Given that, at the beginning the simulation, model is empty and idle and it becomes eventually congested [12]. Therefore, it is necessary to set a warm up period to ensure the reliability of the delivered results. In this case, the replication is relatively long (80.000 minutes). Results of simulations are delivered after 40 replications and a warm up period of 200 minutes for each simulation run

2) The procedure for card number adjustment

In order to adjust the number of cards, we evaluate the use of bottleneck resource and the average length of stay of patients after entering the ISS (LOS). LOS refers to the time between card assignment to a patient and its release. It is equivalent to FT in (1). On one hand, LOS allows us to comment the behavior of the ConWIP system. On the other hand, we use it to determine the mean Output rate of patients in the ISS. This latter refers to the average duration between the discharges of two consecutive patients. This metric is equivalent to TH in (1).

The experiment consists of varying "m" in the simulation model and evaluating the performance indicators. The optimal cards number "m" is the first value that delivers the maximum output rate.

IV. Results

The numerical experiments were carried out on an Intel@Corei-7 processor 2.1 GHz and 8GB Memory RAM. The DES model was implemented in MS Rockwell Arena TM software.

TABLE V and VI present respectively for ConWIP systems 1 and 2 the theoretical number of cards *m*, the capacity, throughput and cycle time of each resource and the flow time of the process. These results show that 6 cards at minimum are needed for the ConWIP system implementation.

As we explained earlier, to adjust the number of cards as response to stochastic processing times, we use the simulation model developed in section III B 2). We varied the number of cards from 1 to 30, for 1000 patients.

Fig. 5 depicts the resources utilization as function of *m*, showing OR as the bottleneck resource. Indeed, the OR reaches its maximum 97% with 11 cards.

Fig.6 depicts the LOS and Output rate curves as function of *m*. The LOS is quasi constant for *m* less than or equal to 6, then it increases regularly to double from 13 cards. On one hand, when the number of ConWIP is less than the capacity of the service (i.e., hungry system) waiting times are short which explains the low LOS (i.e. ≈80 minutes). On the other hand, when the number of cards is greater than the maximum capacity (i.e., overloaded system) an increase in the number of ConWIP increases the average LOS.

TABLE IV. Service time distributions per stage

Stage	Process	Specification	Distribution for service times (min)	p-value
Pre-operative	Preparation	In/out patient	1+GAMMA(2.46, 3.36)	0.169
Per-operative	Anesthesia	General	TRIANGULAR (6, 13.1, 21)	0.164
		Local	1+WEIBULL(3.76, 2.06)	0.337
	Procedure	n=1	TRIANGULAR (14, 27.3, 67)	0.562
		n=2	TRIANGULAR (40, 54, 60)	>0.15
		n=3	UNIFORM (38, 55)	>0.15
	Postoperative treatment	General anesthesia	ERLANG(0.637, 4)	0.194
Post-operative	Recovery	Inpatient	TRIAUGULAR(3, 9.95, 39)	>0.15
	Wake-up	Outpatient	TRIANGULAR(1.76, 1.98, 2.25)	0.138
	Prescription		TRIANGULAR(0.88,0.973,1.13)	>0.15
	Discharge		UNIFORM (20,30)	0.734

TABLE V. Cards count for ConWIP system 1

Parameter (I)	**Value**				**FT (min)**	***m***
M_i	M_1	M_2	$M_{3.1}$	$M_{3.2}$		
Cap_i (1)	4	4	4	2	78,44	6,1
FT_i *(min)(2)*	4,5	51,3	14	8,59		
TH_i *(# patients /h) (3)*	53,3	4,67	17,1	13,9		

TABLE VI Cards count for ConWIP system 2

(I)	**Value**						**FT (min)**	***m***
M_i	***M1***	***M2.1***	***M2.2***	***M2.3***	***M3.1***	***M3.2***	78,44	6.4
(1)	4	2	2	1	4	2		
(2)	4,5	22,2	24,5	4,62	14	8,59		
(3)	53,3	5,4	4,9	12,96	17,1	13,9		

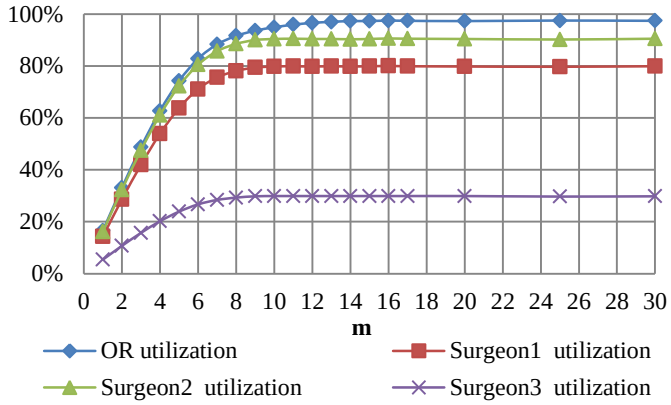


Fig. 4 Resources utilization per number of cards

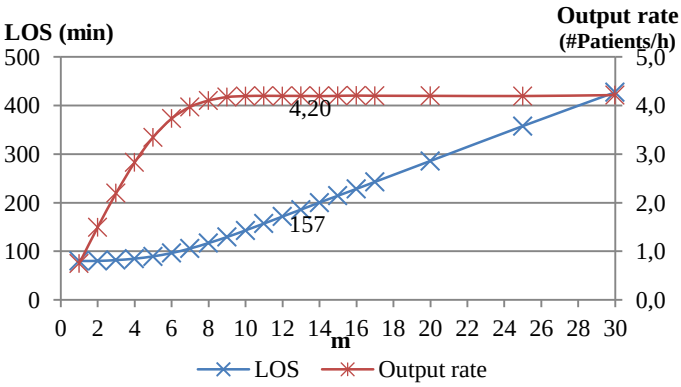


Fig. 5 LOS and output rate per number of cards

The Output rate curve presents also two profiles. Hence, up to 6-8 cards, the output rate increases. It stabilizes from 9 cards (the curve is parallel to the horizontal axis). It reaches its maximum rate (i.e. 4.2 patients per hour) from 11 cards.

These figures allow us to correctly size the number of cards so as to control the LOS and to ensure an average rate of occupancy of resources. In summary, 11 conwips ensure that the bottleneck resource (i.e. OR) will be used on average to 97%.

V. Conclusion

In this paper, we applied a ConWIP approach to manage operations in an ISS. To our knowledge, there is a lack of studies that applied the ConWIP system in non-industrial context especially for surgery departments. This study targets this gap by implementing this production control system for a real case study: a Tunisian ophthalmology ISS.

For the ConWIP system's implementation, we first calculated the number of cards "m" theoretically. For this purpose we took into account the patient types partition and their resource consumptions. The obtained value corresponded to the solution of the deterministic problem. To take into account the variability in process durations, we built a simulation model of the ISS under a ConWIP system. This model allowed us to control and size precisely the number of cards while accounting for randomness in process durations.

Experiments showed that an adequate dimensioning of ConWIPs (i.e. cards) allows maintaining a satisfying performance of the ISS through maximizing the use of bottleneck resource (i.e. the OR) which satisfies health care providers. However, from the point of view of patients, waiting times before beginning the treatment are not investigated. Moreover, patients can experience long waiting times before being assigned a card because their arrivals are not managed using the ConWIP system. This constitutes the main limit of our approach. Future research can be pursued in this direction. Hence, we intend to use ConWIP approach results to manage patients' arrivals to the facility. This will allow satisfying both: healthcare providers and patients, in other words, maintaining an optimal global performance through maximizing the use of bottleneck resources and minimizing patient waiting times before entering the system.

References

[1] M. L. Spearman, D. L. Woodruff, and W. J. Hopp. "Conwip: a pull alternative to Kanban". International Journal of Production Research, 28(5), 879–894, 1990

[2] W. J. Hopp, and M. L. Spearman. "Factory Physics"(Chicago: R. D. Irwin). 1996

[3] Y.Jaegler, P. Burlat and S. Lamouri. "The ConWip Production Control System: a Literature Review". 6th International Conference on Information Systems, Logistics and Supply Chain ILS Conference 2016, June 1 – 4, Bordeaux, France

[4] J. M. Framinan, P. L. González & R.Ruiz-Usano. "The CONWIP production control system: Review and research issues", Production Planning & Control: The Management of Operations, 14:3, 255-265, 2003

[5] F. Crop, T. Lacornerie, X. Mirabel, and E. Lartigau. "Workflow optimization for robotic stereotactic radiotherapy treatments: Application of Constant Work In Progress workflow". Operations Research for Health Care, 6, 18-22, 2015.

[6] A. Macario et al., Where are costs in perioperative? Analysis of hospital costs and charges for inpatient surgical care, Anesthesiology 83 (6) 1138–1144. 1995

[7] A Saremi, P Jula, T El Mekkawy and GG Wang "Appointment scheduling of outpatient surgical services in a multistage operating room department". INT J PROD ECON 141: 646–658. 2013

[8] J H.T Waldhausen, J. R. Avansino, L.Arlene and R.S. Sawin. "Application of lean methods improves surgical clinic experience. Journal of Pediatric Surgery 45, 1420–1425, 2010

[9] Tanya A. Peacock. "Lean Six Sigma: Optimizing Operating Room Utilization at Bayne-Jones Army Community Hospital". Graduate Management Project. U.S. Army Baylor University Graduate Program in Health Care Administration, 2006

[10] Schwarz P. et al. "Lean processes for optimizing OR capacity utilization: prospective analysis before and after implementation of value stream mapping (VSM)". Langenbecks Arch Surg 396:1047–1053, 2011.

[11] J. D. Little. A proof for the queuing formula: L= λ W. Operations research, 9(3), 383-387, 1961.

[12] W D Kelton, R P Sadowski and D A Sadowski (2000). Simulation with arena (2nd ed.). New York: McGraw-Hill.

Multidimensional assignment of services to people with disabilities

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Anderson Fabian MOSQUERA VARELA
Ressourcial
Fondation OVE
69120 Vaulx-en-Velin, France
anderson.mosquera@fondation-ove.fr

Lorraine TRILLING, Thibaud MONTEIRO
DISP Laboratory
Université de Lyon - INSA Lyon
69100 Villeurbanne, France
lorraine.trilling, thibaud.monteiro @insa-lyon.fr

Abstract — Medico-social sector includes various sets of institutions and services to support vulnerable people, such as people with disabilities. This sector must ensure personalized support to comply with the needs of each person. This work is based on a multidimensional characterization of services provided by medico-social institutions. The objective is to propose a mathematical model to deal with this multidimensional assignment problem (MAP). The distinctive characteristic of this assignment comes from that a need can be complied with different services and that a service is able to cover different needs at the same time. The proposed two-step model identifies the optimal set of medico-social services that are able to comply with the needs of people with disabilities.

Keywords—medico-social sector; assignment of services; ILP; multidimensional assignment problem.

I. Introduction

In France according to the Article L. 116-1 to 116-4 of the Social and Family Code, the social and medico-social sector should promote autonomy, protection and avoid the exclusion of people with disabilities. To achieve this, the activities proposed by the government to a person must be adapted to the person's needs. Decision-making in the orientation process of people with disabilities requires assistance.

The orientation process in France is carried out by the departmental office of people with disabilities, as follows: A person with disabilities (PD) makes an accompanying request to the departmental office of people with disabilities, a multidisciplinary team (MT) analyses the request and proposes a list of social and medico-social institutions (SMIS) appropriated to the person's needs. Currently, the orientation provided by the MT to a PD focuses on giving a list of institutions that are more or less adapted to the person's needs.

According to the law, from the moment when a person makes a request until a proposal is sent, the process should not take more than 4 months. Some studies [1] show that, in France, in more than 50% of cases, it takes more than 4 months. Moreover, studies show that from the moment when a person receives a proposal until real admission to an institution, the person can be put on a waiting list for more than one year [2].

During the time that the person remains on the waiting list, without accompaniment, his/her situation may worsen and the proposal made by the MT may not be adapted any more at the time the person is admitted to the institution.

The government has developed several missions which point out some of the most important problems of this orientation process [3]. The main problems that can be withheld from these reports are: (a) Places created in the medico-social sector are not fully adapted with the real need of PD, since there is a lack of tools for analysing the people's needs. (b) Services proposed by institutions may not be adapted to the real person's needs. One of the reasons is that institutions are categorized by types of handicap; (c) Not considering the availability of resources in institutions can lead people to stay on a waiting list for a long time. This wait may worsen the situation of the person and the proposed response will not be adapted to their needs any more. (d) The increase of young adults who stay in a place for children, because they do not have a proposition of adapted places to their situation and these people cannot be left without accompaniment. This practice mobilizes resources that should be dedicated to other needs.

The model proposed in this paper comes from a previous work that analyses and understands the orientation process [4]. This analysis is based on the process modelling with all its actors and the interactions between them. This process model allowed identifying the lack of information on the offer of services in a territory, especially when the MT proposes an assignment for the person. It highlighted major elements, which have to be considered when creating an assignment adapted to the needs of the person. Instead of trying to find a place for a PD in one single institution, we propose here to study the attribution of a list of services provided by one or several institutions to answer the needs of a PD. The objective of this work is to create an optimization model to help MT to propose the most accurate list of medico-social services adapted to the needs of each PD. This list has to find the maximum coverage of PDs' needs, according to the existing service offered by the institutions and the resource availability. Ideally, all needs may be covered, but a partial assignment can be proposed to avoid having PD without accompaniment.

This paper begins with a literature review on the assignment methods and the relationship with the case study. Subsequently, we present the assumptions considered in the model. We conclude with a presentation of the model and an analysis on the experimentation carried out, in a case inspired from field reality.

II. Literature review

The main objective of this work is to propose to a PD with several needs, an adapted list of services of medico-social institutions, while respecting a certain set of constraints. The

model presented in this paper is part of the family of combinatorial optimization problems, specifically to assignment problems, where the aim is matching agents, maximizing or minimizing a specific objective function, subject to costs [5], [6]. If we look at the detail of our problem, we can consider it as a multidimensional assignment problem (MAP), since we are interested in creating a correspondence with several dimensions [7]. This problem is to NP-hard.

Bearing in mind that the current demand for medico-social services exceeds the availability of services, it is clear that not all people with disabilities will have a total response to their needs. This type of situation may be connected with the order acceptance and scheduling decisions in make-to-order systems, Order Acceptance and Scheduling problem (OAS) [8]. The OAS problems seek to accept the orders that maximize the profitability, in a fixed period of time, determined by production scheduling. In our case, the idea is to maximize the number of disabled people accepted in the system with a complete assignment to their needs, based on the best scheduling of the services assigned to the group of people accepted.

The problem of the assignment of services to disabled people is linked with the problem of scheduling with parallel machines [9], [10]. In this problem, a job can be performed in different machines, with the differentiation of time processing and resource consumption given the characteristics of the machines. In our problem, machines are represented by services (different services that can be assigned to respond different needs, with a different impact on needs, given the services characteristics) and the jobs, are represented by the needs of the people (each need requires an assignment, given the situation of the person).

A first scientific contribution of this study is the possibility of proposing one or several services to a person, while these services are offered to other people in the same time that (in parallel). These types of problems can be associated with divergent production systems with co-production (i.e. production of different products at the same time from a single product input) when alternative production processes are available [11]. In the case study, one service can be assigned to several PD according to the resources available in the institution to perform the service. This type of problem can be associated with the production scheduling problem with non-renewable resources [12].

Finally, an important scientific contribution is the application of operations research methods in the medico-social sector (the application of well-known methods in a new sector). It is necessary to highlight the few studies in the medico-social sector, where the benefits of orienting a person to several institutions are studied [13]. A recent study tackles the problem of elaborating optimized action plans in order to improve the overall management efficiency of medico-social institutions, and confirms the lack of use of operations research in the sector medico-social [14]. However, we find in the literature more and more research in the field of elderly people [15], and in the health sector [16].

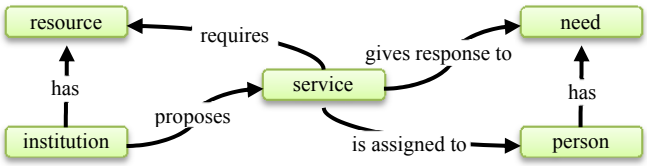
III. Modelling

A. Assumption and notation

The proposed model has the following sets: **SP** is the set of **P** people who looking for an orientation; **SN** is the set of **N** standard needs; **SS** is the set of **S** standard services; **SI** is the set of **I** institutions in a given territory and; **SR** is the set of **R** resources that can be used in a service.

A person $p \in$ **SP** is characterized with various needs. Every need $n \in$ **SN** should receive a response with one or more services. The main objective of the assignment model is to propose to each person p a list of medico-social services that are adapted to their needs. To manage this assignment, a need n could be complied with different services $s \in$ **SS**. Those services are able to give response to different needs in the same time. A service s is proposed by an institution $i \in$ **SI**. The Institution i has got available resources to offer its services. Proposed services must respect the availability of each resources $r \in$ **SR**. Fig. 1 describes the links between all indexes.

Fig. 1.Entities of the model



In order to fulfil this objective, the following assumptions are defined.

A1: We assume there are: (1) a standard list of needs, representing all the possible needs that a person with disabilities can have; (2) and a standard list of medico-social services which represents all the medico-social services that can be offered to a person. This assumption is based on the current situation, since today the French government created these lists for the medico-social sector, which must be integrated in our case [17]. Based on assumption **A1**, the following assumptions are proposed according the information needed in the model.

1) Assumptions related to people with disabilities (PD)

A2: A team evaluates the situation of each PD and determine his/her needs and the criticality of each need. The way to present this information is based on a point system with a scale ranging from 0 to 10, where 0 means that the PD does not have the need and 10 means that the need is at its most critical level and requires more attention, time and resources. Table 1 shows an example where the person 2 has the needs 6 and 9, respectively with 5 points (average criticality) and 9 points (highest criticality).

Notation in the model: $N_{p,n}$ is the number of points that services must accumulate to answer the need n of the person p.

A3: Each person has limitations in relation to the institutions that may come to propose him/her. Notation in the model: DT_p is the maximum distance tolerated by person p, from his reference address to an institution. IT_p is the maximum number of different institutions that the person p tolerates in a proposed assignment.

Table 1. PD's needs

Needs \ People	1	2	3	4	5	6	7	8	9	10
1	0	3	0	0	0	0	0	9	0	2
2	2	0	4	5	0	5	0	0	0	7
3	3	1	9	7	0	5	5	3	3	1
4	0	3	0	1	2	3	5		9	0
5	3	0	0	0	0	0	0	2	5	0
6	4	5	1	4	0	0	0	0	1	3
7	0	5	2	0	2	0	1	1	1	0
8	0	1	1	0	0	4	5		3	1
9	0	9	5	4	0	1	2	3	9	2
10	2	9	1	1	0	1	0	0	0	5
11	1	0	5	9	1	0	0	3	0	1
12	2	0	9	2	4	3	0	0	4	1
13	5	0	0	5	0	1	4	2	5	9
14	9	0	0	0	2	4	0	0	0	3
15	0	0	9	1	9	0	0	0	9	9
16	0	0	9	0	0	0	0	0	0	0
17	0	0	5	3	1	1	2	2	0	1
18	3	0	7	1	0	2	0	0	0	1
19	2	4	9	2	4	0	9	0	1	7
20	2	0	0	0	4	9	1	0	3	3

2) Assumptions related to services

We assume that the characteristics of each service are known beforehand.

A4: A service can contribute to answer one or several needs with a certain number of points. This number of points is represented by a scale of 0 to 10 points, similar to the scale used for the needs of the person. Effectively if a service contributes to 10 points means that the service is very specialized in that need and 0 if the service does not contribute to the need at all.

A contribution of this work is to propose to the person one to several services to answer a need, at the same time that these services can contribute in the answer of other needs. For example, person 1 requires 5 points to have coverage of his/her need 13 (see Table 1). This need can be covered by a single service such as service 9, or by a group of services such as service 1 regrouped with service 7 (see Table 2). Moreover, if service 9 is chosen, this service can contribute, in the same time, to cover need 19 of person 1 with 4 points.

Notation in the model: $\boldsymbol{RSN_{s,n}}$ is the number of points that service *s* contributes to respond to need *n*.

A5: A need can be covered by a combination of several services. There are different ways to support a specific need. So, a need might be complied with a pool of different services. To be able to evaluate how services interact themselves, it is necessary to consider the compatibility between services. If services are compatible for one need, the number of expected points will be the sum of individual contributions.

If services are not compatible for one need, this number will be the maximum value of individual contributions. For example, let's consider that a person is assigned to service 3 and service 8 to support need 4 (see Table 2). If services 1 and 2 are compatible, the person will have 12 points for need 4. On the other hand, if services 3 and 8 are not compatible, the person will have 8 point only.

Table 2. Relationship services-needs

Needs \ Services	1	2	3	4	5	6	7	8	9	10
1	0	0	6	0	0	4	0	0	0	0
2	0	0	0	6	0	0	0	0	0	4
3	6	0	0	0	0	0	0	6	0	0
4	0	0	4	0	0	0	0	8	0	0
5	0	8	0	0	0	0	0	0	10	0
6	0	0	0	6	0	0	0	0	0	4
7	10	0	0	0	0	0	6	0	0	0
8	0	0	6	0	0	0	0	0	0	4
9	0	6	0	0	0	0	0	0	0	4
10	5	0	0	0	0	0	5	0	0	0
11	0	0	0	6	0	0	4	0	0	0
12	0	0	8	0	0	0	0	0	0	4
13	4	0	0	0	0	0	6	0	8	0
14	0	10	0	0	0	0	0	4	0	0
15	8	0	0	0	0	0	0	0	0	10
16	0	0	0	0	0	5	0	5	0	0
17	0	0	0	0	10	0	0	0	0	0
18	0	0	10	0	0	0	0	0	0	0
19	10	0	0	0	0	0	0	0	4	0
20	0	0	0	0	0	8	8	0	0	6

Table 3. Compatibility between services, resources availability in institutions and Relationship services-resources, services-institutions

		Services										Resources					
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6
Services	1	0	1	1	1	0	0	0	1	0	1	0	4	0	3	0	0
	2	1	0	0	0	0	0	1	0	1	0	4	0	0	0	0	1
	3	1	0	0	1	1	1	0	1	0	0	0	0	5	0	0	1
	4	1	0	1	0	0	0	1	0	0	0	4	0	0	0	1	0
	5	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0
	6	0	0	1	0	1	0	0	1	1	0	0	3	0	0	1	0
	7	0	1	0	1	0	0	0	0	0	0	3	0	3	0	0	1
	8	1	0	1	0	1	1	0	0	1	1	0	5	0	0	0	1
	9	0	1	0	0	1	1	0	1	0	1	0	0	0	0	1	1
	10	1	0	0	0	0	0	0	1	1		0	5	0	10	1	1
Institutions	1	0	0	1	1	1	0	1	1	1	1	15	15	15	15	3	4
	2	1	1	1	1	1	1	0	0	0	0	15	15	15	15	3	4
	3	1	1	1	0	1	0	1	1	0	1	15	15	15	15	3	4
	4	1	1	1	1	1	1	0	1	0	0	15	15	15	15	3	4
	5	0	0	1	0	1	0	1	1	1	1	15	15	15	15	3	4

An example of a binary matrix of compatibility is illustrated in Table 3. The compatibility between two services is represented by a symmetric matrix, where 1 means that the two services are compatible and 0 means that the services are not compatible. It is important to note that a service is defined by a specific activity that consumes resources (in hours or in quantity) during a standard period. This is why, a service is not compatible with itself. For example, a service can be: a speech therapy of 2 hours per week. If we propose twice the same service to a person, it means that the person will receive a speech therapy, 4 hours per week. That feature is considered as another service.

Notation in the model: $\boldsymbol{RSS_{\hat{s},s}}$ is equal to 1 if service $\hat{s}$ is compatible with service *s*; 0 otherwise.

A6: A service may require the use of resources, with a quantitative consumption (Table 3 relationship services-resources).

The units of measure of the consumption of resources by a service, depends on the type of resources (Table 3). With human resources the units of consumption will be represented in hours (for example service requires 3 hours working time of a social worker), and for material resources the units of consumption will be represented in quantity (for example a service requires one bed).

Furthermore, some services do not consume any resources like the Service 5 in Table 3. The assignment of this kind of service to a PD will not consume any material or human resources of the institution. However it will help to choose an institution among others to answer a PD's needs (for example, if the institution has an adapted kitchen or offers free access to a green space).

Also, we consider that a type of service could be proposed either with one resource or with another resource. In the data matrix these situations are considered in different rows, as if they were different services, even if they provide the same support to the PD. For example, a service that can be offered with a psychologist or a social worker, in our data tables will be considered as 2 services that offer the same amount of points to the same needs, but they consume different resources.

Notation in the model: $\boldsymbol{RSR_{s,r}}$ is the number of units of resource *r* consumed by service *s*. $\boldsymbol{SP_{s,i}}$ is equal to 1 if service *s* is proposed by institution *i*; 0 otherwise.

3) Assumptions related to institutions

A7: Proposed Services to a PD may belong to one or more institutions (Table 3 relationship services-institutions).

A8: As resources are constrained, it is necessary to know the resource availability in each institution upon decision-making (Table 3 relationship institutions- resources). The unit of availability is the same as in the assumption **A6**.

Notation in the model: $\boldsymbol{RA_{i,r}}$ is the availability of resource *r* in institution *i*.

A9: Parameters that must be known beforehand

$\boldsymbol{D_{i,p}}$ is the distance between institution *i* and the reference address of person *p*. The compatibility between the services presented in the assumption **A5**, implies that the points offered by the services proposed to the person to each need, depends on the compatibility of the proposed service combination. Therefore it is useful to elaborate a pre-treatment which allows to generate all possible combinations of services and calculates, for each of these combinations, the number of points that it contributes to each need.

The number of possible combinations that this pre-treatment must generate is of order $(2^{nS}-1)$, where *ns* represents the number of services. For example, for a total of 3 services (s1, s2 and s3), the total number of possible combinations is $(2^3-1)=7$. c1 (s1), c2 (s2), c3 (s2, s1), c4 (s3), c5 (s3, s1), c6 (s3, s2), c7 (s3, s2, s1). Also it is necessary to calculate the number of points that the combination contributes to each need, This calculation is performed respecting the compatibility of services presented in assumption **A5**. The calculation is made using the parameters $RSN_{s,n}$ and $RSS_{\hat{s},s}$. Note that **SC**, set of all possible combinations of services, with cardinality $(2^{nS}-1)$, where *ns* represents the total number of services.

Notation in the model: $\mathbf{SB_{s,c}}$ is equal to 1 if the service *s* does belong to combination of services *c*; 0 otherwise. $\boldsymbol{RCN_{c,n}}$ is the number of points that combination of services *c* contributes to respond to need *n*.

B. Binary decision variables

We use the following decision variables:

$\boldsymbol{X_{s,i,p}}$= 1 if the service *s* of the institution *i*, is assigned to the person *p*; 0 otherwise.

This decision variables is sufficient to define the allocation problem. In order to facilitate constraint description, we also consider the following binary variables related to variables $X_{s,i,p}$: $\boldsymbol{Y_{i,p}}$ is equal to 1 if at least one service of the institution *i* is assigned to the person *p*, 0 otherwise; $\boldsymbol{U_{c,p}}$ is equal to 1 if the combination of services *c*, is assigned to the person *p*, 0 otherwise; and $\boldsymbol{W_p}$ is equal to 1 if the model proposes a complete assignment to the Person *p*, 0 otherwise.

C. Model

The model considers that the availability of resources is known at the moment of the assignment. The search for a solution is done in 2 steps. The first step tries to maximize the number of people with a complete response (prioritizing people with a critical situation). In a second step, a partial response is looking for the people who do not have a complete response in the first step. In this second step, we propose services that respond to some needs of PD, depending on the availability of remaining resources.

1) Step 1: Find a complete assignment for the maximum number of people regarding their needs.

a) Objective function: Maximize the number of needs (weighted by the number of points) covered by proposing complete answers.

$$\text{Maximize } \sum_{p=1}^{P}\sum_{n=1}^{N}(W_p * N_{p,n}) \quad (1)$$

b) Constraints

$$W_p * DT_p \geq D_{i,p} * Y_{i,p} \;\forall\, i,p \quad (2)$$

$$\sum_{i=1}^{I} Y_{i,p} \leq IT_p \;\forall\, p \quad (3)$$

$$SP_{s,i} \geq X_{s,i,p} \;\forall\, s,i,p \quad (4)$$

$$\sum_{c=1}^{C} U_{c,p} \leq 1 \;\forall\, p \quad (5)$$

$$\sum_{i=1}^{I} X_{s,i,p} \geq \sum_{c=1}^{C}\left(U_{c,p} * SB_{s,c}\right) \;\forall\, s,p \quad (6)$$

$$X_{s,i,p} \leq \sum_{c=1}^{C}\left(U_{c,p} * SB_{s,c}\right) \;\forall\, s,i,p \quad (7)$$

$$Y_{i,p} * \sum_{s=1}^{S} SP_{s,i} \geq \sum_{s=1}^{S} X_{s,i,p} \;\forall\, i,p \quad (8)$$

$$W_p * N_{p,n} \leq \sum_{c=1}^{C}(U_{c,p} * RCN_{c,n}) \;\forall\, p,n \quad (9)$$

$$RA_{i,r} \geq \sum_{s=1}^{S}\sum_{p=1}^{P}\left(X_{s,i,p} * RSR_{s,r}\right) \;\forall\; i,r \quad (10)$$

$$X_{s,i,p}, Y_{i,p}, U_{c,p}, W_p \in \{0,1\} \;\forall\, s,i,c,p \quad (11)$$

Constraints (2) ensure that institutions, which are part of the proposed assignment to the person, are in the perimeter

tolerated by the person. Constraints (3) ensure that a person will not be assigned to more than a maximum number of institutions. Constraints (4) ensure that the assignment includes services proposed by institutions. Constraints (5) ensure that a maximum of one combination of services is proposed to a person. Constraints (6) and (7) ensure that the services offered to people are part of the combination of services. Constraints (8) ensure that the sum of the services available at institution *i* and assigned to person *p* cannot exceed the total number of services offered by institution *i*. Constraints (9) ensure that the combination of proposed services covers all needs of the person. Constraints (10) ensure that the assigned services of each institution do not consume more than the available resources.

2) Step 2: Find a partial assignment for people without a full accompaniment, using the remaining available resources. This means that the assignment proposed to the person does not satisfy all his/her needs entirely.

At this step the model presented above remains mostly the same, but now the index $\boldsymbol{p \in RP}$, where **RP** is a subset of **SP** for people without a full accompaniment in the step 1. The parameter $RA_{i,r}$ is recalculated with the resources available after assignment in the first step. Other changes in this step 2 are with the objective function and constraints (2) and (9). In this step the variable W_p is not used, in its replacement the following variable is proposed. $\boldsymbol{V_{p,n}}$ is equal to 1 if an assignment is proposed for the person *p* for need *n*, 0 otherwise.

a) Objective function: Maximize the number of covered needs among all the people.

$$\text{Maximize } \sum_{p=1}^{P}\sum_{n=1}^{N}(V_{p,n} * N_{p,n}) \quad (12)$$

b) Constraints

$$DT_p \geq D_{i,p} * Y_{i,p} \; \forall \, i,p \quad (13)$$

$$V_{p,n} * N_{p,n} \leq \sum_{c=1}^{C}(U_{c,p} * RCN_{c,n}) \; \forall \, p,n \quad (14)$$

$$V_{p,n} \in \{0,1\} \; \forall \, p,n \quad (15)$$

Constraints (2) are replaced by constraints (13) and constraints (9) are replaced by constraints (14). Constraints (13) ensure that institutions that are part of the proposed assignment to the person, are in the perimeter tolerated by the person. Constraints (14) ensure that the combination of services proposed cover the need *n* of the person *p*.

IV. Computational experiment

The data used in this experiment are presented in Table 1, Table 2 and Table 3. This data have been generated to be representative of a real case but with a reasonable size, allowing to be solved with an exact method. The results presented below were obtained after the implementation of the model in the program IBM ILOG CPLEX Opt. Studio, Version: 12.5.

The instance represents a situation including 10 people with disabilities looking for an assignment of services in a network of 5 institutions. Each institution proposes up to 7 services from a list of 10 standard services (the number of combinations is $(2^{10} - 1) = 1023$). Furthermore, each person can have a maximum of 16 different needs from a list of 20 standard needs. We assume that the five institutions are in the tolerated perimeter of all people and each institution uses up to 6 resources.

Table 4 presents a comparison in the application of the model (with its two 2 steps) in two scenarios: Scenario 1 considers that people can only go to one single institution (which is the current situation). Scenario 2 considers that it is possible for people to go to up to 2 institutions (this scenario shows the advantages when more institutions are allowed to be part of the solution). It is important to note that for each of the scenarios and each stage, CPLEX finds a solution in less than 15 seconds with the data set proposed.

To analyse the performance solutions, 3 indicators have been defined. The first one is related to the entirety of the solution: (A) *number of people with a complete assignment in the first step*. Two indicators are related to the solution proposed to each person: (B) *number of needs covered / number of total needs*; (C) *number of points covered / number of total points*. (A) is the most important indicator in reference to the others, since the medico-social sector currently tries to propose complete assignments to people with disabilities. Results are as follows: with scenario 1; (A) = 4 people (40%); in scenario 2 (A) = 5 people (50%). For people with a complete assignment, in this first step, the indicators B and C are 100%. With this indicator (A), we can see that proposing solutions with more than one institution causes positive effects on the number of people with a complete assignment. Today, the orientation process does not work in this way, since it tries to match PD with places within one institution. In that way, our model, considering the assignment of services to PDs in its first step, presents an interesting contribution for this sector (the possibility of considering more institutions in a complete assignment). In our experimentation related to 10 demands, we went from 4 to 5 people with a complete assignment.

Table 4. Results for scenarios 1 and 2

			Step 1		Step 2	
	Indicators		B	C	B	C
Results scenario 1	People	1	12/12	38/38		
		2			6/9=67%	17/40=43%
		3			9/14=64%	35/76=46%
		4			3/13=23%	12/45=27%
		5	9/9	29/29		
		6			6/12=50%	18/39=46%
		7	9/9	34/34		
		8	8/8	25/25		
		9			7/12=58%	30/53=57%
		10			7/16=44%	20/56=36%
Results scenario 2	People	1	12/12	38/38		
		2			0/9=0%	0/40=0%
		3			8/14=57%	43/76=57%
		4	13/13	45/45		
		5	9/9	29/29		
		6			5/12=42%	13/39=33%
		7	9/9	34/34		
		8	8/8	25/25		
		9			5/12=42%	20/53=38%
		10			4/16=25%	14/56=25%

To improve this solution, decision-maker can modify parameters to identify different alternatives, for example by changing the number of institutions tolerated by PD. Another parameter, in which the decision-maker can act, is the possibility of increasing the geographical perimeter tolerated by PD. In this experiment, all institutions are in the perimeter of PD. In real life the medico-social institutions are found throughout a large territory. That is why extending the admitted perimeter will allow expanding the possibilities of finding a complete assignment.

The other step of the presented model aims to give a partial response to people without answers after the first step. The possibility of proposing a relaxed solution reduces the break of the accompanying path (to prevent that the PD stays without accompaniment, a situation that is sought to avoid in the medico-social sector). Nevertheless, we can see that an improvement in step 1 can reduce possibilities in step 2. For example, in Table 4, PD 2 stays without accompaniment in scenario 2, whereas is he/she has got a partial assignment in scenario 1.

As in the first step, decision maker can act with parameters such as the maximum number of institutions that can participate to one answer, in order to improve the relaxed solution.

We can note that this relaxed solution may be a temporary one. Indeed, this model considers only one period of time. As we said, a person can stay without accompaniment. A multi-period model would improve the assignment. It would be able to find complete assignments for all people over a time horizon and not within a single period.

V. Conclusions and prospects

Currently, the assignment of disabled people is based on a single institution solution and takes into account places only and not real need among its complete aspect. In order to improve assignment decision process, we propose a mathematical model that is able to combine the services of several institutions to propose a more complete solution. Those solutions can be composed by a set of services provided by several institutions. In a second step, the approach is able to find partial assignments with the availability of resources remaining after the first step.

In order to take into account multi-period decisions, development has to be done to propose to people with disabilities who makes a demand a complete assignment to their needs, considering the possibility to offer services in later periods depending on the availability of resources in institutions and in each period.

Acknowledgment

This research was funded by the Agence Régional de Santé (ARS) Auvergne Rhône-Alpes. It undertakes within the research project ACAPELHA (www.ressourcial.fr/acapelha).

References

[1] CNSA, “Dixième année de mobilisation : les MDPH engagent leur transformation,” 2016.

[2] MDPH calvados, “Étude des situations en liste d’attente dans les SAVS et les SAMSAH.” 2013.

[3] D. Piveteau, S. Acef, F.-X. Debrabant, D. Jaffre, and A. Perrin, “« Zéro sans solution » : le devoir collectif de permettre un parcours de vie sans rupture, pour les personnes en situation de handicap et pour leurs proches,” 2014.

[4] A. F. Mosquera Varela, L. Trilling, and T. Monteiro, “Managing the accompanying path of people with disabilities in a medico - social network: a method to create a process model based on autonomous actors point of views,” *International Federation of Automatic Control World Congress*. Toulouse, France, 2017.

[5] S. I. Gass and M. C. Fu, Eds., *Encyclopedia of Operations Research and Management Science*. Boston, MA: Springer US, 2013.

[6] B. Korte and J. Vygen, *Combinatorial Optimization*, vol. 21. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[7] R. E. Burkard and E. Çela, “Linear Assignment Problems and Extensions,” in *Handbook of Combinatorial Optimization*, Boston, MA: Springer US, 1999, pp. 75–149.

[8] C. Og˘uz, F. Sibel Salman, and Z. Bilgintürk Yalçın, “Order acceptance and scheduling decisions in make-to-order systems,” *Int. J. Prod. Econ.*, vol. 125, no. 1, pp. 200–211, 2010.

[9] J. Yoo and I. S. Lee, “Parallel machine scheduling with maintenance activities,” *Comput. Ind. Eng.*, vol. 101, pp. 361–371, 2016.

[10] P. Györgyi and T. Kis, “Approximation schemes for parallel machine scheduling with non-renewable resources,” *Eur. J. Oper. Res.*, vol. 258, no. 1, pp. 113–123, 2017.

[11] J. Gaudreault, J.-M. Frayret, A. Rousseau, and S. D’Amours, “Combined planning and scheduling in a divergent production system with co-production: A case study in the lumber industry,” *Comput. Oper. Res.*, vol. 38, no. 9, pp. 1238–1250, Sep. 2011.

[12] J. E. C. Arroyo and J. Y.-T. Leung, “Scheduling unrelated parallel batch processing machines with non-identical job sizes and unequal ready times,” *Comput. Oper. Res.*, vol. 78, pp. 117–128, 2017.

[13] G. Osorio, T. Monteiro, L. Trilling, and F. Albert, “Multi-criteria assignment policies to improve global effectiveness of medico-social service sector,” *Eng. Appl. Artif. Intell.*, vol. 61, pp. 21–34, 2017.

[14] B. Chabane, M. Basseur, and J.-K. Hao, “R 2-IBMOLS applied to a practical case of the multiobjective knapsack problem,” *Expert Syst. Appl.*, vol. 71, pp. 457–468, Apr. 2017.

[15] H. Liimatta, P. Lampela, P. Laitinen-Parkkonen, and K. H. Pitkala, “Effects of preventive home visits on older people’s use and costs of health care services: A systematic review,” *Eur. Geriatr. Med.*, vol. 7, no. 6, pp. 571–580, 2016.

[16] B. Maenhout and M. Vanhoucke, “An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems,” *Omega*, vol. 41, no. 2, pp. 485–499, 2013.

[17] CNSA, DGCS, and Ministère des Affaires sociales et de la Santé, “Les nomenclatures de besoins et de prestations détaillées.” SERAFIN-PH (Services et Etablissements: Réforme pour une Adéquation des FINancements aux parcours des Personnes Handicapées), 2016.

Planning and Scheduling in Air and Maritime Transportation

Aircraft Maintenance Capacity Planning: A Decision Support Framework

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Duarte Dinis, Ana Barbosa-Póvoa
Center for Management Studies
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
duarte.dinis@tecnico.ulisboa.pt

Abstract— Capacity planning is known to be an important problem faced by maintenance industries, such as aircraft Maintenance, Repair and Overhaul (MRO) organizations, involving large datasets that are not often systematically characterized and effectively used in the decision-making process. Besides presenting a quantitative assessment on the aircraft maintenance capacity problem, in terms of workload for the scheduled and unscheduled maintenance during part of the service life of an aircraft family, this paper proposes a framework for the treatment, recording, and analysis of maintenance data. The framework comprises a set of requirements for the treatment and recording of data and a method for the characterization of the maintenance work, and its purpose is to improve the MROs maintenance capacity planning decision-making process and, ultimately, their overall operational efficiency.

Keywords— maintenance; capacity planning; risk; decision support systems

I. Introduction

Maintenance is formally defined as the 'combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function' [1]. As mentioned in [2], this definition separates the maintenance actions between those directed at retaining the operating conditions of an item and those directed at restoring such operating conditions. The first compose what is defined as preventive maintenance, and the latter compose what is called as corrective maintenance. In aviation, preventive maintenance is referred to as scheduled maintenance, defined as 'all the individual maintenance tasks performed according to the maintenance time limitations' [3]. Corrective maintenance is referred to as unscheduled maintenance and results from discrepancies identified during scheduled maintenance tasks, other unscheduled maintenance tasks, pilot reports, unforeseen events occurred during normal operation such as hard landings, tail strikes, ground damage, lightning strikes, or data analysis [3], [4]. For aircraft Maintenance, Repair, and Overhaul (MRO) companies, which *produce* maintenance services, maintenance can be regarded as a production process that needs to be planned [5]. According to [6], maintenance capacity planning is needed to establish the required resources to perform maintenance, and that its critical aspect is to determine the exact number and types of resources to face the future work demand given its uncertainty. Under this context, the objective of this paper is threefold: (1) based on real maintenance data from a Portuguese independent aircraft MRO, to quantitatively assess the capacity planning problem in terms of workload during part of the service life of an aircraft family; (2) to propose a set of requirements for aircraft maintenance data treatment and recording; and (3) to propose a data analysis method for the characterization of the maintenance work with the aim of helping MROs in the capacity planning decision process and, ultimately, of improving their overall operational efficiency. The remainder of the paper is organized as follows. Section II presents the aircraft maintenance planning process. In section III, a quantitative assessment of the capacity problem using real aircraft maintenance data is presented. In section IV, a set of requirements for maintenance data treatment and recording is established and a data analysis method is defined. Finally, in section V conclusions are drawn on the performed work and future research opportunities for the application of the proposed maintenance characterization framework are presented.

II. Aircraft MRO Planning Process

The aircraft MRO planning process is illustrated in Fig. 1. Upon receiving a request for quotation for a given maintenance intervention, the MRO collects information about the aircraft that is to be intervened and about the maintenance service that is to be performed. Typical collected information includes the aircraft model, its tail number, the client and/or operator, and the type of maintenance intervention. With this information the MRO develops a tactical plan and elaborates a proposal. The proposal contains workload and cost estimations, divided between scheduled and unscheduled maintenance work, and establishes a delivery date. If the client accepts the proposal, a project is formalized. The maintenance project contains a list of scheduled maintenance tasks, whose workload estimations are expected to be relatively accurate regarding the work that is actually performed, and a rough, uncharacterized, workload estimation for the unscheduled maintenance. It is not until the execution of the maintenance intervention, more specifically until the end of the aircraft inspection, that the list of unscheduled tasks is completely known. Two important problems result from this fact: (1) a high degree of uncertainty about the unscheduled maintenance workload remains until late in the planning process, being difficult for an MRO to react if the actual workload exceeds, or falls short, the planned capacity for a given maintenance check; and (2) by not

This work was supported by the Portuguese National Science Foundation (FCT) under Grant PD/BD/52345/2013.

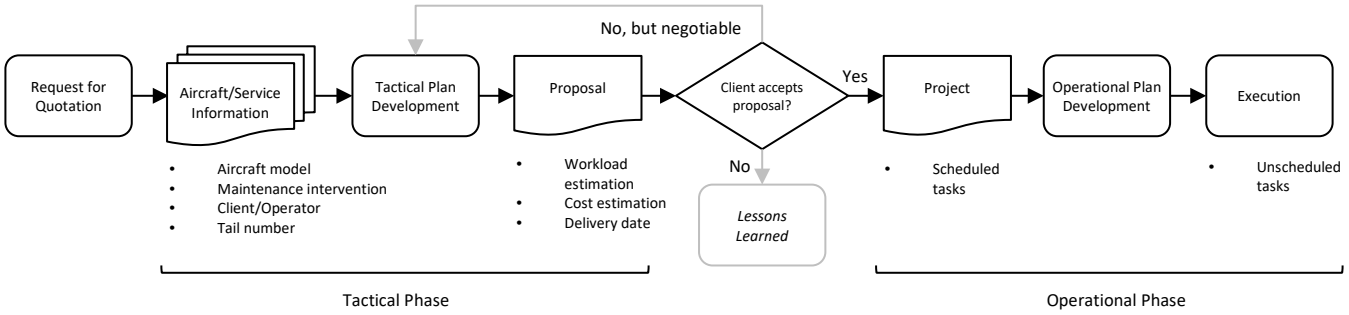


Fig. 1. MRO planning process.

characterizing the maintenance work in terms of where in the aircraft it is expected to appear, when in the check it is more likely to occur, and what technical skills are more likely to be required, current estimations do not allow tactical plans to contain, even as a sketch, a pre-scheduling of the maintenance work. The threat posed by unscheduled maintenance to the capacity planning has been identified in the literature as an important problem to be addressed [7], [8]. These problems are the motivation for the current work, particularly for the development of the framework proposed in section IV. With such framework, the maintenance work can be accurately characterized in terms of workloads per aircraft zone, per work phase, and per technical skill, allowing MRO organizations to gain a detailed picture about what to expect in the maintenance interventions early in the planning process.

III. Aircraft Maintenance Quantitative Assessment

An assessment on the aircraft capacity planning problem in terms of workloads for the scheduled and unscheduled maintenance is hereby presented. The studied sample refers to 203 maintenance projects from a Portuguese independent aircraft MRO, occurred between September 2002 and December 2015, in a total of 61 different construction numbers for a given aircraft family, operating under scheduled commercial services [9]. The maintenance requirements established by the aircraft manufacturer include cycles of four ‘C’ checks, with 5000 flight hours (FH) intervals between each check (x-axis of Fig. 2 and Fig. 3).

Fig. 2 shows the average scheduled and unscheduled maintenance workloads for each of the ‘C’ checks found in the collected data. Results are presented in percentage, having the average scheduled maintenance workload of the 5000 FH ‘C’ check as reference (100%). Regarding specifically unscheduled maintenance, Fig. 3 presents the average unscheduled maintenance workloads for each of the ‘C’ checks. Results are presented in percentage, having the average unscheduled maintenance workload of the 5000 FH ‘C’ check as reference (100%). According to the linear regression in Fig. 3, one can expect a workload increase of 57% in the unscheduled maintenance at each ‘C’ check, from the 5000 FH ‘C’ check to the 40000 FH ‘4C’ check, having the former as reference. Results also show that the maximum value of unscheduled maintenance workload for the considered ‘C’ checks was 593% of the average unscheduled workload of the 5000 FH ‘C’ check. This illustrates the increase of unscheduled workload that occurs throughout the service life of aircraft.

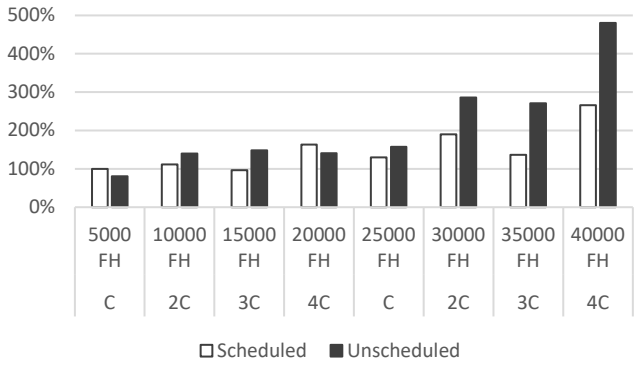


Fig. 2. Average scheduled and unscheduled maintenance workloads for the collected ‘C’ checks.

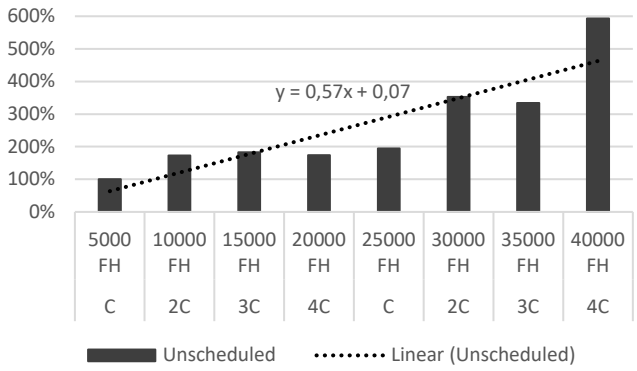


Fig. 3. Average unscheduled maintenance workloads for the collected ‘C’ checks.

When compared with [7] and [8], where is stated that up to 50% of the total maintenance workload is due to unscheduled maintenance, results from the collected data show that this ratio can be as high as 66%.

IV. Framework for Aircraft Maintenance Estimation - FRAME

The Framework for Aircraft Maintenance Estimation (FRAME) consists in the establishment of a set of requirements for the treatment and recording of aircraft maintenance data and in the definition of a method to analyze such data. There are two motivations for the development of FRAME: (1) to characterize the aircraft maintenance workload, particularly the

unscheduled maintenance, as accurately as possible in the presence of scarce information and as early as possible in the tactical phase of the capacity planning process; and (2) to provide MROs with a mean to develop tactical plans that allow a pre-scheduling of the maintenance work.

A. Maintenance Data Treatment and Recording

This section establishes a set of requirements for the treatment of historical data and the recording of current and future maintenance data with the aim of helping aircraft MRO organizations in managing these data and actively and successfully use them for the purpose of capacity planning. Three important shortcomings were found in the collected data: (1) the ambiguous classification of maintenance events by those involved in the planning process; (2) the inability of distinguishing between internal and external zones directly by using the industry standard zoning [11]; and (3) the use of redundant skills denominations resulting from past organizational changes. These issues severely hamper data analysis for the purpose of capacity planning. If treated and recorded according to coherent and unambiguous classification systems, these data prove invaluable in reducing the uncertainty related to capacity planning and, ultimately, in improving aircraft MRO organizations efficiency.

1) Maintenance Events

The established classification system for the aircraft maintenance interventions found in the collected data is presented in Fig. 4. The classification system, although based on the maintenance requirements of the studied aircraft family, is applicable to other aircraft models by considering the appropriate maintenance requirements and is intended to systematize the recording of the type of maintenance interventions for capacity planning. The first level of the classification system is the 'Maintenance Event'. This denomination results from the fact that not every maintenance intervention is a *check*, i.e. a maintenance intervention mainly based on the execution of scheduled inspection tasks. The second level differentiates a 'Maintenance Event' between 'Standard Maintenance Check', 'Non-Standard Maintenance Check', and 'Other Services'. A 'Standard Maintenance Check' is a check composed of a maintenance package established by the aircraft manufacturer, identified by a letter check for 'normal utilization' such as 'A' and 'C' checks, and by multiples of 6 months (MO) for 'low utilization', in the case of the studied aircraft family. A 'Non-Standard Maintenance Check' is a check that is not composed by a maintenance package established by the manufacturer, but is constituted mainly by inspection tasks. By establishing a specific class for these non-standard checks, one avoid to bias subsequent analysis. 'Other Services' refers to any maintenance intervention, limited in scope, referring to a specific maintenance service such as an engine change, an aircraft painting, parking, or even documentation management, material supply, among others.

2) Aircraft Zones

Every aircraft is divided into zones using a three-level, three-digit numbering system, in order to 'facilitate maintenance, planning, preparation of job instructions, location of work areas and components, and a common basis for various

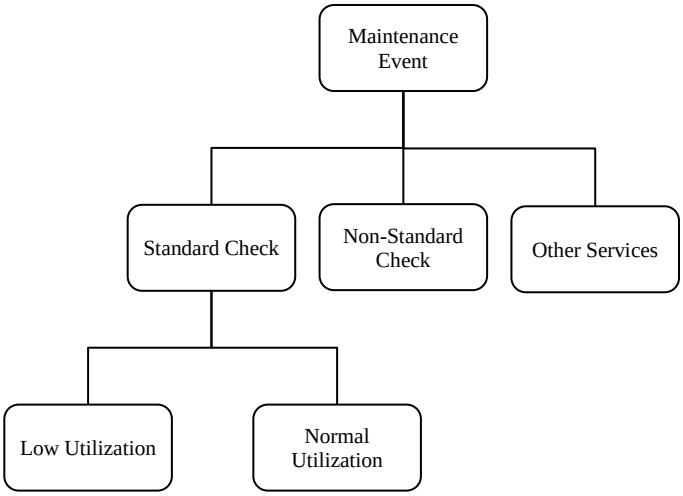


Fig. 4. Maintenance events classification system.

maintenance tasks' according to the aircraft zoning established in [11]. However, the standard aircraft zoning reveals to be too cumbersome for maintenance capacity planning due to its exhaustiveness and does not differentiate between internal and external zones *per se*. To overcome these issues, a complementary aircraft zoning was established and applied to the collected data and its main objectives are: (1) to provide an adequate level of detail for the purpose of capacity planning, while minimizing the number of zones used to increase data manageability in subsequent analysis; and (2) to allow the differentiation between internal and external zones directly. These objectives were accomplished by grouping individual standard zones into one major zone, major sub-zone, or zone. The grouping considers, obviously, the physical proximity of the zones and it aims at helping MRO companies in characterizing the maintenance workload spatially throughout the aircraft. Examples of the established aircraft zoning are presented in Table I.

TABLE I. Examples of the established aircraft zoning

Zone	Description	Standard Zones
100	Lower Fuselage (External)	-
110	Forward Fuselage I (Radome, Nose Landing Gear Compartment and Forward Electronic Compartment)	111, 113, 114, 115, 116, 213, 214, 215, 216
200	Upper Fuselage (External)	-
220	Forward Fuselage II (Cockpit Interior)	121, 122, 123, 124, 221, 222, 223, 224, 225, 226

3) Work Phases

Virtually all aircraft maintenance events unfold in a similar manner: (1) upon arrival, an aircraft undergoes documentation analysis, initial testing, and preparations such as fuel removal and exterior washing – this is called the 'Reception Phase'; (2) after reception, the aircraft access doors and panels are opened and some equipment is removed for maintenance actions to be

accomplished – this is called the 'Accesses Opening/ Disassembly Phase'; (3) aircraft engines, systems, and components are then inspected for anomalies - this is called the 'Inspection/Evaluation Phase'; (4) the anomalies found are corrected to restore aircraft airworthiness – this is called the 'Repairs/Modifications Phase'; (5) after the repairs, the removed equipment is reinstalled and access doors and panels are closed – this is called the 'Assembly/Accesses Closing Phase'; (6) finally, aircraft engines, systems, and components are operationally and functionally tested and after all tests are successful the aircraft is handed over to its operator – this is called the 'Certification and Delivery Phase'. This order under which maintenance events unfold may be used to establish work phases. Assuming that each work phase has pre-established start and end dates, or at least pre-established durations, by assigning one work phase to each maintenance task it is possible to define a moment for their accomplishment. The established work phases are presented in Table II.

TABLE II. Established work phases

Work Phase	Designation
1	Reception
2	Accesses Opening/Disassembly
3	Inspection/Evaluation
4	Repairs/Modifications
5	Assembly/Accesses Closing
6	Certification and Delivery

4) Skills

Each maintenance technician is licensed by the competent Aviation Authority (AA) to perform only a certain type of tasks, in a given aircraft model, according to the skills that she/he possesses [12]. A total of 58 skills designations were found in the collected data, both for hangar and backshop skills, some referring to the same technical skills. After removing redundant designations for the same skills, a total of 35 unique maintenance skills were obtained. From these, it was found that 10 skills accounted for over 98% of the work in the maintenance projects and for this reason the 25 skills least used were omitted from subsequent analysis. The considered maintenance skills are presented in Table III.

TABLE III. Considered maintenance skills

No.	Skill	Hangar/Backshop
1	Systems	Hangar
2	Structures	Hangar
3	Avionics	Hangar
4	Painting	Hangar
5	Generic	Hangar
6	Engines	Hangar
7	Inspector	Hangar
8	Composites	Backshop
9	Cleaning	Hangar
10	B2	Hangar

B. 3-Dimensional Maintenance Data Analysis – 3D-MDA

The 3-Dimensional Maintenance Data Analysis (3D-MDA) consists in the definition of a *space-time-skill* coordinate system used to characterize the expected maintenance work in the aircraft: (1) the space dimension comprises 42 established aircraft zones; (2) the time dimension comprises 6 defined work phases; (3) the skill dimension comprises 10 maintenance skills, presented in Table III, plus 'All', i.e. the analysis of work without differentiating the skill. This approach allows the characterization of the expected maintenance work with an adequate level of detail by calculating the workload distribution of each skill in any given aircraft zone and work phase, hence reducing the uncertainty related to the capacity problem. In addition, by providing MROs with a good insight on what skills are more likely to be required, where in the aircraft are these more likely to be needed, and when in the maintenance intervention are they more likely to be used, 3D-MDA enables a pre-scheduling of the maintenance work and an early identification of potential scheduling conflicts, for example at the aircraft zone level.

1) Statistical Analysis

In order to statistically characterize the maintenance events through the analysis of historical data, the workloads of each maintenance task *t* referring to each of the defined space-time-skill coordinates shall be summed within each project *p* according to (1). Similar calculations are possible for zone totals according to (2), i.e. total workload for skill *i* and zone *j* (without differentiating the phase), and phase totals according to (3), i.e. total workload for skill *i* and work phase *k* (without differentiating the zone). Fig. 5 depicts a template matrix for the representation of the space-time-skill coordinates.

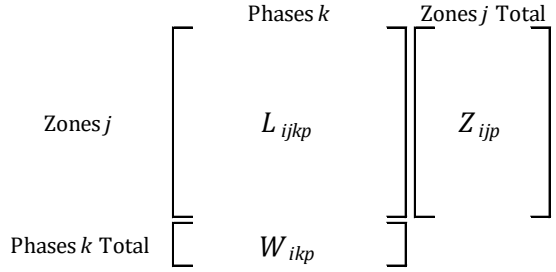


Fig. 5. 3D-MDA template matrix.

Where:

L_{ijkp} – total workload for skill *i* in zone *j* and work phase *k* in project *p* (in man-hours)
Z_{ijp} – total workload for skill *i* and zone *j* in project *p* (in man-hours)
W_{ikp} – total workload for skill *i* and work phase *k* in project *p* (in man-hours)

$$L_{ijkp} = \sum_{tUp} l_{ijkt} \quad \forall\, i,j,k,p \quad (1)$$

$$Z_{ijp} = \sum_{tUp} z_{ijt} \quad \forall\, i,j,p \quad (2)$$

$$W_{ikp} = \sum_{tUp} w_{ikt} \quad \forall\, i,k,p \quad (3)$$

Where:
l_{ijkt} – workload for skill *i* in zone *j* and phase *k* in maintenance task *t* (in man-hours)
z_{ijt} – workload for skill *i* and zone *j* in maintenance task *t* (in man-hours)
w_{ikt} – workload for skill *i* and phase *k* in maintenance task *t* (in man-hours)

After the analysis of the individual projects, these shall be consolidated in subsets according to the maintenance event type. Statistical parameters can then be calculated per maintenance event type *c* and per skill *i*, such as the arithmetic mean according to (4), the standard deviation, the maximum, and the minimum for the workloads, and the probability of occurrence of work in each space-time-skill coordinate according to (5).

$$\bar{C}_{ijkc} = \frac{\sum_{pUc} C_{ijkp}}{N_c} \quad \forall\, i,j,k,c \quad (4)$$

$$O_{ijkc} = \frac{n_{ijkc}}{N_c}\% \quad \forall\, i,j,k,c \quad (5)$$

Where:
$\bar{C}_{ijkc}$ – workload arithmetic mean for skill *i* in zone *j* and phase *k* for event type *c*
C_{ijkp} – total workload for skill *i* in zone *j* and phase *k* in project *p*
N_c– total number of projects of event type *c*
O_{ijkc}– probability of occurrence of work for skill *i* in zone *j* and phase *k* for event type *c*
n_{ijkc}– number of projects with work for skill *i* in zone *j* and phase *k* of event type *c*

2) Risk Analysis

The 3D-MDA enables a risk analysis approach to the maintenance capacity planning problem. Such approach can be used by aircraft MROs to signal potential problematic zone/phase/skill combinations, both in tasks scheduling and in resources allocation [7]. Risk is also defined in this context as a function of the probability of occurrence of a given event, and the impact (or consequence) of that same event [13]. The event in this case is the occurrence of maintenance work in a given aircraft zone, in a given work phase, for a given skill. The probability of occurrence is, as defined before, the probability of performing maintenance work in each of the space-time-skill coordinates. Finally, the impact is the amount of work that is expected be performed in each coordinate, measured in man-hours. Both the probability of occurrence and the workload impact can be defined through ordinal scales according to the decision-maker attitude towards risk [13]. Fig. 6 represents an example of a risk analysis performed using the 3D-MDA.

C. Integration of FRAME in the MRO planning process

The integration of FRAME in the tactical and operational phases of the MRO capacity planning process is presented in Fig. 7. First, the requirements established in section IV-A should be used for the treatment and recording of maintenance data in the data management system. Second, the statistical analysis, risk analysis, and the pre-scheduling of work performed with the 3D-MDA during the tactical phase of the

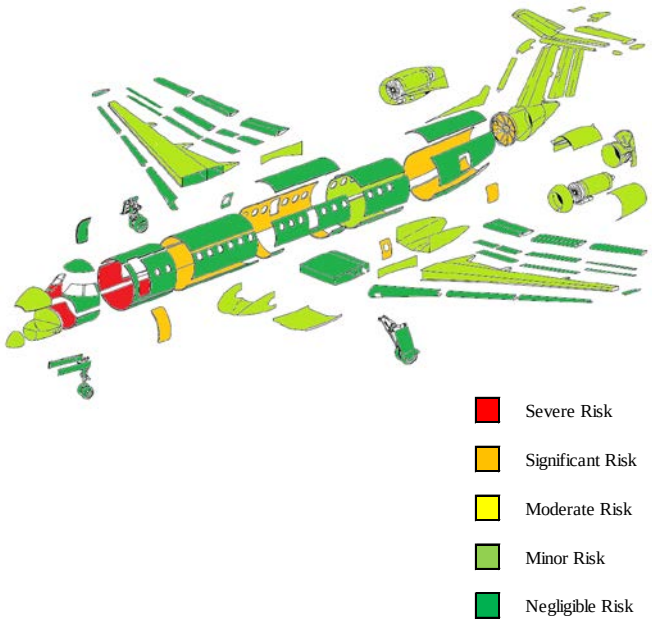


Fig. 6. Risk example for the 'Repairs/Modifications Phase' of the 20000 FH '4C' check and 'All' skills.

planning process are intended to serve as inputs for the development of the operational plan. As soon as the list of maintenance tasks are known in the operational phase, the estimates obtained from the statistical analysis performed in the tactical phase shall be updated with actual workloads and the scheduling of work based on the space-time-skills coordinates shall be updated with actual maintenance tasks. This is intended to ensure the coherence of the decision-making process between the tactical plan and the operational plan.

V. Conclusion

The objectives of this paper are threefold: (1) to quantitatively assess the capacity planning problem in aircraft maintenance regarding workloads, by using real maintenance data from a Portuguese independent MRO; (2) to propose a set of requirements for aircraft maintenance data treatment and recording; and (3) to propose a data analysis method for the characterization of the maintenance work with the aim of helping MRO organizations addressing the capacity planning problem. Regarding the first objective, results show that unscheduled maintenance workloads typically surpass scheduled workloads, and these unscheduled workloads tend to increase over time. This amount of uncertainty represents a serious threat to MROs activity, particularly when multiple aircraft are intervened simultaneously. As for the second and third objectives, FRAME, a framework for the treatment, recording, and analysis of maintenance data is proposed. The collected data revealed important shortcomings in using maintenance data *as is* for capacity planning. In order to overcome these shortcomings, FRAME comprises a set of requirements for the treatment and recording of maintenance data. The established requirements focus on four aspects: (1) the definition of a classification system for the unambiguous classification of maintenance events; (2) the definition of a complementary aircraft zoning to provide an adequate balance

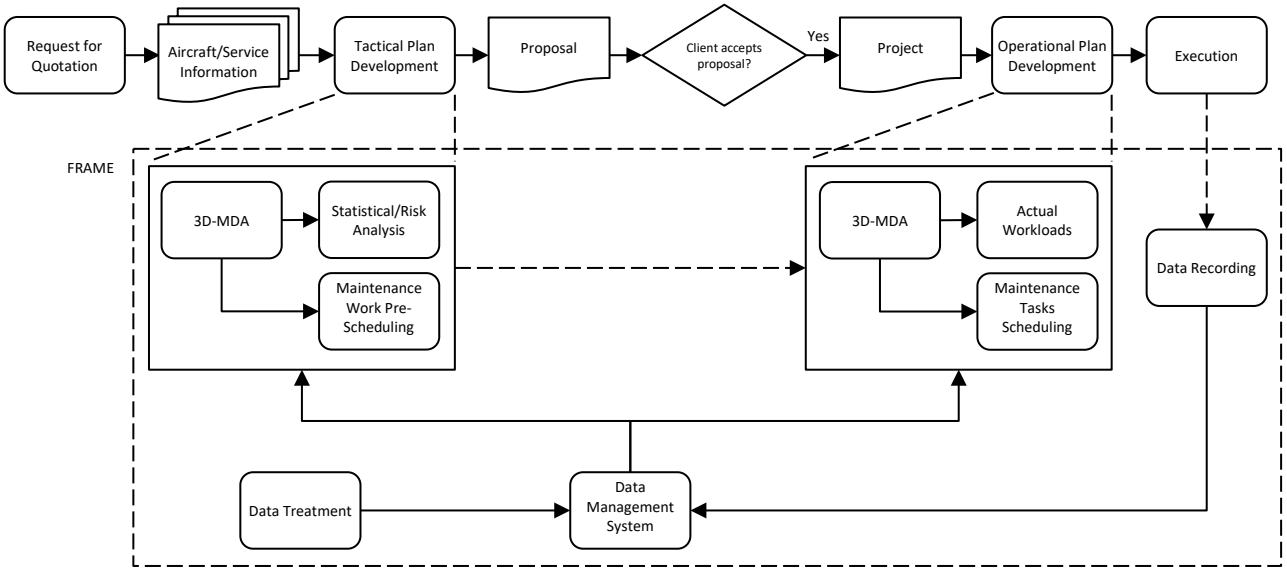


Fig. 7. Integration of FRAME in the MRO planning process.

between data detail and manageability, and to allow the differentiation of internal and external zones; (3) the definition of maintenance work phases for the accomplishment of maintenance tasks within pre-established time frames; and (4) the definition and utilization of unique technical skills. In addition, FRAME includes a data analysis method, the 3D-MDA. The 3D-MDA consists in a space-time-skill coordinate system, based on the established zones, phases, and skills, which allows a detailed characterization of the maintenance work through statistical and risk analysis. The data collected for capacity planning, i.e. the maintenance event, the aircraft zone, the work phase, and the maintenance skill, are transversal to aircraft maintenance and allow the application of the 3D-MDA to other aircraft, or even to other aeronautical products. In addition to the characterization of maintenance workloads, the 3D-MDA can be used with financial, spare parts, materials, ground support equipment (GSE), and tools data to characterize maintenance events in a more comprehensive way. These constitute important research opportunities to further assess the validity and applicability of the 3D-MDA, and in a broader scope of FRAME, to other aspects of the aircraft maintenance activity.

Acknowledgment

This work was supported by the Portuguese National Science Foundation (FCT) under Grant PD/BD/52345/2013.

References

[1] CEN, EN 13306:2001 - Maintenance Terminology. European Standard. Brussels: European Committee for Standardization, 2001.

[2] A. C. Márquez, “The maintenance concept”, in The Maintenance Management Framework, A. C. Márquez, Ed. Springer London, 2007, pp. 68–76.

[3] FAA, Advisory Circular 120-16F. Federal Aviation Administration, U.S. Department of Transportation, 2012.

[4] ATA, ATA MSG-3 Operator/Manufacturer Scheduled Maintenance Development, Revision 2007.1. Washington DC: Air Transport Association of America Inc., 2007.

[5] G. Budai, R. Dekker, and R. P. Nicolai, “Maintenance and production: a review of planning models”, in Complex System Maintenance Handbook, K. A. H. Kobbacy and D. N. P. Murthy, Eds. Springer London, 2008, pp. 321–344.

[6] A. E. Haroun and S. O. Duffuaa, “Maintenance organization”, in Handbook of Maintenance Management and Engineering, M. Ben-Daya, S. O. Duffuaa, A. Raouf, J. Knezevic, and D. Ait-Kadi, Eds. Springer London, 2009, pp. 3–15.

[7] P. Samaranayake, “Current practices and problem areas in aircraft maintenance planning and scheduling – interfaced/integrated system perspective”, in *Proceedings of the 7th Asia Pacific Industrial Engineering and Management Systems Conference 2006 17-20 December 2006, Bangkok, Thailand*, 2006, pp. 2245–2256.

[8] P. Samaranayake and S. Kiridena, “Aircraft maintenance planning and scheduling: an integrated framework”, J. Qual. Maint. Eng., vol. 18, no. 4, pp. 432–453, 2012.

[9] ICAO, Reference Manual on the ICAO Statistics Programme. Doc 9060/5, International Civil Aviation Organization, 2013.

[10] S. C. Eickemeyer, S. Steinkamp, B. Schuster, F. Bodenhage, and P. Nyhuis, “Reliable capacity planning despite uncertain disassembly, regeneration and reassembly workloads by using statistical and mathematical approaches – validation in subsidiaries of a global MRO company with operations in Asia, Europe and North America”, Procedia CIRP, vol. 23, pp. 252–257, 2014.

[11] ATA, ATA Specification 100 - Specification for Manufacturers’ Technical Data, Revision No.37. Air Transport Association of America Inc., 1999.

[12] M. C. Dijkstra, L. G. Kroon, M. Salomon, J. a. E. E. van Nunen, and L. N. Van Wassenhove, “Planning the size and organization of KLM’s aircraft maintenance personnel”, Interfaces, vol. 24, pp. 47–58, 1994.

[13] P. R. Garvey, Analytical methods for risk management: a systems engineering perspective. CRC Press, 2009.

A Two-stage Stochastic Programming for Rescheduling on a Single Yard Crane with Stochastic Release Time of Retrieval Tasks

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Xiaoyi Man†, Feifeng Zheng‡, Chengbin Chu*§
*School of Economics & Management, Tongji University, China
Email:mingliu@tongji.edu.cn
†Glorious Sun School of Business & Management, Donghua University, China
Email:manxiaoyi8996@163.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email:ffzheng@dhu.edu.cn
§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
Email:chengbin.chu@centralesupelec.fr

***Abstract*—Maritime transportation has become the mainstream of the global logistics. The performance of container terminals is the significant impact factor for the entire maritime transportation. However, many uncertainties such as weather variability, task handing time and equipment reliability affect the real-time execution of operational schedules in container terminals. In this paper, we study the rescheduling problem of a single yard crane in an import block. The main contributions of this paper are (1) the practical constraints of stochastic release time of retrieval tasks are considered; (2) a two-stage stochastic programming is first applied in the rescheduling problem. In the framework of the two-stage stochastic programming, the first stage is to make an initial schedule without considering retrieval tasks, while the second stage is to recourse and evaluate the total loss caused by the stochastic retrieval tasks. Both L-shaped method and Sample Average Approximation method will be applied to obtain solutions for the rescheduling problem.**

I. Introduction

Container transportation has become an essential part of the world trade. The performance of seaport container terminals is the key impact factor on the container transportation. The rapidly increase of global economy and containerized trade require the higher efficiency of the container terminals. A container terminal system mainly includes space resources (e.g., berths and yards) and equipment resources (e.g., cranes and trucks) (Vis et al. 2003). Therefore, the management of container terminals may involve the use of space resources which especially refer to container yards and the scheduling of operating equipments which mainly refer to cranes. However, the space utilization is difficult to be improved because of terrain restrictions. Meanwhile, due to the technical development of quay cranes and adequate number of yard trucks, the operating efficiency of the yard crane has become the bottleneck of the entire terminal. It plays a significant role in the development of container terminals.

In a container terminal, when a container vessel arrives, all equipments begin to handle containers in the current vessel. Generally speaking, quay cranes take containers from the vessel to yard trucks, then the containers are delivered to the container yard through yard trucks. There are typically three kinds of container blocks in the yard, i.e., *import block*, *export block* and *transshipment block*. The import block include exclusively import containers. On one hand, the containers in the import block are unloaded from yard trucks. On the other hand, the containers in the import block are delivered to domestic customers (see Figure 1). Thus, the yard crane in the import block both places the containers unloaded from yard trucks on the top of some container stacks and retrieves some containers from the stacks to load onto customer trucks. That is, there may exist two kinds of container tasks concurrently. We refer to the tasks for moving containers from yard trucks to the container stacks in the block as *storage tasks*, and refer to those for moving containers from the container stacks to customer trucks as *retrieval tasks*. The corresponding containers are respectively called *storage containers* and *retrieval containers*.

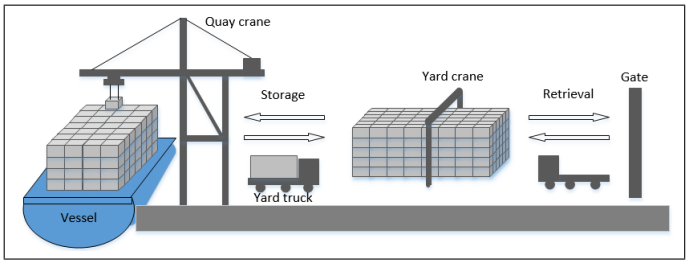


Fig. 1. Illustration of container flows

Motivated by the practical operations of the container terminal (e.g., Shanghai Yangshan Deep-water Port), the predictive vessel plan has been completed to greatly conform with timetables of shipping liners before the container vessel berths. It means that the operational plans for both quay cranes and yard cranes are made in advance. The operational plan of storage tasks from the vessel has been ready. The difference is that although the customers have made advance appointments on retrieval containers, the actual release time for customer trucks to retrieve containers from the import block

is stochastic. For instance, in Shanghai Yangshan Deep-water Port, there are 32.5 kilometers from the inland to the port. Uncertainties such as trucks breakdown, weather variability and approach time of trucks may exist on the way. Therefore, the uncertainties of retrieval tasks make the scheduling of yard cranes complicated. In this paper, we aim to minimize the impact of stochastic retrieval tasks on the initial schedule of yard cranes for storage tasks. First of all, we make an initial schedule for storage tasks without considering retrieval tasks. Secondly, according to the stochastic retrieval tasks, rescheduling is needed to adjust the follow-up plan of yard cranes after the insertion of retrieval tasks. To minimize the impact of stochastic retrieval tasks, a two-stage stochastic programming techniques is employed in this paper.

To the best of our knowledge, we are the first to establish a two-stage stochastic programming model for rescheduling on yard cranes. The contributions of this paper include: (1) the practical constraints of stochastic release time for retrieval tasks are taken into account for the rescheduling problem; (2) the two-stage stochastic programming is applied to reschedule the yard crane with the stochastic insertion of retrieval tasks.

The remainder of this paper is organized as follows. Section II gives a brief literature review. In Section III, we describe the concerned problem and propose a mathematical formulation. Section IV states an example to reveal the impact of stochastic retrieval tasks. The construction of L-shaped method and Sample Average Approximation method are presented in Section V. Finally, Section VI concludes the work.

II. Literature Review

The yard crane scheduling problem in container terminals has caught a vast of interests in the last decades. The literature mainly includes yard crane deployment problem (e.g., Zhang et al. (2002); Chung et al. (2002); Linn et al. (2003); Richard et al. (2003); Guo et al. (2012); Huang et al. (2015); etc.) and yard crane scheduling problem (e.g., Ng et al. (2005); Lim et al. (2007); Li et al. (2009); Wu et al. (2015); Sha et al. (2016); Gharehgozli et al. (2015, 2017); etc.). Most literature pays attention to the static scheduling with considering deterministic constraints of tasks such as release time and processing time. Nevertheless, the scheduling environment of the container yard is full of uncertainties and disruption environment exists. Therefore, this paper is involved in rescheduling problem with stochastic constraint which is rarely considered in the yard crane scheduling problem. Below we briefly review the studies on rescheduling and the two-stage stochastic programming.

A. Rescheduling

Many authors have considered the rescheduling problem in different scenarios such as a single machine, rolling stocks, berths and quay cranes (e.g., Yuan et al. (2007); Acuna-Agost et al. (2011); Hoogeveen et al. (2012); Albrecht et al. (2013); Wagenaar et al. (2017); etc.).

Han et al. (2010) study berth and quay crane rescheduling problems in a simultaneous way with uncertain arrival time and handling time. A mixed integer programming model is proposed to minimize both the expected value and the standard deviation of total service time and weighted tardiness time for all vessels in a time planning horizon. They develop a simulation search procedure based Genetic Algorithm to generate robust berth and quay crane schedule proactively. Zhen et al. (2011) address berth allocation problem under uncertain arrival time or operation time of vessels. Then, Zhen and Chang (2012) extend the work by proposing a bi-objective model for the robust berth allocation scheduling. A heuristic based on the time buffer inserting method is developed to solve the problem.

Expósitoizquierdo et al. (2014) present a study of rescheduling strategies for the quay crane scheduling with random disruptions such as breakdown. They first propose a dynamism management system to follow the implementation of the initial plan according to a completely deterministic environment. However, the unforeseen disruptions exist. Thus, the rescheduling strategies are presented to provide suitable schedules to the disruptions in the working sequences. An evolutionary algorithm is developed for the completely deterministic, then several rescheduling strategies are integrated into the dynamism management system.

Vianen et al. (2015) investigate a simulation-based rescheduling problem of stacker-reclaimer operations in a dry bulk terminal. They develop a simulation tool to reduce the waiting time of cargo trains which are being loaded at the terminal through interrupting ships servicing without significantly affecting the port time of ships. Meanwhile, Li et al. (2015) state the real-time disruption recovery optimization problem for the integrated berth allocation and quay crane assignment. Reactive recovery strategy is proposed to adjust the initial plan to handle realistic disruptions. Both maximizing the service quality and minimizing the recovery cost are the objectives. A heuristic apporoach based squeaky wheel optimization is developed to solve the problem.

Liu et al. (2016) consider the rescheduling problem dealing with the disruption that a quay crane breakdowns in the middle of the execution of a planned schedule. They aim to minimize the negative deviation of unexpected disruptions from the originally planned schedule. The problem is divided into two phases and the behavior perception-based disruption models are proposed for each phase based on prospect theory. An mixed integer programming based relax-and-fix algorithm is proposed to obtain optimal solutions for the first phase and a dynamic programming technique is applied for the second phase.

B. Two-stage Stochastic Programming

Stochastic programming was first proposed by Dantzig (1955) to handle the mathematical programming with uncertainty, then was extended both in theory and computational by subsequent researches (e.g., Huang et al. (2000); Barbarosoglu et al. (2004); Xu et al. (2009); Wang et al. (2015); etc.).

Liu et al. (2009) address the problem of allocating limited retrofit resources over multiple highway bridges to improve the resilience and robustness of the entire transportation network with a two-stage stochastic programming model. The first stage is to make retrofit decisions before the earthquake happens, while the second stage is to evaluate the total loss because of a realized earthquake including repair cost and increased travel delay in the network. The algorithm through extending L-

shaped method with using generalized benders decomposition is developed to handle the proposed model formulation.

Alizadeh et al. (2013) study a two-stage stochastic extension of the bilevel pricing model to improve the position of a product or service through estimating the loss of revenue due to neglecting randomness. The decisions for the first stage are made before observing the random outcome at the second stage. The decision for the second stage corresponds to a "recourse" once all randomness has been removed.

Li et al. (2015) extend the traditional two-stage stochastic programming method by enabling uncertainties presented as probability destiny functions and discrete intervals to be effectively incorporated within the optimization framework. An inexact mixed risk-aversion two-stage stochastic programming model is proposed for the water resources management under uncertainty. The initial decision is generated before the random variable is observed in the first stage. In the second stage, corrective measures for the expected loss caused by random variables are carried out after the random events have happened.

Summing up, in the literature (1) only the rescheduling of berth and quay crane has been addressed for stochastic disruptions, or to say, the rescheduling of the yard crane caused by randomness has been ignored; (2) the two-stage stochastic programming is mainly applied in the schedule of resources (e.g., medical resources; water resources; etc.), it is rarely applied in the rescheduling problem (see Table I). Therefore, to improve the literature studies and make the solution applicable in terminal daily schedules, this paper considers the following scenarios: (1) rescheduling on yard cranes with stochastic release time for retrieval tasks is necessary in the import block; (2) the two-stage stochastic programming is applied to minimize the expected loss caused by randomness for the rescheduling problem.

TABLE I. COMPARISON ON STUDIES OF RESCHEDULING AND TWO-STAGE STOCHASTIC PROGRAMMING

Literature	Object	Uncertainties	Approaches
Liu et al. (2009)	transportation network	earthquake	L-shaped
Han et al. (2010)	berth, quay crane	vessel arrive time, container handling time	Simulation based GA
Liu et al. (2016)	quay crane	quay crane breakdown	relax-and-fix algorithm
This paper	yard crane	release time of retrieval tasks	L-shaped, SAA

III. PROBLEM STATEMENT AND MATHEMATICAL MODEL

A. Problem Statement

The problem tackled in this paper can be described as follows. When a container vessel arrives in the destination port, the container terminal acquires information such as arrive time and the number of container loading/unloading tasks from vessel company. Based on this information, the container terminal generates the predictive vessel operation plans (e.g. quay crane plan, yard truck plan, and yard crane plan). Therefore, after a container vessel is berthed at the designated berth position, all facilities in the container terminal start to handle the containers in the vessel. In this paper, we assume that there is a single yard crane working in an import block which includes both storage and retrieval containers. Without loss of generality, the initial location of the yard crane is at the left end of the import block. For storage containers from the vessel, all of them are stacked in the import block according to the initial schedule. However, the release time for the retrieval containers delivered to domestic customers is stochastic in a time horizon. The real-time operational schedule of the yard crane in the import block is affected by the uncertain retrieval containers. The uncertain factor will cause extra cost and degrade the performance of the initial schedule. Therefore, the objective of this paper aims to minimize the impact of the stochastic release time of retrieval containers on the initial schedule. We denote the above problem as *a Single Yard Crane Rescheduling with Stochastic Release Time of Retrieval Tasks* (SYCR-SRT).

In order to cope with the high uncertainty involved in this problem, we built a model in a two-stage stochastic programming framework, which aims to take into account the expected loss caused by stochastic retrieval containers. The first stage is to make an optimal schedule of the yard crane for all storage tasks before the retrieval tasks arrive. The second stage is to recourse and evaluate the total loss due to the insertion of retrieval tasks. The second stage cost (delay cost) is a random variable dependent on the stochastic release time of the retrieval tasks.

B. Model formulation

1) Basic Notations: Below we propose a two-stage stochastic programming model for the SYCR-SRT problem. We first define parameters and variables, and then present the objective function and relevant constraints.

Indices

i, j : indices of storage tasks;

k : index of retrieval tasks;

t : index of time steps.

Parameters

N : set of storage tasks, and $N = \{1, 2, \ldots, |N|\}$;

R : set of retrieval tasks, and $R = \{1, 2, \ldots, |R|\}$;

T : set of time steps, and $T = \{1, 2, \ldots, |T|\}$;

C_s : the delay cost for a storage task in a unit time step;

C_r : the delay cost for a retrieval task in a unit time step;

l_i : the slot location of storage task i, $1 \leq l_i \leq 40$ in this paper, where $i \in N$;

r_i : the release time of storage task i, where $i \in N$;

h : the process time of a task (storage or retrieval), which is a constant in this paper;

d_i : the due date of storage task i, where $i \in N$;

Ω : set of all possible scenarios for retrieval tasks;

w : the possible scenario of retrieval tasks, where $w \in \Omega$;

P_w : the probability of scenario w, $\sum_{w \in \Omega} P_w = 1$;

l_k^w : the slot location of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

r_k^w : the release time of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

d_k^w : the due date of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

$t_{i,j}$: the gantry time from the location of task i to the location of task j, where $i, j \in N$ and $i \neq j$, similarly, $t_{i,k}$ and $t_{k,i}$, where $i \in N, k \in R$;

M : a sufficiently large integer.

Decision Variables

μ_i : the start time of storage task i, where $i \in N$;

ν_i : the completion time of storage task i, where $i \in N$;

π_i : the delay time of storage task i, where $i \in N$;

μ_i^w : the start time of storage task i in scenario w, where $i \in N, w \in \Omega$;

ν_i^w : the completion time of storage task i in scenario w, where $i \in N, w \in \Omega$;

π_i^w : the delay time of storage task i in scenario w, where $i \in N, w \in \Omega$;

μ_k^w : the start time of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

ν_k^w : the completion time of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

π_k^w : the delay time of retrieval task k in scenario w, where $k \in R, w \in \Omega$;

$\alpha_{i,j}$: =1 if storage task i is processed before storage task j; =0 otherwise, where $i, j \in N$ and $i \neq j$;

$\beta_{i,t}$: =1 if storage task i is started to be processed at time step t; =0 otherwise, where $i \in N, t \in T$;

$\eta_{i,t}^w$: =1 if storage task i is started to be processed at time step t in scenario w; =0 otherwise, where $i \in N, t \in T, w \in \Omega$;

$\delta_{i,k}^w$: =1 if retrieval task k is immediately processed after storage task i; =0 otherwise, where $i \in N, k \in R, w \in \Omega$;

$\rho_{k,i}^w$: =1 if retrieval task k is processed before storage task i; =0 otherwise, where $k \in R, i \in N, w \in \Omega$;

$\xi_{k,t}^w$: =1 if retrieval task k is started to be processed at time step t; =0 otherwise, where $k \in R, t \in T, w \in \Omega$;

2) Mathematical Model: With the above consideration, the *SYCR-SRT* problem is formulated in a two-stage stochastic programming framework as follows.

$$\begin{aligned}\min\ Z &= \sum_{i\in N} C_s \cdot \pi_i + E[F(i,k,w)] \\ &= \sum_{i\in N} C_s \cdot \pi_i + \\ &\sum_{w\in\Omega}[\sum_{i\in N} C_s \cdot (\pi_i^w - \pi_i) + \sum_{k\in R} C_r \cdot \pi_k^w] \cdot P_w\end{aligned}$$

$$\pi_i \geq \nu_i - d_i, \quad i \in N \tag{1}$$

$$\mu_i \geq r_i, \quad i \in N \tag{2}$$

$$\mu_i \geq l_i, \quad i \in N \tag{3}$$

$$\nu_i \geq \mu_i + h - 1, \quad i \in N \tag{4}$$

$$\mu_j \geq \nu_i + t_{i,j} + 1 - (1 - \alpha_{i,j})M, \quad i, j \in N, i \neq j \tag{5}$$

$$\alpha_{i,j} + \alpha_{j,i} = 1, \quad i, j \in N, i \neq j \tag{6}$$

$$\sum_{i\in N} \beta_{i,t} \leq 1, \quad t \in T \tag{7}$$

$$\sum_{t\in T} \beta_{i,t} = 1, \quad i \in N \tag{8}$$

$$\pi_i^w \geq \nu_i^w - d_i, \quad i \in N, w \in \Omega \tag{9}$$

$$\mu_i^w \geq r_i, \quad i \in N, w \in \Omega \tag{10}$$

$$\mu_i^w \geq l_i, \quad i \in N, w \in \Omega \tag{11}$$

$$\nu_i^w \geq \mu_i^w + h - 1, \quad i \in N, w \in \Omega \tag{12}$$

$$\mu_j^w \geq \nu_i^w + t_{i,j} + 1 - (1 - \alpha_{i,j})M, i, j \in N, i \neq j, w \in \Omega \tag{13}$$

$$\mu_i^w \geq \nu_k^w + t_{k,i} + 1 - (1 - \rho_{k,i}^w)M, \quad i \in N, k \in R, w \in \Omega \tag{14}$$

$$\sum_{i\in N} \eta_{i,t}^w \leq 1, \quad t \in T, w \in \Omega \tag{15}$$

$$\sum_{t\in T} \eta_{i,t}^w = 1, \quad i \in N, w \in \Omega \tag{16}$$

$$\pi_k^w \geq \nu_k^w - d_k^w, \quad k \in R, w \in \Omega \tag{17}$$

$$\mu_k^w \geq r_k^w, \quad k \in R, w \in \Omega \tag{18}$$

$$\mu_k^w \geq l_k^w, \quad k \in R, w \in \Omega \tag{19}$$

$$\nu_k^w \geq \mu_k^w + h - 1, \quad k \in R, w \in \Omega \tag{20}$$

$$\mu_k^w \geq \nu_i^w + t_{i,k} + 1 - (1 - \delta_{i,k}^w)M, \quad i \in N, k \in R, w \in \Omega \tag{21}$$

$$\delta_{i,k}^w + \rho_{k,i}^w = 1, \quad i \in N, k \in R, w \in \Omega \tag{22}$$

$$\sum_{k\in R} \xi_{k,t}^w \leq 1, \quad t \in T, w \in \Omega \tag{23}$$

$$\sum_{t\in T} \xi_{k,t}^w = 1, \quad k \in R, w \in \Omega \tag{24}$$

$$\eta_{i,t}^w + \xi_{k,t}^w = 1 \quad i \in N, k \in R, t \in T, w \in \Omega \tag{25}$$

$$\mu_i, \nu_i, \pi_i, \mu_i^w, \nu_i^w, \pi_i^w, \mu_k^w, \nu_k^w, \pi_k^w \in Z, \quad i \in N, k \in R, w \in \Omega \tag{26}$$

$$\alpha_{i,j}, \beta_{i,t}, \eta_{i,t}^w, \delta_{i,k}^w, \rho_{k,i}^w, \xi_{k,t}^w \in \{0,1\}, \quad i \in N, k \in R, t \in T, w \in \Omega \tag{27}$$

The objective function contains two parts. The former part is to minimize the delay cost of storage tasks without considering retrieval tasks, while the second part is the expected value of delay cost which includes both the changes of delay cost of storage tasks caused by the insertion of retrieval tasks

and the delay cost of the stochastic retrieval tasks. Constraints (1)-(8) restrict the conditions to be satisfied for storage tasks without considering retrieval tasks in the first stage. Constraint (1) is the definition of delay time of storage tasks. Constraints (2)-(3) present the start time of the storage tasks. Constraint (4) defines the completion time of storage tasks. Constraint (5) describes the relationship of the start time between storage task i and storage task j if i is processed before j. Constraint (6) states that there is a precedence relationship between two storage tasks. Constraints (7)-(8) ensure the definition of $\beta_{i,t}$. Similarly, Constraints (9)-(16) state the restricted conditions of storage tasks when the insertion of retrieval tasks occurs, and the inserted retrieval tasks are subjected to Constraints (17)-(25).

IV. Example

In this section, we present experiments with different scenarios of stochastic release time of retrieval tasks based on the two-stage stochastic programming model proposed in Section III by Cplex 12.6. In order to reveal the impact of stochastic retrieval tasks, we randomly generate instances with various number of storage tasks and retrieval tasks. The computational experiments are conducted with the software MATLAB R2014b on a PC with Intel Core i5, 3GHz processors and 4 GB RAM.

A. values of parameters

To reveal the impact of the stochastic retrieval tasks on the initial schedule of storage tasks, we present an instance with 10 storage tasks and 2 stochastic retrieval tasks in a given scenario. The values of parameters which are randomly generated are listed in Table II. In Table II, row "Task" reports the indices of the tasks that 1-10 indicate the storage tasks and 1', 2' indicate the retrieval tasks. The second to fourth rows report the slot locations, and the release times, and the due dates, respectively. We also define that a single task processing time h is three time steps when $t_{i,j}$ is one time step if YC moves one slot. Meanwhile, in the practical operations of container yard (e.g., Shanghai Yangshan Deep-water Port), the operators spare no effort to avoid the delay of storage tasks which are unloaded from the current vessel. Therefore, we define $C_s = 2, C_r = 1$.

TABLE II. Values of parameters for Instance 1

Task	1	2	3	4	5	6	7	8	9	10	1'	2'
l_i	5	13	19	27	9	30	15	25	22	35	10	20
r_i	3	10	16	28	12	40	18	30	23	38	8	18
d_i	4	16	22	38	17	47	24	40	30	59	12	22

B. results of instances

According to the result of the instance by Cplex in Figure 2, in the first stage (I), the yard crane handle all the storage tasks 1, 5, 2, 7, 3, 9, 8, 4, 6, 10 in order when ignoring retrieval tasks 1' and 2'. As is shown in Stage I, the yard crane starts to handle task 1 at time step 5 and ends at time step 7, then the yard crane moves to task 5. After completing task 5, the yard crane moves to task 2 to handle it. Similarly, tasks 7, 3, 9, 8, 4, 6, 10 follow the same procedure.

However, when considering retrieval tasks 1' and 2', the schedule of yard crane changes. In Stage II illustrated in Figure 2, task 1' is inserted between task 5 and task 2 at time step 16, and task 2' is in the middle of task 3 and task 9. Because of this stochastic insertions, the start time of all the tasks after the retrieval tasks differs from the result of the first stage. Therefore, the extra cost caused by the retrieval tasks happens.

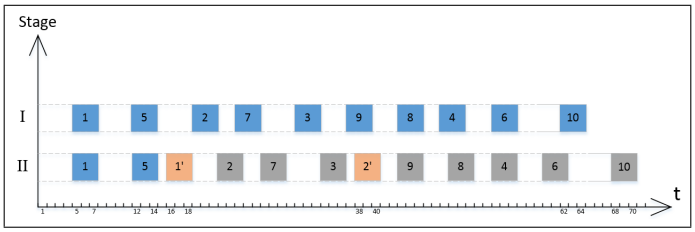


Fig. 2. Result of the instance

Less than two seconds are consumed to obtain the result of the instance by Cplex. In the first stage, the delay cost of storage tasks without retrieval tasks is 116 (e.g., $\sum_{i\in N} C_s \cdot \pi_i = 116$). In the second stage, the changes of delay cost of storage tasks caused by retrieval tasks is 78 (e.g., $\sum_{i\in N} C_s \cdot (\pi_i^w - \pi_i) = 78$); the delay cost of retrieval tasks is 24 (e.g., $\sum_{k\in R} C_r \cdot \pi_k^w = 24$).

For one thing, when there are 12 storage tasks and 3 stochastic retrieval tasks, the time consumed to obtain the solution increases to more than 1000 seconds. For another, many scenarios of retrieval task with different probabilities exist. Thus, more efficient method is necessary to solve the large-scale tasks in practical operations.

V. L-shaped and SAA method

According to the stochastic nature of the SYCR-SRT problem, L-shaped method is applied for the rescheduling problem. L-shaped method based on generalized Bender's decomposition greatly reduces the computational cost to generate a solution. Based on the two-stage stochastic programming model proposed in Section III, the second stage is convex and piecewise linear on a polyhedral domain. Therefore, it can be represented by a finite number of feasibility and optimality cuts. Then it proceeds to generating these cuts by solving successive linear programming problems. Meanwhile, the Sample Average Approximation method (SAA) is built to obtain the expected values with different scenarios of retrieval tasks. Due to the large amount of calculations and the limit of the article, the details of L-shaped and SAA methods are omitted here.

VI. Conclusion

In this paper, we study a rescheduling problem of a single yard crane with considering the stochastic release time of retrieval tasks in an import block. To minimize the impact of the stochastic retrieval tasks on the initial schedule of storage tasks, a two-stage stochastic programming model is proposed. The first stage is to make an initial schedule for storage tasks without considering retrieval tasks, while the second stage is to recourse and evaluate the total loss for all tasks with the insertion of retrieval tasks. Small-scale instances

are carried out to present the rescheduling problem. Both L-shaped method and Sample Average Approximation method are developed for the large-scales.

Future research directions may include (1) interference between more than one yard crane in the import block; (2) reshuffle for retrieval tasks; and (3) development of more efficient algorithms.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (71432007, 71571134, 71571135, and 71428002), and the Cai Yuanpei Program between the French Ministries of Foreign and European Affairs and the Higher Education and Research and the Chinese Ministry of Education (27927VE). The work was also supported by the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program.

REFERENCES

[1] R. Acuna-Agost, P. Michelon, D. Feillet, S. Gueye, *Sapi: statistical analysis of propagation of incidents. a new approach for rescheduling trains after disruptions*. European Journal of Operational Research, 215(1), 227-243, 2011.

[2] A. R. Albrecht, D. M. Panton, D. H. Lee, *Rescheduling rail networks with maintenance disruptions using problem space search.* Computers & Operations Research, 40(3), 703-712, 2013.

[3] S. M. Alizadeh, P. Marcotte, G. Savard, *Two-stage stochastic bilevel programming over a transportation network*. Transportation Research Part B, 58(4), 92-105, 2013.

[4] G. Barbarosoglu, Y. Arda, *A two-stage stochastic programming framework for transportation planning in disaster response*. Journal of the Operational Research Society, 55(1), 43-53, 2004.

[5] R. K. Cheung, C. L. Li, Lin, W, *Interblock crane deployment in container terminals*. Transportation Science, 36(1), 79-93, 2002.

[6] G. B. Dantzig, *Linear programming under uncertainty*. Management Science, 1, 197-206, 1955.

[7] C. Expósitoizquierdo, E. Lallaruiz, B. Meliánbatista, J. M. Morenovega, *A study of rescheduling strategies for the quay crane scheduling problem under random disruptions*. Inteligencia Artificial Revista Iberoamericana De Inteligencia Artificial, 17, 35-47, 2014.

[8] X. Guo, S. Y. Huang, Y. Li, *Dynamic space and time partitioning for yard crane workload management in container terminals*. Transportation Science, 46(1), 134-148, 2012.

[9] A. H. Gharehgozli, G. Laporte, Y. Yu, R. de Koster, *Scheduling twin yard cranes in a container block*. Transportation Science, 49(3), 686-705, 2015.

[10] A. H. Gharehgozli, F. G. Vernooij, N. Zaerpour, *A simulation study of the performance of twin automated stacking cranes at a seaport container terminal*. European Journal of Operational Research, 1-21, 2017.

[11] G. H. Huang, D. P. Loucks. *An inexact two-stage stochastic programming model for water resources management under uncertainty*. Civil Engineering Environmental Systems, 17(2), 95-118, 2000.

[12] X. Han, Z. Lu, L. Xi, *A proactive approach for simultaneous berth and quay crane scheduling problem with stochastic arrival and handling time*. European Journal of Operational Research, 207(3), 1327-1340, 2010.

[13] H. Hoogeveen, C. Lenté, V. T'kindt, *Rescheduling for new orders on a single machine with setup times*. European Journal of Operational Research, 223(1), 40-46, 2012.

[14] S. Y. Huang, Y. Li, X. Guo, *Yard crane dispatching to minimize vessel turnaround times in container terminals*. Simulation Conference, IEEE, 1747-1758, 2014.

[15] R. Linn, J. Y. Liu, Y. W. Wan, C. Zhang, K. G. Murty, *Rubber tired gantry crane deployment for container yard operation*. Computers & Industrial Engineering, 45(3), 429-442, 2003.

[16] A. Lim, B. Rodrigues, Z. Xu, *A m-parallel crane scheduling problem with a non-crossing constraint*. Naval Research Logistics, 54(54), 115-127, 2007.

[17] W. Li, Y. Wu, M. E. H. Petering, M. Goh, R. D. Souza, *Discrete time model and algorithms for container yard crane scheduling*. European Journal of Operational Research, 198(1), 165-172, 2009.

[18] C. Liu, Y. Fan, F. Ordónez, *A two-stage stochastic programming model for transportation network protection*. Computers & Operations Research, 36(5), 1582-1590, 2009.

[19] M. Z. Li, J. G. Jin, C. X. Lu, *Real-time disruption recovery for integrated berth allocation and crane assignment in container terminals*. Transportation Research Record, (2479), 49-59, 2015.

[20] W. Li, B. Wang, Y. L. Xie, G. H. Huang, L. Liu, *An inexact mixed risk-aversion two-stage stochastic programming model for water resources management under uncertainty*. Environmental Science and Pollution Research, 22(4), 2964-2975, 2015.

[21] C. Liu, L. Zheng, C. Zhang, *Behavior perception-based disruption models for berth allocation and quay crane assignment problems*. Computers & Industrial Engineering, 97, 258-275, 2016.

[22] W. C. Ng, K. L. Mak, *Yard crane scheduling in port container terminals*. Applied Mathematical Modelling, 29(3), 263-276, 2005.

[23] W. C. Ng, *Crane scheduling in container yards with inter-crane interference*. European Journal of Operational Research, 164(1), 64-78, 2005.

[24] W. C. Ng, K. L. Mak, *An effective heuristic for scheduling a yard crane to handle jobs with different ready times*. Engineering Optimization, 37(8), 867-877, 2005.

[25] J. L. Richard, C. Q. Zhang, *A heuristic for dynamic yard crane deployment in a container terminal*. IIE Transactions, 35(2), 161-174, 2003.

[26] M. Sha, T. Zhang, Y. Lan, X. Zhou, T. Qin, D. Yu, *Scheduling optimizatin of yard cranes with minimal energy consumption at container terminals*. Computers & Industrial Engineering, http://dx.doi.org/10.1016/j.cie.2016.03.022, 2016.

[27] I. F. A. Vis, R. de Koster, *Transshipment of containers at a container terminal: an overview*. European Journal of Operational Research, 147(1), 1-16, 2003.

[28] T. V. Vianen, J. Ottjes, G. Lodewijks, *Simulation-based rescheduling of the stacker-reclaimer operation*. Journal of Computational Science, 10, 149-154, 2015.

[29] Y. Wu, W. Li, M. E. H. Petering, M. Goh, R. D. Souza, *Scheduling multiple yard cranes with crane interference and safety distance requirement*. Transportation Science, 49(4), 990-1005, 2015.

[30] J. Wagenaar, L. Kroon, M. Schmidt, *Maintenance appointments in railway rolling stock rescheduling*. Transportation Science, 1-23, 2017.

[31] S. Wang, G. H. Huang, *A multi-level taguchi-factorial two-stage stochastic programming approach for characterization of parameter uncertainties and their interactions: an application to water resources management*. European Journal of Operational Research, 240(2), 572-581, 2015.

[32] Y. Xu, G. H. Huang, X. S. Qin, *Inexact two-stage stochastic robust optimization model for water resources management under uncertainty*. Environmental Engineering Science, 26(12), 1765-1776, 2009.

[33] J. Yuan, Y. Mu, *Rescheduling with release dates to minimize makespan under a limit on the maximum sequence disruption*. European Journal of Operational Research, 182(2), 936-944, 2007.

[34] C. Zhang, Y. W. Wan, J. Liu, R. J. Linn, *Dynamic crane deployment in container storage yards*. Transportation Research Part B, 36(6), 537-555, 2002.

[35] L. Zhen, L. H. Lee, E. P. Chew, *A decision model for berth allocation under uncertainty*. European Journal of Operational Research, 212(1), 54-68, 2011.

[36] L. Zhen, D. F. Chang, *A bi-objective model for robust berth allocation scheduling*. Computers & Industrial Engineering, 63(1), 262-273, 2012.

Fuel-oriented aircraft optimization for ground operations and transfer performance

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Qinrui Song†, Feifeng Zheng‡, Chengbin Chu§
*School of Economics & Management, Tongji University, China
Email: mingliu@tongji.edu.cn
†Glorious Sun School of Business & Management, Donghua University, China
Email: qr_s_jonksuk@163.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email: ffzheng@dhu.edu.cn
§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
School of Economics & Management, Tongji University, China
Email: chengbin.chu@centralesupelec.fr

***Abstract*—In this paper, we study the energy-oriented aircraft scheduling as it corresponds with fuel costs and delay costs, which typically incorporates transfer performance and ground operations (including arrival and departure). The above aspects are considered separately in previous works which cannot provide the overall optimization. Our main contributions include: (i) fuel consumption is not only considered as traditional delay costs; (ii) arrival and departure operations are conducted within in the multiple runway system; and (iii) an integer programming model is formulated for this newly proposed problem. We first construct this model and apply CPLEX to solve it in small size instances, numerical experiments have been conducted to demonstrate its correctness. Considering this problem is NP-hard and CPLEX is very time consuming, we devise a genetic algorithm to effectively produce nearly optimal solutions in large size instances. Numerical experiments show that our proposed model and algorithm can be applied to aircraft optimization so as to reduce ground congestion and additional costs.**

I. Introduction

In recent years, expansion of traffic has led to increasing demand; in addition, weather factors have led to delay. The delay caused by the unreasonable dispatch is becoming more and more common. On one hand, runways and other resources are often utilized in difficult situations, which increases the ground-waiting costs. On the other hand, longer wait times in the air increase fuel costs and have a negative impact on the rational use of energy. Currently, in the field of scheduling, Rager et al. (2014), Mansouri et al. (2016) and others have begun to pay attention to the energy issues. Meanwhile, delays in aircraft have brought loss and inconvenience to passengers who would transfer to another airplane; therefore, it limits the improvement in airport service.

As for the airport scheduling, many works investigate the arrival delay costs such as fuel costs. Han and Cheng (2006) minimize the average long-term delay costs of multi-runway airports by using the Markov Decision Process. Lee and Balakrishnan (2008) propose an algorithm for determining the lowest costs arrival schedule for airports, they analyze the tradeoff between average delay and fuel costs, then they take into account the costs of aircraft delay and fuel combustion speed. Based on the characteristics of flight and the optimization of system, Tian and Xu (2012) set up a dynamic optimization model of flight delay scheduling, they take the delay costs as the objective function to minimize the impact of costs and delay. Pan et al. (2015) argue that the costs of maintaining engine power during the flight delay is a major part of the loss, therefore, different models are used to improve the previous flight delay costs model. In addition, Hu and Zhang (2009) and Takeichi (2017) also study the aircraft delay costs problem.

At the same time, the costs of passenger recovery during the delay attract the eyes of many researchers for airport scheduling. Combined with mathematical programming for disruption management in the airline industry (DMAI). Bratu and Barnhart (2006) propose an airline plan recovery model and algorithm to develop aircraft, crew and passenger recovery plans. The objective is to minimize the operating costs of airlines, estimated passenger delay and interrupted costs. Zhao and Zhu (2008) reroute airplanes by minimizing the flight delay costs. The costs of flight delay is subdivided and the passenger disappointment spill over costs is put forward. Mansi (2012) proposes an oscillation strategy heuristicthe goal of this problem is to resume normal operation as quickly as possible, while minimizing the incurred costs and the potential impact on passengers. Pita et al. (2013) argue that airport congestion is the main reason for the major delay affecting the air transport industry, thus they propose a mixed integer linear optimization model assist airlines in conducting flight dispatch and passenger delay costs. Arlkan et al. (2016) put forward the mathematical formulation for the integrated aircraft and passenger recovery problem, they mainly consider the passenger-related costs. Subsequently, Chan et al. (2013) study a similar problem. Sang et al. (2013) and Atkinson et al. (2016) also research on the passenger-related costs in their study.

In summary, although the research on minimization of aircraft delay costs is the most common, they consider the different type of costs separately. The aircraft scheduling based

on one type of costs can reflect the local problem merely, thus the optimization results of this kind of research are different from considering the weight of all costs.

Therefore, comparing with the previous works, our main contributions include: (i) fuel costs, ground-waiting costs and transfer costs are considered in the the optimization process simultaneously, meanwhile, the arrival and departure of the same aircraft are both taken into account in the objective function, (ii) a mixed integer programming model is established to solve this practical problem; (iii) the algorithm is designed by using the idea of genetic algorithm double layer coding.

The rest of this paper is organized as follows. In section 2, based on the problem description, we describe the mixed integer programming model scheduling aircraft with multiple costs. Section 3 introduces the approach to the model solution and the result analysis. Then, detailed description of the algorithm designing is shown in section 4. Finally, in section 5, we summarize this work and put forward the future research direction.

II. Model Formulation

In this section, we mainly introduce the problem background and description, objective function and constraints condition of the model.

A. Problem Description

At the airport's flight area, the set of runways is R, there are n aircraft waiting for landing on runway r. After sliding p_i on runway r, aircraft i lands to the airport. The service time of the aircraft i staying at the airport is D_i, during which passengers finish leaving the plane, crew complete the next flight preparation and the next class of passengers board the plane. After the period of service time D_i, the flight i is ready to take off on runway r which could be different from the arrival sliding runway. In the course of the scheduling process, according to the runway capacity, the number of aircraft i arranged on runway r cannot exceed the maximum capacity of r, it will ensure the sustainable use of airport facilities.

On runway r, the estimated time of aircraft i to start landing is d_{ir}, and the time when the aircraft i actual landing time is s_{ir}. Thus, $d_{ir} - s_{ir}$ is the delay time before the aircraft i arriving. When the aircraft i arrives at the airport, due to the congestion of the runway, i continues flying in the air until runway r is idle. At this stage, the aircraft i consumes excess fuel, it results in unnecessary fuel costs.

According to the demand, the passengers will select the aircraft to transfer. In the optimized aircraft sequence, if the estimated time interval of two aircraft i and j is equal to or more than T_2, it means that i and j have the possibility of transfer. However, The delay of the aircraft causes a loss to the passenger who is preparing for the transfer. Therefore, transfer costs of the passengers is also considered in our optimization model. To reflect the importance of passengers' interests, the per unit time passenger-related delay costs is higher than the ground-waiting costs.

There is another case. When the aircraft i is ready to leave the airport, another aircraft is sliding on the runway r at that time, it means that the departure of aircraft i needs to wait on the ground and then increases the ground-waiting costs. in fact, the per unit waiting time costs are certain and lower than the fuel costs.

B. Conditional Assumptions

To establish the model successfully, the following assumptions are put forward:

- On runway r, the arrival or departure aircraft are the same type, i.e., regardless of the costs of different types of aircraft.
- Each aircraft landed at the airport runway r must take off, i.e., no aircraft strands at the airport.
- We avoid the situation that passengers cannot catch the plane.
- The aircraft can land or take off on any runway within the scope of the dispatch, regardless of the specific runway distribution for the particular aircraft.

C. Objective Description

Our model is a single objective model, the objective function is to minimize the delay costs, which can be divided into three categories: (i) fuel costs, (ii) ground-waiting costs, (iii) passengers' transfer costs.

D. Integer Model

Indices:

- i, j : indices of aircraft;
- r, r_1, r_2 : indices of runway.

Data:

- N : arrival and departure aircraft set;
- N_1 : arrival aircraft set;
- N_2 : departure aircraft set;
- n : the number of aircraft;
- R : runways set;
- p_{ir} : the sliding time on runway r before the aircraft i arriving or taking off, where $i \in N$, $r \in R$;
- d_{ir} : the estimated starting time of the aircraft i on runway r, where $i \in N$, $r \in R$;
- D_i : the service time of the aircraft i at the airport, where $i \in N$;
- M_r : the runway r capacity, where $r \in R$;
- T_2 : the minimum time interval for the transfer;
- H_{ij} : a binary variable. As for aircraft i (arrival) and j (departure), if $d_{jr} - d_{ir} - p_{ir} \geq T_2$, it means that j and i exist the possibility of transfer, then $H_{ij} = 1$, and vice versa, where $i \in N_1$, $j \in N_2$, $r \in R$;
- C_i : the per unit time costs of the aircraft i delay in the air or on the ground, where $i \in N$;

- C : the per unit time costs of the transfer delay;
- M : a sufficiently large integer;

Decisions Variables:

We have the following binary decision variables for our model:

$$X_{ijr} = \begin{cases} 1, & \text{if aircraft } i \text{ enters runway } r \\ & \text{before } j, \\ 0, & \text{otherwise;} \end{cases}$$

$$Z_{ir} = \begin{cases} 1, & \text{if aircraft } i \text{ is assigned to slide} \\ & \text{on runway r,} \\ 0, & \text{otherwise;} \end{cases}$$

$$Y_{ab} = \begin{cases} 1, & \text{if aircraft } i \text{ and } j \text{ are assigned} \\ & \text{to runway } r_1 \text{ and } r_2 \text{ ,} \\ 0, & \text{otherwise;} \end{cases}$$

in which:

$$a = i + n(r_1 - 1), i \in N_1, r_1 \in R$$

$$b = j + n(r_2 - 1), j \in N_2, r_2 \in R$$

- Y_{ab} : an auxiliary variable, which is constrained by the binary variable Z_{ir}, where $i \in N$, $r \in R$.

and an integer decision variable:

- s_{ir} : the actual start time of aircraft i on the runway r, where $i \in N$, $r \in R$.

Based on the start time s_{ir}, we can calculate the completion time, which equals to s_{ir}+p_{ir}.

Objective functions:

The objective function consists of three parts:

$$Obj = Obj_1 + Obj_2 + Obj_3 \tag{1}$$

of which:

$$Obj_1 = min \sum_{i=1}^{N_1} \sum_{r_1=1}^{R} (s_{ir_1} - Z_{ir_1} d_{ir_1}) C_i$$

$$Obj_2 = min \sum_{j=1}^{N_2} \sum_{r_2=1}^{R} (s_{jr_2} - Z_{jr_2} d_{jr_2}) C_j$$

$$Obj_3 = min \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} [(s_{jr_2} - s_{ir_1}) - (d_{jr_2} - d_{ir_1})]$$

$$CH_{ij} Y_{ab}$$

Subject to:

$$\sum_{r=1}^{R} Z_{ir} = 1, \quad i \in N \tag{2}$$

$$\sum_{i=1}^{N} Z_{ir} \le M_r, \quad r \in R \tag{3}$$

$$s_{ir} \le M Z_{ir}, \quad i \in N, r \in R \tag{4}$$

$$s_{ir} \ge d_{ir} - M(1 - Z_{ir}), \quad i \in N, r \in R \tag{5}$$

$$X_{ijr} + X_{jir} \le 1, \quad i \in N, j \in N, i \ne j, r \in R \tag{6}$$

$$Z_{ir} \ge X_{ijr}, \quad i \in N, j \in N, i \ne j, r \in R \tag{7}$$

$$Z_{jr} \ge X_{ijr}, \quad i \in N, j \in N, i \ne j, r \in R \tag{8}$$

$$X_{ijr} + X_{jir} \ge 1 - M(2 - Z_{ir} - Z_{jr}), \quad i \in N, j \in N, i \ne j, r \in R \tag{9}$$

$$s_{jr} - s_{ir} \ge p_{ir} - M(1 - X_{ijr}), \quad i \in N, j \in N, i \ne j, r \in R \tag{10}$$

$$s_{jr_2} + M(1 - Z_{jr_2}) \ge s_{ir_1} + p_{ir_1} + D_i - M(1 - Z_{ir_1}), \quad i \in N_1, j \in N_2, r_1, r_2 \in R \tag{11}$$

$$s_{jr_2} - s_{ir_1} + M(2 - Z_{ir_1} - Z_{jr_2}) + M(1 - H_{ij}) \ge d_{jr_2} - d_{ir_1} + M(2 - Z_{ir_1} - Z_{jr_2}) + M(1 - H_{ij}), \quad i \in N_1, j \in N_2, r_1, r_2 \in R \tag{12}$$

$$Y_{ab} \le Z_{ir_1}, \quad i \in N_1, r_1 \in R, \tag{13}$$

$$Y_{ab} \le Z_{jr_2}, \quad j \in N_2, r_2 \in R \tag{14}$$

$$Y_{ab} \ge 1 - M(2 - Z_{ir_1} - Z_{jr_2}), \quad i \in N_1, j \in N_2, r_1, r_2 \in R \tag{15}$$

Ranges:

$$X_{ijr} \in \{0, 1\}, i \in N_1, j \in N_2, r \in R \tag{16}$$

$$Z_{ir}, Y_{ab} \in \{0, 1\}, \quad a = i + n(r_1 - 1), b = j + n(r_2 - 1), i \in N, r \in R \tag{17}$$

$$s_{ir} \in Z^+, i \in N, r \in R \tag{18}$$

Equation (1) represents the objective function which minimize the total weighted costs. During the process of arrival or departure, constraint (2) indicates that the aircraft i can only be assigned to one runway. Constraint (3) means that the sum of the number of aircraft on the runway r cannot exceed the capacity, i.e., the load rate of the runway should not exceed 100%. Constraint (4) represents that if the aircraft i is assigned to land or take off on runway r, there must be a start time of aircraft i on runway r, and vice versa. Constraint (5) requires that if the aircraft i is assigned to land or take off on the runway r, the actual start time should not be earlier than the estimated start time. Constraints (6)-(9) ensure that only one aircraft can slide on the runway r at every moment, meanwhile, if two aircraft i and j are scheduled to enter or leave on the same runway, the aircraft must exist only one sequence, otherwise if the two aircraft are not arranged on the same runway, we do not have to determine their sequential relationship. Constraint

(10) indicates that if the two aircraft i and j are scheduled to slide on the same runway, the actual start time interval shall be guaranteed lager than the sliding time of the previous aircraft. Constraint (11) indicates that the departure time of the aircraft i should take place after the landing completion time and the service time at the airport. Constraint (12) restrains the aircraft must meet the requirement of the transfer, it avoids the condition that passengers do not catch the plane in the transit process because of the aircraft delay. Constraints (13)-(15) limit that the aircraft i and j are assigned to runway r_1 and r_2 simultaneously. Finally, constraint (16) is the range of the binary variables and constraint (17) is the range of the integer variable.

III. Optimal Solution for Model

As for calculation, we mainly use the commercial solver CPLEX is called in MATLAB R2014a to solve the model accurately. By using the following a group of data, we analyze the optimal solutions of the model and prove its correctness and optimality.

A. Small-size Instance

- The number of aircraft: $n = 8$, and $N_1 = N_2 = n$, $N = 2n = 16$;
- The number of runway: $R = 2$;
- The sliding time: $p_{ir} = [2, 3, 2, 4, 3, 2, 4, 1, 1, 1, 1, 1, 1, 1]$;
- The estimated starting time: $d_{ir} = [1, 2, 2, 5, 3, 4, 7, 5, 8, 8, 12, 14, 12, 13]$;
- The minimum time interval for transfer: $T_2 = 6$;
- The service time of aircraft: $D_i = [2, 3, 4, 3, 2, 6, 2]$; **Note:** In the above instance, a unit time represents 10 minutes.
- The capacity of runway: $M_r = [6, 8]$
- Based on d_{ir} and p_{ir}, We can get the H_{ij}:

$$H_{ij} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

  Note: For example, as for $H_{ij}(1, 4)$ means that the arrival of aircraft 1 and the departure of aircraft 4 meet the transfer requirement.
- The per unit time costs of arrival delay are 5;
- The per unit time costs of departure delay are 1:
- The per unit time costs of passengers transfer delay are 3.

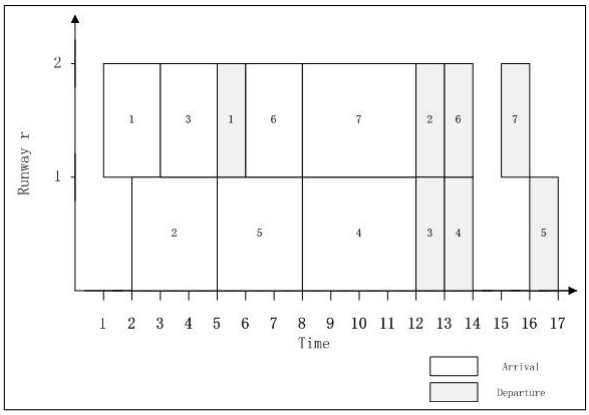


Fig. 1. Scheduling Sequence.

B. Result Analysis

We obtain the optimal result related to input parameters by CPLEX, see Figure 1.

The minimum weighted total costs of the objective function for this optimization are 101, in which the costs of fuel caused by air delays are 35, the ground-waiting costs are 15, and the delay costs in the passengers' transfer are 51. Through Figure 1, runways are fully utilized and the idle time is very small which reduces the waste of resources. Consequently, the model could enhance the operational efficiency of the airport more comprehensive.

In addition, we also test the correctness of the model through multiple sets of random data. Due to space limitation, we don't repeat anymore.

IV. Algorithm Design

As for large-scale instances, CPLEX is very time consuming, even cannot solve the results. Therefore, based on the feasible solution instances, we design the algorithm to solve the model with large-scale instances. According to the two runways and the strong constraint in the practical problem, we adopt the idea of genetic algorithm to obtain the feasible solution. Then, by means of natural genetics, i.e., crossover, mutation and the certain rules of operation, several iterations can produce better approximation solution gradually.

A. Chromosome Representation

By using the two-level coding method, we design the chromosome, each chromosome represents a feasible solution to the problem. The following is our designing ideas: (i) when the total number of aircraft is n, the individual is expressed as an integer string of length $4n$, (ii) the length of the first half of the chromosome of is $2n$, which shows that the arrival or departure of the aircraft among them, at the same time, the order of the genes indicates that the order of aircraft enters the runway, (iii) the length of latter half is also $2n$, it represents the first half of each arrival or departure aircraft is arranged for runway r.

For example, let $R = 2$, $n = 4$, it means the arrival and departure of four aircraft are scheduled on two runways. The length of the chromosome is 16. The first part of the

chromosome shows the eight working process of the four aircraft(including arrival and departure), the first number of the gene shows the aircraft number i and the last two digits of the number represents arrival or departure, in which 01 is the equal of the arrival, 02 is the equal of the departure. From the beginning of the gene 9, it indicates a runway for the arrival or departure of the four aircraft, meanwhile, every gene corresponds to the sequence of the aircraft from the front. As shown in Figure 2, for example, the first gene 301 means that the third plane arrival, the gene 9 indicates that process of gene 1 is scheduled sliding on runway 1. The meaning of other genes is same as it.

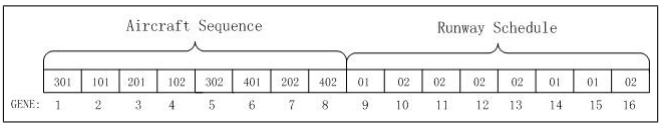


Fig. 2. Chromosome representation.

B. Initialization Population

According to the input parameters and practical problems, we initialize the individual. Followed by the algorithm 1, we determine the initial population scheduling sequence with delay time and next calculate the three types of delay time and costs.

First, we define $Chrom$ to represent the gene population. According to the two parts of the chromosome division, we randomly generate first half of the $Chrom$ to show the aircraft scheduling sequence, then generate the second layer of $Chrom$ to express the runway sequence randomly, see figure 3. Here we add a strong constraint: before randomly allocating the current gene in the second half of the $Chrom$, we compare the number of aircraft allocated to each runway and the runway capacity which we input previously. For example, if the number of aircraft on runway r_1 has reached the upper limit of the r_1 capacity, the random allocation should not be carried out, the runway number of this gene is set to r_2 which don't reach the upper capacity.

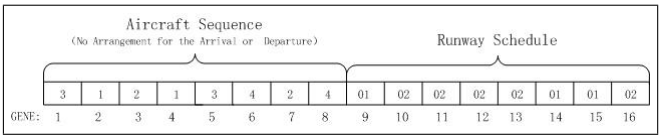


Fig. 3. Initialization population.

Algorithm 1 (Produce a feasible solution):

There are three main points in the individual's fitness calculation: (i) restoring the chromosome to contain the complete arrival and departure aircraft scheduling sequence, (ii) due to the transfer demand, ensuring the arrival and departure aircraft must have a specific order. (iii) calculating the start time of the current process of aircraft, which should consider the completion time of the last process on this runway, the runway idle time, as well as the transfer time interval.

$Step$ 1 : We select the individual as S from the genetic population $Chrom$ to mark the number of aircraft. Next, we add the arrival and departure coding into the first half of the S and define it as P. P is a gene sequence with arrival and departure of aircraft, where arrival is coded as 01 and departure is coded as 02.

$Step$ 2 : Set $i = 1$, $N(i)$ is the first half of the S and the length is $2n$. The initial random sequence is corrected while marking the arrival or departure number of the aircraft sequence. When the tag number is 01, the previous genes should be judged one by one, it determines whether there is the previous presence gene marked 02 satisfies the transfer time interval with the current tagged gene. If it satisfies the above conditions, this gene should advance to the current comparison of the aforementioned departure gene, the other aircraft gene follow by postponement.

See figure 4, when we mark the gene 5 that means the arrival of aircraft 2, then its estimated time would compare with the genes 3 and 4 which represents the departure of aircraft 1 and 3. After that, we find the arrival of aircraft 2 and the departure of aircraft 1 and 3 all satisfy the requirement of transfer, so the gene 5 should be put before the gene 2. When $i \geq 2n$, go to $Step3$.

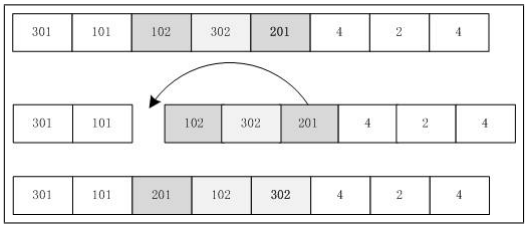


Fig. 4. Revise gene.

$Step$ 3 : $PVal$ is a matrix of $4*2n$, the first row of $PVal$ represents the aircraft scheduling sequence in P, the second row represents the actual time at which the aircraft i starts to slide on runway r, the third row represents the completion time of i sliding on runway r and the last represents the delay time. TF is the time which the last aircraft ends sliding on this runway r. When the estimated time of aircraft on gene i is earlier than the TF of the current assigned runway, the actual start time is TF and vice versa is the estimated start time. Meanwhile, the time interval between the actual arrival and departure of the same aircraft should be equal to or more than the service time of the aircraft staying at the airport.

$Step$ 4 : According to the P, we arrange the actual start time, meanwhile the aircraft start time with the demand for transfer is checked. In order to ensure that the passengers can catch the aircraft after the delay, we take a way to increase the ground-waiting time. Set $i = 1$, when $P(i)$ is the departure of the aircraft, its start time must compare with the comparison with the preceding arrival gene one by one. If the actual start time of Step 3 is earlier than the current comparison gene, the actual start time of this departure process have to delay to the completion time of the comparison gene.

C. Objective Introduction

The results of $PVal$, especially the last row, are used to calculate delay time and total costs. The type of delay costs is determined by the sequence of the $PVal$ first row. In particular, by subtracting actual time difference and estimated

time difference of the two aircraft, we can get the delay in the process of transfer and calculate the costs.

D. Genetic Operations

1) Cross operation: The population acquires new chromosomes by crossing, thus promoting the evolution of the whole population. In this paper, we use integer cross method for the first half of each chromosome. The cross function is randomly selected two individuals from the population and the crossing position. As shown an example in Figure 5, the cross position is 3. The crossing may result in the situation that the number of arrival or departure of aircraft is more than one, or lack of arrival or departure process. In this case, through the fill operation to ensure that each aircraft has two different operating conditions, see Figure 5.

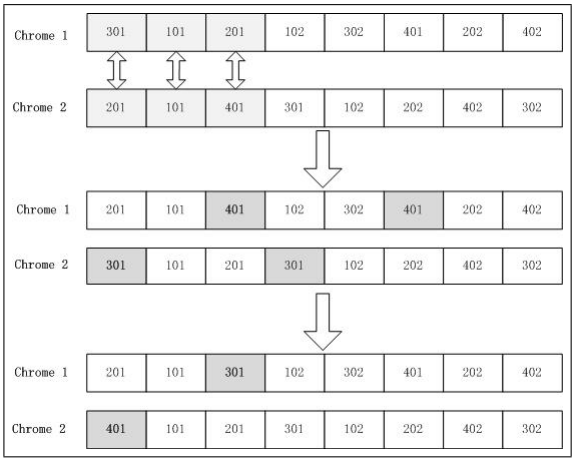


Fig. 5. Cross operation.

2) Mutation operation: The mutation is performed on the second half of each chromosome, select two variations for cross operation randomly. For example, select the gene 4 and 7, as shown in Figure 6.

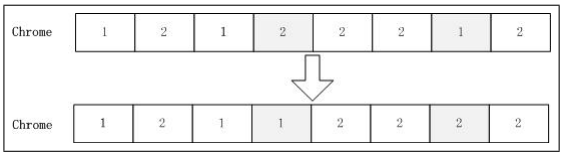


Fig. 6. Mutation operation.

V. Conclusion

In the development of all walks of life, the rational use of energy is a key issue, at the same time, passengers is an important factor restricting the development of the service industry. In this paper, we mainly focuses on the scheduling of minimizing the different weighted costs of the aircraft scheduling, the delay costs of the fuel, aircraft waiting and passengers' transfer loss, the experience of passengers is regarded as an important part of our objective function. By converting the practical problem into the mathematical model, we obtain the accurate scheduling scheme. Furthermore, we use the double-layer coding genetic algorithm which each layer code has different meanings. Double-layer coding collects the solution of the problem completely and expresses the solution of complex problem accurately with a chromosome.

In the future research, we will refine the algorithm to get the results which closer to the exact solution. In addition, we will look for new breakthroughs like airport maintenance into our research. Besides, by adding uncertain factors, we will determine the model into an uncertain model to solve more practical problems.

Acknowledgment

This work is done with the support from the National Science Foundation of China (71571134, 71428002, 71432007 and 71571135) and the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program. Finally, thanks for reviewers reading.

References

[1] U. Arlkan, S. Gürel, M. S. Aktürk, *Integrated aircraft and passenger recovery with cruise time controllability*. Annals of Operations Research, 236(2):295-317, 2016.

[2] S. Atkinson, K. Ramdas, J. W. Williams, *Robust Scheduling Practices in the U.S. Airline Industry: Costs, Returns, and Inefficiencies*. Management Science, 62(11):3372-3391, 2016.

[3] S. Bratu, C. Barnhart, *Flight operations recovery: New approaches considering passenger recovery*. Journal of Scheduling, 9(3):279-298, 2006.

[4] F. T. S. Chan, S. H. Chung, J. C. L. Chow, C. S. Wong, *An Optimization Approach to Integrated Aircraft and Passenger Recovery*. Proceedings of the Institute of Industrial Engineers Asian Conference 2013, Springer Singapore, 729-737, 2013.

[5] S. Han, L. Cheng, *Optimization of dispatch strategy with multi-runway by Markov decision process*. nternational Symposium on Systems and Control in Aerospace and Astronautics, IEEE Xplore, 2006

[6] H. Lee, H. Balakrishnan, *Fuel costs, delay and throughput tradeoffs in runway scheduling*. American Control Conference, IEEE Xplore, 2449-2454, 2008.

[7] R1. Mansi, *Disruptions in the airline industry: math-heuristics for re-assigning aircraft and passengers simultaneously*. European Journal of Industrial Engineering, 6(6):690-712, 2012.

[8] S. A. Mansouri, E. Aktas, U. Besikci, *Green scheduling of a two-machine flowshop: Trade-off between makespan and energy consumption*. European Journal of Operational Research, 248(3):772-788, 2016.

[9] W. Pan, Z. Wang, G. Zhou, Z. Pan, H. Zhang, *Flight Delay Cost Optimization based on Ground Power Supply*. 2015 International Industrial Informatics and Computer Engineering Conference, Atlantis Press, 2015.

[10] J. P. Pita, C. Barnhart, A. P. Antunes, *Integrated Flight Scheduling and Fleet Assignment Under Airport Congestion*. Transportation Science, 47(4):477-492, 2013.

[11] H. K. Sang, E. Feron, J. P. Clarke, A. Marzuoli, D. Delahaye, *Airport Gate Scheduling for Passengers, Aircraft, and Operation*. Computer Science, CoRR abs/1301.3535, 2013.

[12] M. Rager, C. Gahm, F. Denz, *Energy-oriented scheduling based on Evolutionary Algorithms*. Computers & Operations Research, 54:218-231, 2014.

[13] N. Takeichi, *Nominal flight time optimization for arrival time scheduling through estimation/resolution of delay accumulation*. Transportation Research Part C Emerging Technologies, 77:433-443, 2017.

[14] J. Tian, H. Xu, *Optimizing Arrival Flight Delay Scheduling Based on Simulated Annealing Algorithm*. Physics Procedia, 33(1):348-353, 2012.

[15] H. H. Zhang, M. H. Hu, *Multi-runway collaborative scheduling optimization of aircraft landing*. Journal of Traffic & Transportation Engineering, 9(3):86-91, 2009.

[16] X. L. Zhao, J. F. Zhu, *Study on modelling and algorithm of irregular flight delay operation*. Systems Engineering Theory & Practice, 4(28):129-135, 2008.

Arrival and departure joint scheduling in a multi-airport system with capacity restriction based on delay and penalty

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Zhihui Sun†, Feifeng Zheng‡, Chengbin Chu§
*School of Economics & Management, Tongji University, China
Email: mingliu@tongji.edu.cn
†School of Economics & Management, Tongji University, China
Email: monica_szh@163.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email: ffzheng@dhu.edu.cn
§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
School of Economics & Management, Tongji University, China
Email: chengbin.chu@centralesupelec.fr

***Abstract*—To effectively alleviate the increasingly serious situation of flight delays caused by the changes of airport capacity, we investigate a dynamic and stochastic joint scheduling problem of arrival and departure flights for multi-airport system with capacity restriction in this paper. What's more, in order to be closer to the actual operations, we consider the situation that a flight will be given a penalty when landing at one of its alternative airports rather than the predetermined one due to some unforeseen circumstances. To the best of our knowledge, there is no literature considering this situation or studying on the related problems. Then we establish a bi-objective joint scheduling model with capacity restriction under different scenarios, to minimize the total delayed time slots and penalties. At the same time to obtain the exact Pareto front, we develop an epsilon-constraint method, which can be proved to be feasible but time-consuming. Thus, we design a NSGA-II algorithm to seek for the approximate Pareto solutions within short computation time. Simulation experiment shows that the model can optimize the configuration of arrival and departure flights in a busy multi-airport system, reduce the total delayed time slots and penalties significantly, so that to improve the quality of air transport service effectively.**

I. Introduction

Airport capacity, as a main bottleneck resource, can be changed by some uncertain factors like weather condition. And these uncertainties can also change the arrival airport of every flight, which means it may fail to land at its predetermined airport, but an alternative one instead, followed by a penalty. Both of these two changes can bring about difficulties to the optimal allocation of air traffic flow, especially for those multi-airport system with a large number of flights. Therefore, it is essential to obtain the uncertain information about airport capacity so that to make decisions on how to allocate the arrival and departure flights and whether a flight needs to be arranged to land at its alternative airport under different capacity scenarios. There is no doubt an efficient scheduling strategy considering delayed time slots and penalties under different scenarios has an important theoretical and practical significance, which can improve the efficiency of air traffic flow management and guarantee the passenger satisfaction.

There is some literature on arrival and departure joint scheduling. To minimize the total weighted tardiness, Hancerliogullari et al. (2012) model the multi-runway scheduling problem as an identical parallel machine scheduling problem with release times, target times and deadlines, and schedule arrival and departure flights as close as possible to their target times. Through allocated methods of variable priority, Wang et al. (2014) establish a new arrival and departure scheduling optimization model based on the matching of the traffic flow with the capacity to minimize the total delay time. Through Compromise Immune Algorithm (CIA), Ding et al. (2007) optimize the scheduling of departure and arrival flight to decrease delay cost, based on ensuring security and considering airport capacity limit. On account of the methods that CDM theory handling the approach aircraft queues, arrivals and departures are considered jointly and are treated as interdependent operations by Chen et al. (2008). Then capacity conversion model and optimization algorithm are given to minimize the total delay. However, these studies are only concerned about just one objective, such as the delay cost or delay time, and only focus on single-airport scheduling. Aiming at maximizing the throughput and minimizing the flight delay losses as well as robustness, Zhou et al. (2014) build an optimal scheduling model for the arrival and departure flight, and use the genetic algorithm to solve this problem. Based on GHC-GA algorithm, Zhang et al. (2015) set up a dynamic multi-objective optimization model of arrival and departure flights on multiple runways to minimize the control workload and flight delay. Addressing the problem of optimal joint allocation of arrivals and departures at airports with

dynamic tradeoff between arrival and departure capacities, Gupta et al. (2013) set up an integer programming formulation with a discrete combinatorial optimization. However, although these studies build multi-objective models, they still only focus on single-airport arrival and departure joint scheduling, ignoring the multi-airport system.

The literature reviews above show that there is rare literature studying on the dynamic and stochastic joint scheduling problem of arrival and departure flights in a multi-airport system by building a multi-objective programming model, not to mention the fruits of exact algorithm. In this paper, we establish a dual-objective model for the arrival and departure flight in a multi-airport system based on delays and penalty. The optimization objectives are minimizing the total delayed time slots including arrival delay and departure delay, and minimizing the total penalties. In this multi-airport system, we should pay attention to the following points: (1) we only consider arrival flights to the system and departure flights from the system regardless of the connecting flights and internal flights between airports within the system; (2) there will be a penalty to an arrival flight if it lands at an alternative airport rather than its predetermined one because of some uncertain factors like weather condition; (3) there won't be a penalty for departure flights because their departure airports are unlikely to change. By modeling according to these points, we can coordinately analyze the departure and arrival process of the multi-airport system under different capacity scenarios, and the algorithm can produce an exact solution within reasonable computation time, which help the airport managers decide arrival and departure joint scheduling scheme under different scenarios to minimize the total delayed time slots and penalties. The contributions of this paper may include: (1) we simultaneously schedule the arrival and departure flights of a multi-airport system with a dynamic capacity restriction, and consider an actual circumstance where an arrival flight may land at its alternative airport rather than predetermined ones; (2) an epsilon-constraint algorithm is developed and implemented to obtain exact solutions within a huge computation time, which help to trade off the delay time and penalty; (3) a NSGA-II algorithm is developed to obtain approximation of the set of Pareto optimal solutions within a short computation time. The paper is organized as follows. Section II describes the multi-airport joint scheduling problem in details. In Section III we present an integer program model and in Section IV the epsilon-constraint method is used to obtain optimal Pareto front but is time-consuming. Then NSGA-II algorithm is applied to obtain the approximate Pareto front within short computation time, which is given in Section V. Finally we conclude the main contributions in Section VI.

II. Problem Description

During the time period, There are many flights waiting to land at or depart from the airports in the system with capacity restriction, in this situation, a reasonable arrival and departure joint scheduling meeting the capacity restrictions is needed, to ensure the total delayed time slots as few as possible and the total penalties as low as possible under all scenarios. However, because of the uncertain factors such as weather condition, the capacity of the airports within the system are dynamic and stochastic, that is to say, the arrival and departure scheduling have to be adjusted for the current optimal solution when the airport capacity situation changes.

A. Multi-airport System

With the rapid development of air transportation, airport is facing huge challenges. Let's take Beijing capital international airport for example. Its aviation business is undergoing a quick increase, which results in the resource tension and air traffic congestion problem. At the same time, the utilization of other airports within the economic circle should be improved, in which there exist idle facilities and underused resources. In this case, we need to integrate the resources of multiple airports to realize the coordinate operations and improvement of resource utilization at the airports in the area, which motivates the occurrence of multi-airport system. This system can not only service the air transport market more effectively and satisfy the growing demand for air transportation through the functional complementation between airports, but also can avoid the repeated investment and construction of airports, so that can make full use of airport resources.

B. Capacity Conversion Curve

Restrained by bottleneck resources such as runways, the capacity of an airport is limited, which can be given by its arrival capacity and departure capacity. Arrival capacity and departure capacity refers to the number of aircrafts that any given airport can allow to land at and depart from during any given time slice, respectively. On the one hand, the limited number of aircrafts that an airport can accommodate at a time slice, can lead to the mutual restricted relationship between arrival capacity and departure capacity, which accelerates the capacity conversion curve to come into being. On the other hand, due to some uncertain factors such as weather, the capacity of airport is dynamic and stochastic, which means the capacity conversion curve is different under different capacity scenarios, as is shown in Figure 1. We can only seek the optimal solutions on or within the capacity conversion curve.

C. Penalty for Alternative Airport

For the sake of flying security, aircrew have to make decisions on the choice of alternative airports for this flight before taking off. On account of some emergency situations, such as bad weather, changed capacity, broken-down aircraft and unwell passenger, a flight may fail to or are not suitable to fly to its predetermined airport, in this case, pilots can choose a satisfactory alternative airport to land at. However, there must be a penalty for flights landing at alternative airports because of the loss to passengers and airports. And it is impossible for a departure flight to have a penalty because it must depart from the predetermined airport rather than an alternative one.

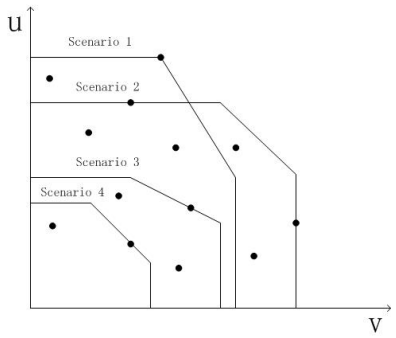


Fig. 1. An example of capacity conversion curve with scenarios

III. Mathematical Formulation

In this section, we present an integer programming formulation for the considered problem.

Indices:

s: scenario index;
i: flight index;
k: airport index;
t: time slot index;

Problem parameters:

S: set of scenarios;
N: set of flights;
B: set of airports;
T: set of time slots;
A_{ik}: 0-1 parameter, equals to 1 means flight i is expected to land at airport k, 0 otherwise;
D_{ik}: 0-1 parameter, equals to 1 means flight i is expected to depart from airport k, 0 otherwise;
a_i: 0-1 parameter, equals to 1 means flight i is arrival flight, 0 otherwise;
d_i: 0-1 parameter, equals to 1 means flight i is departure flight, 0 otherwise;
t_i^A: the expected arrival time slot of flight i;
t_i^D: the expected departure time slot of flight i;
P_s: the probability of scenario s;
p_i: the penalty of flight i when it lands at its alternate airports;
r_s,n_s: the parameters of capacity transition curve under scenario s;
h_s,g_s: the parameters of capacity transition curve under scenario s;

Decision variables:

x_{sikt}: equals to 1 if flight $i \in N$ departs from airport $k \in B$ during time slot $t \in T$ under scenario $s \in S$, 0 otherwise;
y_{sikt}: equals to 1 if flight $i \in N$ lands at airport $k \in B$ during time slot $t \in T$ under scenario $s \in S$, 0 otherwise;
C_{sik}^A: arrival time slot of flight $i \in N$ at airport $k \in B$ under scenario $s \in S$;
C_{sik}^D: departure time slot of flight $i \in N$ at airport $k \in B$ under scenario $s \in S$;
C_{si}^A: arrival time slot of flight $i \in N$ under scenario $s \in S$;
C_{si}^D: departure time slot of flight $i \in N$ under scenario $s \in S$;
u_{skt}: departure flow of airport $k \in B$ during time slot $t \in T$ under scenario $s \in S$;
v_{skt}: arrival flow of airport $k \in B$ during time slot $t \in T$ under scenario $s \in S$;
L_{si} : delayed time slots of flight $i \in N$ under scenario $s \in S$.
I_{sik} : the change of arrival airport of flight $i \in N$ owing to airport $k \in B$ under scenario $s \in S$.

Now we present the mixed integer linear programming formulation as follows:

$$\min \sum_{i \in N} \sum_{s \in S} P_s L_{si} \tag{1}$$

where the objective function is to minimize the total delayed time slots of all flights considering all scenarios.

$$\min \sum_{i \in N} \sum_{s \in S} (\sum_{k \in B} I_{sik}) P_s p_i \tag{2}$$

where the objective function is to minimize the total penalties for flights landing at alternate airports considering all scenarios.

$$\sum_{t \in T} \sum_{k \in B} x_{sikt} \geq d_i, \quad s \in S, \quad i \in N \tag{3}$$

$$\sum_{t \in T} \sum_{k \in B} x_{sikt} \leq 1, \quad s \in S, \quad i \in N \tag{4}$$

where these two constraints require that if flight i is a departure flight, then it must be assigned only one airport to depart from during only one time slot under any scenario s.

$$\sum_{t \in T} \sum_{k \in B} y_{sikt} \geq a_i, \quad s \in S, \quad i \in N \tag{5}$$

$$\sum_{t \in T} \sum_{k \in B} y_{sikt} \leq 1, \quad s \in S, \quad i \in N \tag{6}$$

where these two constraints require that if flight i is an arrival flight, then it must be assigned only one airport to land at during only one time slot under any scenario s.

$$x_{sikt} \leq \sum_{t \in T} x_{sikt} \leq d_i, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{7}$$

$$x_{sikt} \leq \sum_{k \in B} x_{sikt} \leq d_i, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{8}$$

where these two constraints guarantee that if flight i is not assigned to depart from airport k or to depart during time slot t under scenario s, then $x_{sikt} = 0$; and if a flight i is not a departure flight, then $\sum_{t \in T} x_{sikt} = 0$, $\sum_{k \in B} x_{sikt} = 0$.

$$y_{sikt} \leq \sum_{t \in T} y_{sikt} \leq a_i, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{9}$$

$$y_{sikt} \leq \sum_{k \in B} y_{sikt} \leq a_i, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{10}$$

where these two constraints guarantee that if flight i is not assigned to land at airport k or to land during time slot t under scenario s, then $y_{sikt} = 0$; and if a flight i is not an arrival flight, then $\sum_{t \in T} y_{sikt} = 0$, $\sum_{k \in B} y_{sikt} = 0$.

$$x_{sikt} \geq \sum_{t \in T} x_{sikt} + \sum_{k \in B} x_{sikt} - 1, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{11}$$

where this constraint implies that only when flight i is assigned to depart from airport k and to depart during time slot t under scenario s, then $x_{sikt} = 1$.

$$y_{sikt} \geq \sum_{t \in T} y_{sikt} + \sum_{k \in B} y_{sikt} - 1, \quad s \in S, \quad i \in N, \quad k \in B, \quad t \in T \tag{12}$$

where this constraint implies that only when flight i is assigned to land at airport k and to land during time slot t under scenario s, then $y_{sikt} = 1$.

$$\sum_{t \in T} t \cdot x_{sikt} = C^D_{sik}, \quad s \in S, \quad i \in N, \quad k \in B \tag{13}$$

where this constraint defines the departure time of flight i at airport k under scenario s.

$$\sum_{t \in T} t \cdot y_{sikt} = C^A_{sik}, \quad s \in S, \quad i \in N, \quad k \in B \tag{14}$$

where this constraint defines the arrival time of flight i at airport k under scenario s.

$$\sum_{k \in B} C^D_{sik} = C^D_{si}, \quad s \in S, \quad i \in N \tag{15}$$

where this constraint means that the departure time of flight i under scenario s is the one in which it depart from its assigned airport k.

$$\sum_{k \in B} C^A_{sik} = C^A_{si}, \quad s \in S, \quad i \in N \tag{16}$$

where this constraint means that the arrival time of flight i under scenario s is the one in which it land at its assigned airport k.

$$\sum_{i \in N} x_{sikt} = u_{skt}, \quad s \in S, \quad k \in B, \quad t \in T \tag{17}$$

where this constraint calculates the departure flow of airport k during time slot t under scenario s.

$$\sum_{i \in N} y_{sikt} = v_{skt}, \quad s \in S, \quad k \in B, \quad t \in T \tag{18}$$

where this constraint calculates the arrival flow of airport k during time slot t under scenario s.

$$0 \leq u_{skt} \leq r_s, \quad s \in S, \quad k \in B, \quad t \in T \tag{19}$$

$$0 \leq v_{skt} \leq n_s, \quad s \in S, \quad k \in B, \quad t \in T \tag{20}$$

$$u_{skt} \leq h_s + g_s \cdot v_{skt}, \quad s \in S, \quad k \in B, \quad t \in T \tag{21}$$

where these three constraints ensure that the departure flow and arrival flow of airport k during time slot t under scenario s satisfy the restriction of capacity transition curve.

$$L_{si} = (C^A_{si} - t^A_i)a_i + (C^D_{si} - t^D_i)d_i, \quad s \in S, \quad i \in N \tag{22}$$

where this constraint defines the calculation of the total delayed time slots of flight i under scenario s.

$$D_{ik} = x_{sik}, \quad s \in S, \quad i \in N, \quad k \in B \tag{23}$$

where this constraint implies that the departure airport of flight i is always the same as its expected one.

$$C^A_{si} \geq t^A_i, \quad s \in S, \quad i \in N \tag{24}$$

where this constraint shows that the arrival time slot of flight i under scenario s can not be earlier than its expected one.

$$C^D_{si} \geq t^D_i, \quad s \in S, \quad i \in N \tag{25}$$

where this constraint shows that the departure time slot of flight i under scenario s can not be earlier than its expected one.

$$I_{sik} \geq y_{sik} - A_{ik}, \quad s \in S, \quad i \in N, \quad k \in B \tag{26}$$

$$I_{sik} \geq A_{ik} - y_{sik}, \quad s \in S, \quad i \in N, \quad k \in B \tag{27}$$

where these two constraints show whether the arrival airport of flight i is changed owing to airport k under scenario s.

$$x_{sikt}, y_{sikt} \in \{0, 1\} \quad s \in S, i \in N, k \in B, t \in T \tag{28}$$

$$C^A_{sik}, C^D_{sik}, C^A_{si}, C^D_{si}, L_{si}, u_{skt}, v_{skt}, I_{sik} \in \mathcal{Z}^+, \quad s \in S, i \in N, k \in B, t \in T \tag{29}$$

where these constraints give the domains of decision variables.

IV. Solution Example

In order to apply the model above, we design a multi-airport system with three airports, five scenarios, six time slots and thirty-six flights. So there is no doubt that the values of a lot of parameters are known in this model, for example, the planning departure airport and time slot, planning arrival airport and time slot of every flight and so on. In our solution example, we generate these values randomly like TABLE I.

TABLE I
Expected arrival time slot of each flight

i	1	2	3	4	5	6	7	8	9
t_i^A	4	0	3	4	0	4	0	0	0
i	10	11	12	13	14	15	16	17	18
t_i^A	4	0	0	0	4	0	0	2	0
i	19	20	21	22	23	24	25	26	27
t_i^A	0	2	1	3	0	0	0	0	2
i	28	29	30	31	32	33	34	35	36
t_i^A	4	1	4	4	3	4	0	0	0

Based on TABLE I, we can make sense of whether a flight is an arrival flight and its expected arrival time slot. Let's take flight 1 and 2 for example, from the data above, we can conclude that flight 1 is an arrival flight to the system which land at time slot 4, and flight 2 is a departure flight because its arrival time slot is zero.

To solve it exactly, we use Matlab 2014b to compile the epsilon-constraint method to search for the optimal solutions on personal computer, and the set of optimal solutions are shown in TABLE II and the optimal Pareto front is shown in Figure 2.

TABLE II
Set of Pareto optimal solutions

No.	Delay	Penalty
1	0.69	8.08
2	0.84	7.48
3	0.87	6.64
4	1.02	6.04
5	1.05	5.92

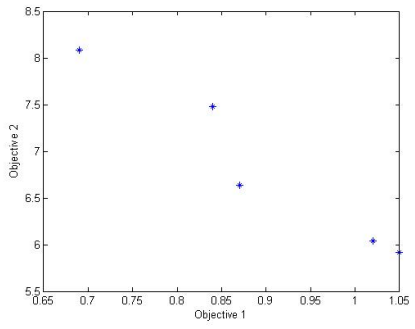


Fig. 2. Epsilon-constraint Pareto front

The optimal Pareto front obtained can assist airport managers to trade off between different objectives and make optimal decisions. However, it takes us more than one day to obtain just five points on the Pareto front. Hence, it is very necessary to have a more efficient heuristic algorithm.

V. NSGA-II Algorithm

As a classical intelligent optimization method, genetic algorithm has been widely used in the field of multi-objective optimization. Thus, to obtain the approximation of the set of Pareto optimal solutions of our bi-objective model within an acceptable computation time, we adopt NSGA-II algorithm in this paper, which is proposed by Deb et al. (2002). This algorithm has the superiorities of ensuring the optimality and diversity of individuals in the population evolution process and reducing the computational complexity when dealing with multi-objective optimal problem.

A. Algorithm Procedure

Step 1: Divide each chromosome into two sets, space genes set and time genes set, whose coding represent the arrival airport or departure airport of each flight and the arrival time or departure time of each flight, respectively, as shown in Figure 3. And initialize the population randomly, which is based on the problem range and constraints. For example, the time slots on each chromosome have to be no more than the expected ones of each flight.

Step 2: Parents are selected from the population by using binary tournament selection based on the rank and crowding distance. And a huge penalty is added to the fitness value once any constraint is violated. For example, capacity constraint is violated when the departure and arrival flow of an airport during one time slot can not meet the constraint of its capacity conversion curve. Then the selected population generates offsprings from crossover and mutation operators.

Step 3: The population with the current population and current offsprings is sorted again based on non-domination, which generates a set of fronts.

Step 4: Calculate the crowding distance of each front, and only the best N individuals are selected to form next population.

Step 5: Generates the new offsprings from crossover and mutation operators to the next population.

Step 6: Judge whether current evolutionary process meets the termination conditions, which is the number of evolution generations in our problem. If meets, then the algorithm is finished and the optimal Pareto solutions are output. Otherwise, go to step 3 to continue iteration.

For a detailed procedure of the NSGA-II algorithm, please see Liu et al. (2016).

The control parameters of NSGA-II algorithm include: the population size is 500, the terminate evolution generation is 50, the chromosome coding length is 360, the probability of crossover is 0.9, the probability of mutation is 0.1. And we use the rule of segmented crossover and mutation to do the genetic operations.

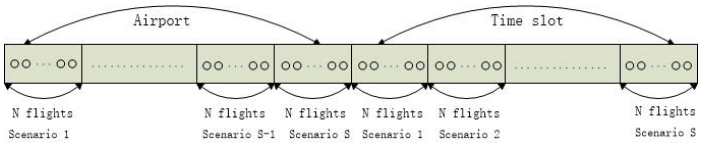


Fig. 3. Chromosome construction

B. Result Analysis

It shows that the individuals are gradually closed to the optimal direction, and can search for the approximation of the set of optimal Pareto solutions in the decision space as quickly as possible. The operation time of the whole algorithm is 0.83min, and we can obtain 7 approximated optimal solutions, which has an obvious advantage compared with more than 1 day to obtain only 5 optimal solutions. The comparison of these two results is shown in Figure 4, which also reveals a trade-off between different objectives and can provide decision supports for airport managers.

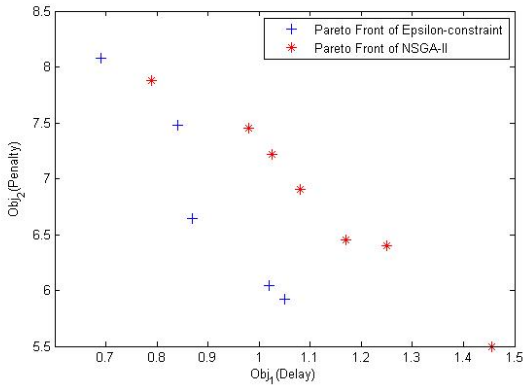


Fig. 4. Comparison of two Pareto fronts

VI. Conclusion

In this paper, we focus on the dynamic and stochastic scheduling problem of arrival and departure flights in a multi-airport system to minimize the total delayed time slots and the total penalties. Then we build an integer programming model, which is solved exactly by epsilon-constraint algorithm to obtain the Pareto front. However, in view of the huge time we spend on the exact solutions by using epsilon-constraint algorithm, we design a NSGA-II algorithm to obtain the approximate Pareto front within a short computation time. Both of these fronts can help airport managers to have a trade-off between different objectives according to the actual situation, and support their decisions on optimal scheduling of arrival and departure flight under different scenarios. It is obvious that compare with the expected planning, our flight scheduling can have a better coordination to the resources in the multi-airport system and boost the running efficiency of the whole system. Our contributions may include: (i) we simultaneously schedule the arrival and departure flights of a multi-airport system with a dynamic capacity restriction, and consider an actual circumstance where an arrival flight may land at its alternative airport rather than predetermined ones; (ii) an epsilon-constraint method is developed and implemented to obtain exact solutions within a huge computation time, which help to trade off the delay time and penalty; (iii) a NSGA-II algorithm is applied to obtain the approximation of a set of Pareto optimal solutions within a short computation time.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 71531011, 71571134, 71101106, and 71428002). The work was also supported by the Cai Yuanpei Program between the French Ministries of Foreign and European Affairs, the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang Program.

References

[1] G. Gupta, Waqar Malik, Y. C. Jung, *Enhanced model for joint optimization of arrival and departure strategies and arrival/departure capacity utilization at congested airports*. Aiaa Modeling & Simulation Technologies, 14(6), 6933-6946, 2013.

[2] G. Hancerliogullari, G. Rabadi, M. Kharbeche, A. Al-Salem, *Metaheuristics for aircraft arrival and departure scheduling on multiple runways*. IIIE Annual Conference. Proceedings, 33(12), 1-8, 2012.

[3] H. Zhou, X. Jiang, *Research on arrival/departure scheduling of flights on multirunways based on genetic algorithm*. Mathematical Problems in Engineering, 16(10), 1-13, 2014.

[4] J. Ding, J. Ji, H. Wang, *Optimization of airport flight arrival and departure based on compromise immune algorithm*. International Conference on Natural Computation, 526(3), 677-681, 2007.

[5] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, *A fast and elitist multiobjective genetic algorithm: NSGA-II*. IEEE Trans. Evol. Comput, 6(2), 182-197, 2002.

[6] L. L. Wang, P. Hu, *Arrival and departure scheduling optimization model and algorithm based on matching of traffic flow with capacity*. Nanjing University of Aeronautics an Astronautics, 47(6), 827-832, 2015.

[7] M. Liu, C. Y. Lee, Z. Z. Zhang, C. B. Chu *Bi-objective optimization for the container terminal integrated planning*. Transportation Research Part B, 93(2016), 720-749, 2016.

[8] Q. Q. Zhang, M. H. Hu, H. H. Zhang, *Dynamic multi-objective optimization model of arrival and departure flights on multiple runways based on RHC-GA*. Journal of Transportation Engineering, 11(2), 359-373, 2015.

[9] S. Chen, M. Hu, H. Zhang, *Research on collaborative decision making system based on conversion of arrival and departure capacity*. Journal of Systems & Management, 21(16), 1-8, 2008.

Production and Design Methods

An industrial perspective on Workload Control: economic evaluation and suitable technologies

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

BERTOLINI, Massimo, ROMAGNOLI, Giovanni, ZAMMORI, Francesco
Dept. of Engineering and Architecture
University of Parma
Parma, Italy
massimo.bertolini@unipr.it; giovanni.romagnoli@unipr.it; francesco.zammori@unipr.it

UCKELMANN, Dieter
Faculty of Geomatics, Computer Science and Mathematics
Hochschule für Technik
Stuttgart, Germany
dieter.uckelmann@hft-stuttgart.de

Abstract—Measuring Work-in-Progress and managing Workload control can be a time-consuming task. While numerous measures are already used, collecting the relevant data and comparing the different measures are a burden thus leading to only few effective monitoring systems. The usage of cyber-physical systems in the context of the fourth industrial revolution on the one hand and harmonised measures on the other promise a huge advantage for higher Work-in-Progress visibility and accountability. This can result in a more useful and efficient evaluation of Workload control systems and settings and, ultimately, in more frequent success stories of Workload control.

Keywords—workload control; work in progress; make to order; auto id; business metrics; performance measures; real time monitoring

I. Introduction

Workload control (WLC) is a production planning and control system that was designed for small and medium-sized enterprises that operate in a Make-to-Order (MTO) logic ([1]). In this sector, in fact, as [2] suggests, it is crucial to have a significant flexibility of production mix to meet high customer's demands and expectations. However, at the very same time, MTO companies must also struggle to control WIP, and thus optimise the trade-off between high throughput rates and short and stable throughput times ([3]).

Despite the concept of WLC has been, and still is, continuously investigated in the scientific literature, its industrial implementations, although increasing in the last years, are still scarce ([4]). This is probably due to several different issues. In 1989, [5] introduced the 'workload control paradox', stating that the implementation of WLC can reduce Work in Process (WIP) and improve queue balance, but this does not lead to an improved overall system performance. More recently, [6] suggests the following reasons for the limited number of WLC success stories:

- differences between theory and practice: common simulation models that can be found in the literature are usually oversimplified – they have been developed theoretically, and have not much in common with real production environments, and reality is frequently more complex than the computer simulated systems;
- literature experiences the lack of a general and comprehensive method to evaluate overall performances of specific production planning and control (PPC) systems.

The aim of this paper is to tackle the second bullet point. In detail, after an overview of relevant literature (Section II) this paper proposes an original approach for evaluating WLC performances from the economic point of view (Section III). As we will discuss, our approach aims at considering, at the same time, several different measures of performance that have been employed in previous studies. Moreover, by means of an economic method, we will propose how to calculate the expected values of all costs and revenues generated by a definite WLC configuration, such as a specific release method or certain values of the norms. By converting the outcomes of the simulated PPC system into costs and revenues, we propose a method to assess and compare the expected economic results of the implementation. Furthermore, to support practitioners and researchers in the pre-implementation and implementation phase, a review of suitable technologies and data processing techniques that can improve both the accuracy and the timeliness of shop-floor information is provided in Section IV. The goal of this section is to summarise available tools to identify objects, measure their quantity, and locate their position. Section V will comment and conclude the paper.

II. Workload control: a brief review of the literature

A. General description

WLC, as the name suggests, is a hybrid push/pull PPC system that aims at controlling WIP and at stabilising queues in front of work centres, to achieve both high throughput rates and stable lead times ([1]). It was proposed by [7] in 1981, to overcome the 'lead time syndrome' in MTO companies ([8]). In these organisations, in fact, often due dates and sales prices must be defined in response to a customer's request, and they are key elements to ensure the customer's order. Due to its characteristics, WLC is particularly suitable for job shop environments, where high levels of production mix flexibility are also required ([9]).

The basic functioning pattern of WLC is relatively simple: to set up a WIP cap, the system uses 3 levels of control, where jobs can be sorted and admitted to the following level:

- *Order entry*: at this level, new orders can be sorted and accepted by the company, and thus enter the system.

Usually, at order entry, both the economic and technological feasibility of the customer's orders are evaluated. Also, several studies in the literature suggest how to deal with due dates at order entry level ([10], [11]).

- *Order release*: one of the major achievement of WLC is the pre-shop pool (PSP), a virtual buffer where orders that have been accepted can be sorted and released to the shop floor. The instrument for releasing orders in the shop floor is called the 'triggering mechanism' ([12]). As [13] reports, the order release phase contains most of the intelligence of the WLC system, and it is a complex phase with several characteristics and processes. Several authors have proposed a wide variety of different order release methodologies, as we will summarise in Section II.B.
- *Priority dispatching*: when released to the shop floor, jobs follow their routing, from the first to the last work centre. As WLC often operates in job shops, a queue will most probably be in front of every work centre. Thus, jobs will be sorted in queues according to some dispatching rule. [14]

B. Order Release phase and WLC evaluation

The order release phase connects production planning and production control, as it links the strategy of a manufacturing company to its daily operations ([15], [16]). Thus, the order release phase is the true core of the WLC model: a good technique for selecting and releasing orders from the PSP to the shop floor is, in fact, the key to a successful WLC implementation ([17]). In their excellent review, [13] proposed a framework with 8 dimensions that can be used to categorise and describe the characteristics of order release procedures. Although we refer to [13] for a complete explanation, we briefly summarise the dimensions to be considered at the order release phase, which deal with releasing the order (i.e. order release mechanism and timing convention), measuring workload values (i.e. workload measure, aggregation of workload measure and workload accounting over time) and controlling workload (i.e. workload control, capacity planning and schedule visibility).

This description makes it clear that the order release phase is relatively complex. Unfortunately, however, despite this great quantity of studies on WLC ([1]), the number of success stories of this system has been increasing in the last years, but it still remains moderate ([17]). According to [6], a major limitation that often afflicts the assessment or comparison of performances of different PPCs in real-world contexts is the wide variety of diverse performance indicators. Researchers, in fact, are often tempted to report the measures that better confirm their hypotheses, often arguing that those performance measures are the most pertinent in the industrial case under analysis. Reference [18] categorises the performance measures used in previous implementation papers into six different categories: time related; dependability; cost-related; shop load-related; market-related and internal co-ordination.

TABLE I. Performance measures used in recent studies

Category	Measures	Studies using this measure
Time-related	Lead time	[17]
	Shop floor throughput time	[19]
	Supply chain time	[20]
Dependability	Lateness	[12]
	Service level	[21]
	Percentage tardy jobs	[22]
	Tardiness	[23]
Shop load-related	WIP level	[8]
	Utilisation rate	[24]
	Parts in queue	[25]
Market-related	Accepted orders	[26]
Cost-related	Inventory level	[27]

Consequently, [18] suggests that 'it is important to ask whether the same set of measures should be used to assess WLC implementation in practice in order to determine appropriate measures for [their] project'. Despite this suggestion, literature in the last 4 years has continued to use a great number of diverse and, sometimes, contrasting measures to evaluate performances of PPC implementation. Table I provides a brief overview of the metrics used in the last 4 years for evaluating performances of PPC systems. We notice that, out of the six categories identified by [18], no attention has been paid to internal coordination (which indeed was only reported once in a 1999 study). The other categories are all present in the table, with a considerable number of studies that focused on dependability, time-related and shop load-related indicators. Metrics like percentage of tardy jobs, lateness, tardiness, shop floor throughput time, gross throughput time (which also consider pre-shop pool waiting time), utilisation rates, and WIP levels were employed in most of the WLC-related studies.

Within the three above-mentioned categories, some new metrics have emerged in the last years. Supply chain time has been introduced by [28], to estimate the processing times of an MTO supply chain with high variation of production mix. Similarly, [21] introduces the service level metric, to measure the percentage of non-tardy jobs. Table I also shows that little attention has been paid to cost-related and market-related metrics. Only a few studies, in fact, have reported performance measures that deal with those two categories: [27] measures the level of inventory into a MTO environment, both in terms of safety stocks and lot sizes. Also, [26] evaluates the volumes of accepted orders to compare and evaluate different PPC settings.

C. Identified criticalities in WLC research

Although our survey of recent literature on the assessment and comparison of different PPC systems is not thorough, we can summarise it in the following points:

- a wide variety of different metrics has been used to evaluate performances of different PPC systems;
- the majority of studies makes use of measures belonging to the following categories: dependability, time-related and shop load-related;
- cost-related and market-related categories are less spread, and measures belonging to these categories are more common in studies that analyse and extend standard push PPC systems (MRP, MRP2, ERP) to the MTO sector;
- several measures are conflicting (e.g. short lead time and high service level, or high volume of accepted orders and small percentages of tardy jobs); thus, when trying to optimise conflicting measures, some form of compromise must be made between those measures.

Due to these issues, and especially to the last bullet point, it is hard to define benchmarks, and the need for standardised measures is evident. Although the literature is aware that the measures we reported ought to be organised in a consistent and comprehensive way, little has been done towards this goal. Looking at the different evaluation criteria, the combining element is their economical business implication – therefore, as the industrial practice suggests, the easiest (and maybe the only) way of considering different and conflicting performance measures at the same time is to convert them all to the same unit of measurement: money.

Recently, a branch of research has tackled the so called problem of 'production planning in manufacturing', which is described by [29] as the 'problem of finding a release plan for jobs that minimizes the total cost (or maximizes the total profit)'. The problem is challenging, due to several sources of variability, such as customers' demand, jobs routings, processing times, machine breakdowns, and many others. The same problem is also addressed by [30], and these studies consider three different types of cost: (i) the holding cost of WIP, (ii) the finished-good holding cost, and (iii) the backlog cost.

While we praise the effort of [29] and [30], we also notice that: (i) to the best of our knowledge, no similar approach has been adopted in the WLC literature, and (ii) the cost structure proposed by those studies is not complete, as it does not consider the important contributions of production costs and sales prices, and thus the huge difference between accepted and non-accepted customer's orders. To consider all the categories listed in Table I, we will propose in the next section a general method for evaluating WLC based upon economic considerations.

III. A new proposal for an economic evaluation of WLC systems

Our proposal is to evaluate performances of different WLC systems, as well as optimising WLC norms by considering the following costs and revenues into the simulation model. The costs and revenues listed in the following subsections are traditional metrics in inventory management literature (see for example [31], [32]). Note that all costs and revenues are strictly positive, apart from III.C and III.D, that are non-negative, and also alternative (if $D > 0$, then $C = 0$; and the other way around). Also, during simulations of the WLC systems, these costs can be calculated in real-time, or approximated accordingly.

A. Production cost

Production cost = Cost of materials + Cost of processing

The cost of production considers all the costs that are borne for manufacturing a customer's order. At minimum, these costs must comprehend materials and manufacturing times on all workstations that process a job. Note that, during WLC simulations, this cost can be calculated in real time. This cost accounts (mostly) for the time-related category of measures in Table II. We note also that the production cost is very important in a MTO company, as both materials and production costs may vary widely from one customer's order to the next.

B. WIP holding cost

*WIP holding cost = Time in the system as WIP * (Cost of materials + Total cost of processing / 2) * $K_{holding}$*

This category of cost considers the cost met to hold WIP inside the production system, and thus it multiplies the time in the system as WIP (i.e. the time spent between raw materials and finished goods inventories) by the average production cost. The average production cost can be approximated as linear; thus, it can be calculated dividing by 2 the sum of the cost of WIP at the beginning of its routing (i.e. cost of materials) and the cost of WIP at the end of its routing (i.e. cost of materials + cost of processing on all work centres). Eventually, this value is multiplied by $K_{holding}$, a constant value that considers the percentage of average production cost to be considered as holding cost per unit of time in the system. WIP holding cost is mainly related to the shop-load category of measures in Table II.

C. Finished good holding cost

*Finished good holding cost = Time in the system as Finished Good * (Cost of materials + Total cost of processing) * $K_{holding}$*

Similarly, the finished good holding cost considers the time in the system as finished goods, multiplied by the total production cost (i.e. cost of materials + cost of processing on all work centres), and again multiplied by $K_{holding}$. We assumed the same $K_{holding}$ for III.B and III.C, but it goes without saying that this is not necessary. Finished good holding costs are linked to the dependability measures of Table II, if jobs are early.

D. Penalty cost

*Penalty cost = Time of delay * Sale price * $K_{penalty}$*

A non-negligible source of costs for MTO companies is the penalty. Whenever customer's orders are shipped late, some kind of cost must be considered, to cover up for reduced sales prices, potential loss of future orders, as well as other indirect costs. We emphasise that it is very hard to exactly calculate this cost (see literature on Out-Of-Stock for some suggestion on how the problem is dealt in the Make-To-Stock sector). However, not considering it would lead to an inaccurate cost structure. Thus, our suggestion is to calculate the penalty cost by multiplying the time of delay (evaluated from the due date agreed upon with the customer) by the sale price (see III.E) and by $K_{penalty}$. $K_{penalty}$ is a constant value that considers the percentage of sale price to be considered as penalty cost per time units of delay. It could be wise to limit the penalty cost to a maximum value (i.e. some

percentage of the sales price). We also remind that costs III.C and III.D are alternative, that means either one or the other must be equal to 0 per every order (which, at least into a simulation model, cannot be early and late at the same time). Like the previous point, penalty costs are also related to the dependability measures of Table II, if jobs are late.

E. Sales price

*Sales price = Production cost * (1 + Mark-up)*

Neglecting sales prices into a MTO company could be extremely misleading if production rates of two different systems are not exactly the same. To avoid this risk, we suggest considering them by multiplying the production cost by 1 plus a given mark-up. The mark-up is the percentage of production cost that covers indirect costs and, eventually, may generate profit. We are well aware that this is a simplification, as mark-up values may vary significantly, but not considering sales prices would mean neglecting market-related measures, as reported in Table II.

TABLE II. Economic measures proposed in this study

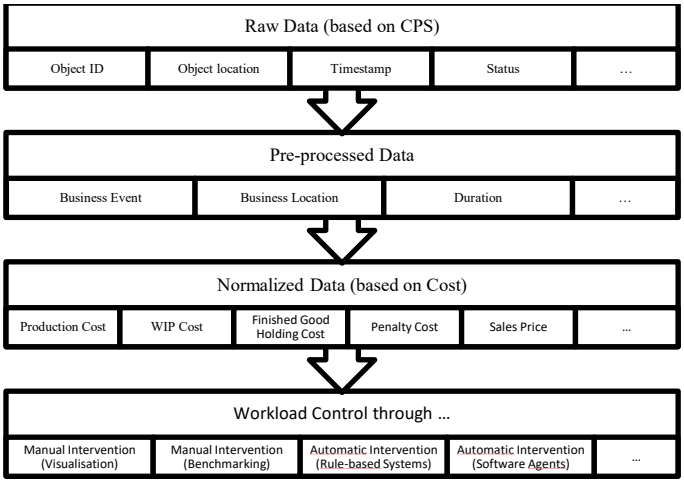
Category	Related economic measure
Time-related	Production cost
Dependability	Finished good holding cost
	Penalty cost
Shop load-related	WIP holding cost
Market-related	Sales price

As it is summarised in Table II, 4 out of the 5 categories reported in Table I are related to an economic measure that we propose, be it a specific cost (e.g. production cost, WIP holding cost, penalty cost etc.) or the sales price. Indeed, we did not report the cost-related category of Table I in Table II, as this category is the common denominator for comparing all other measures on a common ground. In conclusion, we believe that, by considering all the costs and revenues reported in this section, simulations and optimisations of WLC systems, as well as of others hybrid push/pull PPCs, could really take into account many of the distinct and conflicting measures summarised in Table I.

IV. Suitable technologies and data processing

In traditional WLC systems, periodical release methods have been used. However, for a detailed economic evaluation event-based release and control methods are necessary. Modern manufacturing shop-floors require real-time information concerning raw material availability, WIP, and finished goods. Paper-based systems, which are still most dominant in XtO-companies, are prone to errors and slow, as the data reporting and transfer to the IT-systems is usually done hours or days after the actual incidents occur. Therefore, new technologies are investigated to improve accuracy and timeliness of shop-floor information. In the context of Industry 4.0 these technologies (including Auto-ID, sensors, actuators and microcontrollers) and their connection to software-platforms are often referred to as cyber-physical systems (CPS). The main targets are to identify objects, measure their quantity (if no serial number is provided) and locate their position. Traditionally barcodes and 2D-codes are used to track products. The automotive and electronic industries have increased their usage of 2D-codes to track printed circuit boards (PCB) and engine components during assembly in the past years. These technologies however are more efficient in fully automated production scenarios with fixed 2D-cameras and controlled lighting. In manual identification processes, changing illumination may lead to read problems. RFID offer some advantages over optical codes such as no need for a line-of-sight, bulk-reading capability and increased robustness (depending on the housing). In closed-loop scenarios, where the reuse of RFID-tags is possible, RFID has been used for decades. Besides passive RFID, active (battery-powered) RFID-tags or locating systems may be used. Numerous solutions are available in the market [33]. As these solutions are costlier, the number of economically feasible applications is limited, though. In recent years, passive RFID has gained more and more importance, since the technology has matured and reduced prices for RFID-tags and systems have enabled more applications including open-loop scenarios. RFID systems, for example, have enabled WIP-tracking within automotive manufacturing and computer hardware industries [34]. Even the term wireless manufacturing (WM) has been coined to show the importance of RFID and wireless systems in productions scenarios [35]. RFID-tags have even been embedded in products such as PCB [36], and die-cast aluminium housings [37]. RFID-tags can be used to identify material (e.g. in boxes), parts and items. More importantly, it can be utilized for event-based production control [38]. During the identification process time stamps, business locations (e.g. warehouse entry, work station, shelf number), and business steps (e.g. receiving) may be associated and recorded in corresponding databases. The wording for locations may vary, though. The GS1 Core Business Vocabulary Standard [39] provides a guideline for a harmonized usage of locations and business steps. For WIP-analysis an extended vocabulary would be needed, though.

Fig. 1. Workload control based on CPS and financially normalised status tracking



The collected data needs to be processed, analysed and evaluated. The suggested financial normalisation will allow an easy understandable system to evaluate the current WIP-status. WC can be achieved through manual intervention (e.g. based on visualisation, benchmarking) or automatically through rule-based systems or software agents (Figure I). For visualization of the WIP-status, management dashboards applications (e.g. shown on large screens) as well as visual mobile tags may be used. Visual electronic tags (e.g. from Omni-ID, SES-Imagotag) have first been used for shelf labelling. However, they may be connected directly to the object of interest or the corresponding transport unit as well. Visual tags and management dashboards can be used to keep the workers informed about the actual WIP based on planned vs. actual financial data.

V. Discussion and conclusions

Although researchers from different countries have been paying attention to the WLC concept for more than three decades, the number of successful implementations of WLC are still not frequent. As several studies suggest, this is due to several causes. First, a generalised and comprehensive measure has been missing for evaluating and comparing on a common ground performances of different WLC systems. As the literature review has shown, many diverse and conflicting measures are used today. It is very hard, however, for non-experts of the field to understand the pros and cons of different WLC systems or settings, due to the profusion of metrics summarised in Section II. In this paper, we suggested an economic function that considers both costs and revenues that can be generated by a WLC implementation, to summarise into a single measure the contributions of different categories of performance indicators, such as time-related, dependability, shop load-related and market related measures. Moreover, we traced a general outline of available tools and technologies to collect data on WIP based on Auto-ID technologies, and use financial measures for evaluation as well as manual or automatic intervention during production.

A major benefit of using one simple financial measure is an easy internal and external benchmarking possibility. However, the shown model needs to be discussed among peers (e.g. at scientific conferences), verified technically through a prototypic installation and economically through a real-world test scenario. Indeed, the authors are already working on the evaluation of different WLC systems and settings by means of the proposed economic function, and our results will be submitted to relevant journals as soon as they are complete.

References

[1] M. Thürer, M. Stevenson, and C. Silva, “Three decades of workload control research: a systematic review of the literature,” Int. J. Prod. Res., vol. 49, no. 23, pp. 6905–6935, 2011.

[2] M. Stevenson, L. C. Hendry, and B. G. Kingsman, “A review of production planning and control: the applicability of key concepts to the make-to-order industry,” Int. J. Prod. Res., vol. 43, no. 5, pp. 869–898, 2005.

[3] G. Marangoni, G. Romagnoli, and F. Zammori, “Multiple Value Stream Mapping : how to implement Work Load Control in complex systems,” in IESM’2013, International Conference on Industrial Engineering and Systems Management, 2013.

[4] M. Stevenson, Y. Huang, L. C. Hendry, and E. Soepenberg, “The Theory & Practice of Workload Control: A Research Agenda & Implementation Strategy,” Int. J. Prod. Econ., vol. 131, no. 2, pp. 689–700, Jun. 2010.

[5] S. Melnyk and G. L. Ragatz, “Order review/release: research issues and perspectives,” Int. J. Prod. Res., vol. 27, no. 7, pp. 1081–1096, 1989.

[6] M. Bertolini, G. Romagnoli, and F. Zammori, “Simulation of two hybrid production planning and control systems: a comparative analysis,” in Proceedings of 2015 International Conference on Industrial Engineering and Systems Management, IEEE IESM 2015, 2016, pp. 388–397.

[7] J. W. M. Bertrand, “The effect of Workload Control on order flow times,” in Operational Research ’81, J. P. Brans, Ed. Amsterdam, The Netherlands: North Holland Publishing Company, 1981, pp. 779–790.

[8] M. Bertolini, G. Romagnoli, and F. Zammori, “Assessing performance of Workload Control in High Variety Low Volumes MTO job shops: A simulative analysis,” in Proceedings of 2015 International Conference on Industrial Engineering and Systems Management, IEEE IESM 2015, 2016, pp. 362–370.

[9] M. Stevenson and C. Silva, “Theoretical development of a workload control methodology: evidence from two case studies,” Int. J. Prod. Res., vol. 46, no. 11, pp. 3107–3131, Jun. 2008.

[10] M. R. A. Moreira and R. A. F. S. Alves, “A methodology for planning and controlling workload in a job-shop: a four-way decision-making problem,” Int. J. Prod. Res., vol. 47, no. 10, pp. 2805–2821, May 2009.

[11] M. Thürer, M. J. Land, M. Stevenson, and L. D. Fredendall, “On the integration of due date setting and order release control,” Prod. Plan. Control, pp. 1–11, Mar. 2017.

[12] M. Thürer, M. Stevenson, C. Silva, M. J. Land, and M. G. Filho, “Workload control and order release in two-level multi- stage job shops: an assessment by simulation,” Int. J. Prod. Res., vol. 51, no. 3, pp. 869–882, 2013.

[13] D. Bergamaschi, R. Cigolini, M. Perona, and A. Portioli, “Order review and release strategies in a job shop environment: A review and a classification,” Int. J. Prod. Res., vol. 35, no. 2, pp. 399–420, 1997.

[14] H. L. Lu, G. Q. Huang, and H. D. Yang, “Integrating order review/release and dispatching rules for assembly job shop scheduling using a simulation approach,” Int. J. Prod. Res., vol. 49, no. 3, pp. 647–669, Feb. 2011.

[15] G. Zäpfel and H. Missbauer, “New concepts for production planning and control,” Eur. J. Oper. Res., vol. 67, no. 3, pp. 297–320, 1993.

[16] M. K. Malhotra, J. B. Jensen, and P. R. Philipoom, “Management of vital customer priorities in job shop manufacturing environments,” Decis. Sci., vol. 25, no. 5–6, pp. 711–736, 1994.

[17] M. Bertolini et al., “2MTO , a new mapping tool to achieve lean benefits in high-variety low-volume job shops shops,” Prod. Plan. Control, vol. In press, 2017.

[18] L. Hendry, Y. Huang, and M. Stevenson, “Workload control: Successful implementation taking a contingency-based view of production planning & control,” Int. J. Oper. Prod. Manag., vol. 33, no. 1, pp. 69–103, 2013.

[19] M. J. Land, M. Stevenson, and M. Thürer, “Integrating load-based order release and priority dispatching,” Int. J. Prod. Res., vol. 52, no. 4, pp. 1059–1073, Feb. 2014.

[20] K. Kundu and A. P. Staudacher, “Investigation of Effectiveness of Order Review and Release Models in Make to Order Supply Chain,” in MATEC Web of Conferences, 2016, vol. 81, no. 6003, pp. 1–5.

[21] M. Gansterer, “Aggregate planning and forecasting in make-to-order production systems,” Int. J. Prod. Econ., vol. 170, pp. 521–528, 2015.

[22] M. Thürer, M. J. Land, and M. Stevenson, “Card-based workload control for job shops: Improving COBACABANA,” Int. J. Prod. Econ., vol. 147, no. A, pp. 180–188, 2014.

[23] M. Thürer and M. Stevenson, “Workload control in job shops with re-entrant flows: an assessment by simulation,” Int. J. Prod. Res., vol. 54, no. 17, pp. 5136–5150, Sep. 2016.

[24] M. Thürer, M. Stevenson, C. Silva, M. J. Land, L. D. Fredendall, and S. A. Melnyk, “Lean control for make-to-order companies: Integrating customer enquiry management and order release,” Prod. Oper. Manag., vol. 23, no. 3, pp. 463–476, 2014.

[25] P. Renna, “Allocation improvement policies to reduce process time based on workload evaluation in job shop manufacturing systems,” Int. J. Ind. Eng. Comput., vol. 8, pp. 373–384, 2017.

[26] L. Dumetz, J. Gaudreault, A. Thomas, P. Marier, N. Lehoux, and H. B. El-Haouzi, “A simulation framework for the evaluation of production planning and order management strategies in the sawmilling industry,” IFAC-PapersOnLine, vol. 28, no. 3, pp. 622–627, 2015.

[27] M. Gansterer, C. Almeder, and R. F. Hartl, “Simulation-based optimization methods for setting production planning parameters,” Int. J. Prod. Econ., vol. 151, pp. 206–213, 2014.

[28] K. Kundu and A. P. Staudacher, “Investigation of Effectiveness of Order Review and Release Models in Make to Order Supply Chain,” in MATEC Web of Conferences, 2016, vol. 81, no. 6003, pp. 1–5.

[29] J. Liu, C. Li, F. Yang, H. Wan, and R. Uzsoy, “Production planning for semiconductor manufacturing via simulation optimization,” Proc. 2011 Winter Simul. Conf., pp. 3617–3627, 2011.

[30] M. Li, “A Metamodeling-Based Approach for Production Planning,” Proc. 2014 Winter Simul. Conf., pp. 2204–2215, 2014.

[31] W. J. Hopp and M. L. Spearman, Factory Physics, 3rd Editio. Dubuque, Iowa: McGraw-Hill/Irwin, 2008.

[32] W. J. Stevenson, Operations management, 10th Editi. Milan, Italy.: McGraw-Hill/Irwin, 2009.

[33] D. Uckelmann and G. Romagnoli, “RF-based Locating of Mobile Objects,” in 6th International Conference on the Internet of Things (IoT‘16), 2016, pp. 147–154.

[34] A. E. Coronado Mondragon, A. C. Lyons, Z. Michaelides, and D. F. Kehoe, “Automotive supply chain models and technologies: a review of some latest developments,” J. Enterp. Inf. Manag., vol. 19, no. 5, pp. 551–562, Sep. 2006.

[35] G. Q. Huang, Y. Zhang, and P. Jiang, “RFID-based wireless manufacturing for real-time management of job shop WIP inventories,” Int. J. Adv. Manuf. Technol., vol. 36, no. 7–8, pp. 752–764, Mar. 2008.

[36] Murata, RFit Technologies, Schneider Electric, and Sophia-Conseil, “A single and easy traceability solution - From product data to ERP,” in Proceedings of 2010 RFID Journal Live Europe, 2010.

[37] C. Pille, “In-Process Embedding of Piezo Sensors and RFID Transponders into Cast Parts for Autonomous Manufacturing Logistics,” in Proceedings of 2010 Smart Systems Integration (SSI), 2010, pp. 1–10.

[38] P. Engelhardt, System für die RFID-gestützte situationsbasierte Produktionssteuerung in der auftragsbezogenen Fertigung und Montage. Herbert Utz Verlag, 2015.

[39] GS1, “Core Business Vocabulary Standard,” EPCIS and Core Business Vocabulary (CBV), 2017. [Online]. Available: http://www.gs1.org/epcis. [Accessed: 15-Apr-2017].

Methodological problems in defining costs used in industrial management decision models

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Vincent Giard[*, **]

[*]*Emeritus Professor, Paris-Dauphine, PSL Research University, CNRS, 75016, Paris, France (vincent.giard@dauphine.fr)).*
[**]*Affiliate Professor, EMINES, École de Industrial Management, Université Mohammed VI Polytechnique, 43140 Ben Guerir, Maroc (vincent.giard@emines.um6p.ma)*

Abstract—Articles dealing with industrial management decision-making generally rely on a cost system to establish a global indicator and identify the best solution. Cost accounting is based on a number of assumptions as to production system operation that may be quite remote from those used in decision modelling. Our paper points out the origin of this inconsistency and that it may lead to irrelevant decisions.

Keywords— model costing, cost management, supply chain management, operation research.

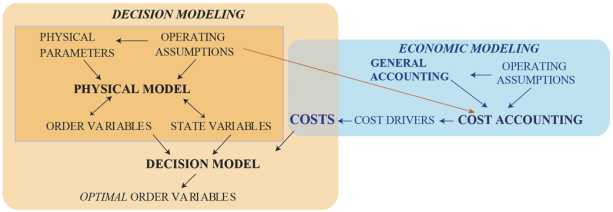
1. INTRODUCTION

An analysis of the impact of operational, tactical and strategic decisions always relies on different physical models of a production system of products (generic term that we use in this paper to refer to both goods and services), whose operation is affected by the decisions to be taken. Such decisions are described by quantitative variables (e.g. quantity to be ordered…) or qualitative variables (transportation routes, alternative investments …) that certain models translate into binary variables. These are called *order variables* and they impact production system operation through a series of more or less complex causal relations that lie at the heart of the chosen model. These interactions determine the value of certain physical parameters (level of inventory, travelled mileage …) chosen for their impact on production system performance. In the more straightforward models that involve a small number of physical parameters, these parameters are often called *state variables*; we use this term in our paper, regardless of model complexity.

The effectiveness criteria (such as minimum service level required…) used in the models are often more stringent in non-profit production systems and they narrow the scope of possible decisions. To choose the most efficient solution, one must refer to an economic criterion because no single solution ever consumes fewer of each production system resource, therefore standing out as the best solution. In certain simple problems, the criterion is physical and the economic objective is implicit: for example, where cost is a linear function of time or distance, one may reason indifferently in terms of a physical indicator or associated cost. Generally, however, economic evaluation relies on a valorization system that stems from cost accounting to calculate a global value indicator being the weighted sum of state variables (and, in certain decision-making issues, of binary order variables). Such weighting approach is a cost system implicitly based on a model of the production system's operations. Such modelling may be quite remote from that used to analyze the physical consequences of the decisions to be taken. This discrepancy may ultimately disqualify the chosen solution as irrelevant.

Let us turn to the foundations of these two types of model to describe physical production systems so as to be able to pinpoint potential sources of inconsistency. To avoid any risk of semantic confusion, we shall establish the distinction between the two types of model by calling them ***decision models*** and ***economic models*** respectively (the latter model underlying cost accounting). Figure 1 shows the relationship between these two approaches

Fig. 1. Relationships between decision modelling and economic modelling.



Our purpose is to show possible inconsistencies between the physical assumptions of these two models. These considerations cannot serve as a basis for a general methodology of cost accounting for decision-modelling purposes; they must be viewed as methodological issues to be kept in mind when defining costs in decision models. After examining physical representations of decision models and economic models (Section 2), we give some illustrations of the implication of those methodological considerations (section 3).

2. PHYSICAL REPRESENTATION CHARACTERISTICS OF DECISION MODELLING VS ECONOMIC MODELLING

2.1 Underlying process models used in Decision Modelling

Decision modelling is chiefly characterized by a mix of three features: the degree of certainty of information used, the degree of reality of the system under review and its spatio-temporal granularity.

- Either the features of the production system and of its environment are considered to be *known with certainty* and the decisional issue is posed within a deterministic context, or

they are not. Most of the time, where they are not, these features are deemed to be *known as random variables*. Such characterization has an impact on the choice of technical model. In a random context and in straightforward fictitious systems, analytical models use variable expectations; where they are complex, they rely on the Monte Carlo approach and involve recourse to spreadsheets or process simulation softwares. In a deterministic context, mathematical programming is quite relevant to the formalization of problems featuring simple causal relations, but a spreadsheet approach may be necessary where causal relations are complex.

- Production systems are either *actual* or *fictitious*. The literature on operational research prefers relatively simple fictitious systems to work out the analytical relation that describes the optimal solution for a type of decision-making issue (e.g. procurement management). But fictitious systems can also be complex and use parametric dimensioning to produce a *generic model* for the problem (for example determination of a delivery route) via mathematical programming or graph theory. In certain cases they take the form of an algorithm of specific resolution. A number of research papers, prepared within the framework of industrial contracts, look into decision-making issues in connection with actual production systems. In this case, the model may lead to *i*) the assessment of decision-making scenarios (what-if analysis), *ii*) the formulation of an optimization problem where the level of complexity is low, *iii*) the formulation of a simulation model or *iv*) a decision support system (DSS) combining a scenario rationale with local optimizations and/or simulation. The proposed approach is often sufficiently generic to be transposable to similar production systems in similar decision-making issues.

- The *granularity* used in the model, in its *spatial and temporal* dimensions, is a key element of decision model characterization. Time has multiple dimensions: mono or multi-period model, length of the period and horizon used in multi-period models. The granularity also involves spatial characterization, with varying levels of detail (or of aggregation) of resources and products. In general, the granularity is fine in operational decision analysis and more aggregated in other cases, except that many strategic decisions imply reference to a detailed model to ensure the robustness of a solution derived from an aggregated model.

2.2 Implicit, underlying, process modelling in Economic Modelling

General accounting involves matching physical and financial flows observed in a business so as to be able to perform periodic assessments of the results of its activity and the value of its assets. Such evaluation must comply with a number of rules as to the booking of certain expenses impacting corporate asset value (Balance Sheet) for the reference period (accounting year). This information is intended for third parties (tax authorities, customers, suppliers, shareholders…) as well as for internal purposes (free cash flow, debt, receivables). This accounting treatment complies with a set of rules to defer investment expenses (depreciation expenses) and allocate to a period the financial flows from products purchased or sold during another period. Expenses are classified by nature e.g. raw materials, and by function e.g. sales, overheads…).

Cost accounting in the widest sense (several approaches are possible) processes the data generated by General Accounting to measure an entity's (department, plant…) operating expenses, to measure the cost for an in-house service (e.g. dispatch of a container) or to measure the cost of a product. The computation of these costs poses two major methodological issues. While we focus on product cost in this paper we note that the same issues arise for both other types of cost.

- The first methodological issue is that of *indirect cost allocation*. These relate to services or inputs shared by several products (workshop energy expenses where there is only one meter, overheads…) and they may be shared (or not) between products.

- The second methodological issue is the *impact of used capacity* on the calculation of both direct and indirect costs (depreciation expenses of a machine dedicated to a type of production). A possible starting point is the actual level of activity – this making it difficult to compare in time and space – or reference to a "standard" level of activity. The latter solution, which is generally preferred, facilitates comparison in time (and space) but raises two issues: there can be no objective definition of a "standard" use of capacity and this could be a source of dispute, or even fraud. The gap between actual use of capacity and that deemed to be standard inevitably poses the problem of consistency between General Accounting and Cost Accounting, as the sum of costs in the latter could be quite different from that booked in accounting expenses.

Several solutions are available to address these issues and enable the calculation of partial costs and full costs. This does not lie within the scope of this paper whose aim is to show *i*) that the costs used in a valorization system stem from a model of production system workflow that applies cost allocation rules over different periods of time, and *ii*) that used capacity (level of activity) assumptions and indirect cost allocation assumptions rely on causal relations that are debatable.

These costs are mainly computed for management control and should not be used indiscriminately for industrial management decision-making purposes. In this regard, some additional considerations are relevant to compute costs for decision-making purposes. The guideline is that the valorization system must provide that the benefits of the proposed decision versus a reference decision (that may be to do nothing) should be reflected in the organization's financial results. To this end, the considerations set forth below should enter into the definition of the cost system used to assess industrial management decisions.

- The distinction between product *direct* and *indirect cost*, referred to above, depends on the level of aggregation used: an expense at a particular site may be indirect at the product level and direct at the product range level. Moreover, one is often led to consider entire product ranges as they all

consume the same resources. The calculation of direct costs for a fictitious product representing an entire family has to be the average weighted cost of the direct costs of the items comprising the family, where the weighting coefficients correspond to an average structure which cannot be deemed to be stable over time.

- Over the last three decades, *activity based costing* (ABC) stood out as a basic approach in cost accounting because it serves to reduce the arbitrary treatment of indirect costs. Its aim is to replace the direct causal relation "products consume resources" by the indirect causal relation "products consume activities that consume resources". This approach encompasses the other cost accounting approaches as these may be deducted from it, which is why we adopt it in this paper. The concept of activity is wide; it can include production of an item as well as the launching of a production order for a set of products on a machine. This approach centered on *cost drivers* produces more relevant definitions of costs for decision modelling purposes.

- Economists have rapidly introduced the distinction between *fixed* and *variable costs*, fixed costs being independent from production volume. This concept, that can be extended to cost drivers, is crucial to decisional analysis, provided one explicitly defines the time horizon used for the decision-making exercise: where one looks at a few days' horizon, previously hired utility workers are allocated to fixed costs (direct in the case of a production worker on a machine dedicated to a given product and indirect in the case of a maintenance worker) but, if the hiring decision has yet to be taken (for example to cope with a hike in activity over the relevant horizon), then this cost becomes variable and must be related to the production increase it would enable.

- Cost accounting focuses on the determination of product costs (in the widest sense, as already noted). A decision-making issue is formulated more in terms of *costs stemming from a decision versus a reference decision* (defined by the values used for the order variables), which is not always clearly set forth. The optimal analytical solution of a fictitious system (e. g. procurement rules) are not pitched against a reference solution. Use of such a reference for an actual case to optimize a recurring decision involves using a cost system whose function is not just to be able to calculate the order variables but also to measure the gain achieved when applying the optimal solution, versus the status quo. This therefore should be reflected in the income statement, failing which the optimization is pointless. In the context of real-life systems, a decisional alternative (covering a set of order variables) will be assessed in relation to a benchmark decision/solution that may actually consist in taking no action (e. g. plant extension) or to continue with what one already has, (e.g. replacement of a machine by a similar one). Such *comparative economic analysis* is called Incremental Cash Flow Report (ICFR) and serves to eliminate all expenses that are constant regardless of the alternative decision under review. Some decisions impact income. We shall leave these aside since they are obviously included in an ICFR.

- Let us add that the alternative decisions under review may impact several accounting periods, which calls for use of *discounting* to summarize the changes in cash flows calculated in the IFCR to compute Net Present Values.

3. EXAMPLES OF CONSEQUENCES OF THE ABOVE METHODOLOGICAL CONSIDERATIONS

A number of costs drawn from cost accounting are useful for decision-modelling purposes provided the representation of the production system workflow is similar to that of the economic model used to determine these costs. Otherwise one faces an inconsistency which is generally overlooked in the literature. Since it is impossible to perform a thorough review of the implications of the above methodological considerations, we have decided to illustrate our discussion through a few typical cases. We shall first look at issues in connection with use of analytical solutions or of generic formulations that relate to operational or tactical decisions. We will then turn to methodological issues that arise in connection with decision models for strategic decision-making.

3.1 Issues raised by use of analytical solutions and by generic formulations of decision-making problems

These issues arise somewhat differently depending on whether one seeks to make concrete use of analytical relations from decision-making models based on fictitious systems, or one seeks to design a generic model.

Use of analytical solutions

We will illustrate our discussion with procurement management where *ordering costs*, *holding cost* and *stock out* costs are typically considered. The former two costs pose similar methodological issues. The latter one raises specific issues. The models referred to here are described in several fairly exhaustive handbooks (for example, Giard, 2003; Silver *et al*., 1998).

Unit ordering cost and unit holding costs

In the deterministic version of the basic model (Wilson's model), the order variables are the ordered quantity and the reorder point and the state variables are the average yearly number of orders and average number of products in stock. In the stochastic version, the averages are replaced by *expectation* and an additional state variable for "expected stock out" has to be added. The optimal analytical solution involves the unit costs associated to the state variable that corresponds to cost drivers. For ICFR purposes, the reference solution is the current solution regardless of whether the product exists or not. The application of these analytical relations to the resolution of concrete issues means that one has to specify correctly the contents of the unit costs to make sure the theoretical gains of switching from the current to the optimized solution show up in the income statement.

Take the example of a centralized order management desk. The new orders are a cost driver. The cost comprises a direct variable cost (stationary, postal expenses…) and a share of the yearly cost K of running this desk which is an indirect cost. Where the desk handles n orders per year (to simplify, let's assume that this is the standard level of activity), the standard cost of handling an order is K/n , which yields a unit

ordering cost of $c+K/n$ and a partial annual ordering cost of $n(c+K/n) = n \cdot c+K$. Where the optimal procurement policy enables a 20 % reduction in the number of orders, the new annual partial cost of orders is equal to $0.8 \cdot n(c+K/n)$ provided that one is able to rapidly reallocate 20 % of resources to the origin of indirect cost K. Where this is not the case, the partial annual ordering cost cut shrinks to $0.2 \cdot n \cdot c$ and the unit ordering cost of this new solution is a $c+K/(n \cdot 0.8)$, a sub-optimal value that calls for choice of a different solution. Where instead of a reduction in the yearly number of orders, the switch to the optimal solution leads to an increase in this number, then the issue is that of order management desk under-staffing and need to cope with the workload increase.

The average number of inventory units is a cost driver that takes two forms, generating two unit cost components of the *holding cost*.

- For a business, an increase in the value of inventory either translates into an increase in working capital requirement or, where bank funding is not available for these assets, into a loss of short-term investment opportunities. This cost driver generates a variable direct cost c' corresponding to an *opportunity cost* proportionate to the value of the product inventory and to the interest rate.

- Inventory also involves yearly cost K' (rental of warehouse, insurance, warden, energy, etc.) corresponding to an indirect cost that again, may be standardized on a volume basis. The definition of volume capacity Θ for the warehouse is tricky as it depends on the standard level of activity and on the degree of offsetting at all times between the high and low levels of inventory for each warehoused items. Where the current solution leads to keep volume θ in the warehouse for the considered product, the holding cost must include an additional *storage cost* $K' \cdot \theta/\Theta$, which leads to a unit holding cost of $c'+K' \cdot \theta/\Theta$. The change in average used capacity resulting from switch to the optimal solution should induce a drop in warehouse nominal capacity in case of a drop in inventory level and therefore induce an inventory unit cost increase (otherwise the sum of contributions from the warehoused items will not cover the annual cost K'). In the opposite case it should induce an increase in warehouse volume capacity Θ, also leading to an increase in storage unit cost.

As above, the change in the "inventory average" state variable induced by the transition from the current solution to the optimal one triggers an immediate change in direct variable cost but raises a similar issue to that encountered with standardization of the indirect cost integrated into the ordering cost.

These issues may lead one to turn to procurement models encompassing a range of product references, subject to capacity constraints (maximum number of orders handled by the desk, maximum aggregate average inventory…). Such approach, workable both in deterministic and random universes, leads to analytical solutions where the Lagrange multiplier related to constraints is interpreted as the marginal cost of change for such constraints.

The decisional hierarchy is such that the sizing of certain resources that are order variables at strategic decision level become a constraint at tactical or operational decision levels. At each level, the physical representations and the cost drivers differ, as illustrated in figure 2.

Unit stockout cost

In random universe models, inventory stockouts are plotted in two ways. We may retain a model integrating only ordering and holding costs but by adding constraints at inventory stockout level. This leads to complex, though operable, analytical formulations. The alternative is to integrate stockout cost in the cost function to be minimized. The issue is different according to whether the customer is corporate or individual and whether demand is lost or deferred. The cost of deferred demand generally corresponds to a standard administrative cost, calculated as a share of indirect fixed costs, which raises the methodological issues noted above. In the case of industrial customers, the stock-out may cause production line stoppage with extremely severe financial consequences that are all the harder to measure as they depend on the duration of stockout. An estimation of the distribution of probability for such duration, used in certain analytical models, leads to analytical solutions that are difficult to implement, are highly hazardous, and have a strong impact on implementation of optimal solutions.

We observed in certain sectors where demand for certain components is random (alternative components for automotive vehicles, for example) that the dysfunctions induced by stockout are so severe that one has to fall back on emergency procurement procedures to avoid imminent risk of stockout. The missing units may be shipped using dedicated exceptional means (aircraft, truck), generating fixed costs regardless of occupancy. Shipment may also be subcontracted to operators who invoice on a unit-basis (variable direct cost). These models lead to analytical solutions that define optimal inventory safety stock levels and enable a choice of emergency means of transportation depending on the cost structure and of demand dispersion (see Sali and Giard, 2015).

Where demand stems from individuals as opposed to firms, it is generally less open to deferment in the case of mass consumer items. In this case, stockout may simply lead the customer to purchase a substitute product from another brand, without any notable impact. The sale may be lost, in which case the stockout leads to a loss equal to the unit margin generated by this item, loss that may be increased if the customer chooses to leave the shop without purchasing. In the case of a regular customer, the repetition of this kind of incident could result in his/her not returning to the shop. This highlights the difficulty of striking a "definite" stockout cost. This may lead to including a specific constraint in procurement policy using a limit of the stockout percentage of demand.

Use of a generic model to design solutions to recurring operational issues

Algebraic Modelling Languages (AML), such as GAMS or Xpress-IVE support recurring operational decision-making thanks to a clear separation between a generic model and the definition of a set of data enabling the instantiation of an operational issue to be solved (delivery route, scheduling…). Such instance is then resolved by a solver to find a "good" solution, provided that the scope of the issue is not too wide (otherwise, heuristics are called for). This type of issues includes physical data (characteristics of shipments to be made or order books, physical constraints…) and, in the more realistic

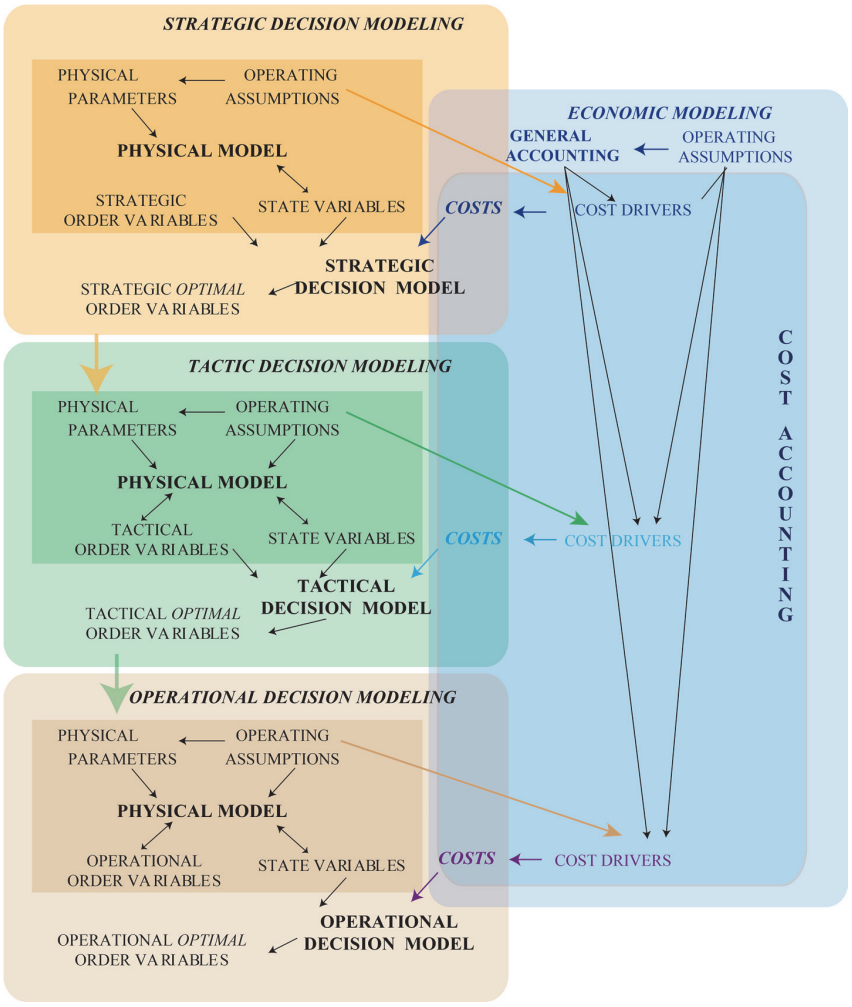


Fig. 2. Costing in decision-making hierarchy

Decision-making models, also includes costs. The question of the relevance of these costs needs to be addressed. Let us refer to the above two examples to spot a number of pitfalls to be avoided

In the periodic determination of transportation routes, the earlier models relied mainly on the time or distance criterion, the cost of the proposed solution being strongly pegged to these parameters. The complexity of the issues to be dealt with and the increase in IS/IT and solver performance have led to switch to the criterion of overall cost. Unit cost related to the "travelled distance" cost driver naturally includes fuel spending and a share of periodic maintenance costs. Where the preferred model precludes solutions that have recourse to overtime, the operational decision-making solution has no impact on the driver payroll that is part of fixed costs over the short term. In the opposite case, the cost of overtime becomes a direct variable cost for the "extra worked hours" cost driver. The issue of allocation of a share of vehicle depreciation costs (or replacement value) must reflect the fact that the chosen solution has no impact on depreciation, which has to be considered an indirect fixed cost, and, therefore, has to be eliminated. Where a mixed vehicle fleet (different truck capacities and performance levels…) is operated for a delivery route, there may be excess loading capacity at certain times, which calls for careful truck selection. Here, the issue of depreciation expense allocation to the different trucks is more complicated but one must remember that fleet sizing issues are part of strategic decision-making. In this exercise, a range of other criteria are taken into account,

including demand fluctuation. Delivery route decisions, therefore, fall within the scope of strategic decision-making, (under the decisional hierarchy referred to above). Accordingly, it seems appropriate to exclude this cost component from operational decision-making.

In many scheduling issues, time is the only aspect to consider. But this is no longer the case when significant material consumption (fluids, energy…) and staffing level also vary (overtime, additional manpower) according to the scheduling option. Since equipment availability is not related to scheduling, it seems preferable, as in the above case, to exclude equipment-related fixed costs in the short term. Take the example of a scheduling issue involving products made on a complex assembly line. Their variety is obtained through a combination of optional components, some of which have an impact on time spent at a number of workstations along the assembly line. This leads to constrain the schedule to respect interval requirements between products that share certain features. These constraints may be alleviated through ad hoc recourse (during one cycle time) to extra manpower at a workstation to cope with the extra workload. In practice, it will be difficult to dedicate the utility worker to this specific workload increase and he/she will probably have to work longer (generally for the full shift working time of the assembly line). The traditional solution proposed in the literature is to use a standard cost calculated as the quotient of an indirect fixed cost (hiring cost) by working time measured in number of cycle times, where it is considered that the cost driver is the request of an utility worker by a workstation during a cycle time. This approach, however, appears misleading for two reasons (Giard & Jeunet, 2010). If the utility worker is used for a single cycle time in the course of a day, the income statement shall record the total expense induced by the entire presence of that utility worker. Additionally, the utility worker may be assigned multiple times to jobs during his/her contract, which considerably increases the number of possible schedules that comply with the constraints of interval requirements. This possible use increase makes sense if the schedule impacts certain variable costs; for example, it may reduce the number of paint gun purges required to change product color at a painting booth. It also makes sense if part of the process may be performed on a processor chosen in a set of heterogeneous parallel processors.

3.2 Decisions as to production system design

This type of strategic decision, always implying investments, may be dealt with under a comparative approach of scenarios or under an optimization model. We refer to the example of the creation or reengineering of a logistics network, bearing in mind that the methodological issues discussed below are very general in scope.

Scenario design implies quite a detailed understanding of the operation of the new target production system. Air transport hub capacity, for example, depends on tangible assets (warehouses, handling equipment…) as well as on human resources and their management style (goods allocation rules, workplace organization…). It also depends on the organization of incoming and outbound means of transport (frequency, aircraft capacity …) as, all other things being equal, a reduction in the frequency of aircraft take-offs linked to payload increase, may generate dysfunctions in connection with insufficient capacity. The scope of the network put at customers' disposal is important for its attractiveness and therefore impacts demand for transportation and, ultimately, network operator profitability. For a given scenario, one may define macro-level cost drivers and define fixed indirect costs and variable direct costs that apply subject to certain conditions. An analysis of an alternative scenario involves posing all these questions and will lead to different traffic, resource and management rule assumptions, ultimately determining a different set of cost drivers. The relevance of ICFR in comparing multiple scenarios to a reference scenario actually depends on consistency of the sets of physical assumptions for each scenario.

One may be tempted to look for a generic formulation of the issues of logistics network design, which is precisely what the vast majority of scholars do. In this context, one implicitly starts from a basic scenario serving to define transportation demand, hub capacity, costs, etc. The problem, however, is that the optimization solution may lead to choosing a set up that is very different from the reference scenario, such that the chosen solution often turns out to be unworkable. This situation is often overlooked in the papers as they tend to focus on model complexity and/or problem-solving lead time.

Let us return to the cost drivers at macro-level. The first point is that, when working at operational level, the cost drivers are physical and linked to specific expenses. However, this is rarely the case when working at tactical and strategic level, where cost-drivers may be proxies (linked to aggregate expenses through relationships generally obtained by linear regression). This approach originated in the 19th century in the cost accounting of railway companies with a ton/miles "driver" used to compare profits between different lines and types of transported goods. Thus, the nature of cost drivers in Figure 2 may differ according to decisional level.

4. CONCLUSION

The above considerations call for more caution when implementing the recommendations derived from theoretical models. They will also encourage researchers to pay closer attention to the relevance of the costs used in their decision models.

REFERENCES

Giard, V. (2003). *Management de la production et des flux*, 3th ed., Economica, Paris.

Giard, V., and Jeunet, J. (2010). Optimal sequencing of mixed models with sequence-dependent setups and utility workers on an assembly line, *International Journal of Production Economics*, 123(2), p. 290-300.

Sali, M. and Giard, V. (2015), Optimal stock-out risk when demand is driven by several mixed-model assembly lines in the presence of emergency supply, *International Journal of Production Research*, 53(11), p 3448-3461.

Silver, E., Pyke, D, Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling* 3th ed., John Wiley and sons, New-York, USA.

Using dominance relation to identify relevant generalized technical contradictions in innovative design
(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Ivana Rasovska
ICUBE-UMR CNRS 735
Université de Strasbourg
Strasbourg, France
ivana.rasovska@unistra.fr

Roland De Guio, Sébastien Dubois
ICUBE-UMR CNRS 735
INSA de Strasbourg
Strasbourg, France
roland.deguio@insa-strasbourg.fr, sebastien.dubois@insa-strasbourg.fr

Abstract—The objective of this paper is to propose a method using optimization principles to identify relevant contradictions within the inventive design problem statement. The proposed method analyzes inventive design problems from the dialectical point of view due to a contradiction model from the theory of inventive thinking (TRIZ). The main concept is to link the conflict between the Pareto solutions from optimization to TRIZ system of contradictions. The trade-offs defined by the Pareto front in the evaluation problem space are suggested to be the best solutions issued from optimization solving approach but not good enough to solve the stated design problem. Thus the inventive solving is necessary to go beyond these Pareto solutions by changing the model linking decision variables to evaluation variables. The starting point for TRIZ solving approach is to describe the underlying inventive design problems corresponding to these Pareto trade-offs through the contradiction model. As many contradictions may exist, the purpose is to rank them according to the dominance relation in order to help the problem solver choosing among them the relevant ones. The method is illustrated and evaluated on the example of electrical circuit breaker.

Keywords—inventive design, contradictions, TRIZ, optimization, Pareto front, dominance relation

I. Introduction

Generally, optimization and inventive problem solving are described as pointing out the differences regarding the two approaches as two separated domains. Thus the inventive design process is studied from the problem solving point of view which represents the common point for both approaches. Based on the information processing theory [1] design is suggested as a process of searching for a solution in a certain problem space in which the problem solving activities take place. Often design problems suffer from very little information about the problem and require a lot of structuring and drawing knowledge to construct problem space. Following [2], problem structuring begins with an interpretation of the problem situation. The goal is to design information systems to support the designer during conceptual design that needs better understanding of the structure of the problem and solution spaces, as well as the process of structuring problems. According to [3] the achieving creativity is only possible by going beyond the bounds of a representation, and by finding a design that could not have been defined by that representation. Exploration of the space shows its limit and identifies the points where the changes can be made. Boden proposes some possible transformations such as dropping a constraint or negating a constraint that allows expanding a search space.

The authors adopted the paradigm of the search of inventive solutions to design problems by browsing a design problem space. The hypothesis is that the best solutions issued from optimization and defined by the Pareto front in the evaluation problem space are not good enough to solve the stated design problem. Thus the use of inventive problem solving is necessary to go beyond these Pareto front solutions. In this paper, the dialectical approach of TRIZ theory (the theory of inventive thinking) reinforced in border of General Theory of Advanced Thinking (OTSM-TRIZ) is used as inventive solving approach, generally well adapted to solve technical design problems [4]. The inventive design problem is tackled through the dialectical principle which aims at a canonical form of problem formulation – a contradiction. The design problem is stated in the shape of a contradiction used for finding a contradiction-free model within the framework of given objectives. One of the goals of this paper is to contribute to the comprehension of this dialectical contradiction model with the aid of optimization solving principles. The purpose is to link the conflict between the Pareto solutions from optimization to the contradictions from TRIZ theory. Also the optimization principles such as Pareto front and dominance relation allows the illustration of some of classical TRIZ contradiction limits. That's why the method proposed in this paper analyzes inventive problems from the general point of view due to a generalized contradiction model proposed in [4]. This generalized contradiction model satisfies the equivalence requirement between the lack of solution from the optimization point of view and the existence of contradiction from the TRIZ point of view. The experience shows that formulating a real robust contradiction is not as obvious as it could appear. Moreover, generally many contradictions may exist as shown in [5] and the questions how to identify and choose contradictions to be addressed arise.

The goal of this paper is to identify the relevant contradictions by introducing optimization principles such as Pareto and dominance relation. Section 2 introduces the notions of the design problem space and the representation of design problem through experiments table. Then different types of TRIZ contradictions are presented and illustrated within the experiments table. Finally, the Pareto dominance relation is presented to delimit design problem space achievable by the optimization and inventive problem solving. Section 3

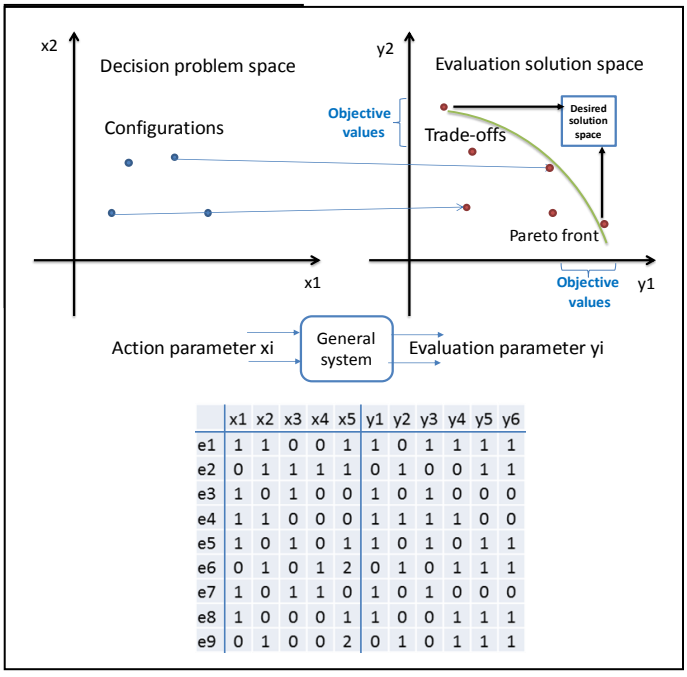
represents how to organize measured situations from the experiments table according to the dominance relation within the evaluation problem space. Then we illustrate how the technical contradictions appear between the different situations and which impact has their solving to the degree of problem invention. The contributions of generalized contradiction model are provided compared with TRIZ classical contradictions solving. Finally, the method to identify the relevant generalized contradictions that describe the underlying inventive design problems corresponding to the Pareto trade-offs is proposed. The method is illustrated and evaluated through the example of electrical circuit breaker.

II. Optimization versus inventive problem solving

A. Optimization problem solving

Optimization and search problems concerning general technical systems are usually described by two types of system parameters. The problem's objectives are described using evaluation parameters, which define the evaluation solution space. In addition, the system variables concerning different system configurations are described by action parameters that define the decision problem space. Thus, any solutions can be described in both evaluation and decision space, and each point of the decision space has a corresponding point in the solution space.

Fig. 1. Illustration of general 2-objective optimization problem

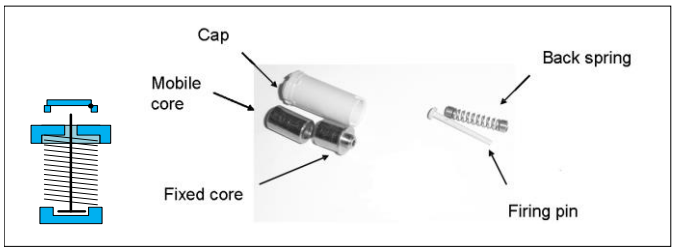


	x1	x2	x3	x4	x5	y1	y2	y3	y4	y5	y6
e1	1	1	0	0	1	1	0	1	1	1	1
e2	0	1	1	1	1	0	1	0	0	1	1
e3	1	0	1	0	0	1	0	1	0	0	0
e4	1	1	0	0	0	1	1	1	1	0	0
e5	1	0	1	0	1	1	0	1	0	1	1
e6	0	1	0	1	2	0	1	0	1	1	1
e7	1	0	1	1	0	1	0	1	0	0	0
e8	1	0	0	0	1	1	0	0	1	1	1
e9	0	1	0	0	2	0	1	0	1	1	1

Fig. 1 illustrates the assertion in the case of a 2-dimensional solution and decision space problem. The goal is defined in the evaluation space by the expected range of values of the evaluation parameters, which defines a desired solution space as a sub-space of the evaluation solution space. The purpose of the optimization methods is, for a given system model, to determine whether there are points in the decision space with corresponding points in the desired solution space. When, as in Fig. 1, no correspondence exists in the desired solution space, optimization techniques based on the well-known concept of the Pareto dominance relation allow definition of the Pareto front. The points on the Pareto front represent the best trade-offs between the evaluation parameters to approach the desired solution. Hence, the purpose of inventive problem-solving methods such as TRIZ is to extend beyond the limit represented by the Pareto front to prompt the evaluation parameters to arrive at a desired value by changing the system model (i.e., changing the action parameters, their interrelationships and their possible value range). Once a new system model is proposed, a new search or optimization problem is stated with the same evaluation solution space as before. To better formulate and represent the design problem for computer oriented tools, the authors use the data on experimental design expressed in the rectangular experiments table as shown on Fig. 1. In general, the experiments are generated in the table with raw data. When possible, a complete design of experiments is performed, or if this is not possible because there are too many variables, the selection of all possible experiments from the research space will be randomized. The values are generally binarized to simplify further data processing.

Let us consider an electrical circuit breaker. When an overload occurs, the overload creates a force (due to magnets and electrical field) which operates a piece called firing pin. The firing pin opens the circuit by pressing the switch, located in the circuit breaker. In case of high overload, the firing pin, this is a plastic stem, breaks without opening the switch. Breaker and components are presented on Fig. 2.

Fig. 2. Example of an electrical circuit breaker



The problem has been studied by experts and the main system parameters and their domains have been defined as: x1: firing pin material (plastic – 1, metal – 0) ; x2: core internal diameter (high – 1, low – 0) ; x3: core external diameter (high – 1, low – 0) ; x4: firing pin diameter (high – 1, low – 0) ; x5: spring straightness (high – 2, medium – 1, low – 0) ; y1: circuit breaker disrepair (satisfied – 1, unsatisfied – 0) ; y2: circuit breaker reusability (satisfied – 1, unsatisfied – 0) ; y3: spring core mounting (satisfied – 1, unsatisfied – 0) ; y4: firing pin bobbin mounting (satisfied – 1, unsatisfied – 0) ; y5: normal mode release (satisfied – 1, unsatisfied – 0) ; y6: firing pin initial position return (satisfied – 1, unsatisfied – 0). In this definition of the problem the xi are the action controlled parameters whereas the yi are the evaluation objective parameters. A set of 8 experiments denoted e1, …, e8, is performed on the system and represent different configuration of the problem. The results of each configuration are provided in a binary shape for the evaluation parameters. The "positive one" values of the y1 to y6 columns mean that the values of yi satisfy the requirements for the given experiment. In contrast,

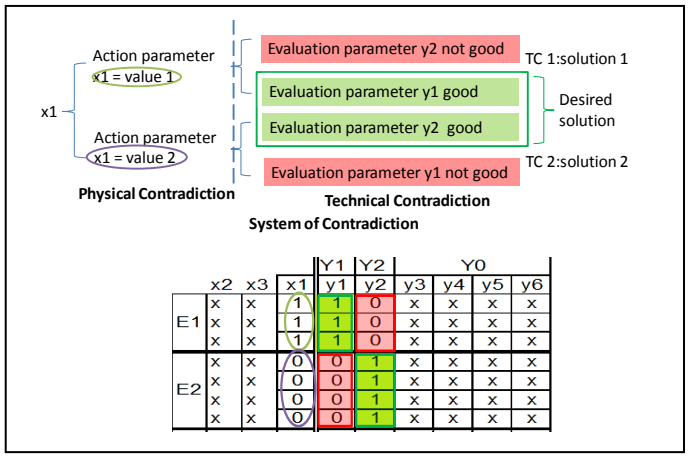
"zero" values of the yi columns mean that the values of the evaluation parameters yi do not satisfy the requirements for the given experiment. Let now consider the Action Parameters. The values 0, 1 and 2 in the matrix refer to 3 different levels of the action parameters during the experiments. The experiment table shown on Fig. 1 correspond to the example presented above.

The authors consider the problems in which no correspondence between the desired solution space and the decision space exists for the existing systems to be inventive design problems. The idea of translating the trade-offs of the evaluation space into contradictions in the decision space is inspired by the TRIZ model of contradictions. The next section provides more information about this link and the TRIZ model of contradictions.

B. Inventive TRIZ problem solving

The authors adopted the dialectical approach of the theory of inventive thinking (TRIZ) proposed by [6], which addresses creative technical problem solving. One of the axioms of TRIZ states that the evolution of technical systems requires identifying and resolving contradictions. These contradictions emanate from the apparent impossibility of meeting the changing requirements in a specific context; they manifest a conflict of values between the parameters of technical systems. Different levels of contradictions are proposed in classical TRIZ, but only two are relevant to the problem resolution: technical and physical contradiction.

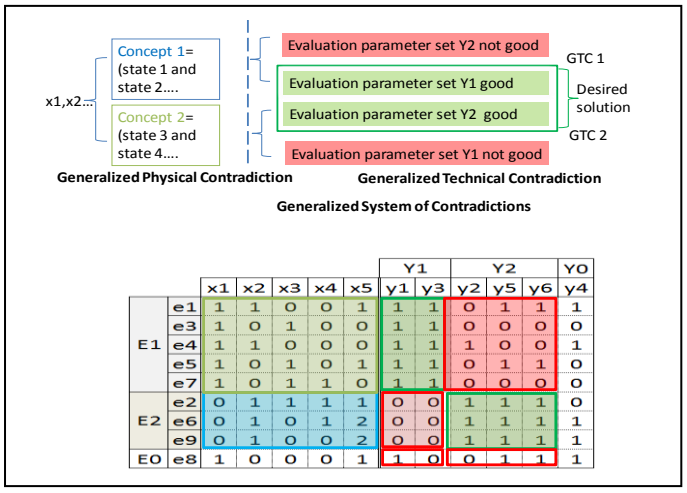
Fig. 3. TRIZ classical system of contradiction and representation within the table of experiments



The technical contradiction expresses conflict in the evaluation problem space; the physical contradiction represents the link between the evaluation space and the decision problem space. Technical contradictions then indicate the obstacles that must be overcome to achieve the solution. They express the opposition between two evaluation parameters of the system and indicate that improving one parameter causes degradation of the second. Physical contradiction defines two contradictory statements of one action parameter required for a single evaluation parameter. The system of contradictions connects a physical contradiction to two technical contradictions that express why the two contradictory statements are required. TRIZ classical technical and physical contradictions are illustrated on Fig. 3 as well as the representation within the experiments table. There is no TRIZ classical technical contradiction within the case of electrical circuit breaker taking into account all the experiments that's why Fig. 3 shows an illustrative example.

The system of contradictions has been extensively used in inventive problem solving. However, there are several reasons for its evolution into the generalized system of contradictions (GSC) [4]. The first motivation is that a TC does not always appear among the possible pairs of evaluation parameters. The second reason is to address multi-objective problems, where more than two evaluation parameters need to be satisfied at the same time. Moreover, regarding the continuous improvement of existing technical systems, more complex situations appear, and designers must handle systems consisting of many design variables and desired outcomes [7]. Hence, the generalized technical contradiction (GTC) was proposed in [8], where the conflict arises between two groups of evaluation parameters: when the requirements for the first group of evaluation parameters are satisfied, the second group is not, and vice versa (Fig. 3). The generalized physical contradiction (GPC) is based on two logical expressions of system states, termed concepts, described by the decision variables (i.e., concept 1 and concept 2 in Fig.4.

Fig. 4. Generalized system of contradictions and description within the table of experiments



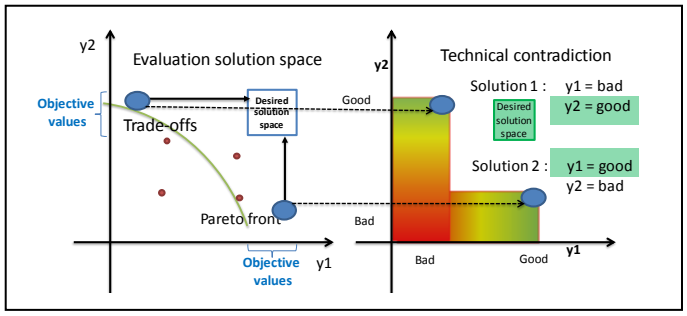
A GPC related to a given GTC appears between two system states in each concept; it formulates the conflict at the level of the action parameter concepts: to achieve the desired results, concept1 and concept2 should hold simultaneously, but unfortunately, no n-tuple of action parameters allows it. This generalized model solves the limitation of classical TRIZ and guarantees multi-objective modeling with respect to the technical contradiction.

C. Optimization versus inventive problem solving

The context of this paper is focused only on the evaluation problem space taking into account measured values of evaluation parameters in each situation. As defined before the solution satisfying all requirements is characterized by all evaluation parameters which are equal to 1. If there is no such solution, the authors consider that the design problem is inventive one and necessities the inventive solving approach to

go beyond the best solution issued from optimization. The best optimization solutions are represented by Pareto front regarding the concept of Pareto dominance. Following this concept, an evaluation vector Y is said to dominate other objective vectors if all components of the considering vector are as good as the components of other objective vectors, and at least one component of Y is better. Accordingly, we can say that one solution Y1 is better than another solution Y2 if Y1 dominates Y2. Optimal solutions, defined as solutions not dominated by any other solution, are on the Pareto front. There may exist several optimal evaluation vectors representing different trade-offs between objectives.

Fig. 5. From optimization to invention through a contradiction



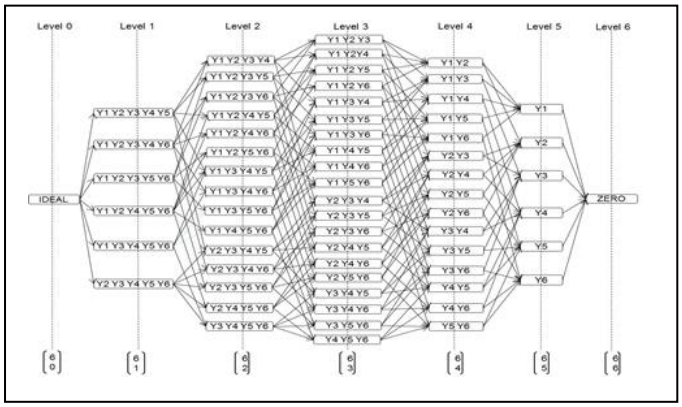
The goal of creative problem solving is to extend beyond this limit represented by the Pareto front and obtain results in the desired solution space. For this purpose, the best solutions issued from the Pareto front are analyzed in terms of the conflicts between their evaluation parameters. In Fig. 5 two arbitrary solutions from the Pareto front on the left side were chosen, and the conflict is described on the right side of the figure. For solution 1, the evaluation parameter y2 is better than for solution 2; however, the evaluation parameter y1 is worse, and vice versa. This conflict in the evaluation space is a limit of the analyzed existing system expressed by technical contradiction. It is also an entry point to use the dialectical approach based on the contradiction corresponding to this conflict. The conflict through the decision problem space in terms of action parameters corresponds to the physical contradiction within the TRIZ theory but is not tackled in this paper.

III. How to identify relevant generalized technical contradictions through dominance relation?

A. Evaluation space structured by dominance relation

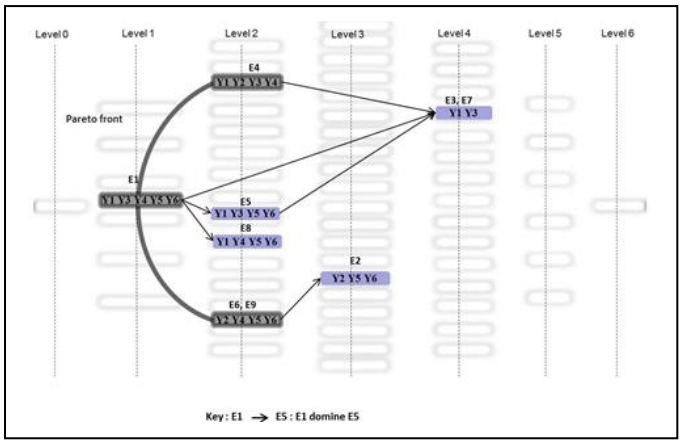
In the case of the combinatorial optimization and using the evaluation parameters issued from the experiments table with binary values, theoretically one disposes all possible situations representing combinations of evaluation parameters' satisfaction as shown on Fig. 6. In the figure each point represents the satisfaction of certain evaluation parameters which are in the title of the represented combination. The number of output points is calculated according to the number of evaluation parameters. In the combinatorial optimization the number of points in this evaluation space is $\sum_{k=0}^{N}\binom{N}{k}$ where N is number of evaluation parameters. Not all of these combinations are achievable in the reality because of different nature of relations between the evaluation parameters. This evaluation space represents only the theoretical knowledge about the system which is independent from the real knowledge. Of course, this binary evaluation space is only available with binarized experiments table, but its equivalent can be built also with real values of parameters.

Fig. 6. Evaluation space structured by dominance relation and levels



One can imagine that these theoretical combinations could be classified according to the dominance relation resulting into different levels of solutions. According to the definition of the dominance relation one case dominates another one if the first case is not worse than the other in all objectives and if this case is strictly better than the other in at least one objective. This classification gives the structured network as shown on Fig. 6; where the ideal case is on the left (all evaluation parameters are satisfied) and the worst case on the right where no evaluation parameter is satisfied. In addition, it is proposed to rank the solutions in different levels defined by the number of objectives reached by each solution (i.e. the number of one values of the binarized solution). Thus, in Fig. 6 the levels reference defines the number of non-satisfied requirements of a solution (i.e. solutions of level 3 do not reach 3 objective values among the whole set of objectives).

Fig. 7. Dominance relation and levels of the circuit breaker problem



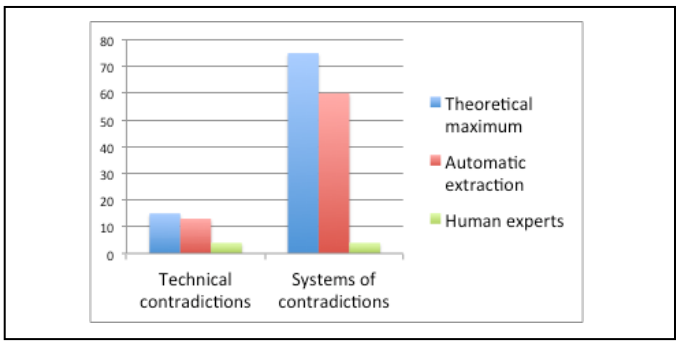
When taking into account real data from the experiments table, our information network becomes much less provided. Indeed, all the points of the evaluation space may not be

reached. Actually, for the case of electrical circuit breaker we have only 7 points in the evaluation space because we have two times two experiments (E6 and E9, E3 and E7) mapped to the same point in the binarized evaluation space. One can identify the dominance relationships between the different experiences that take us to the Pareto set represented by the grey points on Fig.7.

This graphical representation points out different real situations related to the experiments (different configurations of system) but offers also the possibility to represent the contradictions which appear between these situations. The contradictions may appear when we want to move from one point to another point in this evaluation space by acting on the decision variables in the decision space.

Statistically 15 technical contradictions could have been formulated taking into account pairs of lines, and in this table 13 technical contradictions out of 15 and 60 systems of contradictions out of 75 were found by the aid of algorithm proposed in [5]. The human experts, on this problem only formulated 4 systems of contradictions as illustrated in Fig. 8. What we can conclude, is that the human does not have nor time, nor ability to formulate all the possible contradictions, but of course all the contradictions are not necessary to solve the problem, the only thing is to formulate the contradictions which resolution will really lead to the resolution of the problematic situation. The objective of formulating the contradiction is to understand the limits of actual representation of the system and also to help in building new representations.

Fig. 8. Number of different contradictions found by various methods



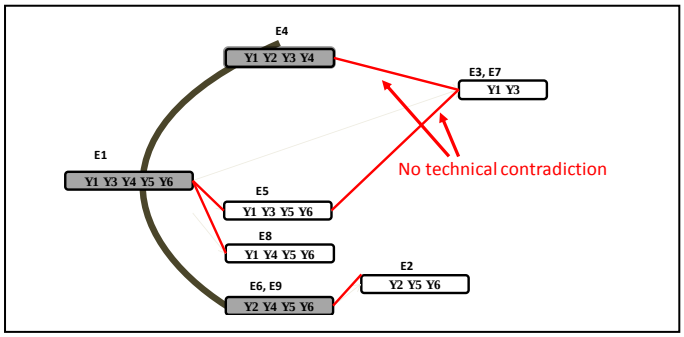
B. Relevant contradiction to be solved

The goal is to study the different types of contradictions that arise between various situations. The contradictions appear in this evaluation space when we want to move from one point to another one except the dominance relation. The contradiction network is created and contradictions can be studied and compared in terms of frequency and importance in relation to their appearance on the Pareto front. There are several possibilities to study the contradictions: throughout the whole experiments table: within on experiment looking to one line, by pair of lines, pair of columns, by the contradiction systems (such as classical TRIZ, generalized contradiction etc.).

Let's analyze what happens if we solve one TRIZ classical technical contradiction in the different levels of evaluation space. The goal is to move in the evaluation space to the ideal point which is on the left side of the evaluation space.

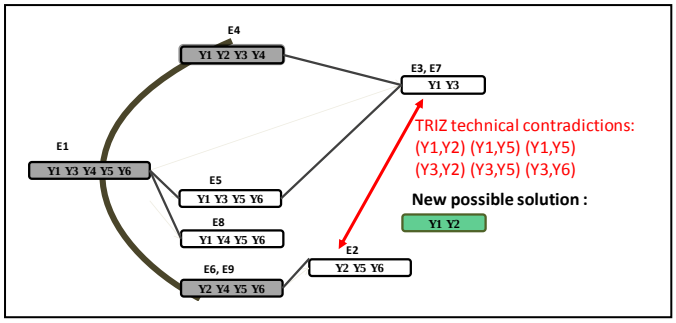
- The first case addresses the situation when we follow the dominance relation and we move from one point to another one according to this relation. This means that no technical contradiction appears as shown on Fig. 9. Thus the solutions on the left with more satisfied evaluation parameters will be naturally better then on the right side. The best solutions with the less possible not satisfied objectives will appear on the Pareto front dominating the other ones.

Fig. 9. No technical contradiction on dominance relation



- Second case concerns TRIZ classical technical contradictions which take into account only two evaluation parameters. Let's consider two points (not situated on Pareto front), we can identify a set of TRIZ classical technical contradictions, and be sure that each of these contradictions will be solved by points that are in upper dominating levels. But this can be possible by deteriorating the other evaluation parameters (because not considered within the case of TRIZ classical technical contradiction). Or if we solve the contradiction we can create a new possible point in the evaluation space which is even worse than the previous one (Fig. 10).

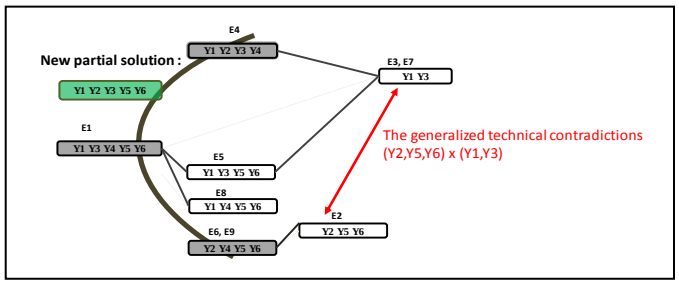
Fig. 10. Solving TRIZ classical technical contradiction



- Third case concerns generalized technical contradictions. If we take into account, the generalized technical contradiction considering all satisfied evaluation parameters between two existing points in the evaluation space, we will always find better partial solution than the considered previous ones (Fig. 11). But one cannot guarantee that the stated Pareto front of available partial

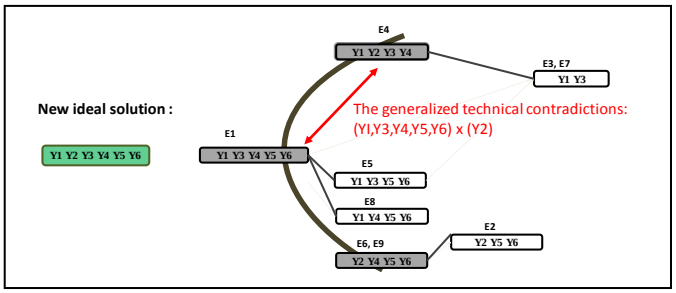
solutions will move towards the ideal point. An another aspect is a simplicity of the studied case with few parameters, based on the model of expert knowledge which is near the ideal, but in more complex situations, or with real value experiments table, it is not systematic that partial solutions exist to the identified contradictions.

Fig. 11. Solving generalized technical contradiction



- Fourth case concerns generalized technical contradiction between two Pareto solutions which will give always the better solution then the previous ones and will move the Pareto front beyond the existing one and towards the ideal point if this is not already a case (see in Fig. 12).

Fig. 12. Solving generalized technical contradiction from Pareto front



IV. Conclusion

The goal of this paper was to identify the relevant generalized technical contradictions by introducing optimization principles such as Pareto front and dominance relation. The dialectical contradiction model first provided new perspectives on the Pareto conflicts and the potential directions for their solutions. Second, it provides input into TRIZ-based inventive problem solving methodologies such as relevant contradictions to be solved.

The proposed methodology is based on three steps: to binarize data to simplify data processing, to analyze a dominance relation of measured situations and to illustrate the trade-offs defined by the Pareto front in the evaluation solution space which should identify relevant contradictions to be solved. The method was illustrated and evaluated on the example of electrical circuit breaker, and the contribution of generalized contradiction model was provided compared with TRIZ classical contradictions solving.

First, the different measured experiments were organized in the solution space according to the dominance relation of their satisfied evaluation parameters. Different situations, what does it change if one contradiction is solved in a certain point of this space, were shown. Our hypothesis was to study the points on the Pareto front because they dominate the others and so if we move from the points which are dominated to the dominating points on the Pareto front, we are able to improve at least one evaluation parameters without deteriorating the others thus no contradiction is meet. Thus the generalized contradictions that describe the underlying inventive design problems corresponding to the Pareto trade-offs are suggested to contribute to make evolve the problem solving model; to go beyond the best solutions issued from optimization within a new representation model.

The Pareto conflict is assimilated to the most generalized technical contradiction. The approach may also help to clarify which action variables of the decision space are involved in the trade-offs defined by the Pareto front in the evaluation solution space. The idea of translating the trade-offs of the evaluation space into generalized physical contradictions in the decision space is inspired by the TRIZ model of contradictions and will be developed in the future research.

References

[1] A. Newell & H. A. Simon, Human problem solving, Englewood Cliffs, NJ: Prentice-Hall, 1972, p. 920.

[2] H. Christiaans & J.D. Restrepo-Giraldo, Problem Structuring and Information Access in Design, *Journal of Design Research of Design Research*, 4(2), 2004.

[3] M. Boden, The Creative Mind: Myths and Mechanism. (T. francis Group, Ed.). London and New York, 2004.

[4] S. Dubois, I. Rasovska & R. De Guio, Interpretation of a General Model for Inventive Problems, the Generalized System of Contradictions. 19th CIRP Design Conference, 2009, 271–276.

[5] L. Lin, I. Rasovska, R. De Guio & S. Dubois, Algorithm for identifying generalised technical contradiction in experiments. *Journal Européen des Systèmes Automatisés*, 24 (1), 2012, pp. 67–85.

[6] G.S. Altshuller, Creativity as an exact science: the theory of the solution of inventive problems. New York, NY: Gordon & Breach Science Publishing, New York, 1984.

[7] F. Joachim & A.F. Peter, Complex problem solving: The European Perspective. Lawrence Erilbaum associates, Inc., 1995.

[8] S. Dubois, T. Eltzer & R. De Guio, A dialectical based model coherent with inventive and optimization problems. *Computers in Industry*, *60*(8), 2009, pp. 575–583.

Industrial Symbiotic Relations as Cooperative Games

(Presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Vahid Yazdanpanah and Devrim Murat Yazan
University of Twente, Drienerlolaan 5, 7522 NB, Enschede, The Netherlands.
Email: {v.yazdanpanah,d.m.yazan}@utwente.nl

***Abstract*—In this paper, we introduce a game-theoretical formulation for a specific form of collaborative industrial relations called "Industrial Symbiotic Relation (ISR) games" and provide a formal framework to model, verify, and support collaboration decisions in this new class of two-person operational games. ISR games are formalized as cooperative cost-allocation games with the aim to allocate the total ISR-related operational cost to involved industrial firms in a fair and stable manner by taking into account their contribution to the total traditional ISR-related cost. We tailor two types of allocation mechanisms using which firms can implement cost allocations that result in a collaboration that satisfies the fairness and stability properties. Moreover, while industries receive a particular ISR proposal, our introduced methodology is applicable as a managerial decision support to systematically verify the quality of the ISR in question. This is achievable by analyzing if the implemented allocation mechanism is a stable/fair allocation.**

I. Introduction

The multi-dimensional concept of *Industrial Symbiosis* focuses on analysis, design, and operation of collaborative relations between traditionally disjoint industrial enterprises with the aim of keeping reusable resources, e.g., recyclable material or waste energy, in their (loosely connected) value chains [1–3]. As *Industrial Symbiotic Relations* (ISRs) aim at the lowest possible discharge of resources, they can be considered as a tool for implementing the concept of *circular economy* [4] in the context of industrial relations. Moreover, ISRs are closely related to *Industry 4.0* paradigm and practice of *Collaborative Networked Organizations* as they all are concerned about the necessity for interrelation between traditionally disconnected industrial firms [5, 6]. Reviewing industrial symbiosis literature, we encounter recent contributions focused on different aspects of this concept. In [3], they present the concept of *perfect industrial symbiosis* and verify the quality of any given ISR by measuring its distance to such a perfect form. Introduced method in [7] focuses on efficiency measuring while [8] studies dynamics of profits in ISRs. Despite contributions that discuss static (multi-criteria) decision analysis in industrial symbiosis (see [9]), one aspect of ISRs that we believe requires more attention is *dynamic decision analysis*. In our view, while we shift from ISR in theory to ISR in practice, two missed elements are 1) applicable decision analysis methods and 2) practical decision support tools that are aware of dynamic operational aspects of ISRs, e.g., methods for analyzing and mechanisms for designing fair and stable collaborations. This asks for formal frameworks tailored to model, verify, and support such decisions (i.e., decision process modeling, decisions verification methods, and decision support tools). In a general view, decisions in ISRs can be categorized in two classes, *selection decisions* and *collaboration decisions*. The former is about choosing among firms and learning about potential ISRs (exploration) while the latter is about getting engaged in (or rejecting) a particular ISR proposal (exploitation). To deal with these two operational decision problems, the mature field of of cooperative game theory [10] and more specifically subfield of Operations Research (OR) games [11] provide vigorous analytical methods and design mechanisms.

In this work, we aim to fill the gap by tailoring analytical tools based on game-theoretical solution concepts to support the second form of decisions, i.e., collaboration decisions, in ISRs. For this purpose, we represent ISRs as market games and model them as cooperative cost-allocation games (see [12, 13]). Accordingly, the focus is on operational aspects of ISRs, analysis of collaboration decisions in ISRs, and tailoring cost-allocation mechanisms that respect the operation of ISRs. Note that in this work we analyze collaboration decisions in bilateral industrial symbiotic relations as the nuclear building blocks for various industrial symbiotic topologies, e.g., Industrial Symbiotic Networks (ISNs) [14].

The paper is structured as follows. In Section II we provide a conceptual analysis of ISRs from an operational point of view. Section III presents the game-theoretical preliminaries and our proposed class of ISR games. In Sections IV we introduce the two tailored solution concepts for allocation of costs in ISRs. Using these notions, firms can reason about stability and fairness of any given ISR. Finally, concluding remarks and future work directions are presented in Section V.

II. Conceptual Analysis

To discuss the intuition behind our proposed operational semantics for the game-theoretical formulation of Industrial Symbiotic Relations (ISRs) and to elaborate nuances of collaboration decisions in ISRs, we present the following running example. Imagine a glass manufacturer firm A and a ceramics manufacturer firm B. Firm A produces glass powder as its excess resource r that (after recycling) can be substituted with i, the primary input of B for its production processes (Figure 1).

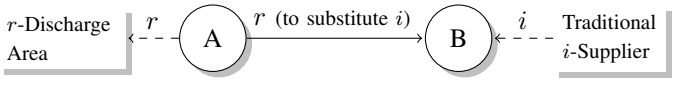


Fig. 1. Schematic industrial symbiotic relation between (glass manufacturer) firm A and (ceramics manufacturer) firm B on resource r (glass powder).

A. Collaboration Decision and Cost-Allocation

In our ISR example, having the potential to form a symbiotic relation over r may enable both A and B to reduce

the costs affiliated with discharging excess r (7 units of util[1]) and purchasing i (11 units of util), respectively. However, establishing such a collaboration has its corresponding operational costs, e.g., costs of transportation and recycling (15 units of util). Intuitively, such an ISR would be feasible if total traditional costs in case of non-cooperation ($7+11$ utils) be more than total operational costs in case of cooperation (15 utils). Such a feasibility condition is necessary but not sufficient for firms to collaborate. One main question is about the method(s) to allocate the operational costs of collaboration such that the contribution of each party will be respected. Should both firms A and B equally pay (7.5 utils) or should firm B pay more as it is enjoying more cost reduction thanks to the collaboration? Although the collaboration may reduce the total cost, how to allocate the total cost of ISR operationalization will be at stake. This relates the practice of ISRs to the concept of *coopetition* [15] in which the players are interested in cooperation but also compete to gain as much as possible out of the benefits created by such a cooperation. Accordingly, studying if an ISR is a stable and fair relation calls for methods to analyze and mechanisms to guarantee its performance in the operationalization stage.

B. Dynamics of ISR-Related Costs

Following [16], we deem that economic profitability can be seen as the main driver for industries to get involved in a potential ISR (*collaboration decision*). In brief, for firms with the potential to establish an ISR, it is reasonable to compare the total cost of ISR-operationalization with their current ISR-related costs, e.g., excess-resource discharge costs. Note that we circumscribe our focus on costs that can be reduced or costs that will be introduced, in case of ISR materialization and not the costs that remain uninfluenced, e.g., costs related to processes that are independent of the ISR in question. In the following, we provide a brief cost analysis for firms on both sides of a potential ISR. This results in two classes of ISR-related *operational* and *traditional* costs. The former refers to costs involved in the process of ISR operationalization while the latter is about traditional costs that firms should take into account if they do not implement the ISR.

In our example, firm B should traditionally purchase i from its i-supplier(s) and firm A has to pay the cost of discharging r. We call the former, *traditional purchasing cost* and the latter, *traditional discharge cost*. On the other hand, the three main ISR-related operational costs are *Treatment*, *Transportation*, and *Transaction* costs [17, 18]. *Treatment Cost*: When a resource based on which an ISR can be established, e.g., glass powder, is out of a production process, it needs to be treated. Depending on the resource type, treatment processes may include sorting, dismantling, liquefaction, etc. [19, 20]. Based on the set of treatment processes required to make the resource usable for the resource-receiver side of an ISR, the implementation of waste treatment facility may change. In general, the treatment process results in a total treatment cost for any particular ISR. *Transportation Cost*: Resource transportation can be done via land vehicles, sea freights, or even combined transportation modes (see [21]) with respect to the resource type and geographical boundaries. Moreover, potential partners may decide to invest in implementation of new infrastructures, e.g., a pipeline system, and paying the investment cost for this. In this work, we abstract from such subtleties in decision-making for the mode of transportation and assume a standard total transportation cost for a given ISR. *Transaction Cost*: The role of transaction costs in establishment of ISRs is studied in the industrial symbiosis literature, e.g., in [22, 23]. According to [24, 25], transaction costs include the costs of market research, contracting negotiations, coordination, and adapting to the use of non-traditional resources, e.g., wastes. As discussed, the former two operational costs are very much dependent on the resource type while the transaction cost merely depends on the administrative aspects of the ISR. As we are focused on industrial symbiotic relations (and not networks), we take into account a single value for transaction cost for a given ISR.

In this work, we consider the two classes of ISR-related operational and traditional costs as the main quantitative parameters for our game-theoretical formalization of ISRs. This is to tailor mechanisms for allocation of *ISR-related operational costs* based on the contribution of the involved firms to *ISR-related traditional costs*. For notational convenience, while discussing about a given industrial symbiotic relation σ between two arbitrary firms A and B, we denote the total σ-related traditional costs for the firm $i \in \{A, B\}$ with $T_i(\bar{\sigma})$ and refer to total σ-related operational costs as $T(\sigma)$. So, the aim is to allocate $T(\sigma)$ to both A and B in a stable and fair manner by taking into account $T_A(\bar{\sigma})$ and $T_B(\bar{\sigma})$. We later discuss about and distinguish between stability and fairness of cost-allocations in ISRs.

C. Game-Theoretical Cost-Allocation Mechanisms

As we discussed in Section II-A, allocation mechanisms can play a key role in the process of ISR operationalization since a fair allocation of operational costs can foster the collaboration. For developing such practical allocation mechanisms, cooperative game theory [10] and Operation Research (OR) games [11] provide theoretical solution concepts to allocate costs to involved players in market games. Notions such as *core of the game* and *Shapley value* guarantee desired properties such that it is reasonable for firms not to deviate from cooperation [10]. In the following, we briefly analyze properties of these two types of game-theoretical solution concepts as two notions that we aim to tailor for allocation of the total operational cost in ISRs.

We first discuss the concept of *core of the game* as the set of all cost-allocations that (1) allocate a cost to each player lower than their traditional cost and (2) guarantee that the total allocated cost is equal to the total operational cost. In our ISR example, the total ISR-related operational cost is 15 utils while firms A and B had to traditionally pay 7 and 11 utils, respectively. In this case, cooperation results in total cost reduction by 3 utils. However, cooperating will be rational for each player, only if they individually pay less than what they had to pay traditionally. When a cost-allocation mechanism satisfies this, it regards the so called *individual rationality (INR)* property [13]. On the other hand, the summation of allocated costs to players should be equal to the total operational cost. Mechanisms that satisfy this property, are called *efficient (EFF)* cost-allocations. The set

[1]A *util* can be any sort of transferable utility, e.g., say that each *util* is one thousand Euros.

of cost allocations that satisfy both INR and EFF form the *core of the game* [10]. The second game-theoretic notion for allocation of costs, is the concept of *Shapley value* [26] as the unique efficient (EFF) mechanism for allocation of a cost among players such that (1) symmetric players pay the same costs, (2) players that their presence in the cooperation results in no cost reduction (referred as dummy players) pay their traditional costs, and (3) if players get involved in another game, the total allocated costs to each player in these two games, can be simply added. These three properties respectively referred to *symmetry (SYM)*, *dummy/null player (DUM)*, and *additivity (ADD)* properties in the cooperative game theory literature [10]. Note that Shapley value is the unique mechanism that allocates a cost value to each player such that all EFF, SYM, DUM, and ADD are satisfied.

III. Industrial Symbiotic Relations as Games

Dynamics of costs and cost-saving values that result from collaboration in market games are often modeled by cooperative games with *Transferable Utility* (TU) games [10, 12]. Such games specify all the possible collaborative agent groups and represent the corresponding cost values. This formal representation enables reasoning about cost-saving as a quantitative outcome of cooperation among agents.

Definition 1 (Cooperative TU Games). [10] A cooperative cost-allocation game with transferable utility (a TU game) is a tuple (N, c) where $N = \{a_1, a_2, \ldots, a_n\}$ is the finite set of agents and $c : 2^N \mapsto \mathbb{R}_{\geq 0}$ is a characteristic cost function that associates a real number $c(S)$ with each subset $S \subseteq N$. By convention, we always assume that $c(\emptyset) = 0$.

In the following definition, we recall two properties that axiomatize the behavior of the cost function in response to structural relations between agent groups.

Definition 2 (Subadditive and Submodular Games). [10] Let $G = (N, c)$ be a cost-allocation TU game. We say G is *subadditive* iff $c(S) + c(T) \geq c(S \cup T)$ for all disjoint agent groups S and T in N (i.e., $S, T \subseteq N$ and $S \cap T = \emptyset$). Moreover, we say G is *submodular* iff $c(S) + c(T) \geq c(S \cup T) + c(S \cap T)$ for all agent groups S and T in N (i.e., $S, T \subseteq N$).

Roughly speaking, in *subadditive* games, agents have rational incentives to cooperate because the total cost will be higher in case of no-cooperation. In most applications, cost-allocation games are usually subadditive. Desirable properties of *submodular* games (also referred as *concave* games) will be elaborated while we focus on cost allocation mechanisms for TU games (in Section IV).

With respect to our scope of application, i.e., bilateral symbiotic relations between industrial agents, we focus on two-person TU games and formalize our Industrial Symbiotic Relation (ISR) games as such. Moreover, following our presented analysis in Section II-A about the feasibility of ISRs (in case they result in the reduction of the total cost), we assume subadditivity as it corresponds to the nature of our application context.

Definition 3 (ISR Games). Let σ be an ISR between firms A and B. Moreover, let $T(\sigma)$ and $T_i(\bar{\sigma})$ respectively represent the total σ-related operational cost and the total σ-related traditional costs for $i \in \{A, B\}$. We say σ-based ISR cost allocation game (ISR game σ) between firms A and B is a subadditive cooperative TU-game (N, c) where $N = \{A, B\}$, $c(N) = T(\sigma)$, $c(\emptyset) = 0$, and $c(\{i\}) = T_i(\bar{\sigma})$ for $i \in N$.

According to Definition 3, the cost function of ISR games characterizes industrial symbiotic relations by associating to each singleton group $\{i\}$[2], the total cost that they will face (individually) in case of no-cooperation. Moreover, it ascribes the total ISR-related operational cost to the two member group N as the amount that members of N have to pay (collectively) in case of cooperation. In addition, the assumed subadditivity of ISR games reflects the feasibility of ISRs, i.e., in ISR games we have that $T(\sigma) \leq \sum_{i \in N} T_i(\bar{\sigma})$. So, regardless of the mechanisms for the allocation of ISR-related operational costs, the total amount to be paid in case of cooperation is at most equal to the sum of the amounts to be paid individually. The following property shows that in general, ISR games are submodular regardless of their particular settings. We later discuss that such a property results in applicability of a large class of game-theoretical cost allocation mechanisms, i.e., mechanisms that are based on the concept of *core of the game*.

Proposition 1 (Submodularity of ISR Games). *Let σ be an arbitrary ISR game. It always holds that σ is submodular.*

Proof: According to Definition 2, a game is submodular iff the $c(S) + c(T) \geq c(S \cup T) + c(S \cap T)$ inequality holds for all possible agent groups S and T in N. In ISR games, by checking the validity of this inequality for all 6 possible combinations of agent groups (for S and T), the claim will be proved. For $S = \emptyset$, we have the following valid inequality $c(\emptyset) + c(T) \geq c(\emptyset \cup T = T) + c(\emptyset)$. For $S = N$, the inequality can be reformulated in the following form that always holds $c(N) + c(T) \geq c(N \cup T = N) + c(N \cap T = T)$. Finally, when S and T are equal to the only two possible disjoint groups, we have the following inequality $c(S) + c(T) \geq c(N) + c(\emptyset)$. This inequality always holds thanks to the subadditivity of ISR games. ■

Note that *submodularity* is not a general property of *subadditive* cooperative games but holds for the class of ISR games. In the following, we recall our ISR scenario between the glass manufacturer firm A and the ceramics manufacturer firm B, and describe the game-theoretical formulation of this scenario.

Example 1 (ISR Scenario as a Game). In the ISR scenario from section II, we assume that the amount of recycled excess r in A completely substitutes the required amount of i in B. Hence, in case the firms operationalize this symbiotic relation, neither of the firms has to deal with associated traditional costs for discharging excess r and purchasing required i. This ISR scenario can be modeled as cooperative game $\sigma = (N, c)$ where $N = \{A, B\}$, $c(A) = 7$, $c(B) = 11$, $c(\emptyset) = 0$, and $c(N) = 15$. This game is both subadditive and submodular. To check subadditivity, we survey all possible couples of agent groups in N. The only two disjoint agent groups are $\{A\}$ and $\{B\}$ for which the cost of the union group

[2]In further references, whenever it is clear from the context that we are referring to a singleton group $\{i\} \subset N$, we use i instead of $\{i\}$.

(c($\{A, B\}$) $= c(N) = 15$) is lower than the summation of the individual costs ($c(\{A\}) + c(\{B\}) = 18$). Thus, the game σ is subadditive. For submodularity, we can rely on Proposition 1.

IV. Allocation Mechanisms and Decision Support

Having Industrial Symbiotic Relations (ISRs) modeled as cooperative games (ISR games), one decision that firms are faced with is either "to get engaged in" or "to reject" a potential ISR. This is mainly to determine the *collaboration decision* (as discussed in Section II-A). Following rational decision-making perspectives presented in [27, page. 210-211], we deem that to support the collaboration decision in ISRs, analyzing two key aspects of collaborations, namely *stability* and *fairness*, results in practical decision support tools for ISRs. In other words, "goodness" of a collaboration can be characterized and validated by checking if it is *stable*, *fair*, or both. In the following, we introduce two methods from game-theoretical literature that axiomatize the *stability* and *fairness* properties. These methods formulate both the stability and fairness of collaborations with respect to distribution of the value that agents can gain thanks to the collaboration. Accordingly, tailoring these mechanisms for ISR games leads to tools for supporting the *collaboration decision* in industrial symbiotic relations. In this case, the way that industrial agents allocate the total ISR-related operational cost specifies the stability and fairness of the ISR.

A. Core Allocations for ISR Games

Core-based mechanisms in cooperative games are mainly concerned about *stability* of possible collaborations [10]. In relation to cost allocation in games, a collaboration is stable iff (1) the summation of allocated costs to individual agents in a collaborative group is equal to the total cost that the group should pay (efficiency) and (2) the allocated costs to individuals is at most equal to their costs in case they defect from the collaborative group (rationality). It is observable that if the agent groups follow a cost allocation method that does not satisfy the two above properties, they end in an *unstable* situation either due to inefficient distribution of costs or as the result of (rational) agents leaving the group. In the following, we define our core-based cost allocation mechanism for ISR games and describe its properties.

Definition 4 (Core Allocations for ISRs). Let σ be an ISR game (as defined in Definition 3) between firms A and B. The core of σ is the set $\Psi(\sigma) := \{\langle T_A^\Psi(\sigma), T_B^\Psi(\sigma)\rangle\}$ such that for $i \in \{A, B\}$ we have that (1) $T_i^\Psi(\sigma) \in \mathbb{R}_{\geq 0}$ (non-negative valued), (2) $\sum_{i \in \{A,B\}} T_i^\Psi(\sigma) = T(\sigma)$ (efficient), and (3) $T_i^\Psi(\sigma) \leq T_i(\bar{\sigma})$ (individually rational).

Following the discussion about stability of collaborations, an ISR σ is stable with respect to the allocation of its operationalization costs, if it implements a cost allocation that belongs to the core of σ. Thus, the presented core allocation for ISRs can be applied as (1) a mechanism for guaranteeing the stability of an ISR (ISR design) and (2) as a verification method to analyze if an ISR is stable (ISR assessment). The set of core allocations for a given ISR, construct a segment representable in two-dimensional space (see Figure 2). Due to the linear formulation of ISRs' core, computing the set of stable allocations is not computationally expensive and can be computed applying linear programming techniques to solve a two-variable system of inequalities [28].

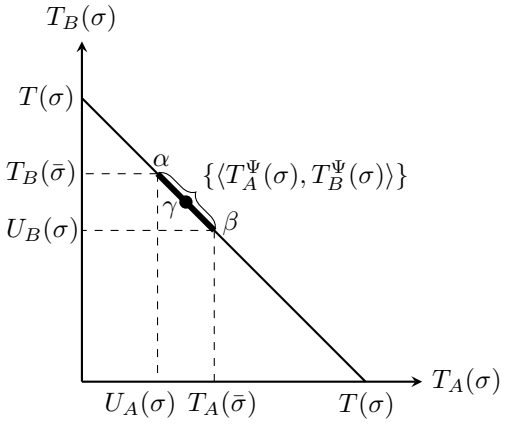


Fig. 2. Schematic core and Shapley allocations for ISR game σ: In this diagram, allocated costs to firms A and B are illustrated on the horizontal and vertical axes, respectively. Moreover, $U_A(\sigma) = T(\sigma) - T_B(\bar{\sigma})$, $U_B(\sigma) = T(\sigma) - T_A(\bar{\sigma})$, and γ represents the Shapley allocation of the game $\Phi(\sigma)$.

Two main concerns with respect to applicability of the concept of *core* are about its *existence* and *fairness*. In other words, is the set of core allocations a nonempty one and does it always provide a fair distribution of costs among involved firms in an ISR game? In the following, we first illustrate that ISR games have a nonempty core and core allocations cannot guarantee the fairness property. This motivates the introduction of a fair cost allocation mechanism, i.e., a Shapley-based allocation mechanism for ISR games.

Proposition 2 (Existence Property of Core of ISR Games). *Let $\Psi(\sigma)$ be the core of ISR game σ between firms A and B. It always holds that $\Psi(\sigma) \neq \emptyset$.*

Proof: According to the *Bondareva-Shapley* theorem (as described in [10]), submodular games have a nonempty core. We already proved in Proposition 1 that ISR games are submodular. Hence, we have that for any ISR game σ, the core is not empty. ■

Note that the concept of *core of the game*, provides a set of allocations that guarantee the stability of the collaboration. Having a range of stable cost distributions is appropriate for ISR scenarios in which firms are allowed to practice their bargaining power (see [29]) in the negotiation process in order to pay the smallest possible share of the ISR-related operational costs. E.g., as illustrated in Figure 2, the most desirable (albeit stable) allocation form for firms A and B occurs in points α and β, respectively. For instance in α, firm A enjoys paying the lowest stable share of the total ISR-related operational cost while B is paying the highest. In this case, B suffers because of this intuitively unfair allocation and pays equal to its total traditional cost. This shows that the concept of core provides a method to verify the stability of an ISR and can be applied as a tool to support the *collaboration decision* in ISRs, i.e., accepting an out of core cost-distribution results in an *unstable* ISR. Nevertheless, it does not grasp the

fairness and neither provides a method for choosing a specific cost allocation to implement among the set of stable cost allocations. In the following, we tailor a solution concept that axiomatize the concept of fairness and provides a single-point allocation that satisfies the fairness property.

B. Shapley Allocation for ISR Games

Regarding the allocation of costs in collaborative groups, there exist various interpretations of the complex notion of fairness (see [30]). In cooperative game theory, the well-established concept of *Shapley value* [26] is a central concept that regards the *fairness* of a cost distribution among members of a collaborative agent group by taking into account their contributions to the collaborative group. In this work, we follow Shapley's view and expect a *fair* allocation to satisfy the following four properties: *efficiency*, *symmetry*, *dummy player*, and *additivity* (as discussed in Section II-C). In brief, if a group G follows an *efficient* method to allocate cost C among its members, the summation of allocated costs to members of group G will be equal to C. The *symmetry* property says that agents that make the same contribution to the total cost, should be allocated the same individual cost shares. The *dummy player* property says that if the presence of an agent A does not result in any cost reduction (in all the possible agent groups), the allocated cost to A should be equal to its individual total traditional cost. Finally, the *additivity* property says that if you combine two games V and U, the allocated cost to an agent A (involved in both the games) should be the sum of the costs allocated to A in the individual games, i.e., playing more than once does not lead to any (dis)advantages for A. For formal axiomatization of these properties, we refer the reader to [26]. In the following, we present our tailored Shapley value for ISR games.

Definition 5 (Shapley Allocation for ISRs). Let σ be an ISR game (as defined in Definition 3) between firms A and B. The Shapley allocation for σ is the tuple $\Phi(\sigma) := \langle T_A^\Phi(\sigma), T_B^\Phi(\sigma)\rangle$ where for $i \in N = \{A, B\}$ we have $T_i^\Phi(\sigma) = \frac{1}{2}[T(\sigma) + T_i(\bar{\sigma}) - T_{N\setminus\{i\}}(\bar{\sigma})]$.

A reader familiar with the notion of Shapley value might expect the two notions of *orders* and *marginal contributions* to be a part of our tailored concept of Shapley value for ISRs. We highlight that due to our domain of application, i.e, bilateral industrial relations, there are two possible orders in ISR games (reflected by the constant value $\frac{1}{2}$). Moreover, the marginal contribution of a given firm $i \in N = \{A, B\}$ can be reformulated in terms of the most desirable stable cost for i and the most undesirable one, i.e., $U_i(\sigma) = T(\sigma) - T_{N\setminus\{i\}}(\bar{\sigma})$ and $T_i(\bar{\sigma})$, respectively. Note that as the Shapley value is defined following a constructive method (in contrast to condition-based definition of core in Definition 4), the existence of the Shapley value for any arbitrary ISR game is guaranteed. Following our discussion about the fairness of collaborations, the next property shows that the Shapley value is the unique fair method for allocation of the total ISR-related operational cost in ISR games.

Proposition 3 (Uniqueness of the Shapley Value). *Let $\Phi(\sigma)$ be the Shapley allocation for ISR game σ between firms A and B. For any fair allocation of costs in σ, denoted by $\Phi'(\sigma)$, we have that $\Phi'(\sigma) = \Phi(\sigma)$.*

Proof: Importing results from [26], we have that for any cooperative game, the Shapley value is the unique allocation method that satisfies all the four properties of fair cost allocations, i.e., efficiency, symmetry, dummy player, and additivity, regardless of the characteristics of the cost function of the game. Accordingly, the uniqueness property holds for ISR games as two-person cost allocation games. ■

Considering core of ISR games and their unique Shapley allocation, the following proposition relates these two forms of solution concepts and shows that in ISR games the Shapley allocation is in the core.

Proposition 4 (Membership in the Core). *Let $\Psi(\sigma)$ and $\Phi(\sigma)$ be respectively the set of core allocations, and the Shapley allocation for ISR game σ between firms A and B. It always holds that $\Phi(\sigma) \in \Psi(\sigma)$.*

Proof: Based on [13, 26], the core of submodular games is nonempty and includes the Shapley value. For ISR games, according to Proposition 2, the core is nonempty. Thus, we have that for any ISR game σ, it holds that $\Phi(\sigma) \in \Psi(\sigma)$. ■

Note that the membership of the Shapley allocation in the core is a property of ISR games and not a general property of the Shapley cost allocation for any class of cooperative games. Accordingly, we have that the Shapley allocation of any ISR game can be illustrated in two dimensional space. More precisely, the Shapley allocation γ (see Figure 2) is the midpoint of the core allocation segment.

Example 2 (Allocations in the ISR Scenario). Considering the presented scenario in Example 1, any cost allocation $\langle T_A, T_B\rangle$ such that $4 \leq T_A \leq 7$ and $8 \leq T_B \leq 11$ and $T_A + T_B = 15$ is a core allocation. Moreover, $\langle 5.5, 9.5\rangle$ is the Shapley allocation.

In general, Shapley allocation does not satisfy the *individual rationality* property. I.e., it might be the case that some agents in a collaborative group should pay higher than their traditional cost in order to guarantee the fairness property. In such cases, we have a fair albeit unstable collaboration because any sacrificing firm has incentive to rationally defect the collaboration. However, the next proposition shows that for ISR games, the Shapley allocation is individually rational.

Proposition 5 (Fairness and Stability). *Let $\Phi(\sigma)$ be the Shapley allocation for ISR game σ between firms A and B. It always holds that for $i \in \{A, B\}$, we have that $T_i^\Phi(\sigma) \leq T_i(\bar{\sigma})$.*

Proof: According to Proposition 4, the Shapley allocation of ISR game σ is also a core allocation. Hence, it also satisfies the *individual rationality* condition in Definition 4. ■

Based on proposition 5, in case the two firms agree to implement the Shapley allocation, it is guaranteed that the relation will be both fair and stable. Although implementing the Shapley allocation seems natural due to its desirable properties, firms may prefer to negotiate among cost allocations in the core with the aim to practice their bargaining power and enjoy more cost reduction. However, industrial agents that suffer from unfair allocations in such cases may defect the collaboration, leave/reject the ISR, and join other ISRs that are practicing fair allocation methods.

V. Conclusions and Future Work

In this work, we present a game-theoretical representation of Industrial Symbiotic Relations (ISRs) and tailor two types of solution concepts for cost allocation in such relations, i.e., *core allocation* and *Shapley allocation* for ISR games. These two notions can also be seen as two approaches for decision support while firms are faced with the collaboration decision, to reject or accept an ISR proposal. This is by enabling firms to systematically reason about and verify *stability* and *fairness* of a particular ISR. Range of stable collaborations, provided by the concept of *core*, allows further negotiation while the *Shapley* allocation leads to a uniquely fair solution. We then show that due to the characteristics of industrial symbiotic relations, ISR games can always be operationalized in both a fair and stable manner. In addition to practical contributions by providing managerial decision support tools, we introduced ISR games as a new class of two-person Operations Research (OR) games. In ISR games, we have the non-emptiness of the core and it is guaranteed that the Shapley value in this class of OR games is an individually rational solution. As a future work, we aim to analyze the validity of presented results using multiagent-based simulations [31]. We also plan to extend our game-theoretical analysis to network level, relate our notions to the concept of *willingness to cooperate* [32], and study Industrial Symbiotic Networks (ISNs). Due to complexities of ISNs (see [14]) guaranteeing *fairness* and *stability* in such networks calls for mechanisms to coordinate agent interactions [33] and governance platforms for, as discussed in [34], "administration of stakeholders by stakeholders".

Acknowledgment

The project leading to this work has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement *No. 680843*.

References

[1] M. R. Chertow, "Industrial symbiosis: literature and taxonomy," *Annual review of energy and the environment*, vol. 25, no. 1, pp. 313–337, 2000.

[2] D. R. Lombardi and P. Laybourn, "Redefining industrial symbiosis," *Journal of Industrial Ecology*, vol. 16, no. 1, pp. 28–37, 2012.

[3] D. M. Yazan, V. A. Romano, and V. Albino, "The design of industrial symbiosis: an input–output approach," *Journal of cleaner production*, vol. 129, pp. 537–547, 2016.

[4] D. W. Pearce and R. K. Turner, *Economics of natural resources and the environment*. JHU Press, 1990.

[5] T. Stock and G. Seliger, "Opportunities of sustainable manufacturing in industry 4.0," *Procedia CIRP*, vol. 40, pp. 536–541, 2016.

[6] L. M. Camarinha-Matos, H. Afsarmanesh, N. Galeano, and A. Molina, "Collaborative networked organizations - concepts and practice in manufacturing enterprises," *Computers & Industrial Engineering*, vol. 57, no. 1, pp. 46–60, 2009.

[7] L. Fraccascia, V. Albino, and C. A. Garavelli, "Technical efficiency measures of industrial symbiosis networks using enterprise input-output analysis," *International Journal of Production Economics*, vol. 183, pp. 273–286, 2017.

[8] T. Jackson and R. Clift, "Where's the profit in lndustrial ecology?" *Journal of Industrial Ecology*, vol. 2, no. 1, pp. 3–5, 1998.

[9] C. Yu, G. P. Dijkema, and M. Jong, "What makes eco-transformation of industrial parks take off in china?" *Journal of Industrial Ecology*, vol. 19, no. 3, pp. 441–456, 2015.

[10] T. S. Driessen, *Cooperative games, solutions and applications*. Springer Science & Business Media, 2013, vol. 3.

[11] P. Borm, H. Hamers, and R. Hendrickx, "Operations research games: A survey," *Top*, vol. 9, no. 2, p. 139, 2001.

[12] L. S. Shapley, "Markets as cooperative games," DTIC Document, Tech. Rep., 1955.

[13] H. P. Young, *Cost allocation: methods, principles, applications*. North Holland Publishing Co., 1985.

[14] V. Yazdanpanah, D. M. Yazan, and W. H. M. Zijm, "Normative industrial symbiotic networks: A position paper," in *Multi-Agent Systems and Agreement Technologies*. Springer, 2016, pp. 314–321.

[15] M. Bengtsson and S. Kock, "Coopetition in business networks: to cooperate and compete simultaneously," *Industrial marketing management*, vol. 29, no. 5, pp. 411–426, 2000.

[16] V. Albino, L. Fraccascia, and I. Giannoccaro, "Exploring the role of contracts to support the emergence of self-organized industrial symbiosis networks: an agent-based simulation study," *Journal of Cleaner Production*, vol. 112, pp. 4353–4366, 2016.

[17] D. C. Esty and M. E. Porter, "Industrial ecology and competitiveness," *Journal of Industrial Ecology*, vol. 2, no. 1, pp. 35–43, 1998.

[18] K. Sinding, "Environmental management beyond the boundaries of the firm: definitions and constraints," *Business strategy and the environment*, vol. 9, no. 2, p. 79, 2000.

[19] S. F. Magram, "Worldwide solid waste recycling strategies: A review," *Indian Journal of Science and Technology*, vol. 4, no. 6, pp. 692–702, 2011.

[20] E. M. Lovelady and M. M. El-Halwagi, "Design and integration of eco-industrial parks for managing water resources," *Environmental progress & sustainable energy*, vol. 28, no. 2, pp. 265–272, 2009.

[21] H. Zijm, M. Klumpp, U. Clausen, and M. Ten Hompel, *Logistics and Supply Chain Innovation: Bridging the Gap between Theory and Practice*. Springer, 2015.

[22] A. C. Carpenter and K. H. Gardner, "Use of industrial by-products in urban transportation infrastructure: argument for increased industrial symbiosis," in *Electronics and the Environment, 2008. ISEE 2008. IEEE International Symposium on*. IEEE, 2008, pp. 1–7.

[23] X. Chen, T. Fujita, S. Ohnishi, M. Fujii, and Y. Geng, "The impact of scale, recycling boundary, and type of waste on symbiosis and recycling," *Journal of Industrial Ecology*, vol. 16, no. 1, pp. 129–141, 2012.

[24] C. J. Dahlman, "The problem of externality," *The journal of law and economics*, vol. 22, no. 1, pp. 141–162, 1979.

[25] O. E. Williamson, "The economics of organization: The transaction cost approach," *American journal of sociology*, pp. 548–577, 1981.

[26] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[27] T. W. Sandholm, "Distributed rational decision making," *Multiagent systems: a modern approach to distributed artificial intelligence*, pp. 201–258, 1999.

[28] G. Dantzig, *Linear programming and extensions*. Princeton university press, 2016.

[29] R. W. Coff, "When competitive advantage doesn't lead to performance: The resource-based view and stakeholder bargaining power," *Organization science*, vol. 10, no. 2, pp. 119–133, 1999.

[30] A. P. Rosenbusch, "Fairness considerations in cooperative games," Ph.D. dissertation, Technische Universität, 2011.

[31] F. Dignum, J. Bradshaw, B. G. Silverman, and W. van Doesburg, *Agents for games and simulations: Trends in techniques, concepts and design*. Springer, 2009, vol. 5920.

[32] D. Yazan, J. Clancy, J. Lovett *et al.*, "Appendix: supply chains, techno-economic assessment and market development for second generation biodiesel." *Advances in biodiesel production: processes and technologies*, pp. 254–280, 2012.

[33] M. Dastani, L. van der Torre, and N. Yorke-Smith, "Commitments and interaction norms in organisations," *Autonomous Agents and Multi-Agent Systems*, pp. 1–43, 2015.

[34] M. P. Singh, "Norms as a basis for governing sociotechnical systems," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 1, p. 21, 2013.

Production Planning

Supply planning optimization for linear production system with stochastic lead-times and quality control

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

O. Ben Ammar, B. Bettayeb, A. Dolgui
SLP/LS2N – CNRS UMR 6004
Automation, Production and Computer Sciences Department.
IMT Atlantique, Bretagne-Pays de la Loire, Campus Nantes, France
{oussama.ben-ammar, belgacem.bettayeb, alexander.dolgui}@imt-atlatique.fr

Abstract— This work consider the problem of supply planning optimization of imperfect production systems with stochastic lead times and quality control. A model for supply planning of the production system and three quality control policies are analyzed. Experimental results highlights the economic advantage of integrating quality control planning at the early phase of supply planning optimization of production systems.

Keywords—Imperfect production systems; Uncertainty; Linear chain production process; Quality control.

I. Introduction and related publications

The growing need for more flexible and responsive production systems leads to several new approaches aiming to endow the production systems of the future with a certain intelligence, agility and resilience. In this context, one of the most important challenges to deal with is uncertainty and its impact on the performances of the system. In fact, uncertainty causes to manufacturers several difficulties in planning production, in regulating inventory, and in meeting demand [1].

In the following, the serial supply chain structures with stochastic lead times are reviewed. We present the case of deterministic demand for one and multilevel serial supply chains. Then, we focus on some relevant works related to production planning integrating quality control.

The simulation study presented by Whybark and Williams [2] was among the first to suggest that safety lead times may perform better than safety stocks in a multi-level linear production-inventory system when the production and replenishment times are stochastic.

A few years later, Weeks [3] proposed a one-stage model where it is assumed a deterministic demand, a stochastic processing time and tardiness and holding costs. This problem has been proved to be equivalent to the standard "Newsboy" formulation.

Another analytical approach is proposed by Yano [4] to model serial production systems in order to optimize the planned lead times. In this model, which is based on a Lot-for-Lot (L4L) policy, it is assumed that procurement and processing times are stochastic and demand is deterministic. The objective function is composed of the holding and tardiness costs. A procedure is developed to generate solutions for two-stage systems. The same author extended this model to three-stage systems in [5]. In this model, a replenishment cost is introduced each time a given component is available after a planned intermediary due date. The performance indicator to be minimized is the sum of the following costs: i) the expected holding and tardiness costs at each stage, and ii) the expected tardiness cost for the finished product. The author underlined the difficulty to model a system composed of more than three stages. It is only after many year that Elhafsi [6] overcome this barrier by proposing an analytical model based on recursivity. The convexity of the objective function is proved and a dynamic programming based heuristic is proposed to find approximate solutions of good quality in an acceptable computer processing time.

In inventory control process, Kim et al. [7] consider one product, a known fixed demand and an Erlang-distributed lead times. The performance indicator is the expected total cost, which is composed of ordering, holding and tardiness costs. Based on the analytical formulation, the authors propose an approximate solution and compare it to optimal ones for the case where the prior information on the lead time distribution is available, and another case where no information exists. The results show the effectiveness and the robustness of this approach and how costs could be reduced under uncertain lead times. Two years later, Kim et al. [8] studied the same problem and analysed the bullwhip effect which is encountered when both demand and lead times are stochastic. An analytical method is used to prove that prior information on lead times variability could be very helpful to control the bullwhip effect in a multi-stage supply chain.

Apparently, the literature contains few publications considering linear chain supply planning with discrete lead times. The most of existing studies investigates the lead time uncertainty for one item or assembly systems. Readers can refer to [9], [10] and [11] where more complete literature reviews are provided concerning supply planning dealing with uncertainty on lead times. For other sources of uncertainty (demand, capacity, cost, etc.) readers can refer to the surveys of Aloulou et al. [12] and Díaz-Madroñero et al. [13].

In the last few decades, there has been a growing interest in integrating quality control in supply and production planning. In fact, the quality should be rigorously checked before delivering the finished product. In this way, the company ensures customer satisfaction and avoid excessive product returns. Nevertheless, it will be necessary to find a good compromise between the inspection costs and the cost of non-conformity. In this regard, Colledani and Tolio [14] investigate the interaction between

quality control system and production system and show how they impact each other's performances. They underline the necessity of jointly considering quality and logistics requirements while designing a production system.

The work of Rosenblatt and Lee [15] studies the relationship and the economic impact of the interaction between production planning, quality of products and the deterioration of the production process. The considered production process has two possible states in which it generates different rates of defectives. The objective is to determine the economic production quantities of this kind of production process while minimizing the total annual cost. The optimal production run time is determined for two deterioration models and multi-state deterioration.

Recently, Bettayeb et al. [16] proposed an integrated model for single item lot sizing and quality control planning. The objective is to minimize the total cost while ensuring a given level of outgoing quality. This cost is composed of holding, setup, production and inspection costs.

In the field of supply planning and quality control few papers analyses the uncertainty of lead time and quality control. In this work, we propose a model of supply planning optimization of imperfect production systems with stochastic lead times and quality control. The rest of the paper is organized as follows. The problem description and the general assumptions are described in the second section. The third one details the analytical models corresponding to three different quality control policies and the optimization approach. Section 4 presents the experimental results, compares the economical performances of each model and analysis the effects of some parameters. The paper ends by giving the conclusion and the perspectives of this work.

II. Problem description and assumptions

The problem under consideration in this paper concerns the optimization of supply planning of an imperfect linear production process where the lead times of the production steps are stochastic. The objective is to evaluate the effect of quality issues on the total cost of such a system and analyze the opportunity of integrating quality control activities in the optimization model of the system.

Before detailing the problem and the general assumptions, the following are the notations used in next sections. All costs are expressed in monetary unit and durations are expressed in time units.

Parameters

C_P Production cost

c_i Unit inspection cost

c_{nc} Unit non-conformity cost

C_Q Quality cost: the sum of inspection and non-conformity costs

SR Sampling rate, i.e. proportion of products inspected

c_h Unit inventory holding cost per time unit

c_t Unit tardiness penalty per time unit

T Due date

q_0 Proportion of non-conforming finished products generated by the whole production process.

m Number of production operations (flowshop)

L_j Lead time of operation j, $\forall j \in \{1,2,\dots,m\}$ L_j is a discrete random variable which varies between l_j and u_j

t_i Unit inspection duration

L Total lead time, i.e. $L = \sum_{j=1}^{m} L_j$

Decision variables

X Order release date

Functions

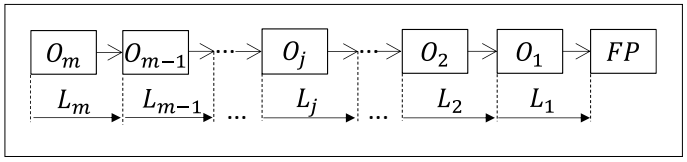
$E[\![.]\!]$ Expected value of $[\![.]\!]$

$F_m(.)$ Cumulative distribution function of L

The considered production process is composed of m process operations (Fig. 1), which are consecutively executed to obtain the finished product FP. Each process step j has a random lead time duration L_j. The product starts to be processed at time X and the fished product is available at $T_{PF} = X + L$, after accumulating the m random lead times $L = L_1 + L_2 + \cdots + L_m$. Note that T_{PF} is a random variable because of the randomness of the total lead time L.

Apart from the processing cost (omitted here), and depending on the due date T and the effective release date of the finished product, the generated production cost either equals the holding cost, if the product is finished before the due date, equals the tardiness penalty if it is available after the due date or equals zero if it is finished just in time.

Fig. 1. Linear production process



III. Analytical models and optimization approach

This section contains three models each of them corresponds to a certain policy regarding the quality control. In the first case, only the non-conformity penalty is assumed and no quality control is performed. In the second case, the quality control is performed at the end of production process but only the lead times are taken into account in the supply and production planning. In the last case, the quality control and the lead time variability are both taken into account while optimizing the supply and production planning.

A. Reference Policy (π_0): No quality control

It corresponds to the case where no quality control is performed during the production process. However, for each non-conforming product delivered to the customer, the producer pays a non-conformity penalty. The total cost is composed of inventory holding cost, tardiness cost and non-conformity cost:

$$TC^0(X,L) = C_P^0 + C_Q^0 \quad (1)$$

Where C_P^0 is equal to the sum of C_H the inventory holding cost and C_T the tardiness cost:

- If the finished products are available before T, it will be stored. The corresponding inventory holding cost C_H is equal to:

$$C_H = c_h \times (T - min(T; X+L)) \quad (2)$$

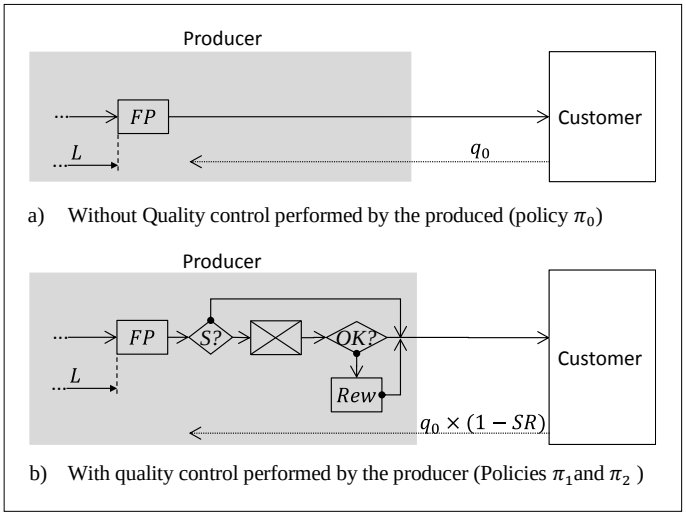
- There will be a stockout of the finished products if they are delivered after the due date T. Then tardiness cost C_T is equal to:

$$C_T = c_t \times (max(T; X+L) - T) \quad (3)$$

And the non-conformity penalty C_Q^0 is equal to $q_0 \times c_{nc}$.

The total cost $TC^0(X,L)$ is a discrete random variable because of the randomness of the lead times L_j. Each of the latter varies between l_j and u_j, so the total cost has a finite number of possible values.

Fig. 2. Quality control policies



a) Without Quality control performed by the produced (policy π_0)

b) With quality control performed by the producer (Policies $\pi_1$and π_2)

Proprety 1. *The expression of the expected value of the total cost, denoted by* $E[\![TC^0(X,L)]\!]$*, is given bellow:*

$$E[\![TC^0(X,L)]\!] = c_h \times \left(T - \sum_{0 \le s \le T-1} \big(1 - F_m(s-X)\big)\right) + c_t \times \sum_{s \ge T} \big(1 - F_m(s-X)\big) + q^0 \times c_{nc} \quad (4)$$

Proof. From expressions (1), (2) and (3), we have:

$$\begin{aligned} E[\![TC^0(X,L)]\!] &= E[\![C_P^0(X,L)]\!] + q_0 \times c_{nc} \\ &= c_h \times (T - E[\![min(T; X+L)]\!]) \\ &\quad + c_t \times (E[\![max(T; X+L)]\!] - T) + q_0 \times c_{nc} \end{aligned}$$

Let Γ a positive discrete random variable (with a finite range). Its expected value can be expressed as follows:

$$E[\![\Gamma]\!] = \sum_{s \ge 0} (1 - Pr[\![\Gamma \le s]\!]) = \sum_{s \ge 0} \big(1 - F_\Gamma(s)\big) \quad (5)$$

Knowing that $X + L$ does not depend on T, and using expression (5), we get:

$$E[\![max(T; X+L)]\!] = \sum_{s \ge 0} (1 - Pr[\![T \le s]\!] \times Pr[\![X + L \le s]\!])$$

Moreover, knowing that $Pr[\![T \le s]\!] = 0$ for all $0 \le s < T$, then:

$$\begin{aligned} E[\![max(T; X+L)]\!] &= T + \sum_{s \ge T} (1 - Pr[\![X + L \le s]\!]) \\ &= T + \sum_{s \ge T} \big(1 - F_m(s - X)\big) \end{aligned} \quad (6)$$

Where F_m is the cumulative distribution function of the total lead time L, which defined as bellow:

- For $\alpha = 0$:

$$F_m(s - X) = Pr[\![X + L_m \le s]\!]$$

- For $1 \le \alpha \le m - 1$; $m' = m - \alpha$:

$$F_{m'}(s - X) = \sum_{\substack{u_\alpha + v_\alpha = s \\ u_\alpha + v_\alpha \in \mathbb{Z}}} Pr[\![L_{m'} = u_\alpha]\!] \times F_{m'+1}(v_\alpha - X)$$

In the same way, we can easily deduce that:

$$E[\![min(T; X+L)]\!] = \sum_{0 \le s \le T-1} \big(1 - F_m(s-X)\big) \blacksquare \quad (7)$$

The expected total cost (4) is the objective function. It is a non-linear function which should be minimized. An exact method based on Newsboy formula is used to solve this problem in polynomial time.

Proposition 1. *The method based on Newsboy formula presented in [17] gives the optimal order release date of the reference case* X_0^**:*

$$F_m(T - X_0^* - 1) \le \frac{c_t}{c_h + c_t} \le F_m(T - X_0^*) \quad (8)$$

Where $F_m(.)$ *is the cumulative distribution function of the total lead time L.*

B. Policy 1 (π_1): Separated afterwards quality control planning

In this case, we suppose that the quality inspection is done just after the production process. A proportion of finished products is randomly sampled and inspected. The non-conforming products detected by inspection are repaired with no-extra cost.

The total cost is composed of production and quality costs, denoted by C_P^1 and C_Q^1, respectively.

$$TC^1(X,L,SR) = C_P^1 + C_Q^1 \quad (9)$$

Note that, in this case, C_P^1 is the real production cost. It is equal to $C_P^{0*} - \Delta_h + \Delta_t$, where:

- C_P^{0*} is the optimal planned production cost associated to optimal order release date X_0^*,
- Δ_h is the holding cost reduction corresponding to the time the product is being inspected :

$$\Delta_h = c_h \times max(min(I; T - X - L); 0) \quad (10)$$

Where $I = i_u \times SR$ corresponds to the total duration of inspection that corresponds to the quality control strategy in place.

- Δ_t is the tardiness cost increase due to the product inspection :

$$\Delta_t = c_t \times \big(I - max(min(I; T - X - L); 0)\big) \quad (11)$$

The quality cost C_Q^1 is composed of the cost of inspection and the penalty paid by the producer to the customer for non-conforming units:

$$C_Q^1 = q_0 \times (1 - SR) \times c_{nc} + SR \times c_i$$

As in the previous case, the total cost is also a random discrete random variable with a finite range. The expected cost is derived by the following property.

Proprety 2. *The expression of the expected value of the total cost, denoted $E[\![TC^1(X_0^*, L, SR)]\!]$, is given bellow:*

$$E[\![TC^1(X_0^*, L, SR)]\!] = E[\![TC^0(X_0^*, L)]\!] - \left[(c_h + c_t) \times \sum_{s \le I-1} F_m(T - s - 1 - X_0^*) - c_t \times I\right] - SR \times (c_{nc} \times q_0 - c_i)$$

Proof. From expression (4) and (7), $E[\![TC^0(X_0^*, L)]\!]$ can be easily calculated. From expressions (10) and (11), we can easily deduce that:

$$E[\![\Delta_h - \Delta_t]\!] = \sum_{s \le I-1} F_m(T - s - 1 - X_0^*) \blacksquare$$

C. Policy 2 (π_2): Integrated production and quality control planning

In this case, we suppose that the quality control is planned in advance. It is made after the production process. A proportion of finished products is randomly sampled and inspected. The non-conforming products detected by inspection are repaired with no-extra cost.

The total cost is composed of production and quality costs, denoted by C_P^2 and C_Q^2, respectively.

$$TC^2(X, L, SR) = C_P^2 + C_Q^2 \quad (12)$$

Where:

$$C_P^2 = [c_h \times (T - min(T; X + L + I)) + c_t \times (max(T; X + L + I) - T)]$$

$$C_Q^2 = q_0 \times (1 - SR) \times c_{nc} + SR \times c_i$$

The total cost $TC^2(X, L, SR)$ is a discrete random variable because of the randomness of lead times L_j. Each of the latter varies between l_j and u_j. Therefore, this total cost has a finite number of possible values. Thus, the mathematical expectation of cost can be determined.

Proprety 1. *The expression of the expected value of the total cost, noted $E[\![TC^0(X, L)]\!]$, is given bellow:*

$$E[\![TC^2(X, L, SR)]\!] = c_h \times \left(T - \sum_{0 \le s \le T-1} \big(1 - F_m(s - X - I)\big)\right) + c_t \times \sum_{s \ge T} \big(1 - F_m(s - X - I)\big) + q_0 \times (1 - SR) \times c_{nc} + SR \times c_i \quad (13)$$

Proof. From expressions (6), (7) and (10), we can deduce:

$$E[\![max(T; X + L + I)]\!] = T + \sum_{s \ge T} \big(1 - F_m(s - X - I)\big)$$

And

$$E[\![min(T; X + L)]\!] = \sum_{0 \le s \le T-1} \big(1 - F_m(s - X - I)\big) \blacksquare$$

Proposition 2. *The method based on Newsboy formula presented in [17] gives the optimal order release date of the reference case X_2^*:*

$$F_m(T - X_2^* - 1 - I) \le \frac{c_t}{c_h + c_t} \le F_m(T - X_2^* - I) \quad (14)$$

Where $F_m(.)$ is the cumulative distribution function of the total lead time L.

IV. Numerical example and discussion

The proposed cases described in Section 3 have been coded in C++. The experiments are carried on computer with 2.32 GHz Intel core i7 and 8 GB of RAM memory.

A. Experiments setting

The experiments are based on an example of a linear production system with three production steps, i.e. $m = 3$. The due date T is equal to 20 and the lead time of each process step is a discrete random variable uniformly distributed over the range of integers from 1 to 5, i.e. $\Pr(L_j = l) = 0.2\ \forall j = 1,2,3\ \forall l = 1,2, \dots, 5$. It is assumed that the system is operating in a steady state, generating a constant proportion of non-conforming finished products $q_0 = 0.05$. The unit holding cost c_h is equal to the monetary unit and the other cost parameters experimented are given in Table 1.

Table 1. Experimented values for each parameter

Parameter	Range
L_j	$[1, 5]$
$\rho_t = c_t/c_h$	$\{0.1, 1, 10, 100, 1000\}$
$\rho_i = c_i/c_h$	$\{0.1, 1, 10, 100, 1000\}$
$\rho_{nc} = c_{nc}/c_h$	$\{0.1, 1, 10, 100, 1000\}$

SR	$\{0, 0.2, 0.4, \dots, 1\}$

B. Results and discussions

The preliminary results presented in this section aim to analyze the effect of integrating quality control in the early stage of supply and production planning. To do so, the following notion are introduced:

- The gap between policy π_1 and policy π_3

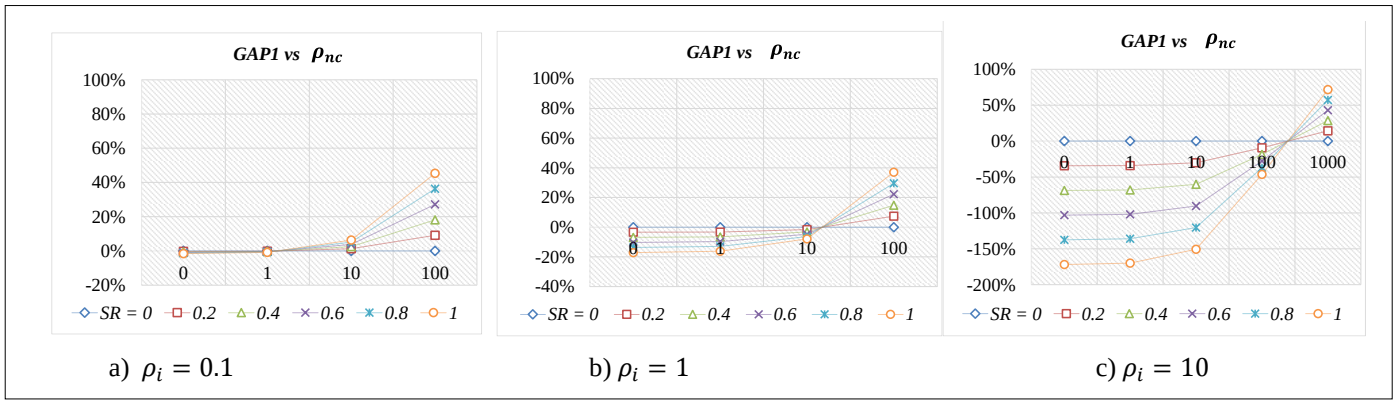
$$GAP1\ (\%) = 100 \times \frac{ETC^1 - ETC^3}{ETC^1}$$

- The gap between policy π_2 and policy π_3

$$GAP2\ (\%) = 100 \times \frac{ETC^2 - ETC^3}{ETC^2}$$

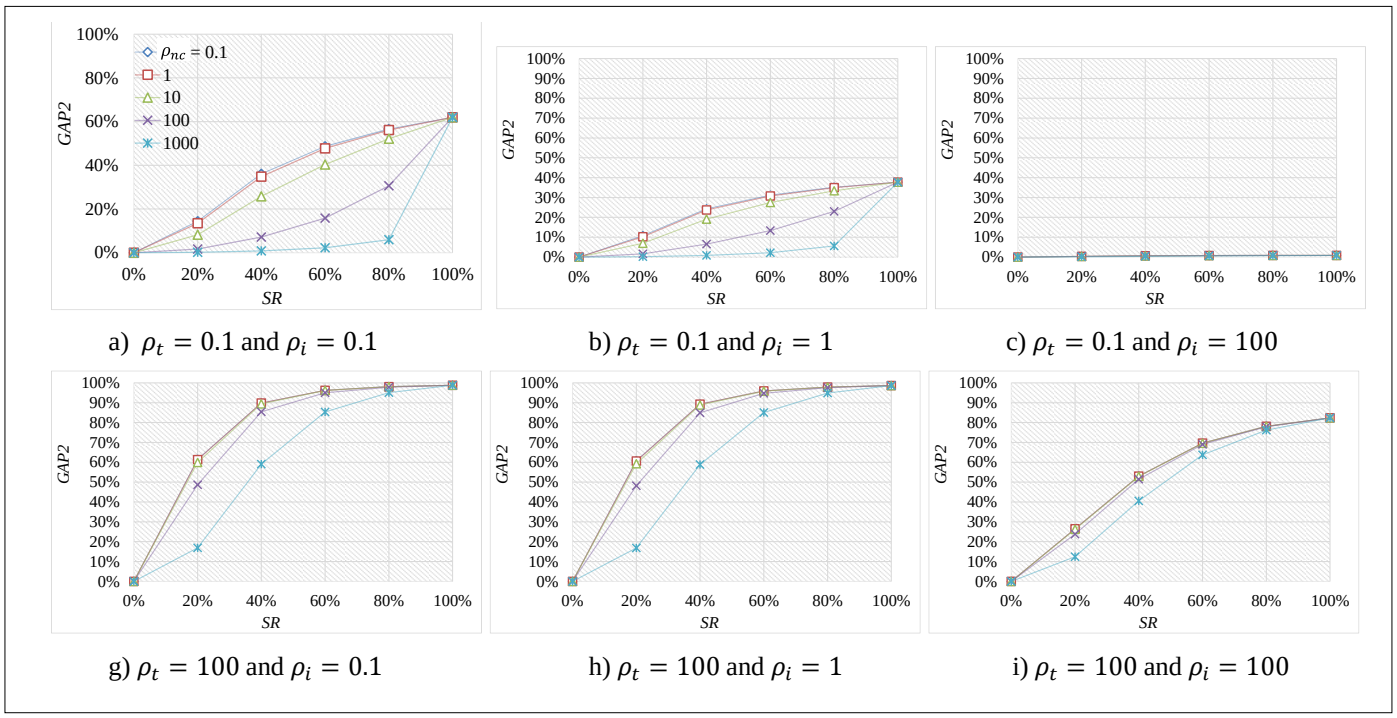
Figure 3 shows the variation of $GAP1$ as function of the cost of non-conformity for different values of sampling rate and for three values of the inspection cost. It can be seen that when the inspection cost is low compared to the holding cost ($\rho_i = 0.1$), $GAP1$ is positive whatever is the cost of non-conformity. Moreover, when the latter increases, this gap increases more significantly with the sample rate. However, when the inspection cost is equal to the holding cost ($\rho_i = 1$), there exists a certain value of non-conformity cost below which the policy 2 becomes more profitable. From Figure 3 c) we can conclude that there exists a certain value of inspection cost beyond which Policy 3 is always dominated by Policy 2.

Fig. 3. Impact of the inspection cost on the Gap between ETC^3 and ETC^1



a) $\rho_i = 0.1$ b) $\rho_i = 1$ c) $\rho_i = 10$

Fig. 4. Gap between ETC^2 and ETC^1 for some combinations of cost parameters



a) $\rho_t = 0.1$ and $\rho_i = 0.1$ b) $\rho_t = 0.1$ and $\rho_i = 1$ c) $\rho_t = 0.1$ and $\rho_i = 100$

g) $\rho_t = 100$ and $\rho_i = 0.1$ h) $\rho_t = 100$ and $\rho_i = 1$ i) $\rho_t = 100$ and $\rho_i = 100$

Figure 4 compares the performances of Policy 2 and Policy 3 by analyzing the gap between them. It proves that integrating quality control in supply and production planning reduces significantly the expected total cost whatever is the number of inspected products. We note that when inspection cost is much higher than the tardiness cost (Figure 4 c)), the policies are equivalent.

V. Conclusion

In this work, we are interested in supply and production planning under uncertainty of lead times and quality control. A model for supply planning of the production system and three quality control policies are analyzed. Experimental results highlights the economic advantage of integrating quality control planning at the early phase of supply planning optimization of production systems.

This exploratory study confirms the opportunity to extend this model by integrating maintenance activities and considering other kinds of production systems such as assembly systems where the finished products are assembled from several components.

Further, the second objective is to extend this model and different proposed technics to multi-period planning, in particular, calculate planned lead times when such a company deals with quality control, uncertainties of production and supply lead times.

References

[1] C. S. Tang, "The Impact of Uncertainty on a Production Line", Management Science, 1990, 36(12), pp. 1518-1531.

[2] D. C. Whybark and J. G. Williams, "Material Requirements planning under uncertainty", Decision Science, 1976, 7(4), pp. 595-606.

[3] J. Weeks, "Optimizing planned lead times and delivery dates", In 21st annual Conference Procceding, Americain Production and Inventory Control Society, 1984, pp. 177-188.

[4] C. A. Yano, "Setting Planned Leadtimes in Serial Production Systems with Tardiness Costs", Management Science, 1987, 33(1), pp. 95-106.

[5] C. A. Yano, "Planned Lead-times for Serial Production Systems", IIE Transactions, 1987, 19(3), pp. 300-307.

[6] M. Elhafsi, "Optimal leadtimes planning in serial production systems with earliness and tardiness costs", IIE Transactions, 2002, 34, pp. 233-243.

[7] J. G. Kim, D. Sun, X. J. He and J. C. Hayya, "The (s,Q) inventory model with Erlang lead time and deterministic demand", Naval Research Logistics, 2004, 51(6), pp.906-923.

[8] J. G. Kim, D. Chatfield, T. P. Harrison and J. C. Hayya, "Quantifying the bullwhip effect in a supply chain with stochastic lead time", European Journal of Operational Research, 2006, 173(2), pp. 617-636.

[9] A. Dolgui, O. Ben-Ammar, F. Hnaien and M. A. Ould Louly, "A State of the Art on Supply Planning and Inventory Control under Lead Time Uncertainty", Studies in Informatics and Control, 2013, 22(3), p. 255-268.

[10] A. Dolgui and C. Prodhon, "Supply planning under uncertainties in MRP environments: A state of the art", Annual Reviews in Control, 2007, 31(2), pp. 269–279.

[11] V. Guide and R. Srivastava, "A review of techniques for buffering against uncertainty with MRP systems", Production Planning & Control, 2000, 11(3), pp. 223-233.

[12] M. A. Aloulou, A. Dolgui and M. Y. Kovalyov, "A bibliography of non-deterministic lotsizing models", International Journal of Production Research, 2014, 52(8), pp.1-18.

[13] M. Díaz-Madroñero, J, Mula and D. Peidro, "A review of discrete-time optimization models for tactical production planning", International Journal of Production Research, 2014, 52, pp. 5171-5205.

[14] M. Colledani and T. Tolio, "Impact of Quality Control on Production System Performance," CIRP Ann. - Manuf. Technol., vol. 55, no. 1, pp. 453–456, Jan. 2006.

[15] M. J. Rosenblatt, and H. L. Lee, "Economic Production Cycles with Imperfect Production Processes", IIE Transactions, 1986, 18(1), pp. 48-55.

[16] B. Bettayeb, N. Brahimi, and D. Lemoine, "Integrated Single Item Lot-Sizing and Quality Inspection Planning," IFAC-PapersOnLine, vol. 49, no. 12, pp. 550–555, 2016.

[17] O. Ben-Ammar, H. Marian, A. Dolgui, and D. Wu, "Reducing the research space of possible order release dates for multi-level assembly systems under stochastic lead times". Advances in Production Management Systems: Innovative and Knowledge-Based Production Management in a Global-Local World, Part III, Springer Series, 2014, vol. 440, pp. 368-374.

Optimization of production and pricing strategy for a profit-maximization model

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Yan-Chun Chen* / Associate Professor
Department of Industrial Management
Tungnan University
New Taipei City, Taiwan
*Corresponding Author: yjchen@mail.tnu.edu.tw

Abstract—In this paper, we formulate a profit-maximization model for a single manufacturer and a single retailer by considering the channel coordination and the manufacturer's imperfect production system with minimal repairs to joint determination of inspection, production, pricing and ordering for a single item. The manufacturer's imperfect production system may shift to the out-of-control state and produces non-conforming items. In the coordination of replenishment, this study is based on a Stackelberg leadership model: the manufacturer is the leader who designs a replenishment cycle considering the maximization of the channel profit, and the retailer is the follower who determines the optimal selling price in accordance with the replenishment cycle. In a production-distribution system with production process deterioration, we derive the optimal solutions for the inspection frequency, inspection interval, economic production quantity and the selling price. And numerical analyses are used to simulate and analyze the effect of changes in the various parameters on the expected unit profit.

Keywords—inventory; inspection; economic production quantity; imperfect production system; channel coordination; integrated model

I. Introduction

In recent years, in the globalized competitive production environment, supply chain management not only requires cost saving and enhances the profitability, but also seeks for the overall optimization of production and distribution system. The decision model of supply chain should be changed from the cost management into the revenue management to satisfy the customer needs timely (Chen & Chen [1]). Consequently, in the context of dynamic environment, it is very important to integrate the resources of manufacturers and retailers. Numerous researches mainly target at the inventory and pricing strategy of retailers and only adopts simple cost model for suppliers. In order to shape the model in accordance with the actual production status, it needs to put the best testing, production and pricing strategy and related costs into cost consideration in the context of maximized profits of production-marketing coordination. As a consequence, this study considers a two-stage channel structure of manufacturers and retailers to seek for the best production strategy, optimal pricing strategy and purchasing strategy for single products.

The classical Economic Production Quantity (EPQ) assumes that the production system is perfect and the processing equipment does not fail and the product quality in production meet standards and will not turn bad over time. Along with the time and the increase of quantity in production, the processing probably is involved in non-controlling status and generates the deficit products, which can be regarded as an imperfect processing. Many researches promoting this traditional economic production batch mode have eased many restrictions to make this model fit in accordance with the actual production status (Rosenblatt & Lee [2], Hariga & Ben-Daya [3]). Freimer et al. [4] studied the effect of the establishment and improvements in cost and processing procedures on the economic production batch mode. Ma et al. [5] considered the impact of an imperfect process with scrap and shortage issues on the economic lot scheduling model and discussed three inspection strategies which include no screening, after-production screening and in-production screening. Chiu et al. [6] explored the economic batch and replenishment strategy during the generation of deficit products and the best economic production batch and replenishment strategy with lowest cost. Peng and Houtum [7] proposed the influence of the maintenance strategy in the best economic batch model. When the cost of unexpected failures is high, the cost savings are quite considerable in the model. The adoption of update theory can trace the long-term average growth ratio of degenerative manufacturing system and find out the best economic production batch and maintenance under the condition of minimized long-term average cost.

Preventive maintenance (PM) can enhance system reliability and reduce the incidence of failure. Pham and Wang [8] argued that the PM of a production system was usually imperfect and reviewed the related literature. Ben-Daya [9] proposed model that integrated EPQ and PM level; the model considered the optimal inspection interval, inspection frequency, and production quantity in the case of a progressively increasing failure rate with a degenerate distribution. Chen [10] extended the Ben-Daya's [9] results to develop an integrated model covering both EPQ and inspection while also exploring how minimal repair, inspection errors and permitted shortage conditions on cost. Saad et al. [11] considered the optimal PM planning and production lot strategy of a single production process to produce a product. Sana [12] studied an imperfect production system with allowable shortages due to regular PM for products sold with free minimal repair warranty. He determined the optimum buffer level and production run time for the imperfect production system with minimum expected

total cost. Chen [13] considered the integrated problem of production, PM, inspection, and inventory for an imperfect production process in which rework and PM error exist. For a deteriorating production system, the paper optimized inspection interval, inspection frequency, and production quantity in terms of maximum profit per unit. Lee and Cha [14] considered the optimal PM problem based on a new counting process. The paper developed two types of PM models and studied detailed properties of the optimal policies which minimize the long-run expected cost rates.

The multi-layer channel coordination mainly targets at the coordination and decision issues of physical flows, funds flow, information flow of production and marketing system. The purpose is to shape the maximized profits for all members in overall channels. Netessine and Rudi [15] explored the optimal pricing strategy needed to maximize profits, taking into consideration the purchasing and transportation costs for suppliers and retailers at the same time. Ho et al. [16] proposed to construct the integrative inventory model of suppliers and retailers to explore the best order quantity, reorder point, lead time and delivery times. Huang et al. [17] explored the problems related to pricing of dual-channel supply chain and production decision when the production cost changes, and carried out researches on revising the original production plan and price when the research costs change. When the research production cost exceeds the threshold, it will change the production quantity and pricing in the context that the purchasing quantity of every retailer is consisted of deficits and forward lead time, which are subject to general distribution and partial out of stock is allowed. Protopappa-Sieke et al. [18] investigated the supply chain of numerous retailers and service standards and proposed a balanced method of service level cooperation and found that the manufacturers and retailers exert significant influence on the profit of the best inventory allocation policy.

Therefore, this study aims at the two-stage channel structure of single manufacturer and single retailer under the condition that the production system of the manufacturer is not perfect and requires minor revisions and in the context of maximized profit of production-marketing coordination and explores the best testing, production and maintenance strategy, the best pricing strategy and purchasing strategy. The purpose of this study is to construct an integrated model in the context of imperfect processing, maintenance and production-marketing coordination and to construct an integrated model to explore the influence of specific parameter changes on the expected profits of channels with numerical analysis.

II. Mathematics Model

Considering a two-stage production and marketing system in this study, the members are consisted of single manufacturer and retailer to sell single item to the ultimate consumer. At the beginning of the production cycle, the system is assumed to be in the in-control state and therefore producing items of good quality. The process is inspected at time t_j , $j = 1,2,....,k$. After each inspection, the process is either in an in-control state or out-of-control state. If the process is in the in-control state, PM is performed after inspection. The age of the system after PM will be correlated with the level of PM. If the process is in the out-of-control state, there are two possible types of process shifts. One is a relatively mild production process shift, where minimal repair can restore the process to the in-control state. The other is a more serious shift, and the process needs to be replaced.

Regarding the integration strategy of production-marketing system, this study adopts Vendor Managed Inventory (VMI) strategy for single shop to manage the replenishment strategy of retailers. However, retailers reserve the rights of retail pricing to realize the maximized profits of production and marketing system.

To facilitate comparison, this study used the same symbols as Chen's [13] paper:

π : Expected channel profit in unit time

T_m : Production cycle of the manufacturer

ρ : Production rate in units per unit of time of the manufacturer

C_s : Setup costs per production cycle of the manufacturer

h_m : Storage cost of the manufacturer for each piece produced in unit time

C_I : Cost of each inspection

C_m : Cost for the manufacturer to produce one product item

C_d : Cost of production of each scrapped, non-conforming item

k : The frequency of inspection during each production cycle

h_j : Arrival interval of the *j*th inspection

t_j : *j*th PM time point, where $t_j = \sum_{i=1}^{j} h_i$

θ : The probability of the second possibility applying when the production process is in the out-of-control state

b_i : Actual age of the system before the *i*th PM

a_i : Actual age of the system before the *i*th PM

$I_m(t)$: Inventory level of the manufacturer at time *t*

C_{apm} : Cost of PM that is actually implemented

C_{mpm} : Cost of PM during the maximal state

p : The probability that the process shifts to the out-of-control state

π_m : Expected profits of the manufacturer in unit time

S_r : Unit order cost for the retailer

h_r : Storage cost of the retailer for each piece produced in unit time

C_r : Cost for the retailer to purchase one product item

P_r : Sales price for the retailer for each item

$I_r(t)$: Inventory level of the retailer at time *t*

$D(P_r)$: Consumption of the retailer in unit time

π_r : Expected profits of the retailer in unit time

Suppose that the production cycle of the manufacturer is equal to the replenishment cycle of the retailer, the two-stage production and marketing channel inventory level is shown in figure 1. The horizontal axis is time and the vertical axis refers to the inventory level ($I(t)$).

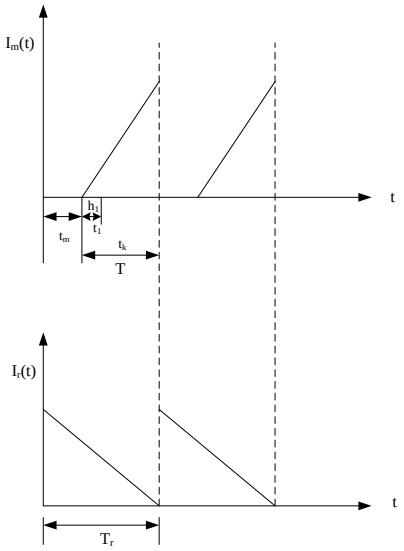


Fig. 1. Inventory level of two-stage supply chain.

For the manufacturer, the total expected cost of each cycle includes set-up cost (C_S), holding cost ($E(HC)$), maintenance cost ($E(M)$), inspection cost ($E(IC)$), cost of manufacturing scrapped non-conforming items ($E(DC)$) and repair delay cost ($E(RC)$). To derive these costs requires first deriving the expected production time ($E(T)$). Therefore, the expected production time is

$$E(T)=\sum_{j=1}^{k} h_j(1-\theta p)^{j-1} \tag{1}$$

The expected storage cost is based on the multiple of the holding cost of unit time inventory and the holding expected time storage quantity of a single inventory cycle, the area under curve from the starting time point of production cycle to the ending time of inventory sales as shown in Figure 1. The expected storage cost is as follows:

$$E(HC)=(\frac{C_h}{2})\rho\left[\sum_{j=1}^{k} h_j(1-\theta p)^{j-1}\right]^2 \tag{2}$$

Following PM, the system will become younger despite its age not returning to a new state. The change in system age is correlated with the maintenance degree. Let

$$\gamma_k=\eta^{k-1}\frac{C_{apm}}{C_{mpm}} \tag{3}$$

The parameter η ($0\le\eta\le1$) is an imperfect factor expressing the deterioration of system age because of the performance of PM. Ben-Daya [19] suggested that the correlation between reduction in age and PM was either linear or nonlinear. This investigation assumes the correlation is linear, yielding:

$$a_k=(1-\gamma_k)b_k \tag{4}$$

This study considers the expected maintenance costs. During the whole production cycle, the maintenance cost is consisted of PM and minor revision cost of production system. After each test, this study carries out PM activities. The system needs to be stopped. The expected maintenance cost ($E(M)$) is:

$$E(M)=C_{apm}\sum_{j=1}^{k-1}(1-\theta p)^j+(1-\theta)pC_{mr}\sum_{j=1}^{k-1}(1-\theta p)^{j-1} \tag{5}$$

From the above $E(M)$ express, the expected times of each production cycle PM is $n_{pm}=\sum_{j=1}^{k-1}(1-\theta p)^j$ and the expected times of minor revision of each production cycle is $n_{mr}=(1-\theta)p\sum_{j=1}^{k-1}(1-\theta p)^{j-1}$. Because the PM is conducted following each inspection and is not executed at the end of the cycle, the inspection cost is $E(IC)=(n_{pm}+1)C_I$. Where $E(N_j^1)$ ($E(N_j^2)$) is the expected number of non-conforming items produced due to the first (second) type out-of-control during the period (t_{j-1},t_j) . These two expected number of non-conforming items are given by

$$E(N_j^I)=\frac{\lambda d_I\rho(1-\theta)}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}\left[\frac{r_{j-1}b_j}{(1-\theta)\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}-\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}+\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-\theta)b_j}\right] \tag{6}$$

$$E(N_j^{II})=\frac{d_{II}\rho\theta}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}\left[\frac{r_{j-1}b_j}{\theta}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}-\frac{1}{\theta^2\lambda}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}+\frac{1}{\theta^2\lambda}e^{-\lambda\theta b_j}\right] \tag{7}$$

The expected number of non-conforming items ($E(N)$) during one cycle is

$$E(N)=p\sum_{j=1}^{k}(1-\theta p)^{j-1}\left\{\frac{\lambda d_I\rho(1-\theta)^2}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}\left[\frac{r_{j-1}b_j}{(1-\theta)\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}-\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}+\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-\theta)b_j}\right]\right.$$

$$+\frac{d_{II}\rho\theta^2}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}[\frac{r_{j-1}b_j}{\theta}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}-\frac{1}{\theta^2\lambda}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}+\frac{1}{\theta^2\lambda}e^{-\lambda\theta b_j}]\}$$

(8)

The cost of manufacturing scrapped non-conforming items is

$$E(DC)=C_d p\sum_{j=1}^{k}(1-\theta p)^{j-1}\{\frac{\lambda d_I\rho(1-\theta)^2}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}[\frac{r_{j-1}b_j}{(1-\theta)\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}$$
$$-\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-r_{j-1})(1-\theta)b_{j-1}}+\frac{1}{(1-\theta)^2\lambda}e^{-\lambda(1-\theta)b_j}]$$
$$+\frac{d_{II}\rho\theta^2}{(e^{-\lambda(1-r_{j-1})b_{j-1}})^{1-\theta}}[\frac{r_{j-1}b_j}{\theta}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}-\frac{1}{\theta^2\lambda}e^{-\lambda(1-r_{j-1})\theta b_{j-1}}+\frac{1}{\theta^2\lambda}e^{-\lambda\theta b_j}]\}$$

(9)

Ben-Daya [9] proposed that during the period of a_{j-1},b_j, the repair delay cost of the second failure RC_j can be assumed as one-time linear function of delay repair time. That is to say, $RC_j=R(b_j-t)=r_0+r_1(b_j-t)$, in which r_0 refers to fixed repair cost, r_1 refers to the variation repair cost of delay repair, which are stable constants. The expected revision delay cost of a production cycle is

$$E(RC)=\theta p\sum_{j=1}^{k}(1-\theta p)^{j-1}\{(r_0+r_1b_j)\frac{e^{-\lambda\theta(1-r_{j-1})b_{j-1}}-e^{-\lambda\theta b_j}}{e^{-\lambda\theta(1-r_{j-1})b_{j-1}}}-r_1\frac{\theta}{e^{-\lambda\theta(1-r_{j-1})b_{j-1}}}[\frac{(1-r_{j-1})b_{j-1}}{\theta}$$
$$\times e^{-\lambda\theta(1-r_{j-1})b_{j-1}}+\frac{1}{\lambda\theta^2}e^{-\lambda\theta(1-r_{j-1})b_{j-1}}-\frac{b_j}{\theta}e^{-\lambda\theta b_j}-\frac{1}{\lambda\theta^2}e^{-\lambda\theta b_j}]\}$$

(10)

The expected profit function of each cycle for the manufacturer refers to multiple of the products quantity for sales (the total production quantity minus the scrapped defective products) and sales minus the total cost. The expected profit function of the unit time for the manufacturer is:

$$\pi_m=\frac{C_r[P\times E(T)-E(N)]-[C_s+E(HC)+E(M)+E(IC)+E(DC)+E(RC)]}{E(T)}$$

(11)

The expected profit function of unit time of the retailer is as shown:

$$\pi_r=(P_r-C_r)D(P_r)-\frac{S_r}{E(T_r)}-\frac{h_r}{2}D(P_r)E(T_r)$$

(12)

In which, the expected replenishment cycle time is $E(T_r)=[\rho E(T)-E(N)]/D(P_r)$. Therefore, the expected profit for the whole channel in unit time is

$$\pi=\frac{C_r[P\times E(T)-E(N)]-[C_s+E(HC)+E(M)+E(IC)+E(DC)+E(RC)]}{E(T)}$$
$$+(P_r-C_r)D(P_r)-\frac{S_r}{E(T_r)}-\frac{h_r}{2}D(P_r)E(T_r)$$

(13)

III. Optimal Solution

The optimal solution refers to the strategy when the expected profit of the whole channel is maximized. Because this strategy is the Stackelberg leadership model of the game theory, the manufacturer is the leader and the retailer is the follower. Therefore, the manufacturer will set out the optimal inspection frequency and the optimal replenishment cycle for the retailer on the maximum profits. The retailer finds out the optimal price. Assuming that the relationship between the consumption in unit time by the retailer and retail price is linear, that is $D(P_r)=a-bP_r$, a,b are constants. When the concepts of Banerjee and Rahim [20] are applied, each inspection has the same cumulative risk ratio. The relationship formula of the time interval of each inspection is as follows:

$$\int_{t_{j-1}}^{t_j}\boldsymbol{r(t)dt}=\int_0^{t_1}\boldsymbol{r(t)dt},\ j=2,3,....,k \tag{14}$$

When the processing procedure time in the controlled state is controlled by Exponential distribution, the probability density function is $f(t)=\lambda e^{-\lambda t}$, $t>0,\lambda>0$, and the time interval of the inspection is

$$h_j=h_1,\ j=2,3,....,k \tag{15}$$

The expected profit of the unit time of this channel can be rewritten as:

$$\pi=\frac{1}{h_1\times\sum_{j=1}^{k}(1-\theta p)^{j-1}}\times\{C_r[\rho\times h_1\sum_{j=1}^{k}(1-\theta p)^{j-1}-p\sum_{j=1}^{k}(1-\theta p)^{j-1}$$
$$\times[d_I\rho((1-\theta)r_{j-1}h_1-1+e^{-\lambda(1-\theta)r_{j-1}h_1})+\frac{d_{II}\rho}{\lambda}e^{-\lambda h_1(\theta-(1-\theta)(1-r_{j-1}))}]]$$
$$-[C_s+(\frac{C_h}{2})\rho[h_1\times\sum_{j=1}^{k}(1-\theta p)^{j-1}]^2+C_{apm}\sum_{j=1}^{k-1}(1-\theta p)^j+(1-\theta)pC_{mr}\sum_{j=1}^{k-1}(1-\theta p)^{j-1}$$
$$+(\sum_{j=1}^{k-1}(1-\theta p)^j+1)C_I+pC_d\sum_{j=1}^{k}(1-\theta p)^{j-1}[d_I\rho((1-\theta)r_{j-1}h_1-1+e^{-\lambda(1-\theta)r_{j-1}h_1})$$
$$+\frac{d_{II}\rho}{\lambda}e^{-\lambda h_1(\theta-(1-\theta)(1-r_{j-1}))}]$$
$$+\theta p\sum_{j=1}^{k}(1-\theta p)^{j-1}[(r_0+r_1h_1)\frac{e^{-\lambda\theta(1-r_{j-1})h_1}-e^{-\lambda\theta h_1}}{e^{-\lambda\theta(1-r_{j-1})h_1}}-\frac{r_1\theta}{e^{-\lambda\theta(1-r_{j-1})h_1}}(\frac{(1-r_{j-1})h_1}{\theta}e^{-\lambda\theta(1-r_{j-1})h_1}$$
$$+\frac{1}{\lambda\theta^2}e^{-\lambda\theta(1-r_{j-1})h_1}-\frac{1}{\theta}e^{-\lambda\theta h_1}-\frac{1}{\lambda\theta^2}e^{-\lambda\theta h_1})]\}+(P_r-C_r)D(P_r)$$
$$-\frac{S_rD(P_r)}{\rho h_1\sum_{j=1}^{k}(1-\theta p)^{j-1}-p\sum_{j=1}^{k}(1-\theta p)^{j-1}[d_I\rho((1-\theta)r_{j-1}h_1-1+e^{-\lambda(1-\theta)r_{j-1}h_1})+\frac{d_{II}\rho}{\lambda}e^{-\lambda h_1(\theta-(1-\theta)(1-r_{j-1}))}]}$$
$$-\frac{h_r}{2}\{\rho h_1\sum_{j=1}^{k}(1-\theta p)^{j-1}-p\sum_{j=1}^{k}(1-\theta p)^{j-1}[d_I\rho((1-\theta)r_{j-1}h_1-1+e^{-\lambda(1-\theta)r_{j-1}h_1})$$
$$+\frac{d_{II}\rho}{\lambda}e^{-\lambda h_1(\theta-(1-\theta)(1-r_{j-1}))}]\}$$

(16)

When $\pi_{k-1}\le\pi_k$ and $\pi_{k+1}\le\pi_k$, we can find the optimal value of k^*.

IV. Numerical Analysis

To illustrate the critical aspects of the developed model, one numerical example is used to explore this model, and the parameters are as follows:

$\rho=1000$, $C_S=150$, $S_r=50$, $h_m=0.5$, $h_r=0.5$, $C_I=10$, $C_m=5$, $C_r=10$, $C_d=20$, $C_{mpm}=30$, $r_0=10$, $r_1=0.5$, $d_I=0.3$, $d_{II}=0.9$, $\eta=0.99$, $\lambda=5$.

TABLE 1. Effect of PM level on expected profit[a].

$\frac{C_{apm}}{C_{mpm}}$	0.0	0.25	0.5	0.75	1.0
π	207	2394	3129	3349	3380

[a] $k=3, h_1=0.2450, \theta=0.5$.

Table 1 shows the effects of various PM levels. The results show that expected profit increases with the PM level. That is, as C_{apm} equals to C_{mpm} yields the maximal profit. This suggests that the maximal PM activities must be implemented when the other parameters are fixed.

Figure 2 shows the effects of the value of θ on the expected profit per unit. The expected profit when performing minimal repair (θ <1.0) is more than the expected profit without performing minimal repair (θ =1.0). This inference is applicable that enterprises must probe into the optimal probability of minimal repair according to different conditions.

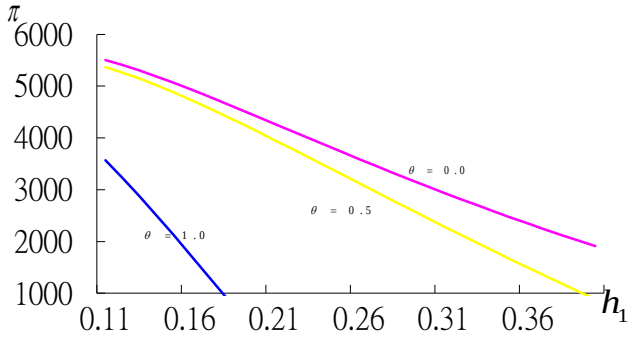


Fig. 2. Expected profit per unit for different minimal repair condition. ($k=4$)

Table 2 shows the frequency of inspection, the first inspection interval, the EPQ, and the expected profit per unit under different values. The expected profit per unit is larger when minimal repair is performed ($\theta < 1.0$) compared to that when minimal repair is not performed ($\theta = 1.0$).

TABLE 2. Optimal solutions under different conditions.

		$\theta=0.0$		$\theta=0.5$		$\theta=1.0$	
k	h_1	Q	π	Q	π	Q	π
5	0.2000	1000	4346	538	4123	314	198
6	0.1667	1000	4758	509	4651	293	1506
7	0.1429	1000	5040	489	5002	278	2424
8	0.1250	1000	5237	473	5242	267	3088
9	0.1111	1000	5375	461	5408	259	3576
10	0.1000	1000	5472	451	5524	252	3942

V. Conclusions

This study integrates the coordinated production, inspection and sales strategy between the manufacturer and the retailer and expects to maximize the profits of the whole channel. Regarding the processing production, maintenance, processing minor repair, testing and the order and inventory strategy of the retailer, this study seeks the maximum best testing times, testing cycles and expected profits of the whole channel. For most papers, the individual the manufacturer and the retailer only consider their own profits and do not take the profits of channels into consideration. However, in the context of such highly competitive environment, the enterprise profits have been shrinking. The environment and strategy can not strive for win-win opportunity between the manufacturer and the retailer. One party seeks for the maximized profit and the other party definitely enjoys less profit, resulting in the vicious competition and circle of the whole channel. Consequently, we should have good integration mechanism to generate mutual cooperation between the manufacturer and the retailer for win-win outcome. The numerical analysis can understand the changes of parameters exert significant influence on the optimal profit and offer good reference towards the coming decision for enterprise managers as well. This study finds out that in the condition of different processing and minimal repairs, there are significant influence on the testing times, testing cycle and expected profits of the whole channel.

References

[1] J.M. Chen, and T.H. Chen, "The profit-maximization model for a multi-item distribution channel", Trans. Res. Part E, vol. 43(4), pp. 338-354, 2007.

[2] M.J. Rosenblatt, and H.L. Lee, "Economic production cycles with imperfect production process", IIE Trans, vol. 18, pp. 48-55, 1986.

[3] M. Hariga and M. Ben-Daya, "Note: the economic manufacturing lot-sizing problem with imperfect production process: bounds and optimal solutions", Naval Res. Logist., vol. 45, pp. 423-432, 1998.

[4] M. Freimer, D. Thomas and J.Tyworth, "The value of setup cost reduction and process improvement for the economic production quantity model with defects", European Journal of Operational Research, vol. 173, pp. 241-251, 2006.

[5] W.N. Ma, D.C. Gong and G.C. Lin, "An optimal common production cycle time for imperfect production processes with scrap", Mathematical and Computer Modelling, vol. 52, pp. 724-737, 2010.

[6] S.W. Chiu, C.L. Chou and W.K. Wu, "Optimizing replenishment policy in an EPQ-based inventory model with nonconforming items and breakdown", Economic Modelling, vol. 35, pp. 330-337, 2013.

[7] H. Peng and G.J. Houtum, "Joint optimization of condition-based maintenance and production lot-sizing", European Journal of Operational Research, vol. 253(1), pp. 94-107, 2016.

[8] H. Pham and H. Wang, "Invited review imperfect maintenance", European Journal of Operational Research, vol. 94, pp. 425-438, 1996.

[9] M. Ben-Daya, "The economic production lot-sizing problem with imperfect production process and imperfect maintenance", International Journal of Production Economics, 76(3), pp. 257-264, 2002.

[10] Y.C. Chen, "Optimal inspection and economical production quantity strategy for an imperfect production process", International Journal of Systems Science, vol. 37(5), pp. 295-302, 2006.

[11] M.A. Saad, N. Suliman and S.H. Jawad, "Optimization of preventive maintenance schedule and production lot size", International Journal Production Economics, vol. 137, pp. 19-28, 2012.

[12] S.S. Sana, "Preventive maintenance and optimal buffer inventory for products sold with warranty in an imperfect production system", International Journal of Production Research, vol. 50(23), pp. 6763-6774, 2012.

[13] Y.C. Chen, “An optimal production and inspection strategy with preventive maintenance error and rework”, Journal of Manufacturing Systems, vol. 32, pp. 99-106, 2013.

[14] H. Lee and J.H. Cha, “New stochastic models for preventive maintenance and maintenance optimization”, European Journal of Operational Research, vol. 255(1), pp. 80-90, 2016.

[15] S. Netessine and N. Rudi, “Supply chain choice on the internet”, Management Science, 52(6), pp. 844-864, 2006.

[16] C.H. Ho, S.K. Goyal, L.Y. Ouyang and K.S. Wu, “An integrated vendor-buyer inventory model with defective items and partial backlogging”, International Journal of Logistics Systems and Management, vol. 8(4), pp. 377-391, 2011.

[17] S. Huang, C. Yang and H. Liu, “Pricing and production decisions in a dual-channel supply chain when production costs are disrupted”, Economic Modelling, vol. 30, pp. 521-538, 2013.

[18] M. Protopappa-Sieke, M.A. Sieke and U.W. Thonemann, “Optimal two-period inventory allocation under multiple service level contracts”, European Journal of Operational Research, vol. 252(1), pp. 145-155, 2016.

[19] M. Ben-Daya, “Integrated production maintenance and quality model using the imperfect maintenance concept”, IIE Transactions, vol. 31, pp. 491-501, 1999.

[20] P.K. Banerjee and M.A. Rahim, “Economic design of $\bar{X}$ -chart under weibull shock models”, Technometrics, vol. 30, pp. 407-414, 1988.

Impact of lead-time sensitive cost on lead-time quotation and pricing problems

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Abduh Sayid Albana
Univ. Grenoble Alpes, CNRS,
Grenoble INP*, G-SCOP,
F-38000 Grenoble, France
abduh-sayid.albana@grenoble-inp.org

Ramzi Hammami
ESC Rennes School of Business,
35065 Rennes, France
hammami.ramzi@gmail.com

Yannick Frein
Univ. Grenoble Alpes, CNRS,
Grenoble INP*, G-SCOP,
F-38000 Grenoble, France
yannick.frein@g-scop.grenoble-inp.fr

***Abstract*—Some research has been done on M/M/1 queueing systems with lead-time- and price-dependent demand. However, most of these works assume a constant production cost. It is known that the longer the quoted lead-time the better the firm can manage the production and reduce the cost. The idea of this paper stems from this observation. Indeed, we investigate the lead-time quotation decision in an M/M/1 make-to-order queue while assuming the production cost to be a decreasing function in lead-time. We consider two settings: (1) lead-time is variable but price is fixed, and (2) price and lead-time are both decision variables. For each setting, we provide an approach to find analytically the optimal lead-time and price (if it is a variable). We use the optimal solutions to conduct experiments and derive some insights.**

***Keywords*—lead-time dependent demand, lead-time dependent cost, lead-time quotation, M/M/1**

I. Introduction

For a large number of firms, the quoted lead-time and price have great effect on the customers' demand. In particular, the delivery lead-time has become a key factor of competitiveness for companies and an important purchase criterion for many customers. The joint effect of lead-time and price has received the attention of many research works. In a pioneer paper, Palaka et al. [4] studied the lead-time setting, pricing decisions, and capacity utilization of a profit-maximizing firm that faces a linear price- and lead-time-sensitive demand while modeling the firm as an M/M/1 queue. Costs related to congestion (holding cost) and late deliveries (penalty cost) are considered. Production cost in [4] was assumed to be constant. The authors firstly considered the capacity utilization as a constant and then as a decision variable. So and Song [5] developed an analytical framework to understand the strong interrelationships among pricing, delivery time guarantees, demand, and the overall profitability in the service context. Similar to [4], So and Song [5] considered constant production cost. However, So and Song [5] used a log linear function to model the demand as a function of price, delivery time, and delivery reliability level. Ho and Zheng [8] studied market share competition on the basis of delivery time for service firms. Many firms choose to set customer expectation by announcing their maximal delivery times. Customers will be satisfied if their perceived delivery times are shorter than their expectations. In [8], no production cost was incurred. Ray and Jewkes [7] focused on customer lead-time management where the demand is a function of price and lead-time, and where the price itself is sensitive to lead-time. In their research, both demand (in function of price and lead-time) and price (in function of lead-time) are modeled as linear functions, which finally led to a classical linear demand function. Liu et al. [3] studied a decentralized supply chain with a supplier and a retailer facing price- and lead-time-dependent demand. They used Stackelberg game to analyze the price and lead-time decisions of the supplier (as a leader) and the retailer (as a follower). The authors assumed constant unit production cost. Pekgün et al. [6] studied the centralization and decentralization of pricing and lead-time decisions of an M/M/1 queue. They used the same framework of [4] for their centralized model but without considering the holding and penalty costs. Pekgün et al. [6] assumed that unit production costs were constant.

In all of the aforementioned papers, the firm's operations are modeled as an M/M/1 queue and the production cost is assumed to be a constant. Our research investigates the lead-time quotation and pricing decisions in an M/M/1 make-to-order queue from a new perspective as we consider the production cost to be a decreasing function in lead-time. It is well known that the unit production cost depends on the quoted lead-time in many situations. Firms, that quote short lead-time to their customers, are exposed to high risk of not respecting their quoted lead-time. Thus, they have to think different decisions to reduce the risk. Examples include: buying items from quick-response but expensive suppliers, increasing the stock of semi-finished products, using faster but more costly transportation modes, and rethinking the production/delivery process by moving the trade-off toward shorter lead-time instead of cheaper production. In all these examples, the shorter the quoted lead-time the more difficult the required actions become and, consequently, the higher the unit production cost.

We consider two settings: (1) lead-time is variable but price is fixed, and (2) price and lead-time are both decision variables. For each setting, we provide an analytical approach to find the optimal lead-time and price (if it is a variable). We use the optimal solutions to conduct experiments and derive some insights.

The organization of the paper is as follows. Our models are explained in section II. In section III, we deal with model of variable lead-time with a fixed price. We dedicate section

*Institute of Engineering Univ. Grenoble Alpes

IV to solve the model with both lead-time and price as decision variable. The insights of results are given in their related section. And we finally conclude and give future work directions in section V.

II. The model

We model the firm as an M/M/1 Make-to-Order queue. Customers arrive and are served in a first-come-first-served fashion. The demand follows a Poisson process of mean arrival rate λ, which cannot be larger than the amount of demand $d(p,l)$ obtained when price, p, and lead-time, l, are quoted to the customers. As usually assumed in the literature (see the aforementioned papers), we consider that the demand linearly decreases with price and quoted lead-time, $d(p,l) = a - b_1 p - b_2 l$ where b_1 and b_2 are the sensitivity coefficients to price and lead-time,respectively. The production capacity is constant (μ) and the service time is exponentially distributed.

Unlike most of the published works, we do not assume a constant production cost but consider the unit production cost (m) to be decreasing function in quoted lead-time (l). Indeed, m is given by the following non-linear function: $m = C_1 + \frac{C_2}{l}$. This function implies that the increase in unit production cost that results from one unit decrease in lead-time is not constant (as in linear functions) but is growing with smallest values of lead-time. In other words, the smaller the value of the lead time the more difficult (and, consequently, the more expensive), its reduction becomes. Clearly, the production cost usually used in the existing literature is a particular case of our cost function with $C_2 = 0$.

The firm's objective is to maximize the total expected profit, given by $\lambda(p - m)$ while satisfying the quoted lead-time with a given service rate, s. The general model is given below.

$$\underset{l,p,\lambda,m}{\text{Maximize}}\ \lambda(p-m) \tag{1}$$

$$\text{Subject to } \lambda \leq a - b_1 p - b_2 l \tag{2}$$

$$1 - e^{-(\mu-\lambda)l} \geq s \tag{3}$$

$$\lambda \leq \mu \tag{4}$$

$$m = C_1 + \frac{C_2}{l} \tag{5}$$

$$\lambda, l, p, m \geq 0 \tag{6}$$

Decision variables:

λ	: mean arrival rate (demand),
l	: quoted lead-time,
p	: price of the goods/service,
m	: unit production cost.

Parameters:

a	: market potential,
b_1	: price sensitivity of demand,
b_2	: lead-time sensitivity of demand,
μ	: mean service rate (production capacity),
s	: service level defined by company ($s \in [0,1]$),
C_1, C_2	: production cost parameters.

The objective function is given in Equation (1). Equation (2) ensures that the mean demand rate received by the firm cannot exceed the demand generated by the quoted price and lead-time. Equation (3) guarantees that the probability of meeting the quoted lead-time, given by $1 - e^{-(\mu-\lambda)l}$ (since $e^{-(\mu-\lambda)l}$ is known to be the probability that a job is late in M/M/1 queue), must not be smaller than the required service level. Equation (4) guarantees a steady state of the M/M/1. Equation (5) defined the production cost function. The non-negativity constraints are given in Equation (6).

As stated before, we consider two settings: (1) lead-time is variable but price is fixed, and (2) price and lead-time are both decision variables. Before moving to the solving approach of each setting, we point out the following general result, which is verified under all settings.

Lemma 1. *Demand constraint (2) is binding and we have $\lambda = a - b_1 p - b_2 l$ at optimality.*

Proof: Suppose that the optimal solution is given by quoted lead-time l^*, price p^*, and demand rate λ^* such that $\lambda^* < a - b_1 p^* - b_2 l^*$. The revenue increases when l increases as an increase in l leads to decreasing the production cost ($m = C_1 + \frac{C_2}{l}$). An increase in l also leads to decreasing the probability of having late delivery ($e^{-(\mu-\lambda)l}$). Thus, by keeping price p^* and demand rate λ^* constant, one could increase the lead-time from l^* to l' until $\lambda^* = a - b_1 p^* - b_2 l'$. This change will increase revenues, while the service level constraint remains satisfied. Therefore, we would get a new solution: p^*, λ^*, and l' with a better profit than the supposed optimal solution, which is impossible. ■

III. Model with variable lead-time and fixed price

In this section, we consider the model when price is fixed (not a decision variable).

A. Optimal policy with variable lead-time and fixed price

We solve analytically the model when price is fixed (not a decision variable). By considering the result of Lemma 1 and integrating the expression of production cost (m) into the objective function, the problem can be formulated as a constrained non-linear optimization model with single variable l. Indeed, we respectively replace λ and m by $a - b_1 p - b_2 l$ and $C_1 + \frac{C_2}{l}$. Thus, the model with fixed price becomes equivalent to the following model.

$$\underset{l \geq 0}{\text{Maximize}}\ f(l) = (a - b_1 p - b_2 l)\left(p - C_1 - \frac{C_2}{l}\right) \tag{7}$$

$$\text{Subject to } (\mu - a + b_1 p)l + b_2 l^2 \geq \ln\left(\frac{1}{1-s}\right) \tag{8}$$

$$l \leq \frac{a - b_1 p}{b_2} \tag{9}$$

Equation (8) represents the service constraint after rewriting $1 - e^{-(\mu-\lambda)l} \geq s$ as $(\mu - \lambda)l \geq \ln\left(\frac{1}{1-s}\right)$. In the rest of the paper, we let z denote $\ln\left(\frac{1}{1-s}\right)$. Equation (9) guarantees that $\lambda \geq 0$, which is equivalent to $l \leq \frac{a-b_1 p}{b_2}$. Note also that constraint (4) is immediately satisfied if the service constraint is satisfied and the lead-time is positive. Therefore, constraint (4) can be removed. In order to solve the model, we firstly consider the following lemma.

Lemma 2. *The profit function $f(l)$ is concave in l.*

Proof: The concavity of $f(l)$ is proven by the negative second derivative function $\frac{\partial^2}{\partial l^2} f(l) = -\frac{2C_2(a-b_1p)}{l^3} < 0$. ■

In the published papers where the production cost is assumed to be constant, the service constraint is binding, which simplifies the solving approach. In our model, the service level constraint is not necessarily binding. Indeed, if we have a given lead time l for which the service constraint is not tight, we could reduce l without violating this constraint. By decreasing l, we increase the demand and, consequently, the revenue λp. But, we also increase the cost. Thus, this does not necessarily improve the overall profit. This trade-off may lead to a non-binding situation.

In the following Lemma, we identify the candidates for optimality in each of binding and non-binding situations.

Lemma 3. *The service level constraint is not necessarily binding, and we have:*

- *In the non-binding situation, the candidate for optimality of the base model, if satisfying the service constraint, is* $l^{NB} = \min\left(\sqrt{\frac{(a-b_1p)C_2}{(p-C_1)b_2}}, \frac{a-b_1p}{b_2}\right)$,
- *In the binding situation, the candidate for optimality of the base model is* $l^B = \frac{a-b_1p-\mu+\sqrt{(a-b_1p-\mu)^2+4b_2z}}{2b_2}$ *with* $z = \ln\left(\frac{1}{1-s}\right)$.

Proof: **Non-binding situation:** If we ignore the service level constraint (8), the problem becomes: Maximize $_{0\le l\le \frac{a-b_1p}{b_2}}$ $f(l) = (a - b_1p - b_2l)\left(p - C_1 - \frac{C_2}{l}\right)$. We have $\frac{\partial}{\partial l} f(l) = \frac{(a-b_1p)C_2}{l^2} + (C_1 - p)b_2$. Thus, $\frac{\partial}{\partial l} f(l) = 0 \Leftrightarrow l^2 = \frac{(a-b_1p)C_2}{(p-C_1)b_2}$. We only consider the positive root, which is $l = \sqrt{\frac{(a-b_1p)C_2}{(p-C_1)b_2}}$. Given that $f(l)$ is concave as shown in Lemma 2, and $0 \le l \le \frac{a-b_1p}{b_2}$, the candidate for optimality in the non-binding situation, if satisfying the service constraint, is l^{NB} (as expressed in the Lemma).

Binding situation: In this case, the service level constraint (8) is binding, implying that $(\mu - a + b_1p)l + b_2l^2 = z$ at optimality. Thus, we can obtain the lead-time l directly from the service level equation. The quadratic equation $b_2l^2 + (\mu - a + b_1p)l - z = 0$ has positive discriminant $\Delta = (\mu - a + b_1p)^2 + 4bz$, implying that we have two roots, with only one positive root $l_1 = \frac{a-b_1p-\mu+\sqrt{(\mu-a+b_1p)^2+4bz}}{2b_2}$. Thus, in the binding situation, the candidate for optimality of the base model is $l^B = l_1$. ■

Based on the results of the previous lemma, we can now announce the optimal lead-time of the base model with variable lead-time and fixed price.

Proposition 1. *The optimal solution of the problem with variable lead-time and fixed price is the following:*

- *If* $l^B \le \frac{a-b_1p}{b_2}$ *then* $l^* = \max(l^B, l^{NB})$, *(where* l^B *and* l^{NB} *are as defined in Lemma 3)*
- *Otherwise, the problem is infeasible.*

Proof: Firstly, we demonstrate that if $l^B > \frac{a-b_1p}{b_2}$ then the problem is infeasible. Note here that according to the proof of Lemma 3, the values of lead-time l that satisfy the service level constraint are such as $l \ge l^B$. Thus, if $l^B > \frac{a-b_1p}{b_2}$ then there are no feasible values of l that can satisfy the service constraint (i.e., $l \ge l^B$) while giving a positive demand (i.e., $l \le \frac{a-b_1p}{b_2}$). Therefore, the problem is infeasible if $l^B > \frac{a-b_1p}{b_2}$.
Now, assuming that $l^B \le \frac{a-b_1p}{b_2}$. We have two cases: $l^B \le l^{NB}$ or $l^B > l^{NB}$.
Case of $l^B \le l^{NB}$**:** In this case, l^{NB} satisfies the service constraint as it is greater than l^B. Given in addition that $f(l)$ is concave and reaches its maximum in $\sqrt{\frac{(a-b_1p)C_2}{(p-C_1)b_2}} = l^B$, which is equal to or greater than l^{NB}(see Lemma 2), we deduce that $f(l^B) \le f(l^{NB})$. Consequently, the optimal solution is l^{NB}.
Case of $l^B > l^{NB}$**:** In this case, l^{NB} doesn't satisfy the service constraint. Given the concavity of $f(l)$ and its maximum in $\sqrt{\frac{(a-b_1p)C_2}{(p-C_1)b_2}} = l^B$, which is equal to or greater than l^{NB}, we deduce that the best feasible value of lead-time is l^B. ■

B. Experiments and insights with fixed price

In order to illustrate our analytical results and understand better the behavior of the model, we conduct experiments with the following base example with parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $\mu = 10$, $s = 0.95$, $C_1 = 2$, and $C_2 = 3$. In particular, we study the effect of the customers' sensitivity to lead-time (b_2) and the cost sensitivity to lead-time (C_2). We do experiments by varying one parameter and keep the other parameters constant. In each case, we provide the optimal values of quoted lead-time, demand, revenue, total cost, profit (i.e., revenue – total cost), and observed service level (which is not necessarily the service level s, initially defined by the company).

1) Effect of customers' sensitivity to lead-time (b_2) with fixed price:

We fix the price to 8.81 (price is obtained from the optimization with price as decision variable and $b_2 = 2$, see table IV) and vary the value of b_2 from 2 to 20. The results are reported in Table I.

TABLE I. EFFECT OF DEMAND SENSITIVITY TO LEAD-TIME (b_2) WITH FIXED PRICE

b_2	Lead-time	Demand	Total cost	Profit	Serv. lev. realized
2	2.89	8.96	27.23	51.77	95%
8	0.98	6.93	35.16	25.94	95%
10	0.83	6.41	35.87	20.61	95%
12	0.74	5.92	35.98	16.19	95.03%
16	0.64	4.55	30.56	9.58	96.89%
20	0.57	3.35	24.35	5.19	97.74%

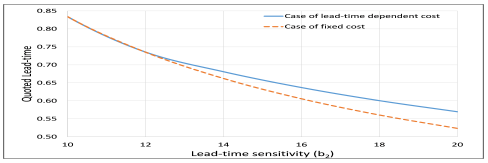


Fig. 1. Variable cost vs Fixed cost:quoted lead-time

Observing Table I, we derive the following insights:

- As expected, an increase in customers' sensitivity to lead-time (b_2) leads to a decrease in quoted lead-time. Note also that, despite the decrease of quoted lead-time, which favors attracting more customers, the demand and profit always go down when b_2 goes up.
- Unlike the demand and profit, the total cost is not a monotonous function in lead-time sensitivity (b_2). Indeed, for the values of b_2 ranging from 2 to 12, the total production cost increases with an increase in b_2. This region corresponds to the cases where the service constraint is binding (observed service level $= s$). Then, an increase in b_2 from 12 to 20 leads to decreasing the total cost. This region corresponds to the non-binding situations (observed service level $> s$). In this case, the effect of demand decrease on reducing the total cost becomes more important than the effect of lead time decrease on increasing the total cost.
- It is important to note that when b_2 becomes significantly large (here, $b_2 \geq 12$), the model chooses to do better than the required service level (non-binding situation). Indeed, in this case, the model has to quote a short lead-time in order to maintain a certain amount of demand. However, this implies a high unit production cost, which may annihilate the gain obtained by reducing the lead-time. Consequently, the model reacts by reducing the quoted lead-time but not as much as allowed by the service constraint, which implies a non-binding situation. Clearly, this behavior cannot be captured by the existing models where the unit production cost is constant.
- In order to show the impact of taking into account the sensitivity of cost to lead-time (as we do in our model), we re-solve the model while considering a constant unit production cost. Indeed, we calculate the average unit cost ($C_1 + \frac{C_2}{l}$) over all instances of Table I and take the obtained value as a constant cost in these new experiments. In other words, as the average unit cost in Table I is 5.55, we consider in these experiments $C_1 = 5.55$, and $C_2 = 0$ (cost does not depend here on lead-time). We firstly compare the results of our base model to the results of the model with constant cost in terms of quoted lead-time. The results, reported in Fig. 1, show that both models leads to the same optimal lead-time for low values of b_2, which corresponds to the binding situation. In this case, we can check analytically that the quoted lead-time does not depend neither on C_1 nor on C_2 (lemma 3). However, when the customers are very sensitive to lead-time (non-binding-situation), we see that the lead-time quoted by our base model is greater than the lead-time quoted by the model with constant cost.
- In order to evaluate the impact of not quoting the right lead-time when the cost is assumed constant, we take the optimal lead time in this case and inject it in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 2$ and let $\Pi 1$ denote the optimal profit of our base model (given in Table I). Then, we calculate the percentage of gain $\Pi = \frac{100 \times (\Pi 1 - \Pi 2)}{\Pi 2}$ for the different values of b_2. We obtain an average gain $\Pi = 4.81\%$ in non-binding situations. Over all instances, the average gain is $\Pi = 1.93\%$. Of course, such a gain may represent a very significant amount if one considers real data.

2) Effect of cost sensitivity to lead-time (C_2) with fixed price:

We fix the price to 8.35 and vary the value of C_2 from 2 to 10. The results are reported in Table II.

TABLE II. EFFECT OF COST SENSITIVITY TO LEAD-TIME(C_2) WITH FIXED PRICE

C_2	Lead-time	Demand	Total cost	Profit	Serv. lev. realized
2	1.44	7.93	26.83	39.37	95%
4	1.44	7.93	37.80	28.39	95%
6	1.62	6.89	39.38	18.20	99.34%
8	1.87	5.39	33.91	11.14	99.98%
10	2.09	4.07	27.67	6.35	99.99%

Observing Table II, we deduce the following:

- At first, an increase in C_2 does not impact the quoted lead-time (this corresponds to the binding situation). Then, when we are in the non-binding situation, if the cost is more sensitive to lead-time (C_2 increases) then the model quotes longer lead-time.
- For relatively high values of C_2, it is important to note that the realized service level is very high (close to 1). In this case, the model chooses to quote a relatively long lead-time (with comparison to what the existing capacity allows to do) in order to reduce the cost. Although this implies less demand, the loss of revenue is here offset by the saving in cost.

IV. MODEL WITH BOTH LEAD-TIME AND PRICE AS DECISION VARIABLES

In this section, we consider a more complex setting with price p as decision variable.

A. Optimal policy with both lead-time and price as decision variables

Given that demand constraint is binding (as demonstrated in Lemma 1), we can see the link between l, p, and λ. Thus, the problem can be simplified into two variables (either l & p, p & λ, or l & λ). Here, we deduce that $p = \frac{a - b_2 l - \lambda}{b_1}$ and formulate the problem with l and λ. The objective function is given in Equation (10). The service constraint expressed as a function of l and λ is given in (11). As the price must be positive, we must have $a - b_2 l - \lambda \geq 0$, which is guaranteed by constraint (12).

$$\underset{l,\lambda>0}{\text{Maximize}}\ g(l,\lambda) = \lambda \left(\frac{a - b_2 l - \lambda}{b_1} - C_1 - \frac{C_2}{l} \right) \tag{10}$$

$$\text{Subject to } \lambda \leq \mu - \frac{z}{l} \tag{11}$$

$$\lambda \leq a - b_2 l \tag{12}$$

We determine the optimal solution of the problem above in Proposition 2. Because of page limitations, we cannot show the detailed demonstration of proposition 2 in this paper.

Proposition 2. *We denote:*
$\Delta = (a - C_1 b_1 - 2\mu)^2 + 4b_2(2z - C_2 b_1)$,
$M = \frac{a - C_1 b_1 - 2\mu + \sqrt{\Delta}}{2b_2}$, $g_1(l) = \frac{\left(a - C_1 b_1 - b_2 l - \frac{C_2 b_1}{l}\right)^2}{4b_1}$, *and* $g_2(l) = \left(\mu - \frac{z}{l}\right)\left(\frac{a - b_2 l - \mu + \frac{z}{l}}{b_1} - C_1 - \frac{C_2}{l}\right)$.
If $\Delta \leq 0$ *then the optimal lead-time* $l^* = \sqrt{(C_2 b_1)/b_2}$ *and optimal demand* $\lambda^* = \frac{1}{2}(a - C_1 b_1 - b_2 l - \frac{C_2 b_1}{l})$,
If $\Delta > 0$ *then the the optimal lead-time can be obtained as follows:*

- *Consider the cubic equation* $Q(l) = \mu b_2 l^3 - (az - 2\mu z + \mu C_2 b_1 - C_1 b_1 z)\, l - 2z(z - C_2 b_1) = 0$. *This equation has at maximum 2 positive roots. We denote them by* R_1 *and* R_2.
- *Three potential values of optimal lead-time are:* $max\left(\sqrt{(C_2 b_1)/b_2}, M\right)$, R_1, *and* R_2. *If* R_1 *or* R_2 *(or both)* $\notin [\frac{z}{\mu}, M]$ *then we remove it (them).* R_1 *or* R_2 *can be obtained explicitly.*
- *The optimal lead-time is the one that leads to the maximum of* $g_1\left(max\left(\sqrt{(C_2 b_1)/b_2}, M\right)\right)$, $g_2(R_1)$, *and* $g_2(R_2)$.
- *if* $l \geq M$ *then optimal demand* $\lambda^* = \frac{1}{2}(a - C_1 b_1 - b_2 l - \frac{C_2 b_1}{l})$, *otherwise* $\lambda^* = \mu - \frac{z}{l}$

In all cases, the optimal price can be easily deduced from this expression $p^* = \frac{a - b_2 l^* - \lambda^*}{b_1}$.

B. Experiments and insights with variable price

We recall that we use a base example with parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $\mu = 10$, $s = 0.95$, $C_1 = 2$, and $C_2 = 3$. The objective of this section is to illustrate the model behavior when both price and lead-time are variables. In particular, we will focus on the comparison between the case of fixed price and this case of variable price.

1) Effect of demand sensitivity to price (b_1) with variable price:

We vary the demand sensitivity to price (b_1) and report in Table III the optimal lead-time, price, demand, total cost, profit, and realized service level.

TABLE III. EFFECT OF DEMAND SENSITIVITY TO PRICE (b_1)

b_1	Lead-time	Price	Demand	Total cost	Profit	Serv. lev. realized
2	1.45	16.70	7.93	32.31	100.08	95%
4	1.59	8.08	8.12	31.54	34.08	95%
6	1.74	5.21	8.28	30.81	12.33	95%
8	2.00	4.13	5.00	17.50	3.13	99.99%
10	2.24	3.50	1.58	5.29	0.25	99.99%

The main observations are summarized hereafter:

- An increase in demand sensitivity to price leads, as expected, to a decrease in offered price. The model also reacts by quoting a longer lead-time in order to reduce the cost and offset the effect of offering a small price.
- In order to show the impact of taking into account the sensitivity of cost to lead-time, we re-solve the model while considering a constant unit production cost. Indeed, we calculate the average unit cost $(C_1 + \frac{C_2}{l})$ over all instances of Table III and take the obtained value (namely, 3.70) as a constant cost in these new experiments (i.e., $C_1 = 3.70$, and $C_2 = 0$). Note that the cases of $b_1 = 9$ and 10 are not considered here as they lead to non-feasible problems. We firstly compare the results of our base model to the results of the model with constant cost in terms of quoted lead-time. The results, reported in Fig. 2, show that an increase in the value of b_1 leads to an increase in quoted lead-time in our model while it has the opposite effect when the cost is assumed to be constant. Indeed, with an increase in b_1 in case of constant cost, the lead-time decreases as much as allowed by the service constraint in order to offset the effect of the price increase (triggered by the increase in b_1) on reducing the amount of demand.
- In order to evaluate the impact of not quoting the right price and lead-time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 2$ and let $\Pi 1$ denote the optimal profit of our base model (given in Table III). Then, we calculate the percentage of gain $\Pi = \frac{100 \times (\Pi 1 - \Pi 2)}{\Pi 2}$ for the different values of b_1. In some cases ($b_1 = 7$ and 8), we obtain a negative profit when we consider the optimal solution given by the model with constant cost. For the other cases which give positive profits, our model leads to an average gain $\Pi = 32.73\%$.

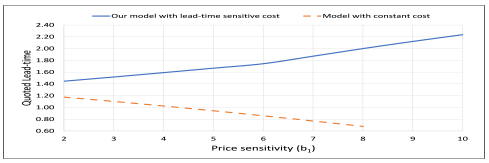


Fig. 2. Effect of b_1 with variable and constant cost

2) Effect of demand sensitivity to lead-time (b_2) with variable price:

We have studied the effect of b_2 in case of fixed price. Now, we conduct the experiments in case of variable price and report the results in Table IV.

TABLE IV. EFFECT OF DEMAND SENSITIVITY TO LEAD-TIME (b_2) WITH VARIABLE PRICE

b_2	Lead-time	Price	Demand	Total cost	Profit	Serv. lev. realized
2	2.89	8.81	8.96	27.23	51.77	95%
8	1.35	7.85	7.78	32.85	28.26	95%
16	0.88	7.31	6.61	35.66	12.68	95%
18	0.82	7.25	6.30	35.77	9.93	95.11%
20	0.77	7.25	5.51	32.35	7.58	96.92%

- It is firstly important to note that an increase in b_2 leads not only to a decrease in lead-time but also to a decrease in price. We expected that the model will react by quoting a shorter lead-time (in order to guarantee a profitable amount of demand) and to increase

the price (since decreasing the lead-time implies a higher cost), surprisingly, as customers become more sensitive to lead-time, the model prefers not to quote a very short lead-time (with comparison to the case of fixed price), but to react by decreasing lead-time and price simultaneously in order to find the best trade-off between cost and demand. A comparison between the cases of fixed price and variable price shows that the demand is always greater in case of variable price although a longer lead-time is quoted in this case.

- In Fig 3, we report the variation of lead-time for increasing values of b_2 in three following situations: variable price with variable unit cost ($C_1 + \frac{C_2}{l}$), variable price with fixed unit cost (taken as the average value of unit costs obtained in case of variable cost), and fixed price with variable cost (discussed previously). We can see the longest lead-time is quoted when the price is variable and the cost depends on lead-time. The shortest lead-time is quoted when the price is fixed and the cost depends on lead-time
- In order to evaluate the impact of not quoting the right price and lead-time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 2$ and let $\Pi 1$ denotes the optimal profit of our base model (given in Table IV). Then, we calculate the percentage of gain $\Pi = \frac{100 \times (\Pi 1 - \Pi 2)}{\Pi 2}$ for the different values of b_2. We found that our model leads to an average gain $\Pi = 22.75\%$. This shows once again the interest of our model with comparison to existing models where the cost is assumed to be constant.

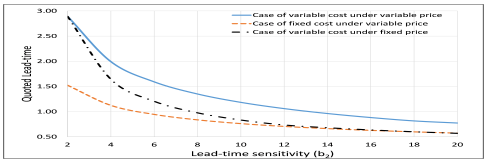


Fig. 3. Variation of lead-time for increasing values of b_2

3) Effect of cost sensitivity to lead-time (C_2) with variable price:

We vary the value of C_2 from 2 to 10. The results are reported in Table V.

TABLE V. EFFECT OF COST SENSITIVITY TO LEAD-TIME (C_2) WITH VARIABLE PRICE

C_2	Lead-time	Price	Demand	Total cost	Profit	Serv. lev. realized
2	1.44	8.35	7.93	26.83	39.37	95%
4	1.74	7.83	8.28	35.61	29.15	95%
6	2.01	7.35	8.51	42.38	20.19	95%
8	2.31	7.25	7.14	39.03	12.76	99.86%
10	2.58	7.25	5.51	32.35	7.58	99.99%

Observing Table V, we deduce the following:

- Unlike the case of fixed cost, for which a first increase in C_2 (from 2 to 4, see Table II) does not impact the quoted lead-time, we see that the lead-time is sensitive to C_2 in case of variable price even for small values of C_2.
- An increase in C_2 implies a longer lead-time which favors a decrease in demand. In order to offset the effect of lead-time on demand, the firm reacts by decreasing the price as it favors a greater amount of demand. Thus, surprisingly, when the production cost is more sensitive to shorter lead times, the firm reacts by offering a smaller price.
- As we observe in the figure 4, the firm quotes a longer lead-time when the price is variable. Indeed, in case of fixed price, the firm is obliged to keep a relatively short lead-time even with an increase in C_2 in order to guarantee a sufficient amount of demand. However, when the price is variable, the firm can quote a longer lead-time (as this implies a smaller cost) and offset its impact on demand by decreasing the price as we discussed in the previous point.

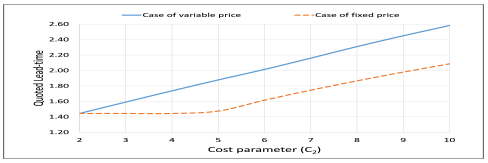


Fig. 4. Effect of C_2 with variable and constant price

V. CONCLUSION

We analytically solved the problem of lead-time quotation in an M/M/1 make-to-order queue while assuming the production cost to be a decreasing function in lead-time. We considered two settings: (1) lead-time is variable but price is fixed, and (2) price and lead-time are both decision variables. We conducted experiments and derived interesting insights. For instance, we showed that our models lead to significant gains with comparison to existing models where the cost is assumed to be constant. Our work can be extended by considering the holding and penalty costs.

REFERENCES

[1] F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, 7th ed. New York: McGraw Hill, 2001.

[2] R. Irving, *Beyond the quadratic formula*, Washington, DC: Mathematical Association of America, 2013.

[3] L. Liu, M. Parlar, and S. X. Zhu, "Pricing and Lead-time Decisions in Decentralized Supply Chains", Management Science, vol. 53, pp. 713-725, 2007.

[4] K. Palaka, S. Erlebacher, and D. H. Kropp, "Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand", IIE Transactions, volume (30), pp. 151-163, 1998.

[5] K. C. So, and J. S. Song, "Price, delivery time guarantees and capacity selection", European Journal of Operational Research, vol. 111, 28-49, 1998.

[6] P. Pekgün, P. M. Griffin, and P. Keskinocak, "Coordination of marketing and production for price and leadtime decisions", IIE Transactions, vol. 40, pp. 12-30, 2008.

[7] S. Ray and E. M. Jewkes, "Customer lead-time management when both demand and price are lead-time sensitive", European Journal of operational research, vol. 153(3), 769-781, 2004.

[8] T. H. Ho and Y. S. Zheng, "Setting Customer Expectation in Service Delivery: An Integrated Marketing-Operations Perspective", Management Science, vol. 50, pp. 479-488, 2004.

A MILP for strategic planning of production-inventory network manufacturing bulk materials

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Sabah Belil[a,c]
[a]EMINES, Mohammed VI Polytechnic University, Ben Guerir, Morocco
sabah.belil@emines.um6p.ma, s.belil@isima.fr

S. Kemmoé-Tchomté[b]
Université Clermont Auvergne
[b]CRCGM EA 38 49
Clermont Ferrand, France
sylverin.kemmoe_tchomte@uca.fr

N. Tchernev[a,c]
LIMOS UMR 6158
[c]Université Clermont Auvergne
Aubière, France
(LIMOS, UMR CNRS 6158),
nikolay.tchernev@uca.fr

Abstract— This paper deals whit strategic planning of a multi facility plant manufacturing bulk materials. The plant is composed of many manufacturing lines producing different families of products and many storage facilities. These lines and storage facilities are situated at different locations in a large multi-facility plant forming a production-inventory network. Moreover, the multiple products manufactured need to be segregated during the manufacturing, storage and transportation to be able to satisfy the manufacturing and storage constraints. The objective is to evaluate the plant capacity to satisfy the total demand. The production-inventory network planning problem is modelled as a network flow problem. Model takes the form of a Mix Integer Linear Program (MILP) which objective is to maximize demand satisfaction while minimizing storage. Using this MILP for capacity allocation and resources sizing, an optimal solution is obtained which determines the strategic planning in order to satisfy the total demand versus a given production, storage and distribution capacities. The efficacy and usefulness of the integrated model is illustrated with an example involving six manufacturing lines and six varieties (families) of bulk products.

Keywords— strategic planning, graph theory, maximum flow, linear programming.

I. Introduction

Production planning is an important aspect and needs to be addressed while planning future production strategy to manage capacity and demand. According to the length of the planning horizon and the importance of decisions to be made, planning tasks are usually classified into three different planning levels, it means long-term, mid-term and short-term planning [1][2]. The strategic planning is responsible for creating the enterprise future facility and determining capacity design, product allocation and resource sizing. Considering strategic decisions, the tactical planning has to specify the optimum use of various resources, including manufacturing plants, warehouses, transports, whereas the third referred as operational level is related to detailed scheduling definition, sequencing, lot sizing, assigning loads and vehicle routes etc. Although that resource sizing and capacity allocation decisions are in the scope of long term planning, they must be taken regarding tactical and operational level. Indeed they have a huge impact on operational efficiency which must be considered in order to avoid to inefficient system. For example, inventory level and consequently inventory costs are linked to available capacity as ability to satisfy customer's demand.

The scope of this paper is the strategic decision making process. Typically, the planning horizon is 1 - 5 years and the planning period is a month. Once the facility network design is performed, the most important strategic decisions are capacity allocation and resources sizing decisions, inventory management and distribution decisions within the supply network, e.g. deciding the aggregate quantities and material flows for manufacturing and distribution of products. The relevant issues treated at the strategic capacity planning are related to determining when and by how much the capacity levels should change. The input to long-term capacity management is a sales plan, based on a demand forecast. The sales plan can be satisfied by a corresponding capacity plan.

The paper deals with aggregate production planning which encompasses the simultaneous determination of company's production, inventory levels over a finite time horizon during the strategic decision making process. We propose a model for strategic capacity allocation and resources sizing of a multi-facility plant manufacturing system tackling resources constraints and integrating storage constraint impacts. The manufacturing system under study is a production-inventory network from the process industry and it is modelled as network flow problem. Model takes the form of a Mix Integer Linear Program (MILP) whose the objective is to maximize demand satisfaction while minimizing storage. Remain of the paper is organized as follow. A motivation and background is given in the next section. The third section formulates the multi commodity network graph model and develops the strategic level multi-period planning model as a MILP. In the next section an experiment is given along with a numerical example to illustrate both the model's applicability and the solution procedure. Section 5 concludes the study.

II. Motivation and background

A. Motivation

In this study, we deal with the implementation of an aggregate production planning model for a decision support

during the strategic decision process for medium-sized multi-product company from the process industry. Throughout the discussion, we avoid excessive details of the case to keep it illustrative, so that a lot of companies with continuous production systems can associate themselves with the situation in this study. The company produces different bulk products, which differ in recipe (many types). The processing stage consists of several steps: mixing of raw materials according to recipe and a number of subsequent processing steps that are executed without interruption or intermediate storage, i.e. continuous process. The manufacturing equipment is organized in flow line and it is considered as a manufacturing line. There are many heterogeneous parallel lines grouped and located in different facilities. The finished bulk product may be stored in one of the storage sheds or can be transported to the distribution facility where trucks or bulk freight wagons are loaded for a customer delivery. At each storage shed, many products can be warehoused, however during the storage or retrieval process only one product can treated. The material handling process to the storage sheds and to the distribution facility is performed by a material transport system composed of many belt conveyors organized in a multi-line pipe line mode.

For many years for the company it was a common policy to produce to order since there are sufficient capacity. As a consequence of the huge increase in product variety and in volume, the company is forced to shift a part of its production system from make-to-order (MTO) to make-to-stock (MTS) and has to operate under a hybrid MTO–MTS strategy. Therefore the capacity design and planning become a very important base for the long term decision making. Indeed they have a huge impact on operational efficiency which must be considered in order to avoid an inefficient system.

B. Background

Many researches have been made about strategic capacity allocation and resources sizing problem. In this article we focus on capacity allocation and inventory management problem. Capacity allocation and inventory management problem may be defined as the determination of optimal use of given resources in order to meet final demand. This scope has been a well-established research area within operations research (OR). For recent reviews we refer to [4] and [5]. [6] presents an overview of the current state-of-the-art of planning and scheduling problems in the chemical process industry. From the best of our knowledge one of the first studies concerning capacity allocation and optimization problem is propose by [7]. In this research no inventories are considered between periods. A lot of others models have been made in the case of mono-period situation (for example see [8]). These articles do not consider necessary evolution of the system under consideration. It consists in producing new products during future periods, transferring production from one factory to another and adjusting production volume according to demand variation from one period to the next.

Aggregate production planning (APP) involves the simultaneous determination of company's production, inventory and employment levels over a finite time horizon [9]. It aims to maximize profit through optimal levels of production, capacity, and inventory while satisfying the demand over a specified time horizon. Aggregate planning, as the term suggests, pertains to aggregate decisions concerning product family rather than stock-keeping unit level decisions in relation to individual products. In other words, aggregate planning will determine the total production level in a plant for a given time period and a given product family, rather than the quantity of each individual product family member. For reviews we refer to [11][12]and [13].

In this paper an aggregate planning approach for process industry is identified and addressed. We propose a linear optimization model for strategic production-inventory network planning of a plant manufacturing bulk materials. The planning problem is formulated as a multi commodity network model that takes the form of a MILP which objective is to maximize demand satisfaction while minimizing inventory. As explain in the next section, the scope of our study is a multi-facility production, inventory and distribution system belonging to the process industry.

Network models assume that the location allocation problem is embedded in a network composed of links and nodes. Demands typically arise on the nodes, though some research has been done on the case in which demands arise on the links and nodes. One practical example in which demands arise on both the nodes and the links is the demand for emergency highway services. Recently [14] used single period network model for flexibility evaluation in manufacturing network strategic planning.

In the next paragraph we present a network model based on graph theory. The main contribution of our modelling approach consists in the fact that the proposed model can be solved with commercial solver and optimal solution is obtained for industrial size problems.

III. Multi commodity network strategic model

A. Problem setting: definitions and assumptions

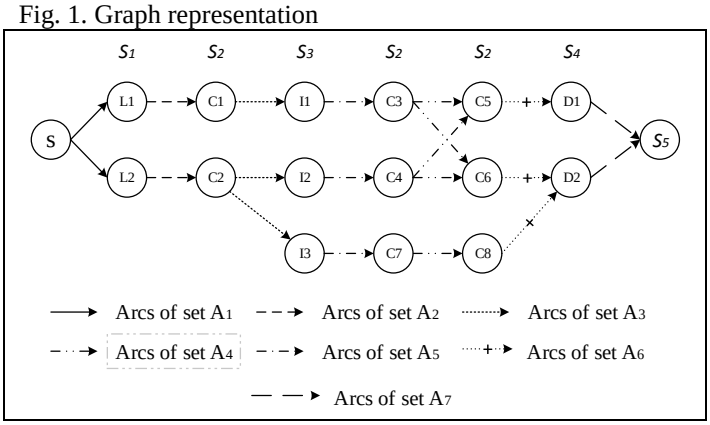
Consider the system presented before, manufacturing system is composed of M manufacturing facilities which can potentially produce any product distributed by the company. The capacity design is made during the factory design, i.e. long term strategic planning. Each production facility has a set of production lines capable of producing any product family. At the long term planning level, the units of analysis are: (*i*) a product families F consisting of a set of products variants, each family $f \in F$ has its own demand; (*ii*) a manufacturing system consisting of the set of facilities with known capacity; (*iii*) material handling system composed of belt conveyors organized in a multi-line pipe line mode which transport finished products to distribution points. At the facility level, the units of analysis are: (*i*) the set of items (end-products) grouped into product family produced by the facility depending on the process plans; (*ii*) a set of manufacturing lines L which can produce one or more product families; (*iii*) a family inventory (storage sheds) which represent the cyclic or safety stock; (*iv*) material handling system composed of belt conveyors which relay manufacturing lines with storage sheds and distribution points. During the aggregate capacity planning these product families should be allocated to different lines belonging to facilities to be produced.

All families have a given demand for each period of the time horizon under consideration, as well as all manufacturing lines have its own capacity. Moreover, it is allowed to change the family manufactured allocation to facility. Indeed, transfers are acceptable between facilities if product families which are manufactured in these facilities have manufacturing compatibilities. It means that product family realized in one manufacturing line can be manufactured in the other and reciprocally. This aspect stands for the mix flexibility of the system. Excess production can be made in order to be stocked and used in the next period. Regarding the demand of product family, the aim of our work is to provide a model permitting to test the capacity of the manufacturing system and to propose a production plan maximizing the demand satisfaction rate while minimizing inventory.

B. Network graph model

We model the manufacturing system by using a network flow approach. The goal of such representation is to evaluate the maximum flow (demand satisfaction) able to circulate in the graph and then to find the configuration which minimizes inventory. Of course, restrictions, which must be satisfied by the flow, are introduced, such as equality between the flow reaching a node and the flow going out, or respect of maximum and minimum capacity. In our case, this last restriction imposes that the value of a flow circulating on an arc must be greater than a minimum value or null. Here, this flow represents the charge of the system which is produced in the planning time horizon. Maximum arc capacities model three items as follows: (*i*) the manufacturing capacity of a given manufacturing line used for manufacturing a given product family; (*ii*) conveyor capacity; (*iii*) the product family demand which is produced by a given manufacturing line. The positive minimum arc capacities allow standing for profitability limits for each family production.

Fig. 1. Graph representation



Nodes in the network model the following items: (*i*) regular manufacturing line capacity; (*ii*) storage capacity in storage sheds; (*iii*) a distribution node modelling links between manufacturing system and family demands. Finally two dummy nodes S and S_5 are introduced, which are respectively called the source and the sink. They are not submitted to any constraint and are used to evaluate the flow in the graph. Indeed, it is calculated from the first to the last. Source is the start of arcs going to nodes standing for manufacturing lines. Their maximum capacities correspond to the line capacity represented by the node. So a flow circulating on one of these arcs stands for manufactured part of the demand of the family assigned to the node ending this arc that is produced. All these elements of the proposed network model are summarized on Fig. 1 using the notation presented in the next section.

C. Notations

Notations are classified into five categories, namely set, indexes, parameters, continuous decision variables and binary decision variables, as illustrated below (see also Fig.1):

Set / Subset:

N	Set of nodes $N = S \oplus S_1 \oplus S_2 \oplus S_3 \oplus S_4 \oplus S_5$, $\forall i, j \in N; \quad S_i \cap S_j = \emptyset$
S	Source node;
S_1	Subset of nodes modelling manufacturing lines (L1 and L2 on Fig.1);
S_2	Subset of nodes modelling transport conveyors (C1…C8 on Fig.1);
S_3	Subset of nodes modelling inventories (I1, I2 and I3 on Fig.1);
S_4	Subset of nodes modelling distribution processes (D1 and D2 on Fig.1);
S_5	Sink node;
T	Set of periods;
T_0	The period preceding the beginning of the planning horizon, $T \subset T_0$;
F	Set of product families;
A	Set of arcs such that: $A = A_1 \oplus A_2 \oplus A_3 \oplus A_4 \oplus A_5 \oplus A_6 \oplus A_7$, $\forall i, j \in A$; $A_i \cap A_j = \emptyset$
A_1	Subset of arcs going from S (source node) to an element of S_1 (manufacturing lines);
A_2	Subset of arcs going from S_1 to an element of S_2 (transport conveyors);
A_3	Subset of arcs going from S_2 to an element of S_3 (storage conveyors);
A_4	Subset of arcs going from S_2 to an element of S_2 during the transport by transfer conveyors;
A_5	Subset of arcs going from S_3 to an element of S_2 (retrieval conveyors);
A_6	Subset of arcs going from S_2 to an element of S_4 (distribution processes);
A_7	Subset of arcs going from S_4 to an element of S_5 (distribution processes).

Indexes

i	Node $i \in N \backslash S_5$;
j	Node $j \in N$;
i, j	Arc$(i, j) \in A$;
t	Period $t \in T$;
f	Family $f \in F$.

Parameters

$capab_{f,j}$ Binaries parameters whose value is equal to one if production line $j \in S_1$ is capable to produce the product family $f \in F$, and zero otherwise;

$Capa_{i,j}^{max}$ Maximum capacity of arc$(i,j) \in A$;

Pro_j^{min} Minimum quantity produced by production line $j \in S_1$, and transported on the arc$(i,j) \in A_1$;

$Inv_{i,}^{max}$ Inventory capacity at node (facility)$i \in S_3$;

$Inv_{i,f}^{init}$ Initial inventory at node (facility) $i \in S_3$ of family $f \in F$ at periodT_0;

$D_{f,t}$ Demand of family $f \in F$ at period $t \in T$;

H Large positive penalty constant.

Decision Variables

Continuous

$\varphi_{i,j,f,t}$ Integer flow circulates on the arc $(i,j) \in A$, of product family $f \in F$ at the period $t \in T$;

$Inv_{i,f,t}$ Inventory carried at node (facility) $i \in S_3$ of product family $f \in F$ at the period $t \in T \cup T_0$;

$\varphi_{f,t}^{max}$ Amount of the flow of product family $f \in F$ at the period $t \in T$, able to move on the set of arcs$(i,j) \in A_1$;

$DS_{f,t}$ Demand satisfaction rate of product family $f \in F$ at the period $t \in T$.

Binary

$X_{i,j,f,t}$ Binary variable equal to 1 if product family $f \in F$ is produced at the period $t \in T$ by line production $j \in S_1$ and transported on the arc$(i,j) \in A_1$; 0 otherwise.

D. Integer linear programming model

The model for strategic production-inventory-distribution network planning takes the form of MILP. In the linear formulations introduced below, the objective (1) is to minimize the inventory while maximizing the demand satisfaction rate.

$$\sum_{i,f,t}(Inv_{i,f,t} - H^2 \times DS_{f,t}),\ i \in S_3, f \in F, t \in T \quad (1)$$

Subject to

$$Inv_{i,f,T_0} = Inv_{i,f}^{init}, \forall\, i \in S_3, \forall\, f \in F, T \subset T_0 \quad (2)$$

$$\sum_{f \in F} \varphi_{i,j,f,t} \leq Capa_{i,j}^{max}, \forall\, (i,j) \in A, \forall t \in T \quad (3)$$

$$\sum_{i \in S} \varphi_{i,j,f,t} = \sum_{k \in S} \varphi_{j,k,f,t}, \forall\, t, f \quad (4)$$

$$\forall (i,j) \in A \backslash (A_3 \cup A_7); \forall (j,k) \in A \backslash (A_1 \cup A_5)$$

$$\forall j \in N \backslash S_3, j \geq 1, j \leq |S|, Inv_j^{max} = 0$$

$$Inv_{j,f,t-1} + \sum_{i \in S} \varphi_{i,j,f,t} = \sum_{k \in S} \varphi_{j,k,f,t} + Inv_{j,f,t} \quad (5)$$

$$\forall (i,j) \in A_3; \forall (j,k) \in A_5$$

$$\forall j \in S_3, j \geq 1, j \leq |N|, Inv_j^{max} > 0, \forall\, t \in T, f \in F$$

$$\sum_{f \in F} Inv_{,i,f,t} \leq Inv_i^{max} \quad \forall\, i \in S_3, t \in T \cup T_0 \quad (6)$$

$$\sum_{i \in S} \varphi_{i,j,f,t} = SD_{f,t} \times D_{f,t} \quad (7)$$

$$j = |N|, j \in S_5;\ \forall t \in T, f \in F, \forall\, (i,j) \in A_7$$

$$SD_{f,t} \leq D_{f,t}\ \forall t \in T, f \in F \quad (8)$$

$$\varphi_{f,t}^{max} = \sum_{j \in S} \varphi_{0,j,f,t}, \forall t \in T, f \in F \quad (9)$$

$$\varphi_{0,j,f,t} \leq capab_{f,j} \times Capa_{0,j}^{max}, \forall t \in T, f \in F, j \in S_1 \quad (10)$$

$$\varphi_{i,j,f,t} \geq X_{0,j,f,t} \times Prod_{i,j}^{min}, \forall t \in T, f \in F, (\mathrm{i,j}) \in A_1 \quad (11)$$

$$\varphi_{i,j,f,t} \leq X_{i,j,f,t} \times Capa_{i,j}^{max}, \forall t \in T, f \in F, (\mathrm{i,j}) \in A_1 \quad (12)$$

$$X_{i,j,f,t} \leq \varphi_{i,j,f,t}\ \forall t \in T, f \in F, (\mathrm{i,j}) \in A_1 \quad (13)$$

$$\varphi_{i,j,f,t} \geq 0, \varphi_{f,t}^{max} \geq 0,\ Inv_{i,f,t} \geq 0,\ DS_{f,t} \in (0..1) \quad (14)$$

$$X_{i,j,f,t} \in [0,1] \quad (15)$$

Constraints (2) ensure that the quantity of product family $f \in F$ held in warehouse $i \in S_3$ at the beginning of T is equal to the initial inventory. This set of constraints models the inventory at the beginning of the planning horizon. Constraints (3) are capacity ones, which guarantee that the flow circulating on the arc $(i,j) \in A$, of product family $f \in F$ at the period $t \in T$ cannot exceed the capacity of the arc. Constraints (4) are the flow conservation at nodes $N \backslash (S_1 \cup S_3 \cup S_5)$. Constraints (5) are the flow conservation at warehouses. Constraints (6) ensure that the sum of the stored quantity at node (facility) $i \in S_3$ of product family $f \in F$ at the period $t \in T$ is lower or equal to the inventory capacity at node i. Constraints (7) ensure that the multi-flow arriving at the sink correspond to the demand of product family $f \in F$ at period $t \in T$. Constraints (8) verify that for a product family demand $D_{f,t}$ equal to zero, the satisfaction demand rate $SD_{f,t}$ is null. Constraints (9) ensure that the flow circulating on the arc $(i,j) \in A_1$ is equal to the quantity of product family $f \in F$ produced by the production line $j \in S_1$ at the period $t \in T$. Constraints (10) verify arc capacity constraints as well as the production line capability to produce à given family $f \in F$. Constraints (11), (12) and (13) define lower and upper bounds of flows circulating on arcs with given minimum capacities and impose the bounds on arcs belonging to set A_1. So, by using binaries variables $X_{i,j,f,t}(i,j) \in A_1$, we restrict value of each flow circulating on such arcs to be upper than the minimum capacity (11) and lower than the maximum capacity (12) or null (13), i.e. each production line $j \in S_1$ does not produce more than the capacity of the arc $(i,j) \in A_1$. Constraints (14) and (15) define the domain for the decision variables.

IV. Numerical example

In order to illustrate our model we propose an easy understandable real numerical example issue from company mentioned above. It is composed by a multi facility system consisting of 2 factories comprising three manufacturing lines each. Such a system has a portfolio of six product families. Two product families are allocated to one of two facilities and others can be allocated during the planning process on any facility, it means that two facilities have the ability to produce each of them. All manufacturing lines ($|S_1| = 6$) belonging to a given

facility have the ability to make any product family allocated to this facility. There are 49 ($|S_2| = 47$) conveyors which can transport, storage and retrieval the product, 9 ($|S_3| = 9$) storage sheds, and 2 ($|S_4| = 2$) distribution facilities. Table I summarize the sets of arcs which are used in the flow network model.

TABLE I. Number of arcs in each subset

From	To					
	S	S_1	S_2	S_3	S_4	S_5
S		$\|A_1\|$=6				
S_1			$\|A_2\|$=6			
S_2			$\|A_4\|$=94	$\|A_3\|$=9	$\|A_6\|$=3	
S_3			$\|A_5\|$=9			
S_4						$\|A_7\|$=2
S_5						

The product family demand is strongly seasonal as shown in Fig. 2. and Fig. 3.

Fig. 2. Demand forecast for product families F1, F2 and F3

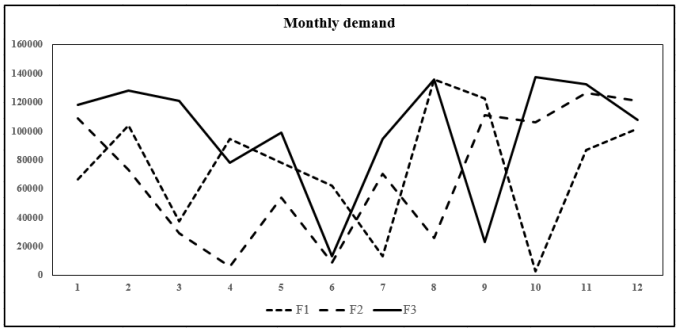
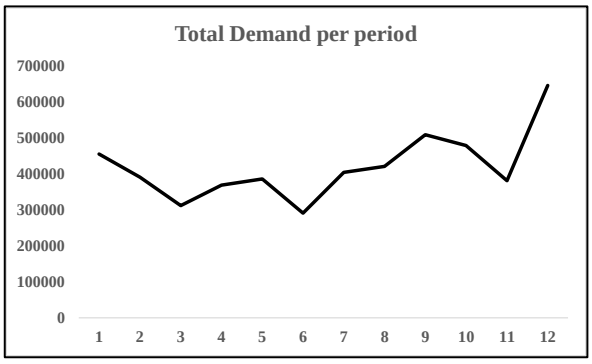


Fig. 3. Demand forecast for all families



This problem has been solved in less than 5 seconds using Cplex. Corresponding results are reported in TABLES II, III, IV and TABLE V. The total demand has been satisfied. As illustrated in table II the capacity of manufacturing line 1 is fully utilized. Table III underlines that during the sixth period, manufacturing lines are used in average 70% except line 6 which is used 100%. The same conclusion holds for period 3.

TABLE II. Capacity of manufacturing line 1 in tons/month

Period	Family	Capacity used per family	Total Capacity used per period
1	2	64800	64800
2	3	50587	64800
2	4	9081	
2	6	5132	
3	6	64800	64800
4	1	64800	64800
5	3	64800	64800
6	2	8743	64800
6	6	56057	
7	2	10387	64800
7	3	29489	
7	6	24924	
8	1	64800	64800
9	3	22854	64800
9	4	41946	
10	6	64800	64800
11	1	10932	64800
11	2	53868	
12	4	25827	64800
12	6	38973	

TABLE III. Capacity used during period 6

Manufacturing line	Family	Capacity used	% of capacity used
1	2	8743	13%
1	6	56057	86%
2	4	35788	55%
2	6	15165	23%
3	1	62176	48%
3	3	13138	10%
5	5	13850	16%
6	4	86400	100%
Average manufacturing line utilization			70%

TABLE IV. Capacity used during period 11

Manufacturing line	Family	Capacity	Production
1	1	64800	10932
1	2		53868
2	3	64800	39684
2	6		25116
3	2	129600	64800
3	3		64800
4	1	86400	27770
4	2		7762
4	3		50868
5	1	86400	75954
5	4		10446
6	3	86400	27200
6	5		50000
Total			509200

Analyzing the demand at period 11 and 12, it is obvious that demand increase in a huge way. To satisfy the demand it is necessary to anticipate the production and to store finished products in period 11 in order to satisfy the demand at period 12. As shown in TABLES IV and V, the production has been anticipated in period 11. Never mind that the whole demand in period 11 is 381 430 tons, the production is 509 200 tons and the inventory is 127 770 tons. During the period 12 the whole demand is 646 170 tons when the production is 518 400 tons. The inventory of the period 11 enables to satisfy the demand of period 12.

TABLE V. Capacity used during period 12

Manufacturing line	Family	Capacity	Production
1	4	64800	25827
1	6		38973
2	1	64800	64800
3	4	129600	64800
3	6		64800
4	1	86400	8905
4	2		77495
5	2	86400	43284
5	5		43116
6	3	86400	57490
6	4		22319
6	5		6591
		Total	518400

Through this example two mains aspects have been shown. First, we have highlighted the importance of strategic capacity allocation process as a way to evaluate investment while keeping operational efficiency. This relevance is also demonstrated by the fact that the optimal solution is obtained in few seconds in the study case with real industrial size.

V. Conclusion

In this paper an aggregate planning approach for process industry have been proposed. A linear optimization model for strategic production-inventory network planning of a plant manufacturing bulk materials is presented. The planning problem is formulated as a multi commodity network model that takes the form of a MILP which objective is to maximize demand satisfaction rate while minimizing inventory. The scope of the study is a multi-facility production, inventory and distribution system belonging to the process industry.

The experimental studies investigated relationships between capacity planning and the impact of inventory management during the strategic decision making. The proposed model is demonstrated by a comprehensive example of a real industrial problem size. The experimental results prove that the proposed model is relevant in the scope of manufacturing literature. Indeed, it offers a model that is easy to use for capacity design through a "what-if process" during strategic planning, comparing to complex existing tools. Moreover, it gives a model able to face to any real unexpected peaks in demand in a short time in operational context.

The future work consists of designing decision tool for strategic planning in order to include Monte Carlo simulation and by this way to take better into account demand uncertainty.

References

[1] R. Anthony, Planning and Control Systems. A Framework of Analysis, Boston, 1965.

[2] B. Fleischmann, H. Meyr, and M. Wagner. Advanced Planning, In Stadtler, H., Kilger, C. (ed.), Supply chain management and advanced planning. Springer Verlag, Berlin, 2002.

[3] H. Li, H. Linda, and T. Ruud, A strategic capacity allocation model for a complex supply chain: Formulation and solution approach comparison, International Journal of Production Economics, vol. 121, p. 505-518, 2009.

[4] C. Vidal and M. Goetschackx, Strategic production distribution models: A critical review with emphasis on global supply chain models. European Journal of Operational Research vol. 98, p. 1–18, 1997.

[5] M. Melo, S. Nickel, F. Saldanha-da-Gama, Facility location and supply chain management – A review, European Journal of Operational Research, vol. 196, p. 401-412, 2009.

[6] J. Kallrath, Planning and scheduling in the process industry. OR Spectrum vol. 24, p. 219–250, 2002.

[7] G. Eppen, R. Martin, and L. Schrage. A scenario approach to capacity allocation, Operations Research, vol. 37, p. 517-527, 1989.

[8] V. Jayaraman and H. Pirkul. Planning and coordination of production and distribution facilities for multiple commodities, European Journal Operational Research, vol. 133, p. 394-408, 2001.

[9] S. Nam and R. Logendran, Aggregate production planning - A survey of models and methodologies, European Journal of Operational Research vol. 61, p. 255-272, 1992.

[10] S. Chopra, and P. Meindl, Supply Chain Management, Pearson Education, Inc., Addison Wesley Longman Pte. Ltd., Singapore, 2001.

[11] H. Luss, Operations Research and Capacity Expansion Problems: A Survey. Operations Research, vol. 30(5), p. 907-947.

[12] J.A. Van Mieghem, Capacity Management, Investment, and Hedging: Review and Recent Developments, Manufacturing & Service Operations Management, vol. 5(4), p. 269-302.

[13] (10) S.D. Wu, M. Erkoc and S. Karabuk, S. Managing Capacity in the High-Tech Industry: A Review of Literature. The Engineering Economist, vol. 50, p. 125-158.

[14] S.Kemmoe, P-A. Pernot, and N. Tchernev, Model for flexibility evaluation in manufacturing network strategic planning, International Journal of Production Research, vol. 52,(15), pp. 4396-4411, 2014.

Project Management

Development of professional competences for industry 4.0 project management

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Cerezo-Narvaez, Alberto; Otero-Mateo, Manuel; Pastor-Fernandez, Andres
Dept. mechanical engineering and industrial design
University of Cadiz
Puerto Real, Spain
alberto.cerezo@uca.es

Abstract—In the context of industry 4.0, this research highlights the importance of people. Based on the premise that technology alone cannot guarantee the success of projects undertaken, the human factor should not be forgotten, because Industry 4.0 is not only based on technology but also on the creativity of people applying technology. Thus, organizations have to develop the ability to accept, adopt and exploit the technological vanguard, data integration and automation, improving their levels of productivity and efficiency, as well as their health, safety and environmental standards. In response to this weakness, project management (PM) responds, thanks to administrate a battery of knowledge, attitudes, aptitudes and experiences, both specific and generic and human, facilitating the achievement of objectives and strategies.

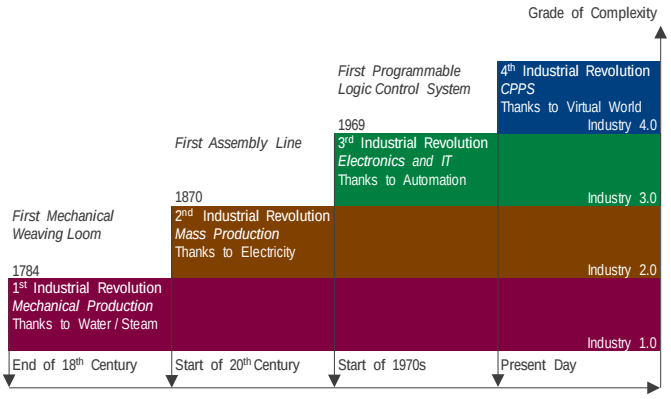
The International Project Management Association (IPMA) competence baseline (ICB), focused on people, analyzes the training needs of Industry 4.0 workers, whose roles evolve from craft and mechanical works to supervision and management jobs, proposing a professional development plan, for a better adaptation to change. To assess the PM importance in Industry 4.0, a structural equation model is proposed, providing strong causal relationships among PM skills, from change and transformation, intrinsic dimension characteristics and value drivers of Industry 4.0.

Keywords—Industry 4.0; professional competence; project management; IPMA ICB

I. Introduction

After a period in which Industry 3.0 was based on Information Technology (IT) and on a high level of automation, a giant leap was made towards Industry 4.0, from a need to update the role of industrial sectors for global economy growth. Over the last few years, Industry 3.0 has allowed manufacturing processes to be optimized. International guidelines, such as ISO 9001 [1], have allowed continuous improvement systems to be established within organizations, with approaches focusing on main stakeholders and the use of IT tools to support the Organization Management System. Nevertheless, as shown in Fig. 1, the industrial paradigm obsolescence based on manufacturing processes optimization as an isolated mechanism, without any influences from outside the production system, with a high level of product standardization, is a hurdle that must be overcome in order for Industry 4.0 to be effectively implemented.

Fig. 1. Industrial Evolution. Based on [2] and [3]



The Spanish Government, through the Ministry of Economy, Industry and Competitivity (MINECO), led by the Spanish Association for Standardization (AENOR), which holds the Secretariat of the Technical Committee CEN/TC 389 "Innovation Management" of the European Committee for Standardization (CEN), has set up the “Connected Industry 4.0” initiative, whose objective is to augment the contribution of the industrial sector to the GDP, increase employment and bring about a positive trade balance, by driving the digital transformation of Spanish industry.

Connected Industry 4.0 has been launched as a joint initiative coordinated by both the public and private sectors in Spain, similar to initiatives in other European Union countries. The European Commission, together with its Member States and industry itself, established a governance framework in the first semester of 2016, and more recently, in March 2017, one of the key milestones has been established: the launch of the European platform for national initiatives. The objective of this platform is to create a collaborative space to gather together information on the different initiatives in existence.

This new industrial revolution has a series of characteristics in which technology is the key tool to articulate change. In Industry 4.0, data is processed efficiently, effectively and rapidly, to meet value chain needs, from product or service design to customer usage. This has a direct impact on concurrent engineering, as all stakeholders are involved in each stage of product or service life cycle. It is a way of meeting both project and product or service needs and expectations, from its conception to its final delivery, optimizing costs, setting deadlines, dealing with risks and addressing mandatory and agreed quality requirements.

In this context, the existence of collaborative networks is encouraged, increasing the exchange of knowledge, both horizontally and vertically, so that channels are created to identify improvements, problems, errors and solutions, linked directly to the final product or service.

With the emergence of organizations in Industry 4.0, "connected people" or "professionals 4.0" appear, who are distinguished by being more efficient, more flexible, faster and, consequently, more competitive, thanks to the total connectivity among machines, systems, and people. This is essential to change, as it favours value chain generation and reduces waste. Likewise, this results in a greater connection between processes and decision-makers, which means tasks can be quickly and correctly assigned, at the same time encouraging collaboration with stakeholders, improving transparency and information transfer. To this end, means are established to gather data automatically, facilitating global traceability.

Qualified human capital is, therefore, fundamental, as this is very recently-developed technology, making adequate training essential for the new industrial revolution to be sustainable and develop its full potential [4]. Workers must accept moving to new profiles where new roles will require, as a priority, PM competences intrinsic to Industry 4.0. From a human point of view, success will lie in being able to communicate, share, and use information to solve complex problems, adapt and innovate in response to new demands and changing circumstances, marshal and expand the power of technology to create new knowledge and expand human capacity and productivity [5].

Industry 4.0 workers have to acquire new responsibilities, develop decision-making skills and the ability to apply a global vision of all processes [6]. Besides, continuous training and individual appraisals are necessary to manage change and organizations must anticipate training demands in order to face these challenges, and be aware of needs for training throughout workers' career [7]. If organizations identify management skills and key functions required, then they can establish a competence framework for promoting performance excellence [8]. In this regard, there are many studies that research the relationship between PM skills and project success [9-14].

In the development of competences, the IPMA approach can be used as a reference. IPMA is the World's first PM association (federation of 70 national PM associations), founded in 1965, that offers unique, role-specific competence development guidelines, for improved project success since its first ICB in 1999 [15]. This approach was included by the Project Management Institute (PMI), the World's leading professional membership association for PM (with certification holders in 210 countries), founded in 1969, recognizing in its first project manager competency development framework (PMCDF) in 2002 that good practices are not enough to be effective, being required to domain competency dimensions [16]. In 2012, with the international standard ISO 21500 [17], management by competences is definitively consolidated.

This research identifies the importance of human capital using the IPMA ICB 4.0 standard on project manager competences, and its influence on the adaptation of people to the change and transformation required for this new industrial revolution, both in its dimensions and value drivers.

II. Transition to Industry 4.0

Industry 4.0 is built on firm pillars like innovation, processes automation and sophistication, using new technology, but without forgetting that the main asset of an organization is its people. The Internet of Things (IoT) will be the vehicle that joins machines, people and data, integrating information systems in such a way that communication among all these elements will speed up decision-making, according to standardized criteria, eliminating the human subjectivity, in the field of spontaneity.

Industry 4.0 is full of challenges and opportunities, as providing clients with rapid responses, fully-adapted to their needs, means having a competitive advantage. In order to do this, it needs to be globally integrated within the supply chain, but placing clients at the center of the process, so that organizations can obtain a personalized made-to-measure product. This vision is a concept that was proposed by P. Kotler in 1989 [18], even before the World Wide Web (www) was developed in the 1990's, focusing products and services production personalized to clients, towards low cost, high quality and supplying large production volumes.

The reason mass personalization concept has not had a great impact on industry up until now is mainly due to difficulties in obtaining and assimilating information (big data), but this change has been imposed with globalization and demand change to a faster rhythm, which requires flexibility in the production process that was not possible to achieve with industry 3.0.

A big challenge of Industry 4.0 is, therefore, the wide variety of data and the difficulty of its acquisition, as it is not all homogeneous, and needs to be integrated for good decision-making. For this reason, the independent systems must interconnect and interoperate. Another challenge is to face online threats and, therefore, an effective control must be carried out by means of cybersecurity tools and protocols.

Legislation based on the identification and designation of European Critical infrastructures and an evaluation of the need to improve their protection (Directive 2008/114/CE), establishes, among other things, that the main and ultimate responsibility to protect critical infrastructures falls on to the Member States and their operators, calls for the implantation of a series of initiatives and actions on the part of the States for its transposition to national legislations, and tries to measure critical infrastructures protection, which is developed in the community initiative for the Global European Strategy in 2013.

In Spain, for instance, a Cybersecurity Code of Law (modified on the 9th of March 2017) includes the classification of critical infrastructures that can be divided into these strategic areas: Energy, Nuclear Industry, IT, Transport, Water Supply, Food Supply, Health, Financial Sector, Chemical Industry, Space, Resources and Administration. Currently, there are 3,500 sensitive installations and infrastructures registered in Spain and any project carried out in this type of infrastructure must meet legal requirements referred.

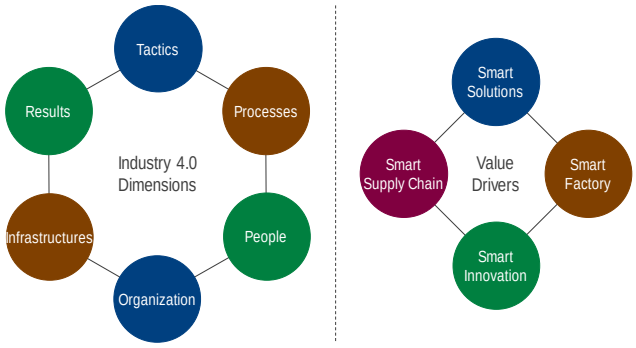
All the afore-mentioned dimensions are compiled in the Advanced Digital Auto-diagnosis Tool (HADA) [19], launched in March 2017 by the MINECO, defining a digital maturity model in industry 4.0, with the aim of boosting understanding and an approximation to new business transformation paradigm.

This digital maturity model has been used as a base to draw up a specific business model, shown in Fig. 2, via the different key dimensions in business strategy for industry 4.0:

- Business model and tactics:
  Evaluating the adaptation capacity of the organization to its surroundings and to the market
- Processes:
  Analyzing the digital capacities of the operating model
- People:
  Identifying the abilities and qualifications of people
- Organization:
  Relating the capacities of the business structure with other agents
- Infrastructures:
  Identifying the transformation capacity permitted by the cyber physical means
- Results:
  Evaluating the level of incorporation of technology to existing products and services, as well as its digitalization potential

In order to validate the level of maturity in the implantation of Industry 4.0, studies by Capgemini Consulting [20], Deloitte AG [2] and McKinsey Digital [3] on Industry 4.0 have been taken as a reference. The importance of digital technology is highlighted to increase customer value, which strongly affects innovation (intelligent innovation and collaboration models). With the aim of better understanding, core value drivers are shown in Fig. 2. On one hand, "Smart Solutions" and "Smart Innovation" allow companies growth to be taken advantage of, and on the other hand, "Smart Supply Chain" and "Smart Factory" drive organizations efficiently. These drivers allow to subsequently establish a causal relationship with "Key Success Factors", by means of the questionnaires proposed.

Fig. 2. Industry 4.0 (Dimensions and Value Drivers)



III. Professional competences in Industry 4.0

The organization's culture has a great influence on the human resources' strategy, affecting everyday aspects, thoughts and behaviours of general managers and employees. Therefore, it is only when the culture, the organization and the human resources strategy work as a coherent whole that strategies are effective and contribute to creating competitive advantages [21].

Knowledge can be defined as the combination of experience, know-how, values, information, perception and ideas that create a framework of mind that helps people to assess and generate new ideas, knowledge and experience [22].

Human capital (HC) plays a key role in knowledge since it adds value to the knowledge and competences of the organization's people, and the capacity to generate them is useful to achieve the organization's mission [23]. HC refers to the knowledge acquired by people increasing their productivity and adding value to their contribution. HC includes employee's personal contacts and relationships, as well as other individual qualities such as reputation, loyalty, multi-tasking or flexibility [24]. From a business perspective, there is a strong relationship between HC and success. This relationship becomes stronger when tasks to be performed are complex and require highly qualified employees [25]. Learning plays a very important role in developing HC in all its dimensions, both individual and professional [26].

A holistic framework is useful in identifying the combination of competences that are necessary for particular occupations and to promote labour mobility [27]. Table I shows how the most prestigious PM associations and scientific researchers have organized competences around common areas:

TABLE I. Areas of competence

Standards & Authors	Human	Specific	Generic
[17] ISO 21500	Behavioural	Technical	Contextual
[15] IPMA ICB 4	People	Practice	Perspective
[28] PMI Talent Triangle	Leadership	Technical	Strategy&Business
[16] PMI PMCDF 3	Personal	Performance	Knowledge
[8] Cheng, Dainty & Moore	Attitudinal	Occupational	Understanding
[27] Le Deist & Winterton	Social	Functional	Cognitive
[5] Binkley et al.	Living in the world	Tools for working	Ways of thinking
[29] Onisk	Behavioural	Professional	Compliance

This research focuses on the IPMA's model on PM, examining the competence elements in the Industry 4.0. With a PM approach, it is drawn on theoretical notions, as well as professional experiences of stakeholders and mainly on trainers.

The work of a project manager in industry 4.0 involves optimizing raw materials, reducing energy costs, reducing waste and therefore making the project more efficient and environmentally-friendly. An organization that follows an Industry 4.0 philosophy will be better positioned against its international competition. As is indicated in the ICB 4.0 standard, in its competence change and transformation: "People usually do not object to change – they object to being changed. Successfully addressing their objections can be done by, for example, building support, addressing resistance and developing the required knowledge and ability to implement the change…

…Projects usually deliver new capabilities. However, it is only when these capabilities are put to use that value is added and benefits can be achieved. Organizational or business changes often affect or alter processes, systems, organizational structure and job roles, but most of all they influence people's behaviour. Changes can be quite small or they can require a complete transformation."

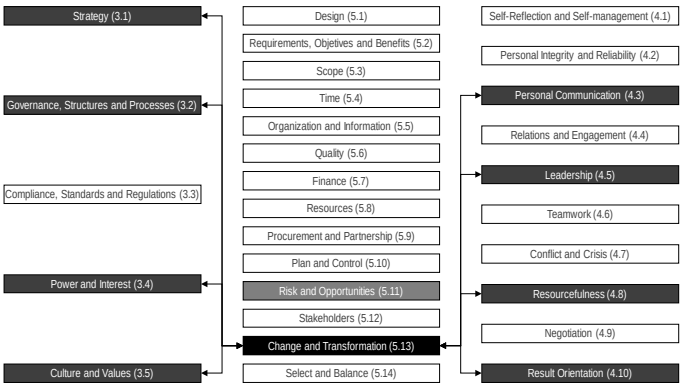
The appearance of industry 4.0 means, therefore, a change in business models and a jump in the way of managing projects, as all stakeholders have an active role in decision-making. Workers, in their new role as project managers, as well as having more tools at their disposal for managing the industrial process, have to have greater human and generic skills to satisfy the needs and expectations of this new industrial revolution.

IV. Relational model of PM competences with Industry 4.0

IPMA establishes in its ICB 4.0 standard, competences for project managers, structured in three areas: perspective, practical and people. Change and transformation management belong to these groups of practical competences. This competence provides the process with the tools and techniques that can be used to help people and organizations on the way to success in carrying out this change and transformation.

On analysing each of the ICB 4.0 (International Project Management Association 2015), competences, relationships can be seen between the different competences. The conceptual model of the relationships existing between the people-practice-perspective competences is shown in Fig. 3.

Fig. 3. Influences on Change and Transformation Management

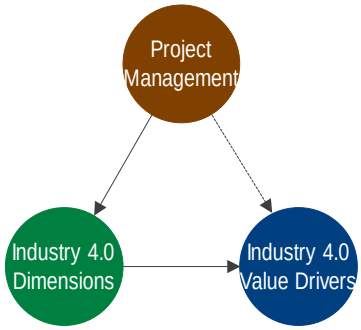


Change management is directly related to perspective competences (oriented towards results, governance, structures and processes, power and interest, culture), and people ones (communication, leadership, ingenuity and goals, objectives and benefits), and therefore the analysis of the group of related competences allows a conceptual model to be established, for its subsequent statistical validation.

For the case being studied, it is necessary to identify the core competences, which is done in two levels, in order to establish a knowledge framework on Project Management competences in Industry 4.0. A collection of essential and fundamental skills are selected from literature [27,30-34], as well as the relationships included in the ICB 4.0, as shown in Fig. 3.

Taking into account the literature review, the model proposed and based on competences in PM and Industry 4.0 that establishes the relationship among Project Management (PM), Industry 4.0 Dimensions (ID), and Industry 4.0 Value Drivers (VD) constructs, is validated in accordance with the Fig. 4.

Fig. 4. Conceptual Model Influences PM with Industry 4.0



V. Validation

Research shows the influence of PM Competences on the Dimensions of Industry 4.0, and these dimensions on the Value Drivers of Industry 4.0, establishing a relationship, both direct, and of second level between PM Competences and Industry 4.0 Value Drivers, as shown in Fig. 4.

Observation by questionnaire is the method used to confirm raised hypothesis. The collection and prosecution of data is managed by the free platform Google Forms. Based on the analysis of responses received, model under study is characterized and the relationships among interest variables are contrasted statistically. To achieve it, a specific scale of five items, from 1 to 5, is adopted, allowing to measure attitudes and determine the grade of compliance with proposed claims, and helping professionals in motivating their choices to capture their intensity towards those affirmations.

To select items, a structured two-step approach is adopted. From literature's primary sources, a list is composed by characteristics of Change and Transformation, Project Management and Industry 4.0; and success criteria for Value Drivers Industry 4.0. From literature's secondary sources, previous list is refined with critical success factors, defined as characteristics or variables that cause an essential and direct impact, in Dimensions and Value Drivers of Industry 4.0; and critical success factors in projects, business and organizations. Based on the analysis of information sources, the questionnaire is prepared in order to ask respondents about the degree of importance given to each of items that are exposed in Table II:

TABLE II. Critical Issues

PM.Project Management		
CT.Change&Transformation	PV.Perspective	PS.Personal
1.Change Capability 2.People Behaviour 3.Assistance to Change 4.Risk and Opportunity	1.Strategy 2.Governance & Structures 3.Power and Interest 4.Culture and Values	1. Personal Communication 2. Leadership 3. Resourceffulness 4. Result Orientation
ID.Industry 4.0 Dimensions		
TT.Tactics	PC.Processes	PP.People
1.I+D Proposals 2.Sustainability 3.Investment Plans	1.Digitization 2.Data Integration 3.Automation	1.Digital Skills 2.Digital Training 3.Qualifications
OG.Organization	IT.Infraestructures	RS.Results
1.Motivation to Transform 2.Collaboration 3.Digital Roles	1.Technological Capacity 2.Business Solutions 3.Collaborative Platforms	1.Digital Features 2.Interconections 3.Data Analysis

VD.Industry 4.0 Value Drivers			
SS :Smart Solutions	SI.Smart Innovation	SF.Smart Factory	SC.Smart Supply
1.New Opportunities 2.Client Relationship	1.Time to Market 2.Customization	1.Production 2.IT Infraestructure	1.Efficiency 2.Value Proposition

Once the questionnaire is fully defined, it is distributed by social networks, through official councils of professionals and thanks to PM associations. After the exhibition period is ended, 355 responses are received, of which 35 are rejected (almost 10%) for not presenting respondents a demonstrable competence in PM and Industry 4.0 (neither experience nor expertise). To describe the final sample of 320 individuals, a series of control questions are made, in order to characterize the sample population, related to personal issues, with the organization where respondents provide their services and about PM, which are summarized in Table III.

TABLE III. Characteristics of Sample Population

Organization Size	U's	%	Organization Role	U's	%	Training in PM	U's	%
Self-Employment	65	20,3	Technician	67	20,9	Neither	78	24,4
Micro	92	28,8	Member Team	60	18,8	University Grade	61	19,1
Small	52	16,2	Project Manager	135	42,2	Postgrade	106	33,1
Medium	38	11,9	Program Manager	31	09,7	Master	60	18,7
Large	73	22,8	Portfolio Manager	27	08,4	Doctorate	15	04,7
Industrial Sector	**U's**	**%**	**Project Complexity**	**U's**	**%**	**Experience in PM**	**U's**	**%**
IT	63	19,7	Very Simple	33	10,3	< 01 year	27	08,4
Automotive	42	13,1	Simple	107	33,4	01-05 years	51	15,9
Aerospace	34	10,6	Normal	83	26,0	06-10 years	93	29,1
Research	33	10,3	Complex	70	21,9	11-20 years	112	35,0
Others	148	46,3	Very Complex	27	08,4	> 20 years	37	11,6

Then, it is necessary to guarantee the accuracy and reliability of the measuring instrument, in order to ensure the results from answered surveys. With a heterogeneity of 19,55%, the sample population incurs a statistical error of 4,45% for a confidence interval of 95,45%. It also has a reliability of 90,55%, according to Cronbach test. As a descriptive summary of the sample population, in Table IV, the mean ($\dot{x}$), standard deviation (σ), and homogeneity (r_{i-t}) are presented, grouping into constructs.

TABLE IV. Results of Statistical Study

PM	$\dot{x}$	σ	r_{i-t}	ID	$\dot{x}$	σ	r_{i-t}	VD	$\dot{x}$	σ	r_{i-t}
CT	3,93	1,00	0,78	TT	4,44	0,64	0,77	SS	4,28	0,75	0,72
CT1	3,93	0,93	0,55	TT1	4,61	0,57	0,56	SS1	4,09	0,80	0,78
CT2	3,93	0,93	0,62	TT2	4,52	0,59	0,61	SS2	4,46	0,64	0,51
CT3	4,02	1,02	0,65	TT3	4,18	0,68	0,54				
CT4	3,84	1,08	0,61	PC	4,33	0,71	0,88				
				PC1	4,01	0,71	0,61				
				PC2	4,48	0,64	0,63	SI	4,36	0,70	0,88
				PC3	4,49	0,67	0,71	SI1	4,41	0,68	0,76
PV	4,36	0,72	0,81	PP	4,20	0,73	0,94	SI2	4,32	0,72	0,65
PV1	4,04	0,82	0,61	PP1	4,04	0,70	0,64				
PV2	4,32	0,68	0,70	PP2	4,35	0,72	0,83				
PV3	4,63	0,57	0,57	PP3	4,22	0,74	0,74				
PV4	4,43	0,65	0,55	OG	4,26	0,74	0,82	SF	4,29	0,71	0,87
				OG1	4,33	0,77	0,66	SF1	4,29	0,76	0,70
				OG2	4,44	0,65	0,64	SF2	4,28	0,66	0,76
				OG3	4,02	0,74	0,62				
PS	4,48	0,66	0,82	IT	4,06	0,84	0,93				
PS1	4,46	0,63	0,61	IT1	4,03	0,84	0,72				
PS2	4,35	0,71	0,63	IT2	4,23	0,75	0,75	SC	4,49	0,63	0,90
PS3	4,62	0,62	0,59	IT3	3,90	0,89	0,58	SC1	4,54	0,61	0,72
PS4	4,47	0,67	0,62	RS	4,05	0,88	0,93	SC2	4,45	0,65	0,82
				RS1	3,89	0,84	0,80				
				RS2	4,28	0,82	0,75				
				RS3	3,97	0,92	0,73				
PM	4,25	0,84	0,99	ID	4,22	0,77	0,99	VD	4,36	0,70	0,99

To build the structural equation model (SEM), items that make up every dimension are grouped in latent variables. Subsequently, these constructs are related to each other, forming a network of interrelations that recognize their complexity and multidimensionality. However, SEM exhibits suggested adjustments, thanks to Lagrange multipliers test, in which improvement is obtained by adding a series of covariances to errors of the theoretical model, being proposed only those that are statistically significant, establish relationships among factors errors from the same latent variable and own a theoretical baseline justification. Scores on each dimension, do not only depend on the main constructs, but they are influenced by terms of (residual) errors. In this way, the constructs are free of measurement error, because these parameters are different from the variables which are associated: PM, ID and VD.

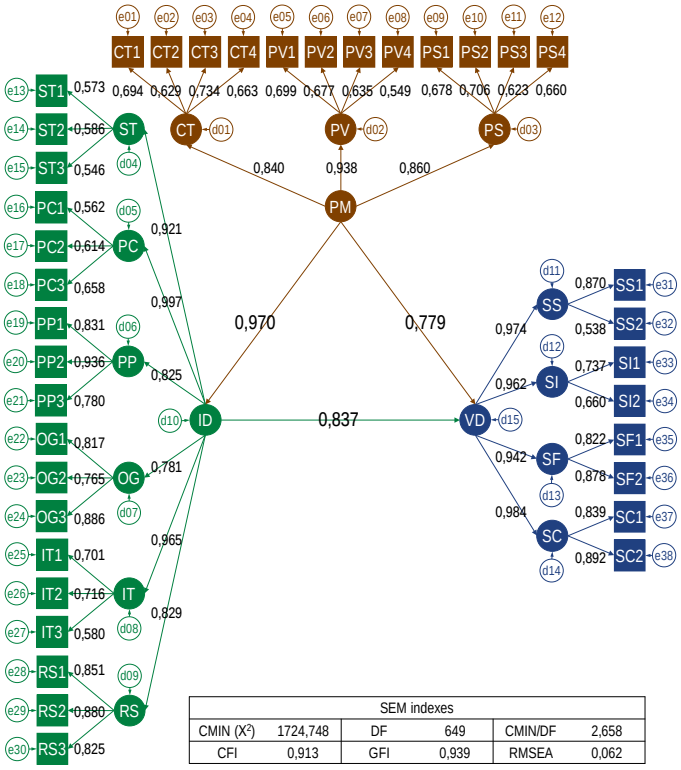
Thanks to obtained results in the statistical model, as can be checked in Fig. 5, causal analysis can be validated, because in all studied cases, each of the following requirements occur:

- CMIN/DF: Chi-square fit index divided by its degrees of freedom, lower than 3
- CFI: Comparative fit index, confronting the fit between the independent and the estimated model and being located between 0 and 1, higher than 0,9
- GFI: Goodness of fit index, expressing the variability explained by the estimated proposed model and being located between 0 and 1, higher than 0,9
- RMSEA: Root mean square error of approximation, predicting how the model fits the covariance matrix of population sample, lower than 8%

Once constructs are modeled, the relationship that binds PM, IM and VD, as main goal of this research, must be specified. Fig. 5 shows the structural equation model (SEM) that links them:

- Relationship between ID-VD is, for sample population, 83.7%, so that Industry 4.0 Dimensions justifies the 70.1% of the variance of Industry 4.0 Value Drivers
- Relationship between PM-ID is 97.0% and between PM-VD is 77.9%, so that Project Management justifies the 60.7%, by direct effect, and the 66.3%, by indirect effect, of the variance of Industry 4.0 Value Drivers, bringing its total effect to 81.2%

Fig. 5. Structural Equation Model (IBM®SPSS Amos v22.0)



However, this model must be understood as a first justifiable step to analyze, with greater precision, the relationships between PM areas with Industry 4.0 Dimensions and Value Drivers, in order to prioritize those PM domains in which to act, thanks to its impact on this new emerging industry critical factors.

VI. Conclusions

From the model described, based on the IPMA ICB 4.0 standard, there is a body of competences, directly and significantly related to Industry 4.0 dimensions. Competences affecting the professional development of organizations' human capital are critical success factors to move forward securely in the transition towards Industry 4.0 value drivers, confirming Dulewicz & Higgs [9], Koutsikouri, Dainty and Austin [10], Müller & Turner [11], Turner, Müller & Dulewicz [12], Kandelousi, Ooi & Abdollahi [13] and Awan, Ahmed & Zulqarnain [14] conclusions. PM by competences allows to lead people, manage processes and administrate resources.

Practical skills, although they are factors defining success in the implantation of projects, paving the way for a transition towards the new industrial paradigm, are not enough in themselves to tackle this change, but also require the opening of perspective competences (strategy, governance & structures, power & interest and culture & values) and people ones (communication, leadership, resourcefulness and results orientation), as well as the central competence (change and transformation), enriching the Le Deist & Winterton [27], Stevenson & Starkweather [30], Chipulu et al., [31], Omidvar, Samar & Alias [32], Su, Goluvich & Robbins [33] and Zadeh et al. [34] models, in the search for success in the implantation of projects in the current context of information technology.

Industry 4.0 means a new challenge for PM practitioners, as it has been demonstrated that there are highly relevant relationships between behavioural, technical and contextual competences. This means it will be necessary for project managers to adapt to this new industry needs and expectations.

Once the importance of people on Industry 4.0 is confirmed, most critical competences should be emphasized and prioritized in future researches, according to their Industry 4.0 impact.

References

[1] Otero-Mateo M, Pastor-Fernández A. Impact of the ISO 9001:2015 standard in the engineering field. Integration in SMEs. DYNA Ingeniería e Industria 2016; 91(1): 118-121.

[2] Schlaepfer RC, Koch M, Merkofer P. Industry 4.0. challenges and solutions for the digital transformation and use of exponential technologies. Zurich, 2015.

[3] Wee D, Kelly R, Cattell J, et al. Industry 4.0. How to navigate digitization of the manufacturing sector. Munich, 2015.

[4] Rodriguez D, Patel R, Bright A, et al. Developing competency models to promote integrated human resource practices. Human Resource Management 2002; 41(3): 309-324.

[5] Binkley M, Erstad O, Herman J, et al. Defining 21st century skills. Draft White Paper 1. Melbourne, 2010.

[6] Knowles-Cutler A, Lewis H. Talent for survival. Essential skills for humans working in the machine age. London, 2016.

[7] van Laar E, van Deursen AJAM, van Dijk JAGM, et al. Relation between 21st-century and digital skills: A systematic literature review. Computers in Human Behavior 2017; 72: 577-588.

[8] Cheng MI, Dainty ARJ, Moore DR. What makes a good project manager? Human Resource Management Journal 2005;15(1):25-37.

[9] Dulewicz V, Higgs M. Assessing leadership styles and organisational context. Journal of Managerial Psychology 2005; 20(2):105-123.

[10] Koutsikouri D, Dainty A, Austin S. Critical success factors for multidisciplinary engineering projects. In: 22nd Annual ARCOM Conference. Birmingham: ARCOM, 2006, pp. 219-228.

[11] Müller R, Turner JR. Matching the project manager's leadership style to project type. International Journal of Project Management 2007; 25(1): 21-32.

[12] Turner JR, Müller R, Dulewicz V. Comparing the leadership styles of functional and project managers. International Journal of Managing Projects in Business 2009; 2(2): 198-216.

[13] Kandelousi NS, Ooi J, Abdollahi A. Key success factors for managing projects. International Scholarly and Scientific Research & Innovation 2011; 5(11): 1185-1189.

[14] Awan MH, Ahmed K, Zulqarnain W. Impact of project manager's soft leadership skills on project success. Journal of Poverty, Investment and Development 2015; 8: 27-47.

[15] IPMA. Individual competence baseline for project, programme & portfolio management. 4th ed. Zurich: IPMA, 2015.

[16] PMI. Project manager competency development framework. 3rd ed. Newtown Square: PMI, 2017.

[17] International Organization for Standardization. ISO 21500: Guidance on Project Management. Geneva: ISO, 2012.

[18] Kotler P. From mass marketing to mass customization. Planning Review 1989; 17(5): 10–47.

[19] MINECO. HADA. Herramienta de autodiagnóstico Digital avanzada. Madrid, 2017.

[20] Bechtold J, Kern A, Lauenstein C, et al. Industry 4.0 - The Capgemini Consulting view. Sharpening the picture beyond the hype. Paris, 2014.

[21] Buckley S, Giannakopoulos A. The contribution of human capital in gaining a competitive advantage in an enterprise. In: 8th European Conference on Knowledge Management. Barcelona: Academic Conferences Limited, 2007, pp. 157-162.

[22] CEN. European Guide to good Practice in Knowledge Management - Part 1 to 5. Geneva: CEN, 2004.

[23] Palacios-Marques D, Gil-Pechuán I, Lim S. Improving human capital through knowledge management practices in knowledge-intensive business services. Service Business 2011; 5(2): 99-112.

[24] Kase R, Zupan N. Human capital and structural position in knowledge networks as determinants when classifying employee groups for strategic HR management purposes. European Journal of International Management 2009; 3(4): 478-494.

[25] Unger JM, Rauch A, Frese M, et al. Human capital and entrepreneurial success: A meta-analytical review. Journal of Business Venturing 2011; 26(3): 341-358.

[26] Afiouni F. Human Capital Management, What does it really mean? In: 1st European conference on Intellectual Capital. Haarlem: ACPI, 2009, pp. 10-19.

[27] Le Deist FD, Winterton J. What is competence? Human resource development international 2005; 8(1): 27-46.

[28] PMI. PMI's pulse of the profession®. The Competitive Advantage of Effective Talent Management. Newtown Square, 2013.

[29] Onisk M. Is measuring soft-skills training really possible? Sidney, 2011.

[30] Stevenson DH, Starkweather JA. PM critical competency index: IT execs prefer soft skills. International Journal of Project Management 2010; 28(7): 663-671.

[31] Chipulu M, Neoh JG, Ojiako U, et al. A multidimensional analysis of project manager competences. IEEE Transactions on Engineering Management 2013; 60(3): 506-517.

[32] Omidvar G, Samad ZA, Alias A. A framework for job-related competencies required for project managers. International Journal of Research in Management & Technology 2014; 4(1): 1-10.

[33] Su R, Golubovich J, Robbins SB. Bridging science and practice: Toward a standard, evidence-based framework of 21st century skills. Industrial and Organizational Psychology: Perspectives on Science and Practice 2015; 8(2): 289-295.

[34] Zadeh MT, Dehghan R, Ruwanpura JY, et al. An index to assess project management competencies in managing design changes. International Journal of Construction Engineering and Management 2016; 5(1): 11–24.

Coping with New Business Challenges: Analysis of the Elements that Affect the Sustainability of Continuous Improvement Programs

(Presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Orlando Chirinos
Université Savoie Mont Blanc
SYMME
Annecy, France
Orlando.Chirinos-Colmenares@univ-smb.fr

Georges Habchi
Université Savoie Mont Blanc
SYMME
Annecy, France
Georges.Habchi@univ-smb.fr

Magali Pralus
Université Savoie Mont Blanc
SYMME
Annecy, France
Magali.Pralus@univ-smb.fr

Zahir Messaoudène
ECAM Lyon
Lyon, France
zahir.messaoudene@ecam.fr

Abstract— Companies have evolved over time and this includes their efforts to overcome obstacles. In this article, we discuss sustainability as synonymous with success, and evaluate the sustainability of Continuous Improvement Programs (CIP) and the elements that affect it. The objective of this article is to highlight and analyse the results from a self-assessment investigation. We appraise the impact of a set of elements over sustainability and their maturity (absence or their degree of presence) while being deployed within the organization. This appraisal is done by assigning different weights and scores to the answers of the self-assessment investigation. We were able to observe that those elements with a higher impact score are heavily linked with management approaches. On the other hand, those same elements are not necessarily the more mature or appropriately displayed, and by the use of the maturity scoring system, we investigate the different stages of each element. This opens a pathway for further research into the behaviour of particular elements with the potential of enhancing or diminishing the sustainability of Continuous Improvement Programs.

Keywords— *Sustainability, Continuous Improvement, Self Assessment Investigation, Elements, PDCA.*

I. Introduction and Context

Operational excellence practices are borne from solutions to industry demands as a way of achieving desired performance and "*ensure a continuous delivery of positive results*" [1 p. 143]. At the beginning of the 20th century, the industry was marked by mass production, and so improvement methods such as standardization, time measurement analysis, worker motion study, and ergonomics were the solutions to overcome the obstacles at that time.

Adaptability is the key to longevity [2]. According to Bonebeau *et al*. [3 *p. 9*] some organization systems "*can be understood as a set of dynamic mechanisms in which global structures appear from a system of interactions between components at different levels*". Moreover, since companies are formed by people, they tend to be in a state of constant change. Jaca *et al*. [4] suggested the idea of flexibility and adaptation to satisfy customers, meaning that the idea of constant change is distinctly present. Companies launch initiatives that increase successes and reduce failures [5], and become what is known today as a culture or philosophy [6] of **Continuous Improvement** (**CI**). This is related not only to the above (deviation reduction), but also to *quality management programs* [6], *reduction of production costs* [7], *effective resource management, employee empowerment* [8], and *customer satisfaction* [9], among others.

There are three main features regarding CI philosophy, worth highlighting. The **first** is CI considers not only technical tools but also different management approaches [10]. CI is capable of handling the symbiosis between these two styles: a purely technical aspect, which leverages operational excellence, whilst a managerial aspect is less visible but has high impact, especially on long-term and medium-term strategies. **Secondly**, providing feedback is the core of CI since it builds mutual understanding and teamwork [11]. Mistakes are inevitable in any process; they are not seen as a negative aspect but as a useful tool for re-analysing strategies based upon previous experiences. **Thirdly**, the complex structures created by today's companies can also affect CI [12]. In order for CI to be successful, it must consider all agents within the organization, which are disseminated through different levels and with different evaluation approaches.

Being sustainable can translate into corporate success [13], and yet despite proving that the application of those **Continuous Improvement Programs** (**CIP**) can drastically upgrade operational performance [14], firms experience problems in gaining acceptance for CIP and sustaining them [10]. In terms of the issues associated with the lack of sustainability of CIP, authors cite lack of direction, motivation, structure, among others [15]. According to other authors, there is a lack of research defining **sustainability** [15], which contributes to our research motivation. Therefore, it is necessary to study the concept of sustainability that can be deployed within a CIP framework.

This article is presented as follows: in section two, we present our approach and methods, and explain the tool used to collect the data from firms surveyed; in section three, we explore the results obtained. Section four discusses the question formerly posed, while the final section presents the conclusions and suggested further research.

II. Sustainability of CIP

According to Prajogo *et al.* [16 *p. 207*], a definition of the term sustainability within a CI context can be described *"as the ability of an organisation to adapt to change in the business environment, to capture contemporary best practice methods and to achieve and maintain superior competitive performance"*. In this case, it is possible to transfer it into our framework. However, there are limitations, such as the lack of capacity to translate the definition within the objectives of the organization.

The challenge dwells on how companies can understand its profound meaning and the innate incongruences that surrounds it [17]. From previous works [18 *p. 4*], and by analysing the same terminology from different research fields, we define sustainability as:

"The capacity to maintain a high level of performance achieved gradually over time, that is measurable, according to well-defined processes, limited by finite resources (financial, material, human among others), and with the particularity of being able to adapt itself to context evolution"

Sustainability can be seen as an ability to adapt and change its environment [19]. It is imperative that companies understand the factors that characterize the sustainability of CIP [4]. From previous studies [20], and based on a literature review e.g. *scientific journals, case studies, reviews articles, interviews, books and books reviews,* we identify 15 elements that influence the sustainability of CIP (see Table 1). From that exercise, we were able to source approximately 231 references, of which 60 revealed *Tools and Methods for CI* as having the greatest influence over the potential for sustaining CIP. Only six references were found for *Diagnosis*, suggesting it has the least influence by volume of sourced literature.

Constant adaptation is a big part of the sustainability definition [19]. Therefore, we propose a concept that *Continuous Improvements Programs balance themselves between a two-state dynamic, one stable and other fluctuating.* This should enhance our research process thought.

Basing ourselves on that idea, we do believe that a set of elements (already mentioned) (Table 1) helps to create this balance. This leads us to an appreciation that the maturity of elements can be seen as triggers for change, or factors that can act as stabilizers. Having developed a definition of sustainability pertaining to CIP, and the elements that characterize it, two questions arise:

Is the perception of the elements impacting the sustainability of CIP in agreement with current theories?

Can the maturity level of each element be measured or evaluated?

Table 1. Elements impacting the sustainability of CIP and their frequency of occurrence in the sourced literature.

	Abbreviation	Sources	Rank	% Occurrences
Tools and methods for CI	TOOL METH CI	60	#1	26%
Change management	CH MGMT	26	#2	11%
Leadership	LDSHP	20	#3	9%
Organizational routines	ORG R	18	#4	8%
Performance of CI initiatives	PER CI	17	#5	7%
Employee engagement	EMP ENG	15	#6	6%
Risk management	RISK MGMT	14	#7	6%
Knowledge management	KNOWL MGMT	10	#8	4%
Decision-making	DEC-MAK	9	#9	4%
Training on methods and tools for CI	TOOL METH CI	8	#10	3%
Organizational culture	ORG CE	8	#11	3%
Communication management	COMM MGMT	7	#12	3%
Time management	TIME MGMT	7	#13	3%
Resources	RES	6	#14	3%
Diagnosis	DIAG	6	#15	3%

Our objective is to evaluate **the impact of those 15 elements** upon the sustainability of CIP **and their degree of maturity**, then later compare those two outcomes. In order to measure the maturity or level of deployment of those elements, we are using the **Plan-Do-Check-Act (PDCA)** cycle as a ruler to quantify that maturity. Introduced in 1951 by Dr W. Edwards Deming at the Japanese Union of Scientists and Engineers (JUSE), this approach has become one of the bases of CI. A tool capable of being applied to all types of organizations, promoting knowledge, benefiting the use of team-working and empowering organizations to improve their endeavours [21]. With these results, we intend to answer the questions already established and respond to the ideas previously presented.

III. Approach and Methods

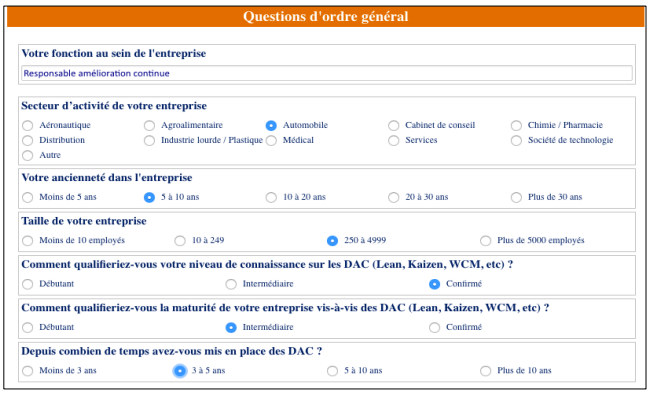
The approach used to measure the impact and the degree of maturity of the elements was a self-assessment investigation aimed at companies experienced in the deployment of CIP. For this endeavour, Mont-Blanc Industries[1], a business cluster from the Haute-Savoie department in France, supported us. We designed a self-assessment investigation composed of 113 questions to be completed over a three-month period from December 2016 to February 2017. Mont-Blanc Industries has 315 members who were approached obtaining 80 results. The assessment tool was electronic-based (the companies were approached

[1] http://www.montblancindustries.com/
[2] http://www.lesphinx-developpement.fr/

directly via e-mail) by using the software SPHINX[2], an electronic tool for data collection and analysis.

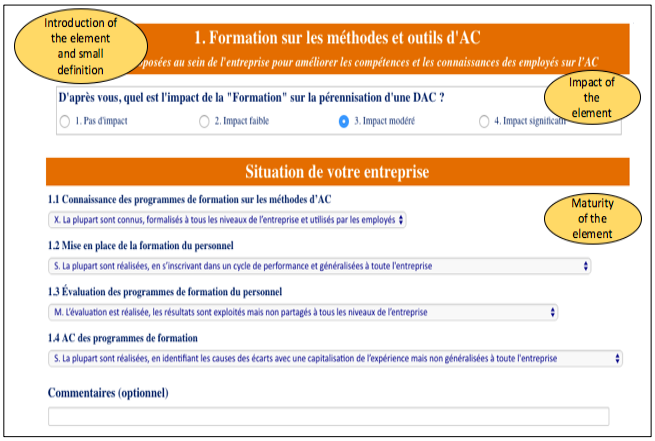
In the assessment, we have two sections. **Section 1**, was general positioning, made of seven questions, one of which was open, and the others closed. We covered: the company sector, respondent's role within the organization, years of experience, company size, their perception of their own knowledge concerning CIP as well as how they describe the maturity of the companies they work for regarding CIP. These latter two questions were evaluated with a three-scale selection: *Beginner*, *Intermediate* and *Confirmed*, from Figure 1.

Figure 1. First section of the assessment self-assessment investigation.



The **second section** (Figure 2) was composed of **closed questions**. We examined two major streams: element impact over the sustainability of CIP and their degree of maturity (with a PDCA approach).

Figure 2. Second section of the assessment self-assessment investigation.



In order to illustrate that section, Figure 2 shows a screenshot for the element *Training on methods and tools of CI.* The first question regarding the impact is, "**According to you, what is the impact of *Training* over the sustainability of CIP?**" With four options: *No-Impact, Low-Impact, Moderate-Impact* and *Significant-Impact*. The question was designed to garner the respondents point of view, their **knowledge and personal experience**.

For element maturity, the next four questions were directed to the **work experience** of the respondent. We used an original approach based on the PDCA cycle. By doing that, we position the element within a problem-solving scope [22]. Using the same *Training on methods and tools of CI* element as an example, the PDCA approach was divided as follows: **P** (1.1 Knowledge of the training programs concerning CI methods), **D** (1.2 Application of employee training), **C** (1.3 Evaluation of employee training programs) and **A** (1.4 CI of employee training programs). For each step of the PDCA cycle, the respondent selected how the element is being deployed in the organization. The scale used was: *Not present, Present, Methodical, Systematic, Exemplary* (IEMSX by its French initials) (Table 2). Each of these scale levels represents how the element is being deployed.

Table 2. Subdivision of the PDCA cycle by element.

(I) Not present	This part of the element is not applied within the company
(E) Present	The company has taken into account this part of the element
(M) Methodical	The practice of improvement is processed according to a method which can be generalized
(S) Systematic	The practice is processed with method and the application field is effective and systematic (sustained in time)
(X) Exemplary	The method, its application and their results deserve to be communicated externally because they are effective and simple

In this **second section** of the self-assessment investigation, we applied an evaluation scale based on weights (Table 3), enabling us to obtain the respondents' perception, by scoring the element global impact over sustainability of CIP, as well as their perception of the element deployment throughout the company (element maturity).

Table 3. Weight of the impact of an element over the sustainability of CIP and weight of the maturity of an element.

Element Impact Degree	**Weight**
No-Impact	1
Low-Impact	2
Moderate-Impact	3
Significant-Impact	4
Maturity of the Element with PDCA approach	**Weight**
(I) Not present	1
(E) Present	2
(M) Methodical	3
(S) Systematic	4
(X) Exemplary	5

IV. Results of the Self-Assessment Investigation

The first result from the assessment showed that 95% of the respondents place themselves between *Intermediate* and

Confirmed regarding their knowledge of the CI philosophy. In the case of the companies' maturity regarding CIP, 34% of them belong to the *Beginner* tier, 46% *Intermediate* and finally 20% as *Confirmed*. As we can see from Figure 3, the presence of *Significant-Impact* is high among the majority of elements, with an average of 62%. In the case of *Employee Engagement*, it is high as 95% while for *Risk Management* it is only 35%.

In order to obtain an overall score for the impact of elements, we used the weights assigned to every impact degree, and then calculated the arithmetic mean for all.

$$Ie_i = \frac{\sum_{j=1}^{N} W_{i,j}}{N}$$

Where:

Ie_i: Arithmetic mean of the impact of the element i

N: Number of terms (80 companies)

W_i: Weight (1 to 4) of the impact degree of the element i for firm j

Considering the global score from the 80 companies, *Employee Engagement*, obtained the highest score with 3.91. On the other hand, the lowest score among the whole elements is *Risk Management* with 3.13. A deeper analysis regarding these results is explored in the **Discussion** section, to obtain results over the maturity of the element, we used a similar approach. However, in this particular case, each element was divided into four steps (PDCA) and each of those steps had a deployment score (IEMSX).

$$Me_i = \frac{\sum_{j=1}^{N}(W_{P_j} + W_{D_j+} + W_{C_j} + W_{A_j})}{N}$$

Where:

Me_i: is the Arithmetic mean of the maturity of the element i

W_{Pj}: is the weight (1 to 5) of the maturity of the Plan step

W_{Dj}: is the maturity of the Do step

W_{Cj}: is the maturity of the Check step

W_{Aj}: is the maturity of the Act step

The results are interesting (Figure 4) since the elements do not follow the order of the previous exercise, meaning: *Tools and Methods for CI* is the most mature with 12.41. On the other extreme of the chart, the element less mature concerning a PDCA approach is *Time Management* with a score of 8.58 points.

Figure 3. Elements impact and their global scores concerning the sustainability of CIP.

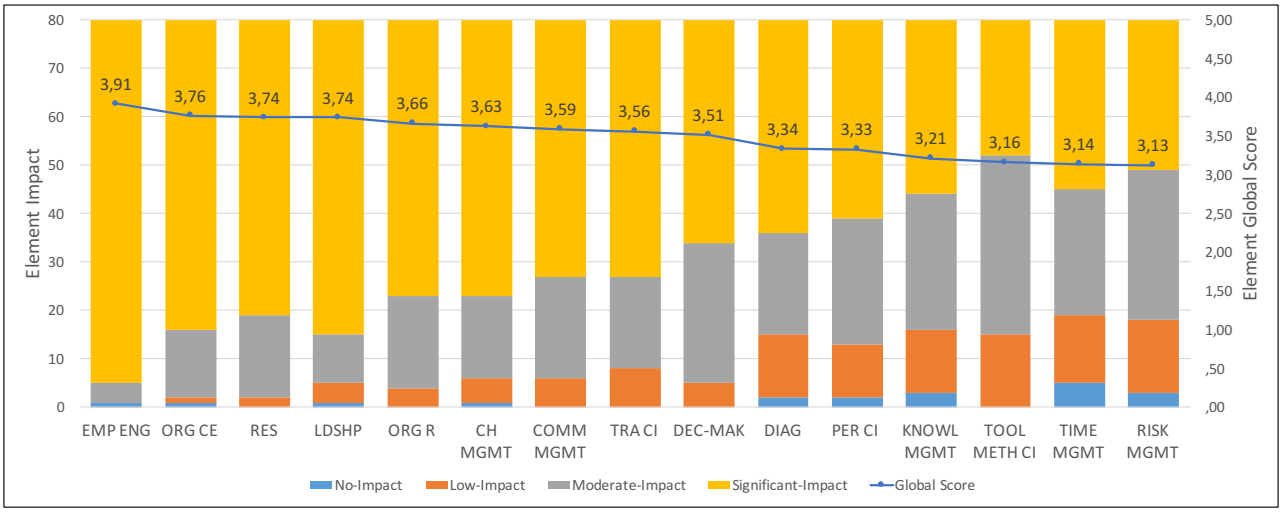
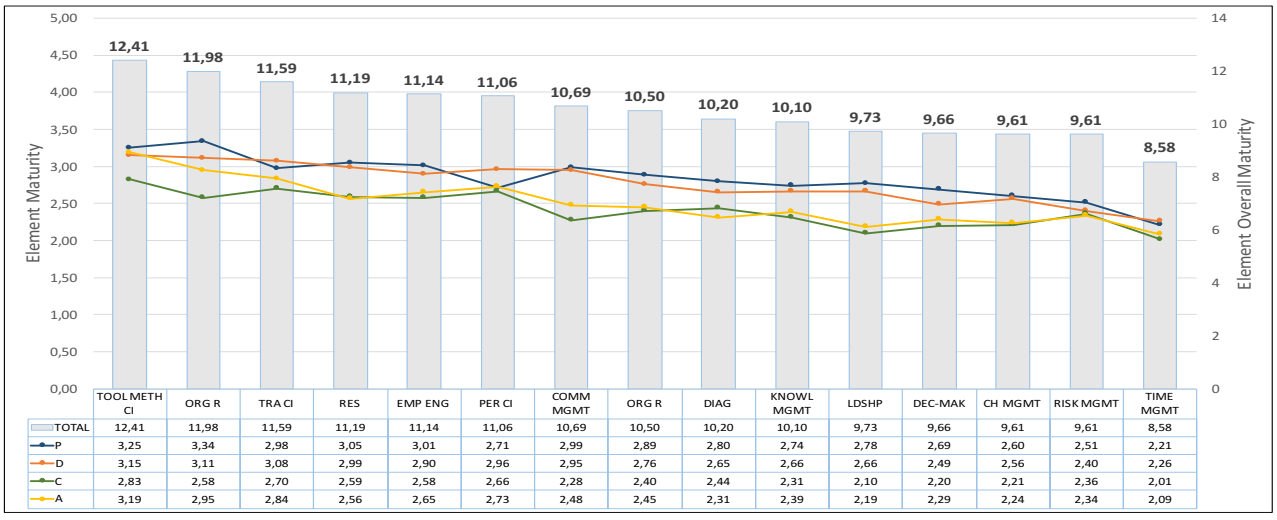


Figure 4. Maturity of the elements and their overall scores using a PDCA approach.



	TOOL METH CI	ORG R	TRA CI	RES	EMP ENG	PER CI	COMM MGMT	ORG R	DIAG	KNOWL MGMT	LDSHP	DEC-MAK	CH MGMT	RISK MGMT	TIME MGMT
TOTAL	12,41	11,98	11,59	11,19	11,14	11,06	10,69	10,50	10,20	10,10	9,73	9,66	9,61	9,61	8,58
P	3,25	3,34	2,98	3,05	3,01	2,71	2,99	2,89	2,80	2,74	2,78	2,69	2,60	2,51	2,21
D	3,15	3,11	3,08	2,99	2,90	2,96	2,95	2,76	2,65	2,66	2,66	2,49	2,56	2,40	2,26
C	2,83	2,58	2,70	2,59	2,58	2,66	2,28	2,40	2,44	2,31	2,10	2,20	2,21	2,36	2,01
A	3,19	2,95	2,84	2,56	2,65	2,73	2,48	2,45	2,31	2,39	2,19	2,29	2,24	2,34	2,09

V. Discussion

The first question that we posed in this article was related to; verify if the perception of the elements impacting the sustainability of CIP is in concordance with current CI theories. To tackle that question we focused on the respondents' knowledge and experience: 95% of the respondents were between *Intermediate* and *Confirmed* regarding knowledge over CI theories, which can be seen as a positive outcome in relation to the possible prevention of any misinterpretation of the questions posed.

It is interesting that the quantity of *Significant Impact* across the 15 elements is high (Figure 3), because this suggests that the respondents placed significant importance upon them, and that these elements were well selected for study.

As stated before, CI has the particularity of taking into account two approaches, one managerial and other technical. From Figure 3, we can note that the elements with a minimum of 50 "votes" on significant impact, are among those with a strong management character, such as *Employee engagement*, *Organizational Culture, Leadership*, *Change management*, *Organizational routines* and *Communication management*. The nature of these elements makes them difficult to measure [13]; nevertheless, this does not detract from the importance of their role. In the case of *Employee engagement*, not only is it the element with the highest impact, but also has the highest quantity of *Significant–Impact* appreciation. This denotes the importance of the human factor, showing how passion and commitment of the employees impact upon their productivity, while also becoming agents of innovation in order to solve problems.

Additionally, the last three elements with global scores smaller than 3.16 are: Tools and methods for CI, Time management and Risk management. It is also noted that they are more diversified regarding the participants' perception.

These elements are highly involved with CIP, leading us to believe that their impact is relative to the ease of their exploitation. As essential parts of CIP, these elements are easily approachable, consequently their impact on sustain the CI endeavours is less relevant.

The second question focused on the potential for evaluating the maturity of the elements. While addressing this issue, the questions are embedded within the respondent perception of the organization, meaning that we are dealing with another point of view, which can be seen as more practical. In this case, we analysed two angles, the first examined how the elements are being deployed within a **PDCA framework**, and the second concerned the **overall maturity** score of each element.

By analysing the arithmetic mean of each step of the PDCA, it is noticeable that the stage **PLAN** is the highest followed by **DO**, **ACT** and **CHECK**. This leads us to believe that a PDCA framework is useful to analyse the way the company embrace the CI.

The stages **PLAN-DO** are always the easiest stages to achieve and put into practice, they are highly linked with strategies and are tangible. Meanwhile the levels **CHECK-ACT** that are linked with improvement, feedback and adaptability to change [25], are less developed.

In our particular case the stage **CHECK** is the less developed, which generates an interesting analysis, and new questions to answer. For example, *do we need to consider other factors in order to approach this framework?* It is important to remark that in this article, and for the sake of a first analytical attempt, we are considering the whole population instead of separating the samples per maturity, which can give us new information.

While studying the **element overall maturity** score, there are some valuable impressions to discuss. The first deals with the fact that the elements considered mature are not the same as those measured with high relevance within their impact over sustainability of CIP. For example, in the case of *Tools and methods for CI*, it is the third lowest element within the impact appreciation (with a global score of 3.16), however, it is the more mature within a company environment (with an Element overall maturity of 12.41). The same analogies can be seen with other elements, for example *Leadership* which is considered to have a *Significant-Impact*, but in our study population its maturity is not well displayed.

This comparison shows us a dichotomy rarely studied: the perception of the elements impacting the sustainability (in our case with theories related to sustainability of CIP) is not faithfully translated to an actual working environment (in our case of study their deployment over a PDCA cycle). The difficulty in deploying a specific element within a CI framework can be the reason for its impact, and vice versa.

These behaviours invite us to explore even further, and even improve certain characteristics of our methodology. For example, only one person from each organization completed the self-assessment investigation, which can create some misconceptions or misinterpretations regarding certain elements, and expose the results to a single response bias. In addition, our population was almost entirety formed by people working in high rank positions that can generate a skewed view of the organization regarding the CIP, therefore it would be interesting to include other profiles to add different points of view over the matter, the use of personal interviews may have altered the outcome of our self-assessment investigation. With respect to the evaluation carried out and noting that the difference of the global score (Figure 3) between the first element (Employee Engagement)

and the last one (Risk Management) is small, it would be wise to reconsider the scores scale given in order to visualize it better, on the same idea it will be recommendable to run statistical treatments to measure the significance of the difference already mentioned.

VI. Conclusion and Further Research

Continuous Improvement philosophies have gone head-to-head with the industrial demands, and it considers different factors across the organization in order to be in optimal setup. Since being sustainable is synonymous with success, it is important to analyse a group of elements that have potential to affect the sustainability of CIP.

We set out to identify whether the perception of the impact of the elements that affect the sustainability of CIP can be evaluated and be in concordance with CI current theories. We demonstrated that they could be evaluated and by the use of an electronic self-assessment investigation, we observed that the majority of elements that have high relevance regarding their impact are linked with human factors, such as *Employee engagement*, *Leadership*, *Culture* and *Communication*.

Our second question regarding the evaluation of the maturity of elements was answered by using a **PDCA** approach (where each element was analysed relating to its deployment within the organization). We found that high impact elements are not necessarily the more mature, and while analysing each element by its cycle we can observe how the steps **CHECK** and **ACT** are the least mature regarding their deployment.

As prospective research, and since we are dealing with a highly rich database, we are in a position to analyse the impact of the elements by respondents' knowledge level and compare their conducts. The same exercise could be conducted with enterprises and measure each PDCA cycle separately and study possible coincidences between elements, or even between different types of companies. Moreover, as future research, it is planned to address the analysis of the whole maturity population and appraise those elements that are affecting the overall performance in order to obtain behavioural patterns that influence the sustainability of CIP.

References

[1] J.M. Sánchez, "Assessing sustainability of the continuous improvement process through the identification of enabling and inhibiting factors", PhD diss., Virginia Tech, 2012.

[2] R.A. Burgelman and A.S. Grove, "Let chaos reign, then rein in chaos—repeatedly: Managing strategic dynamics for corporate longevity", Strategic Management Journal, Vol. 28, no. 10, pp. 965-979, October 2007.

[3] E. Bonabeau, M. Dorigo, and G. Theraulaz, "Swarm intelligence: from natural to artificial systems", Oxford university press, 1999. Oxford, England.

[4] C. Jaca, E. Viles, R. Mateo, and J. Santos, "Components of sustainable improvement systems: theory and practice", The TQM Journal, Vol. 24, no. 2, pp. 142-154, 2012.

[5] T. Juergensen, "Continuous improvement: Mindsets, capability, process, tools and results," The Juergensen Consulting Group, 2000.

[6] S.K. Breja, D.K. Banwet, and K.C. Iyer, "Towards sustainable excellence: strategic analysis of Deming Prize winning companies", The TQM Journal, Vol. 28, no. 3, pp. 390-410, 2016.

[7] N. Bhuiyan and A. Baghel, "An overview of continuous improvement: from the past to the present", Management Decision, Vol. 43, no. 5, pp. 761-771, 2005.

[8] P. Hines, P. Found, G. Griffiths and R. Harrison, "Staying Lean: thriving, not just surviving", CRC Press, January 2011.

[9] T.A. Boyle, A.C. Bishop, K. Duggan, C. Reid, T. Mahaffey, N.J. MacKinnon and A. Mahaffe,. "Keeping the "continuous" in continuous quality improvement: exploring perceived outcomes of CQI program use in community pharmacy", Research in Social and Administrative Pharmacy, Vol. 10, no. 1, pp. 45-57, 2014.

[10] J. Bessant, S. Caffyn, J. Gilber., R. Harding and S. Webb, "Rediscovering continuous improvement", Technovation, Vol. 14, no. 1, pp. 17-29, February 1994.

[11] T. Bortolotti, S. Boscari and P. Danese, "Successful lean implementation: Organizational culture and soft lean practices", International Journal of Production Economics, Vol. 160, pp. 182-201, 2015.

[12] E. Keating and R. Oliva, "A dynamic theory for sustaining process improvement teams in product development", Advances in Interdisciplinary Studies of Teams, Vol. 5, pp. 245-281, 2000.

[13] M.F. Marin and G. Varoni, "Competencias directivas requeridas por los CEO ante la complejidad de las organizaciones del siglo xxi", Suma de Negocios, Vol. 7, no. 16, pp. 113-124, 2016.

[14] N. Bateman, "Sustainability: the elusive element of process improvement", International Journal of Operations & Production Management, Vol. 25, no. 3, pp. 261-276, 2005.

[15] T. Cooke-Davis, "The "real" success factors on projects", International Journal of Project Management, Vol. 20, no. 3, pp. 185-190, 2002.

[16] D. Prajogo and A. Sohal, "The sustainability and evolution of quality improvement programmes–an Australian case study", Total Quality Management & Business Excellence, Vol. 15, no. 2, pp. 205-220, 2004.

[17] P.J. Benghozi, "La pérennité: un lest ou un gyroscope pour l'entreprise?" Revue Française de Gestion, Vol. 2, pp. 177-181, 2009.

[18] O. Chirinos, M. Pralus, G. Habchi and Z. Messaoudène, "Company sustainability, an overview of the existing literature: in search of a common definition and its influence over continuous improvement endeavours", The 20th World Congress of the International Federation of Automatic Control, 9 -14 July 2017, accepted to appear.

[19] M. Zairi, "Total quality management sustainability: what it means and how to make it viable", The International Journal of Quality & Reliability Management, Vol. 19, no. 5, pp. 502, 2002.

[20] O. Chirinos, M. Pralus, G. Habchi and Z. Messaoudène, "Continuous Improvement Programs and the key elements characterizing their Sustainability, a first attempt", 12th International Conference on Industrial Engineering, 3-5 May 2017, accepted to appear.

[21] R. Moen and N. Clifford, "Evolution of the PDCA Cycle", Paper presented at the The Asian Network for Quality Congress, Tokyo, Japan, 2009.

[22] G. Gorenflo and J.W. Moran, "The ABCs of PDCA", viewed 30th January 2017, http://www. phf. org.

Measuring the Success of Collaborative University-Industry R&D Funded Contracts

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Gabriela Fernandes*, Madalena Araújo,
Pedro Magalhães, Ricardo J. Machado
Centre ALGORITMI
University of Minho
Guimarães, Portugal
*g.fermandes@dps.uminho.pt

Eduardo B. Pinto
CCG/ZGDV Institute
University of Minho
Guimarães, Portugal
eduardo.pinto@eng.uminho.pt

Abstract— This paper describes a method specially devoted to measure quantitatively the success of collaborative university-industry R&D funded contracts, which could be managed as a singular project or a program. The method aims to measure the success throughout the program/project lifecycle, combining both retrospective (lagging) and prospective (leading) indicators. It identifies the phases, process components, or factors that influence the program/project success, as well as the main key performance measurement indicators for each process component.

The proposed method was achieved by conducting a thorough review of current research in this area through published literature, and the analysis of two consecutive R&D collaborative funded contracts case studies – named HMIExcel and IC-HMI, between University of Minho and Bosch Car Mutimedia, together accounting to a total investment of over 70 million Euros, from 2012 to 2018.

Keywords— Project Success; program management; R&D University-industry collaborations.

I. Introduction

Firms seek innovation by making collaborations with universities [1-3]. Collaborative university-industry R&D projects are among the main public policy actions to promote innovation [4]. Research and development (R&D) is increasingly viewed as an important input to innovation [5]. Frascati Manual [5] defines R&D as "…. creative and systematic work undertaken in order to increase the stock of knowledge – including knowledge of humankind, culture and society – and to devise new applications of available knowledge." To provide guidance on what is and what is not an R&D activity, the Frascati Manual identifies five criteria, the activity should be: novel, creative, uncertain in its outcome, systematic and transferable and/or reproducible.

The attractiveness of universality-industry collaborations is recognized as having positive impact on both partners and also on the countries where those collaborations occur [4, 6, 7]. University-industry collaborations are considered to be an engine towards knowledge-based societies and economies, threatened by increased global competition, with ongoing economic issues and high levels of unemployment. The university-industry collaborations are on the agenda of the political decision makers. The different states, including EU, support initiatives such as collaborative R&D project through public's funds and thereby stimulating innovation [3, 8].

Industry embraces such collaborations for a number of reasons, namely to increase their R&D investments through public funding, to obtain higher performances in their innovations initiatives, to share risks and uncertainties [9, 10] at lower cost [6, 11, 12] and to increase the resources skills and competences to outperform on the global competition [7, 13]. University seeks, for example, for acquisition of funds to hire human resources and purchase cutting-edge equipment [3, 12], develop new case study teaching material [6] and to increase the capacity to attract new students [12].

It is a demanding task to achieve these benefits due to the existing "cultural gap" between both partners, requiring well organized and managed R&D collaborations [14]. To overcome these challenges the literature suggests the presence of some factors such as trust, prior experience of working together, predefined objectives, quality of communication and commitment [6, 15], to ensure the success of university-industry R&D collaborations.

With the increasing of collaborative R&D programs/projects between industry and university[14], the necessity of monitoring and assessing the outcomes of such initiatives has increased as well [4]. How to measure the success of such university-industry R&D collaborations is also an important challenge. It is necessary a two-fold systematic measurement: to assess programs/projects *a posteriori*, to find out which benefits are provided by the collaboration matching them with the initial benefits expectations [4, 14] and, not least important, to monitor the on-going process to enable adjustment and improvement [7, 14].

Based on Perkmann, Neely and Walsh [14] suggested performance measurement system for measuring the success of university-industry alliances, a method was developed to assess the success of collaborative university-industry R&D funded contracts, combining both retrospective (lagging) and prospective (leading) indicators.

The method uses tangible/specific outcomes as indicators, like patents, publications or licenses [1, 4, 14], as well as intangible/subjective indicators such as social relationships, organizational arrangement or motivations [1, 14], although, being complex to have general agreement on the criteria used to evaluate these subjective indicators. The method also includes variables that allow to judge the success of the program/project not only *a posteriori* but also *a priori*, during the program/project lifecycle [4].

The paper has the main objective of presenting the method to measure the success of collaborative university-industry R&D

This research is sponsored by the Portugal Incentive System for Research and Technological Development. Project in co-promotion nº 002814/2015 (iFACTORY 2015-2018).

funded programs/projects. The case studies used as research method were the HMIExcel and IC-HMI R&D collaborative funded programs, between University of Minho (UMinho) and Bosch Car Multimedia in Portugal (Bosch). The case studies were selected by convenience, in fact, this research work resulted from UMinho and Bosch partnership necessity to have a quantitative method to measure and compare the success of their R&D collaborative programs/projects.

HMIExcel, closed in 2015 (IC-HMI stills in progress), was considered a successful university-industry R&D collaborative funded program as reported on Pinto el al.[16]. However, its success was mainly subjectively perceived by the different program stakeholders. A well established method to measure quantitatively its success was missing, preferably a method that could be used to compare the success between different initiatives. Such a method could also be adopted by funding entities to assess the success of the programs/ projects that they support financially.

This paper is organized as follows. The second section summarizes the existing literature in measurement frameworks on university-industry alliance context. The third section gives a background of the two case studies used in this research study. The fourth section discusses the proposed method for measuring the success of collaborative R&D funded programs/projects, based mainly on the performance measurement system developed by Perkmann *et al.* [14]. All the adaptations made, in both process components and performance indicators, are explained. Finally, the main reflective conclusions that emerged from this study, as well as the suggestions for future work are discussed.

II. Relation to Existing Theories and Work

Success can be judged through project success criteria or critical success factors (CSF): 1) criterion: "*a standard by which you judge, decide about, or deal with something"*; and 2) factor: *"a fact or situation that influences the result of something"*. Some examples of criteria are the iron triangle (time, costs and quality), functional and technical specifications while some examples of factors are project mission, top management support, project schedule/plan, client consultation and acceptance, monitoring and feedback, troubleshooting and communication. The definition of success criteria and CSFs depend on projects, thus, it is not possible to develop a list that meets the needs of all projects [17].

Being measurements considered as management tools, they are an inherent part of a business process. Their own process is guided by objectives and their generic process is known as E-4 (Establish, Extract, Evaluate and Execute), based on the Deming cycle (Plan, Do, Check and Act). E-4 is one way of performance management that implements improvements through performance measurement in order to achieve success. Their actions meanings are [18]:

- Establish: measurement program and preparation, activities support and objectives control.
- Extract: performance measurement defined by control needs.
- Evaluate: evaluation of measurements in terms of background of actual status and objectives.
- Execute: implementation of improvements in order to decrease the differences between status and objectives

The development of a measurement method with qualitative as well as quantitative key performance indicators to assess the success on collaborative university-industry R&D funded programs/projects is important for all programs/projects' stakeholders. A key performance indicator (KPI) points out what should be done to increase performance [19]. Grimaldi and Von Tunzelmann [4] assessed the UK government´s LINK program aimed at promoting pre-competitive collaboration between firms and universities with the elaboration of indicators that take into account both the direct and the indirect outcomes of the partnership. Particularly, they proposed measuring direct outcomes, such as patents and publications, and indirect outcomes represented by commercial exploitation of patents and scientific results and follow-up opportunities generated by alliances. Grimaldi and Von Tunzelmann [4] also assessed the degree of matching between the initial objectives and its final achievements. They concluded that it is necessary to develop a more reliable and exhaustive measurement framework that could help government to control their project success, not only *a posteriori* but also *a priori*.

Based in case studies and published research, Barnes, Pashby and Gibbons [6] identified the management 'success' factors from both parties and combined those factors into a single framework tool tested through a further case study. The framework developed fosters an awareness of the key issues affecting the success of university-industry partnerships and stimulate the manager to take appropriate and timely action to prevent problems before they arise. Barnes *et al.* [6] argue that strong collaborations are based on strong personal and organizational relationships.

Perkmann *et al.* [14] proposed a 'success map' to measure the performance of university-industry alliances, based on the existing literature on university-industry relationships. The 'success map' distinguishes between four different process stages: 'inputs', 'in-process activities', 'outputs' and 'impacts', and proposes to measure process components for each of the four process stages, providing guidance for designing both prospective and retrospective indicators for the performance management system and also the ingredients by which positive outcomes are achieved within university-business alliances.

For example, in the 'input' stage as first process component/ factor to achieve success, Perkmann *et al.* [14] suggest that the collaborations need to mobilize sufficient and capable resources. Firms can share the R&D cost, allowing economies of scale, and, in many cases, also leveraging via public funding when they work with university. The involvement of highly qualified and motivated academic researchers, are important to ensure the success, since the 'cultural gap' between both parties generally reduce the enthusiasm for academic researchers to work with industry. Those factors- availability of resources and the presence of highly qualified and motivated academic researchers- are requirements for the execution of 'in-process activities' stage. Perkmann *et al.* [14] point out that to pursue a quality research project it is necessary the co-presence of industry resources and qualified researchers in the project team. The pursuit of good quality and relevance creates training and learning opportunities that constitute the process component

'relevant research'. The result of the activities of the in-process stage generates the outputs. At the 'outputs' stage the first factor pointed is the new scientific knowledge generated from the good quality research project. Relevant research project should allow the development of new technologies and complement the in-house R&D activities executed by the industrial partners. The result of training and learning opportunities should lead to more trained and skilled staff from both parties. Finally, the 'impact' stage - the new scientific knowledge can generate new ideas that can be developed in new R&D projects. The availability of new technologies can be a source of new products or process improvements innovation. The new scientific knowledge and new technologies can result on new concept solutions. Finally, the availability of a more skilled and trained staff within the industrial partner should lead to accumulating human capital that can increase the in-house expertise. Perkmann *et al.* [14] also indicate several performance indicators - metrics that measure each process component of each process stage of the 'success map'. Seppo and Lilles [20] based on the Perkmann *et al.* [14] work proposed a set of performance indicators and their usage to assess the inputs, outputs and impact stages of university-industry alliances.

III. Case Study

The selected case studies, named HMIExcel and IC-HMI, resulted from a strategic partnership established between University of Minho (UMinho) and Bosch Car Multimedia (Bosch) in Portugal, regarding the development and production of advanced car multimedia solutions.

UMinho is currently among the most prestigious institutions of higher education in Portugal, and is in the top 100 universities under 50 years old (75th position) worldwide. Founded in 1973, UMinho is engaged in the valorization of the knowledge-research chain, development and innovation. UMinho stands out by the volume of publications and by the number of patent requests, as well as by the high level of collaboration with industry, with around 250 R&D contracts signed annually with industry players.

Bosch is located in Braga, Portugal, since its foundation in 1990. Over the years, Bosch became one of the automotive industry's biggest suppliers and the leading plant of the Car Multimedia division of Bosch Group. Presently, Bosch produces a wide portfolio of products, such as navigation systems, instrumentation systems, car radios, steering angle sensors, and electronic controllers. In 2015, Bosch achieved a turnover of around 516 million Euros, 99% for exports, and employed around two thousand collaborators.

HMIExcel managed as a program, had an investment of about €19 million, ran from May 2013 to June 2015, was partly funded through the Portuguese incentives for R&D and involved around 300 researchers and collaborators, from UMinho and Bosch, respectively. The IC-HMI program foresees an investment of €54.7 million, from July 2015 to June 2018, with the admission of 94 new staff dedicated to R&D in Bosch and 173 new researchers in UMinho.

During the case studies conduction, three research methods were applied: document analysis, observation, and several unstructured focus groups in order to discuss the suitable method to measure the success of collaborative university-industry R&D funded programs/projects. The focus groups were conducted with different HMIExcel and IC-HMI's stakeholders, namely the Program manager, two Project Managers, and four PMO Officers. There is a high proportion of PMO Officers participation, for two reasons: (1) more availability - they are more aware of the importance of measuring the success of the programs/projects that are responsible to support its management; and (2) their role, since they are the main elements responsible for improving and supporting program and project management practices, including how to measure its success.

The focus group moderator (researcher) used as auxiliary material, the performance measurement system proposed by Perkmann *el al.* [14] to measure university-industry alliances to stimulate the opinion of the focus groups participants.

IV. Proposed Method for Measuring Success of Collaborative R&D Funded Contracts

The proposed method for measuring success of collaborative R&D funded contracts was achieved by conducting a thorough review of current research in this area analyzing published literature. Perkmann *et al.* [14] was selected as the main theoretical foundation for the development of the method. It was selected taking into account similitude of objectives and robustness. The work from Seppo and Lilles [20] also brought light to the development of the measurement method, by pointing out key performance indicators to successful collaboration laid on how well its organized and managed [9].

As mentioned at the introduction, this research work resulted from a necessity of UMinho and Bosch partnership to have a quantitative method to measure and compare the success of their R&D collaborative programs/projects; therefore it was decided to link the measurement method and the Program and Project Management lifecycle adopted by UMinho and Bosch to manage their collaborative R&D programs, based on the Fernandes et al. [21, 22] work divided in four phases: 'Program Preparation'; 'Program Initiation'; 'Program Benefits Delivery'; and 'Program Closure'. Therefore, the four phases suggested by Perkmann *et al.* [14] for building the 'success map' stages were linked to the Program and Project Management lifecycle from Fernandes *et al.* [21, 22] as shown in Table I.

TABLE I: Linkage Perkmann *et al.* and Fernandes *et al.* phases

Perkmann *et al.* stages	**Fernande *et al.* phases**
Inputs	Program preparation
Inputs	Program initiation
In-process activities	Program benefits delivery
Outputs	Program closure
Impacts	Post-program

For each program management phase (Table II) one or more important 'process components' was defined and one or more 'key performance indicators' was identified to measure each

TABLE II: Method for measuring the success of collaborative university-industry R&D funded programs/projects

Phase	Phase Weight	U	I	Component Weight	Process Component	Indicator Weight	Key Performance Indicator
Program Preparation	*25%*	X		*15%*	Researchers' capability	*40%*	Scientific impact (Nº citations/average)
						30%	**% of researchers involved with past experience in UI collaborative R&D projects**
						30%	**% of senior researchers (not research assistants)**
		X		*15%*	Researchers' motivation	*100%*	% of research income from industry
			X	*15%*	**Industry collaborators' capability**	*50%*	**% of Industry team collaborator's with a postgraduate degree or a higher level of education (postgraduate, master or PhD)**
						50%	**% of collaborators involved with past experience in UI collaborative R&D projects**
			X	*15%*	**Industry collaborators' motivation**	*100%*	**Degree of pressure to innovate**
		X	X	*20%*	**Opportunities/ challenges**	*100%*	**Nº of opportunities/challenges**
		X	X	*20%*	Applied research	*100%*	% of projects ideas with joint objective setting
*	*5%*	X	X	*100%*	**Governance established**	*100%*	**Joint governance model setting**
Program benefits delivery	*50%*		X	*25%*	Collaboration intensity	*15%*	**% (Nr. Steering meetings/ Nr. planned)**
						35%	**% (Nr. WP meetings/ Nr. planned)**
						15%	**% (Nr. Progress meetings/ Nr. planned)**
						35%	**% (Nr. Sharing events/ Nr. planned)**
		X	X	*15%*	Technology	*100%*	Nº of complete standard patent applications or other IP
		X	X	*15%*	New knowledge	*50%*	Nº of publications
						50%	**Nº of joint publications**
		X	X	*10%*	**Management and organization quality**	*100%*	**% Technical Deliverables (reports or prototypes) executed on time**
		X	X	*10%*	**Governance embedment**	*100%*	**Degree of governance embedment**
		X	X	*25%*	Human capital	*20%*	**% Recruitment of PhD's get hired in industry**
						20%	% Recruitment of research assistants (without PhD) in industry
						20%	**Perception of the impact of the program in the development of the academic or professional career**
						20%	**Structure of collaborators' habilitation**
						20%	**Number of master's degree and PhD degree obtained**
Program closure	*15%*		X	*50%*	Innovations	*50%*	Nº New Products
						50%	Nº New Process improvements
			X	*35%*	Solution concepts	*50%*	Nº of new solution concepts
						50%	**Increase level of TRLs**
		X	X	*15%*	New project ideas	*100%*	Nº new project ideas
Post-program	*5%*	X	X	*50%*	Technology achievement	*100%*	Nº of patents granted
		X	X	*50%*	**Partnership sustainability**	*100%*	**Value of new collaborative research projects generated**

*Program Initiation / U – University / I – Industry

component as shown in Table II. One focus group was conducted to identify the weight of each 'program phase', 'process component' and 'key performance indicator'. However, these weights can be changed for the context of each specific program/project that might adopt this method to measure its success. The text highlighted in bold in Table II indicates the 'process components' and 'key performance indicators' introduced as new to the Perkmann et al. [14] performance measurement system, from the result of the analysis of the two case studies used, HMIExel and IC-HMI.

The method combines both retrospective (lagging) and prospective (leading) indicators, allowing to be used during the whole program management lifecycle. For example, at the end of the 'program preparation', the KPIs of this phase can be assessed to measure the success until that program stage, but during 'program benefits delivery' the indicators of the 'program preparation' and 'program initiation' should be again reassessed, and the KPIs from the 'program benefits delivery' are assessed for the first time.

The KPIs might be assessed quantitatively through the use of bibliographic search engines and the distribution of surveys to program/project members, (e.g. to identify the researcher past experience in UI collaborative R&D projects), or subjectively by the program/project manager through the university/industry reports.

A. Program Preparation

'Program preparation' aims to design the R&D program and to strive for the financial support for the program [21]. This program management phase is composed of six process components: 1) Researchers' capability; 2) Researchers' motivation; 3) Industry collaborators' capability; 4) Industry collaborators' motivation; 5) Opportunities/challenges; and 6) Applied research. Of these process components, 3 of them (1, 2 and 6) are referred on Perkmann's study, and the 3 remaining are suggested by the present study.

Concerning 'Researchers' capability', Perkmann *et al.* [14] suggest measuring the 'scientific impact' by the citations counts record, but alternatively, the researchers' H-index can be used

[23]. Nevertheless, two additional indicators are suggested to measure 'researches' capability': researchers past experience in UI collaborative R&D projects and percentage of involvement of senior researchers (often too many young assistant researchers are hired for these programs/projects).

The main input factor for any type of R&D activity is the R&D costs [14, 20, 24]. R&D cost is critical into the innovation process by allowing to buy proper equipment, staff recruitment and contract services to achieve desired output. Perkmann *et al.*[14] point the benefit of sharing the R&D cost in University-industry collaborations, allowing firms to achieve economies of scale, additionally can gain by leverage via public funding. In this sense, it is suggested "Researchers' motivation" as a process component that has as KPI the percentage of research income from industry – the higher this value, the higher is the researchers' motivation.

Perkmann *et al.* [14] only suggested the 'researcher's capability' and 'researchers' motivation', however this is about effective collaborative R&D program/project and therefore, 'industry collaborators' capability' and 'industry collaborators' motivation' is also important, namely there are certain features of a firm that have effect in its ability to utilize the knowledge transferred from university [21]. For example, the firm's collaborators qualification, as well as both prior experience of working together and prior experience on broadest sense of university-industry collaborations, are important factors in the success of university-industry R&D programs/projects [6].

Other important aspect observed and discussed during the focus groups is the 'opportunities/challenges'. For a R&D initiative to be a success there is necessarily a certain number of new project ideas to study, otherwise there will not be the necessary input to design the program/project.

Perkmann *et al.* [14] also suggest the importance of developing 'applied research', and therefore the importance of taking into account the degree to which detailing objectives and potential solutions of each initial project idea are jointly determined by industry and university members.

B. Program Initiation

'Program Initiation' aims to guarantee the initial planning of the program and the alignment of the program objectives and outcomes with the stakeholders that will effectively get involved into the program execution and establish the governance model of the program/project [21]. This program management phase is composed by only one process component 'Governance established'.

Due to the unpredictable nature of R&D programs/projects, and the existent "cultural gap" between university and industry, it is of upmost importance to establish strong agreements between all parties for the achievement of the program/project objective [6, 9].

The establishment of a governance model to support the overall program management and the management of the constituent projects, that clearly define the objective, roles and responsibilities and program work plan, can be seen as good practice that will facilitate program management, commitment, trust, team spirit, communications on both sides of the university-industry alliance [9, 15, 25].

C. Program Benefits Delivery

Program Benefits Delivery aims to manage in an integrated way constituent projects of the program in order to deliver the expected program benefits [21]. This phase is supported by six process components: 1) 'Collaboration intensity'; 2) 'Technology'; 3) 'New knowledge'; 4) 'Management and organization quality'; 5) 'Governance embedment'; and 6) 'Human capital'. Of these process components, 4 of them (1, 2, 3 and 6) are referred on Perkmann's performance measurement system, and 4 and 5 are suggested through the present study.

Perkmann *et al.* [14] suggest to assess the 'collaboration intensity" by a constructed metric obtained by operationalizing the intensity of interaction between the partners such as the periodic exposure of intermediate results to all parties' members. The interaction between academic researchers and industry through a wider set of channels is more likely to build capabilities necessary to fill the gap between science and application [26]. Therefore, the frequency and variety of the types of interactions between the partners enable the transmission of know-how [14, 26].

Perkmann et al. [14] also suggest that the high-quality research should make it possible to generate new scientific knowledge. Whilst, conducting relevant research should allow the development of new technologies. Therefore, the number of patents applications submitted and the number of publications and joint publications are the KPIs suggested to measure the components 'technology' and 'new knowledge' respectively.

A new process component 'management and organization quality' is considered. As argued by Barnes *et al.* [9] the key to successful collaborations lies on how well it is managed, and an easy way to manage it is through the percentage of technical deliverables (reports or prototypes) executed on time. This KPI is also an indicator if the governance model is well established and embedded in the consortium organization.

In the 'human capital' process component four new KPIs are introduced: 1) the percentage of recruitment of PhD's that get hired by the industry – the creation of scientific employment, 2) the perception of the impact of the program members in the development of their academic or professional career, 3) the structure of collaborators' habilitation – if it increased or not, and 4) the number of master's degree and PhD's degree obtained through the counting of dissertations and thesis obtained under the R&D program context.

D. Program Closure

Program Closure aims to execute a controlled closure of the program and determine if the collaboration can be sustained [21], namely if the program/project was a success and new programs/projects can be initiated in a near future. All the process components proposed for the program closure are supported by Perkmann's study [14]; it is only introduced a new KPI. The availability of new technologies can be a source of product (service) innovation or process innovation. Finally, the new scientific knowledge and new technologies can generate new solutions, which represent an intermediate step between ideas and innovations commercially exploitable, which can be evaluated by the increased level of Technology Readiness Level (TRL) attained, since the project initiation until closure. Lastly,

in order to guarantee university-industry collaboration, it is also important to assess the process component 'New project ideas' generated during the on-going R&D program and can be used to be the foundation of a proposal for a new R&D funding program.

E. Post-program

In the 'Post-program' phase it is suggested to assess two process components: 1) 'Technology achievement' and 2) 'Partnership sustainability'. Perkmann *et al.* [14] propose the number of patents actually deployed or licensed, here it is proposed the number of patents granted to measure 'technology achievement'. For the process component 'partnership sustainability' it is suggested to evaluate the value of new collaborative research programs/projects generated as indicator for this process component.

V. Conclusions

The research reported in this paper builds knowledge in the domain of program/project management in the context of collaborative university-industry R&D funded contracts. A detailed method to measure quantitatively and track the success of university-industry R&D collaborations throughout the program/project lifecycle is presented, which could be used to compare the success between different initiatives with university and industry partners, but also could be adopted by funding entities to assess the success of the programs/ projects that they support financially. For future research, it would be useful to apply the method by exemplifying it through its implementation in real case studies, trying to answer the following research question: How to operationalize the application of the method for measuring the success of a collaborative University-Industry R&D Funded program/project? The application of the method, in order to discuss the process of implementation and its implications, tend to get lengthy and almost impossible to report on a single paper; therefore, this paper focuses on the development of the method.

References

[1] M. Perkmann, and K. Walsh, "University–industry relationships and open innovation: Towards a research agenda", I. J. Man. Reviews, vol. 9(4), pp. 259-280, 2007.

[2] P. D'Este, and M. Perkmann, "Why do academics engage with industry? The entrepreneurial university and individual motivations", The Journal of Technology Transfer, vol. 36(3), pp. 316-339, 2011.

[3] M. Perkmann, Z. King, and S. Pavelin, "Engaging excellence? Effects of faculty quality on university engagement with industry", Research Policy, vol. 40(4), pp. 539-552, 2011.

[4] R. Grimaldi, and N. Von Tunzelmann, "Assessing collaborative, pre-competitive R&D projects: the case of the UK LINK scheme", R&D Management, vol. 32(2), pp. 165-173, 2002.

[5] OECD, Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Reaearch and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities, 7th ed., Paris: OECD publishing, 2015.

[6] T. Barnes, I. Pashby, and A. Gibbons, "Effective University – Industry Interaction: A Multi-case Evaluation of Collaborative R&D Projects", European Management Journal, vol. 20(3), pp. 272-285, 2002.

[7] A. Healy, Perkmann, M., Goddard, J., and Kempton, L., Measuring the impact of university–business cooperation, Final Report, Luxembourg: Publications Office EU, 2014.

[8] H. Etzkowitz, "Innovation in Innovation: The Triple Helix of University-Industry-Government Relations", Social Science Information, vol. 42(3), pp. 293-337, 2003.

[9] T.A. Barnes, I.R. Pashby, and A.M. Gibbons, "Managing collaborative R&D projects development of a practical management tool", International Journal of Project Management, vol. 24(5), pp. 395-404, 2006.

[10] K. Cravens, N. Piercy, and D. Cravens, "Assessing the performance of strategic alliances: matching metrics to strategies", European Man. J., vol. 18, pp. 529-541, 2000.

[11] S. Ankrah, and O. Al-Tabbaa, "Universities–industry collaboration: A systematic review", Scandinavian Journal of Management, vol.31(3), pp. 387-408, 2015.

[12] C. De Fuentes, and G. Dutrénit, "Best channels of academia–industry interaction for long-term benefit", Research Policy, vol. 41(9), pp. 1666-1682, 2012.

[13] A.L. Sherwood, and J.G. Covin, "Knowledge Acquisition in University–Industry Alliances: An Empirical Investigation from a Learning Theory Perspective". J. Product Innovation Manag., vol. 25(2), pp. 162-179, 2008.

[14] M. Perkmann, A. Neely, and K. Walsh, "How should firms evaluate success in university–industry alliances? A performance measurement system", R&D Management, vol. 41(2), pp. 202-216. 2011.

[15] E.M. Mora-Valentin, A. Montoro-Sanchez, and L.A. Guerras-Martin, "Determining factors in the success of R& D cooperative agreements between firms and research organizations", Res. Policy, vol. 33(1), pp. 17-40, 2004.

[16] E. Pinto, Fernandes, G. Oliveira, J., Araújo, M., Pontes, A., and Machado, R.J., "Managing a Successful University-Industry Collaborative Funded Innovation Programme", in XXVII ISPIM Innovation Conference, pp. 1-13, 2016.

[17] L.A. Ika, "Project success as a topic in project management journals", Proj. Man. J., 40, pp. 6-19, 2009.

[18] C. Ebert, "Software Project and Process Measurement", Journal Defense Software Engineering, pp. 22-27, 2009.

[19] D. Parmenter, Key Performance Indicators: Developing, Implementing and Using Winning KPIs, New Jersey: John Wiley & Sons, Inc., 2010.

[20] M. Seppo, and A. Lilles, "Indicators measuring university-industry cooperation", Discussions on Estonian Economic Policy, vol. 20(1), pp. 204-225, 2012.

[21] G. Fernandes, Pinto, E.B., Machado, R.J., Araújo, M., and Pontes, A., "A Program and Project Management Approach for Collaborative University-industry R&D Funded Contracts", Procedia Computer Science, vol. 64, pp. 1065-1074, 2015.

[22] G. Fernandes, Pinto, E.B., Araújo, M., Pontes, A.J., and Machado, R.J., "Perceptions of Different Stakeholders on Managing Collaborative University-Industry R&D Funded Contracts", Proc. Comp. Sienc.,vol.100, pp.878-887, 2016.

[23] L. Czarnecki, M.P. Kaźmierkowski, and A. Rogalski, "Doing Hirsch proud; shaping H-index in engineering sciences", Bulletin of the Polish Academia Science Technical Sciences, Vol. 61 (1), pp. 5-21, 2013.

[24] R. Adams, J. Bessant, and R. Phelps, "Innovation management measurement: A review", International J. Management Reviews, vol. 8(1), pp. 21-47, 2006.

[25] V. Chiesa, Frattini, F., Lazzarotti, V., and Manzini, R., "Performance measurement in R&D: exploring the interplay between measurement objectives, dimensions of performance and contextual factors", R&D Management, vol. 39(5), pp. 487-519, 2009.

[26] P. D'Este, and P. Patel, "University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry?", Research Policy, vol. 36(9), pp. 1295-1313, 2007.

Improving Project Management Practice: A Car Industry Case Study

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Daniela Gomes, Gabriela Fernandes*, Madalena Araújo
ALGORITMI Research Centre
University of Minho
Guimarães, Portugal
*g.fermandes@dps.uminho.pt

Abstract—This paper presents a research project aimed at improving project management practices of new products development projects in a car industry case study. Four important initiatives were proposed and implemented: (i) redefinition of an integrated PM process; (ii) definition of a template for project planning supported with Work Breakdown Structure (WBS) template; (iii) customization of a software tool to support the PM process; and (iv) development of a project performance measurement system.
The customized software managed to improve project planning, time tracking, prioritization of work and communication, as well as decreased unproductive time and misunderstandings through organization of all project teamwork. The software also allowed data extraction and its visual appreciation by means of the creation of project performance measurement indicators graphics derived from the performance measurement system developed.

Keywords— Project Management; NPD; WBS; Performance Measurement System

I. Introduction

Project management (PM) is an area which has been growing very fast, becoming increasingly attractive for companies due to the value that it can add to organizational performance [1, 2]. According to the PMBoK, PM is "the application of knowledge, competences, tools and techniques to project activities to meet project requirements" [3]. The car industry case under study considers PM as a core competence, having the organization a Central Directive of PM based on the PM Body of Knowledge (PMBoK) of the Project Management Institute [3]. Therefore, PMBoK becomes one of the most authoritative sources of this research project.

The development of new products is a complex process that integrates knowledge and many functional areas [4]. New product development (NPD) is an activity essential to promote evolution and better business opportunities to companies [5]. This type of activities impose to make decision under uncertainty pushing to a better project planning and controlling [5]. NPD projects are integrated in the organizations' innovation process. The particularity of NPD projects are their inherent uncertainty and complexity that makes their success a difficult accomplishment. According to Tidd, Bessant and Pavitt [6], the best way to manage innovation is through the routines of each organization that creates an intuitive response for new challenges through experience and lessons learned. Nevertheless, it is essential to let the routines be adapted and improved, and not become barriers [6]. In 2008, Pons [5] defined a project life cycle for NPD project as being constituted by idea generation, concept design, feasibility study, detailed design, test product, finalize design, set up production, arrange marketing, arrange distribution, market growth, market maturation, market decline, product withdrawal and closure. The standardization of PM processes during the project lifecycle is up most of importance to project success [7, 8].

Project planning is also essential to project success; it is one of the most important activities of PM, since no project should initiate execution without a plan [9]. One of the most useful tools and techniques of project planning is the Work Breakdown Structure (WBS) [10, 11]. The WBS technique can be implemented with the support of the WBS standard of PMI [12], and through the implementation of WBS templates, contributing to the PM standardization approach, one important strategy to achieve success as argued by Milosevic and Patanakul [8].

The concept of success is subjective, having generated many theories [13]. Some authors distinguish the project success from PM success believing that the PM success is an internal concern and project success is a more embracing concern that involves specific project objectives and constraints [14]. Success can be judged through criteria or critical success factors (CSF) and should be measured by a performance measurement system.

The performance measurement system is "a balanced and dynamic system that is able to support the decision-making process by gathering, elaborating and analyzing information" [15].

PM practices can be supported by PM software [16], such as an enterprise PM system, agile software and social tools. Some of their useful features are resource management, time tracking, reporting, charts, portfolio management, collaboration, process/methodology and so on.

Summarizing, for improving PM practice in the Car Industry Case Study it was proposed: (i) the redefinition of an integrated PM process; (ii) the definition of a WBS template to support project planning; (iii) the customization of a software tool to support the PM process; and (iv) the development of a project performance measurement system.

This research is sponsored by the Portugal Incentive System for Research and Technological Development. Project in co-promotion nº 002814/2015 (iFACTORY 2015-2018).

The paper follows a common structure. The second section discusses different concepts linked to the four initiatives proposed. The third section describes the research methodology applied in this study. The fourth section specifies and discusses the results obtained for each research goal. Finally, the conclusions as well as the limitations and suggestions for future work, are discussed.

II. Literature Review

Projects have been increasingly used through the years and, nowadays, they are a common way of work. It is important to clarify the meaning of project because, sometimes, it is difficult to distinguish it from activities/processes [13].

According to PMBoK of PMI [3], a project is a temporary endeavour undertaken to create a unique product, service or result with clear-cut objectives, a starting point, an ending point and (usually) a limited budget. It is considered an art to achieve project goals of scope, cost, time, quality and satisfaction through direction and coordination of all kinds of resources during the project life cycle [17].

The projects growth seem to have led to the restructuring of companies' organization. According to an automotive case study [18], the key for that restructuration was the hierarchy change through the project manager emphasis and the increase of team involvement, autonomy and knowledge. It was also essential the adoption of a middle-up-down logic by enforcing the product development performance with middle management, creating a combination of tacit with strong knowledge (involving suppliers as well) as a strategy to settle restrictions in the beginning, with all the team, and decreasing the time spent and the freedom along the rest of the project.

Project planning is one of the main functions of PM. It consists on the process definition, preparation and coordination of all smaller plans of the project and integration of all in one [3]. To plan a project, it is essential to have the objectives, restrictions and requirements of the project from the client or sponsor/responsible.

Planning should be supported with some tools and techniques, Work Breakdown Structure (WBS) is one the most useful tools and techniques [10, 11]. WBS is a hierarchical decomposition that subdivides deliverables and work into more manageable components [3]. According to PMI [12], the WBS levels reflect the major deliverables divided logically by the major phases of project life cycle in order to improve assessing the team and the individual performance through communication, cost measurement and schedule performance.

There are different methods to support the creation of WBS. They are important in terms of promoting consistency and repeatability; promoting and enforcing the principals of organization's WBS guidelines; reducing the development effort and simplifying WBS process. Some examples of these methods are [12]:

- Top-Down: Starts by defining the major deliverables until decomposing to a detailed level.
- Bottom-Up: Starts with all deliverables and works backwards until grouping them into upper levels, having all work packages aggregated to a single parent (the project).
- WBS standards: Set of principles to create a WBS, including numbering, naming convention or required elements.
- WBS templates: Sample of WBS with hierarchical elements filled in some level of detail or generic information to be completed with the project information.

Independently of the method chosen, WBS has to describe all the project work hierarchically organized, and has to consider some factors, such as [12]:

- WBS element represents only one single deliverable and is compatible with organization and accounting structures. It is associated to only one parent and its child elements represent 100% of its work.
- Deliverables can be tangible (all reporting and others) or intangible (information/communication, integration, administration, training, process management, procurement and others). They are unique and distinct and have to be all included on WBS structure with only one team responsible.
- WBS development process has to be iterative, reviewed and revised along project planning progress. It provides flexibility.

A high-quality WBS is constructed in order to satisfy all the project requirements, representing the full scope of work. It supports the project manager on managing the varying degrees of scope, budget, schedule and risk, contributing to achieve project success.

The success does not have the same meaning for everyone and depends on its context [13]. There are many studies about the understanding of project success because of its complexity and ambiguity.

Some authors argue that project success should be distinguished from PM success. One of the most common meaning to PM success is the ability to fall within time, cost and quality (iron triangle) [14]. PM success is an internal concern that cares about doing the activities, and project success is a more embracing concern that involves specific project objectives and its constraints as time, costs, quality and satisfaction and cares about achieving business results [14].

Success can be judged through project success criteria or critical success factors (CSF). Criteria - a standard by which you judge, decide about, or deal with something (example: iron triangle); factor - or situation that influences the result of something (example: project plan, communication).

During the years, many studies and theories were created about what should define the project success. However, in the end, all the conclusions conduct us to the same: "project success continues to be in the eyes of the beholder" [19]. That perception, usually, depends on individual personality, culture, project type, contract type and lead us to a discussion about efficiency and effectiveness [19].

One way to measure success is through Performance Measurement Systems. As Ebert [20] said "in almost every area of life, setting goals and monitoring achievement are the foundations for success".

Performance measurement systems are essential to lead an organization [21] by evaluating and controlling the overall business operation to check if the performance is according to their targets in terms of goals and objectives [22]. However, measuring the performance is not self-sufficient to fulfill organization success [23]; it helps on a better understanding, management and improvement of what the organization does [24].

III. Research Methodology

Case study is a qualitative methodology that supports research on studying complex phenomena within their contexts [25]. Case Study strategy was selected, because contextual conditions are relevant to the phenomenon under study. PM is dependent on context [26], as well as it is difficult to separate the phenomenon and context [25].

The Car Industry organization under study is growing really fast, everything is continuously changing, what makes the following of that growth a difficult task. Therefore, the company should commit to continuous improvement. In particular a section involved in NPD projects, has been chosen by convenience. The organization was established in 1990. It develops and produces a wide portfolio of products, such as navigation systems, instrumentation systems, car radios, steering angle sensors, and electronic controllers, exporting 99% of its production. In 2016, the total number of employees was around two thousand.

IV. Case Study Analysis

A. Descriptive Analysis and Data Gathering

In this particular organization's section involved in NPD projects, the main problem to tackle was the poor organization owing to its fast growing of resources in recent years and the increased interest in PM. In 10 years the collaborators grew from 3 to 30, so the necessity to separate functions and responsibilities became relevant, as well as the need to create new functions and responsibilities and clarify the division of all subsections. PM is an example of a new function and responsibilities that appeared recently due to the emergence of project requests with more rigorous PM practices. The company started to receive project requests every year, having an average of 23 running simultaneously. Usually, a project emerges under the scope of a new product development. Each project has a duration of approximately one year and half with an average cost of 72.000 euros.

Therefore, to intervene according to the study requirements, it was important to understand the section's under study actual situation and its real necessities. This was achieved through four semi-structured interviews with responsible people of each subsection and the head of the section, coupled with observation and documents analysis. The main issues were detected as being:

- The lack of clarification of the division of subsections;
- Difficulties on the flow of information (sharing);
- The lack of standardization of processes;
- Difficulties on planning;
- Difficulties on justifying the needs (e.g. new resources, more budget) and in how to make visible the work done;
- Few collaboration and communication between the project teams.

Therefore, it was possible to evaluate and propose some important PM improvements.

B. Improvements Proposal

After the analysis of the actual situation through interviews, observation and documents analysis, the problem origins were quickly detected – weak definition of processes with the respective person responsible. Having identified the problems and matching them with the general objective of this dissertation, the improvement of PM practices, specific objectives were defined: 1) redefinition of an integrated PM process; 2) definition of a WBS template to support project planning; 3) customization of a software tool to support PM processes and 4) the development of a performance measurement system. To achieve these four goals, the following steps were proposed:

1. Structuring of every process (Table I) performed in the section, identifying inputs, outputs and responsibilities for each process's task – standardizing the processes for everyone (also helping in new collaborators integration) and defining who does what.

TABLE I: 1st Improvement Proposal

Problem	Solution	Results/ Benefits
• The lack of clarification of the partition of subsections • Difficulties on the flow of information (sharing) • The lack of standardization of processes	**Processes Structuring (Obj. 1)**	• Identification of inputs, outputs and responsibilities for each task • Standardization of the processes • Better understanding of the work/information flow • Better integration of new collaborators (by better understanding of processes)

2. Standardization of PM (Table II) namely by creating templates, allowing a better planning of all the activities and its performance evaluation.

TABLE II: 2nd Improvement Proposal

Problem	Solution	Results/ Benefits
• Difficulties on the flow of information (sharing) • The lack of standardization of processes • Difficulties on planning	**Standardization of PM through WBS template (Obj. 2)**	• Creation of a WBS template • Attribution of responsibilities for each project task • Standardization of PM • Easier and faster project planning • Better understanding of the work/information flow

3. Customizing an enterprise PM tool (Table III) to organize all the activities and to help in the planning and tracking.

TABLE III: 3rd Improvement Proposal

Problem	Solution	Results/ Benefits
• Difficulties on the flow of information (sharing) • Difficulties on planning • Few collaboration and communication between the project teams • Difficulties on justifying needs and in how to show the work done	**Customization of an enterprise PM tool (Obj. 3)**	• Support of activities • Faster work • Easier Communication • Automation • Total Vision • Portfolio management • Easier planning and tracking • Resource allocation support • Integration of all the information in a single tool

4. Identifying performance measurement indicators (Table IV) to guide evaluation and to obtain relevant results with an automatic performance measurement system – to obtain indicators which describe the work done and company' section vision without increasing the work to get them.

TABLE IV: 4th Improvement Proposal

Problem	Solution	Results/ Benefits
• Difficulties on justifying needs and in how to show the work done	**Development of performance measurement system (Supported by EPM tool) (Obj. 4)**	• Definition of success criteria and critical success factors • Definition of the key tasks • Vision of the section work status • Way of justification of needs or results • Implementation of improvements • Automatic obtainment – do not increase work • Increasing project success

V. Results and Discussion

The results were cross checked with the objectives, dividing this chapter accordingly, by the four subsequent objectives of this paper's section on:

A. Redefinition of an integrated PM process

The first objective is the redefinition of an integrated PM process in order to standardize it, defining responsibilities, inputs and outputs.

The structuring of processes should be the first step, since an overall picture of the whole process, dependencies and outcomes would facilitate real time control and monitoring, fostering increased performances. The focus was centered on NPD PM. Along this line, the project process was redefined attributing inputs, outputs and responsibility for each task based on company directives.

Mapping structuring was the technique used, which is a visual representation that connects tasks (can also connect inputs and outputs), shows the workflow and clarifies tasks, providing a total vision, resulting in a flowchart.

The flowchart was divided into project phases according to the project's requirement: concept evaluation, feasibility evaluation, producibility evaluation, machinery and equipment (MAE), preparation for sample production, preparation for series production and process release. Usually, multinational companies present the restructuration mentioned previously, the main project is divided into different projects suited to each area in order to fortify the knowledge, not forgetting the PM that strategically route the results for the good of the overall project. This specific case study, addressed a specific type of project of center of competences in order to define the specifications to develop new products.

In the end of each phase, an audit should be performed to evaluate if the project should continue or not. Its main objective is to assure the production process development, from the beginning to the end of the project, contributing to preventive quality assurance.

B. Definition of a WBS template to support project planning

Since NPD PM is a new approach for the firm's section, it seemed appropriate to improve its practices starting from the beginning – planning. Therefore, based on the redefinition of the integrated NPD PM process done, a WBS template was created.

The work is divided into work packages in order to help to evaluate if the project plan matches the project's objectives. Project's WBS was divided in two main work packages: PM work and process development work, i.e. development of the project result. PM work package groups all the tasks done by project manager in order to manage the project, known as PM life cycle, which includes the following phases: initiating, planning, project following and closing.

Process development work package groups all the tasks that have to be done by the team to achieve the specific project outputs. It matches with the product life cycle that, in this case, are divided in the following phases: concept evaluation, feasibility evaluation, producibility evaluation, machinery and equipment (MAE), preparation for sample production, preparation for series production and process release.

The template was developed with the support of Microsoft Project. For each work package a responsible is named as well as an estimate of the time duration, with the purpose of improving the organization, allocation and control of work, developing also a better communication and motivation through team work.

Nevertheless, the project manager still has power to change and adapt the WBS template to better adjust to each specific project.

C. Customization of a software tool to support PM processes

The objective was to find an internal tool with functionalities of an enterprise PM to support PM but also portfolio management.

The more adequate tool found, taking into account the software tools availability restrictions, was Track&Release (T&R) – software from Atlassian Enterprise. It was possible to customize this tool in order to adapt it for activities and projects.

T&R is a software tool to manage activities. T&R also fits in portfolio management principles, allowing the projects selection based in resource allocation.

Due to common problems in portfolio management, such as resource struggles and lack of the overall projects' and activities' overview, to take better decisions [27], it was proposed to include every activity (i.e. project's as well as operation's activities) in T&R. Usually the operations work

prevails over project work (roughly 85% and 15%, respectively) and the solution found to decrease this resource struggle was the insertion of both type of activities in this T&R software, since T&R does not allow to allocate percentages of work occupation to collaborators. Thereby, this improvement proposal also improved resource management, one of the firm's section main difficulties, as well as portfolio management. The customization is summarized on Table V.

Several customizations were made, for example, more typologies of activities were defined with other fields to embrace all activities:

- Meetings – Allows to track open point list resulting from the meeting and to extract a meeting minute;
- Request – Allows to receive requests from other departments and to register them;
- Cost – Allows to register all purchases done;
- Analysis – Allows to receive requests for firm's section laboratory from other departments and to register them;

Component "processability" release (CPR) – Allows to register component process ability release (CPR, one activity of firm's section with specific characteristics).

TABLE V: T&R customization.

Features	Benefits
More typologies of activities	Register of all activities
Kanban board	Better vision and self-organization
Time tracking	Measure the effort of each activity
Files attachment	Better communication and collaboration
Security level	Restriction of information
WBS template for projects	Standardization of project WBS and simplification of work (easier and faster)
Gantt Chart	Improving time management through Gantt chart (time lines)
Approval status	Allow workflow approval

To help on the adoption of this software, work instructions were prepared to collaborators, and individual meetings with everyone took place with the aim of explaining and testing T&R. After some tests, focus groups dividing collaborators by knowledge area and activities type, were organized, in order to receive feedback for new ideas.

D. Development of a performance measurement system

Firstly, to develop the performance measurement system it was necessary to match the indicators found on the literature with the company context in order to define what should be measured. As Ebert [20] found in his consulting services', the company studied has also problems, namely the lack of historic database to decide on release criteria as well as the unclear definition of how they can get work status.

The indicators definition was not an easy task since it is difficult to translate the work done in indicators, as well as making them visually clear. Other difficulty was about concepts. In practical situations, indicators are not distinguished through performance or result type. Companies usually define every indicator as a performance indicator, but key indicators are distinguished from others through importance assigned.

Therefore, the indicators were proposed based on difficulties found on justifying the needs (e.g. new resources, more budget) and in how to show the work done. Therefore, the objectives were knowing the work load status and if the number of resources were adequate, in order to anticipate and decrease the difficulties found, identify potential problems and be able to implement improvement actions.

To measure capacity and work status, the more reasonable unit found was the hours of effort because, sometimes, units of activities (ex: reports) do not tell anything about how was the work, it can last hours or months to do one single report. Table VI presents an example of how each performance indicator is established.

The target line is based on the plan and it is a way to evaluate the performance against the expectation in order to rate a smile with the visual status of the activity. The main concern on the definition of this structure was making its interpretation easy and intuitive.

It can be concluded if the target was completed and can be compared with the past year, thus improvements can be decided if necessary.

Matching and crossing the company projects context and existing KPIs in literature, four KPIs were implemented: quantity of audits done and with which result; the alignment of audit 4 with the second quality gate customer of product project; cost plan fulfillment and the ratio of plan vs non plan projects. The four KPIs are calculated extracting automatically information from their source in the T&R, requiring no extra work for the production of KPIs graphics, thus avoiding non-productive activities.

TABLE VI: KPIs

KPI	Metric	Periodicity	Objective	Target
Quantity of audits done	Unit	Monthly	- Requirement status & volatility - Schedule & time	Resource capacity
The alignment of audit 4 with QGC2 of product project	%	Monthly	- Quality	TGs plan
Cost plan fulfillment	€	Monthly	- Requirement status & volatility	Budget
The ratio of plan vs non plan projects	% and unit	Monthly	- Effort/planned capacity	n.a.

The performance measurement system provides a useful visualization of the project portfolio firms' section status and allows a more efficient performance evaluation and management. It is essential to plan improvements and follow an E-4 (Establish, Extract, Evaluate, Execute) process to implement them, all supported by the KPIs' graphical information provided [20].

VI. Conclusions and Future Work

The main contribution of this research study is to practice, by improving PM practice in a car industry case study, and giving guidance to practitioners on how they can improve the management of their projects. In summary four important initiatives were implemented:

- Redefinition of PM process in order to standardize it while reducing the time spent, and increase the project

quality, efficiency, planning, communication and team work;

- Creation of a template for project planning subjective in order to standardize it and increase the project success and knowledge as well as decrease the time spent on doing the planning;
- Customization of the software T&R to support PM practices, organization and communication as well as a record database.
- Development of a performance measurement system in order to identify problems and have a clear vision of the work status.

T&R software integrates the WBS template developed and other improvements as time tracking, Kanban board, Gantt chart and collaborative work through document attachment, project prioritization, self-organization, data registering, faster and easier communication and easier access to information of everything that it is done.

T&R software is essential for the overall visualization of all in one, allowing the improvement not only of PM but also of portfolio management. The implementation of the four initiatives had a good impact on the firm's section improvement but they also had some difficulties and limitations associated. The main limitation of the T&R software implementation was the collaborators resistance to change and the daily discipline required to actualize the project information. The major limitations of the performance measurement system are the requirement of a constant update of the project information in the T&R tool and the time (more than one hour) necessary to update the graphics and project status.

Since the implementation of the four improvement initiatives is very recent, it was not possible to obtain a quantitative value of these improvements in the company's section. However, the project stakeholders' perception is positive concerning the implementation of these four improvement initiatives. It was possible to receive positive feedback relatively to several results - their daily emails decreased; the software application and the set of improvements enable easier and better tracking, clearer vision of work status, easier integration of new collaborators, increase their available time for other jobs due to the better planning, ease the due date definition and tracking and allows clarification of the objectives.

These four improvement initiatives have resolved some important problems but there is still room for further developments, under a continuous improvement philosophy. We focused on general PM initiatives. Other improvement ideas could have been derived from specific automotive industry literature. However, we believe that a major research avenue for future work will be to invest on social PM since one evident problem is the communication and sharing inside the firm's section.

References

[1] M. Aubry, and B. Hobbs, "A fresh look at the contribution of project management to organizational performance", Project Management Journal, vol. 42(1), pp. 3-16, 2011.

[2] L. Zhai, Y. Xin, and C. Cheng, "Understanding the value of project management from a stakeholder's perspective: Case study of mega-project management", PMJ, vol. 40(1), pp. 99-109, 2009.

[3] Project Management Institute, A Guide to the Project Management Body of Knowledge, 5th ed, Pennsylvania: PMI, Inc., 2013.

[4] L. Almeida, and P. Miguel, "Managing new product development process: a proposal of a theoretical model about their dimensions and the dynamics of the process", in Complex Systems Concurrent Engineering: Collaboration, Technology Innovation and Sustainability, G. Loureiro and R. Curran, Eds. 2007, London: Springer, pp. 239-246.

[5] D. Pons, "Project management for new product development", Project Management Journal, vol. 39(2), pp. 82-97, 2008.

[6] J. Tidd, J. Bessan, and K. Pavitt, Managing Innovation: Integrating Technological, Market and Organizational Change. 3rd ed, John Wiley & Sons, Ltd., 2005.

[7] G. Fernandes, S. Ward, and M. Araújo, "Improving and embedding project management practice in organisations — A qualitative study", International J. of Project Management, vol. 33(5), pp. 1052-1067, 2015.

[8] D. Milosevic, and P. Patanakul, "Standardized project management may increase development projects success", IJPM, vol. 23(3), pp. 181-192, 2005.

[9] H. Kerzner, Project Management - A systems approach to planning, schedulling and controlling, New York: John Wiley & Sons, Inc, 2013.

[10] Besner, and B. Hobbs, "The Perceived Value and C. Potential contribution of Project Management Practices to Project Success", Project Management Journal, vol. 37(3), pp. 37-48, 2006.

[11] G. Fernandes, S. Ward, and M. Araújo, "Identifying useful project management practices: A mixed methodology approach", International Journal of Information Systems and PM, vol. 1(4): p. 5-21, 2013.

[12] Project Management Institute, Practice Standard for Work Breakdown Structures. 2nd ed, Pennsylvania: PMI, Inc., 2006.

[13] Jugdev, and R. Muller, "A Restrospective Look at Our K. Evolving Understanding of Project Success", PMJ, vol. 36(4), pp. 19-31, 2005.

[14] L. Ika, "Project success as a topic in project management journals", Project Management J., vol. 40, pp. 6-19, 2009.

[15] P. Garengo, S. Biazzo, and U. Bititci, "Performance measurement systems in SMEs: A review for a research agenda", I. J. Management Reviews, vol.7,pp. 25-47, 2005.

[16] G. Fernandes, S. Ward, and M. Araújo, "Developing a Framework for Embedding Useful Project Management Improvement Initiatives in Organizations", Project Management Journal, 45(4), pp. 81-108, 2014.

[17] P. Fewings, Construction Project Management: An Integrated Approach, 2nd ed. New York: Routledge, 2013.

[18] C. Middler, "Projectation of the Firm: the Renault Case", Scandinavian Journal of Management, 11(4), pp. 363-375, 1995.

[19] R. Muller, and K. Jugdev, "Critical Success factors in projects Pinto, Slevin, and Prescott – the elucidation of project success", International Journal of Managing Projects in Business, 5(4), pp. 757-775, 2012.

[20] C. Ebert, "Software Project and Process Measurement", J. of Defense Software Engineering, pp. 22-27, 2009.

[21] A. Krishman, "The Evolution of Performance Measurement System (PMS) and Linkage to the Environmental Uncertainty and Strategy; a Review of Literature", J. Global Business Management, vol. 4, 2008.

[22] M.I. Bhatti, H.M. Awan, and Z. Razaq, "The key performance indicators (KPIs) and their impact on overall organizational performance", Quality & Quantity, vol. 48(6), pp. 3127-3143, 2014.

[23] M. Bourne, and P. Bourne, Handbook of Corporate Performance Management, UK: John Wiley & Sons, 2011.

[24] D. Parmenter, Key Performance Indicators: Developing, Implementing and Using Winning KPIs, New Jersey: John Wiley & Sons, Inc., 2010.

[25] P. Baxter, and S. Jack, "Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers", The Qualitative Report, vol. 13(4), pp. 544-559, 2008.

[26] C. Besner, and B. Hobbs, "Contextualized Project Management Practice: A Cluster Analysis of Practices and Best Practices", Project Management Journal, 44(1), pp. 17-34, 2013.

[27] B.S. Blichfeldt, and P. Eskerod, Project portfolio management–There's more to it than what management enacts. IJPM,26(4), pp.357-365, 2008.

A Performance Evaluation Methodology for Decision Support in Industrial Projects

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Fan LI, Alain ETIENNE, Ali SIADAT
LCFC, Arts et Métiers ParisTech
Metz, France
lifan026@gmail.com, {alain.etienne, ali.siadat}@ensam.eu

François VERNADAT
LGIPM, University of Lorraine
Metz, France
vernadat30@gmail.com

Abstract—Considering the requirement to simultaneously satisfy multiple stakeholder objectives in industrial projects, performance should be comprehensively evaluated for the purpose of decision support. Due to the proliferation of evaluation criteria and data to be taken into account in performance management, it is necessary to develop an operational methodology for efficient performance management and decision making. The objective of this paper is to propose a benefit-cost-value-risk based methodology to help decision makers in performance evaluation at any stage of an industrial project.

Keywords—project management; performance measures; performance management; decision support

I. Introduction

Customer demand is now rapidly changing in terms of diversity of products and services [1]. In addition, the time to develop and distribute a product or a service becomes much shorter in today's global competitive markets. Furthermore, new concept such as Industry 4.0 will drastically change the management of industrial systems (projects or processes). Consequently, companies must rely on integrated, dynamic and relevant performance information to ease efficient and effective decision making to promote a pro-active management style regarding simultaneous accomplishment of multiple stakeholder objectives.

Although the problem of performance measurement and management is not new and many methodologies and approaches have already been proposed in this field, such as Activity-Based Costing [2], Balanced Scorecard [3], Failure Mode, Effects and Criticality Analysis [4], Analytic Hierarchy Process [5], ECOGRAI [6], [7], QMPMS [8], Performance Prism [9] or the risk oriented model [10], just to name a few, there is still room for improvement to help practitioners and researchers in industrial engineering and management sciences to make better decision in a more systematic manner.

The paper claims that performance of an industrial system can be comprehensively evaluated by following four dimensions: benefit, cost, value and risk [11]. In this context, *Benefit* is defined as a qualitative list of potential advantages or gains for a stakeholder compared to an objective that is set beforehand for the realisation of an industrial system. *Cost* refers to total expenses for the production, distribution and acquisition of the final result (a product, service or deliverable) of an industrial system. *Value* is described as the degree of satisfaction of a set of stakeholder expectations or needs, expressed by the appreciation level of a number of performance indicators. *Risk* means the likelihood and consequence of an event occurrence impacting the achievement of some stakeholder objectives. In addition, it is assumed that the four dimensions are independent at a broad level because they are analysed and evaluated separately.

Based on these four selected evaluation axes, the aim of this paper is to propose a complete performance measurement and management framework as well as operational methodology to apply recommended supporting methods and tools for decision support in industrial systems.

Firstly, the analytical framework for performance expression, which is a key element of the proposed methodology, is presented. Then, the global structure and main operations of the multi-phase methodology is briefly introduced. Finally, an experimental application to construction projects during the implementing stage is involved to demonstrate the use of the proposed methodology for decision support in an industrial project.

II. Performance expression

The performance of an industrial system is multidimensional (multi-viewpoint, multi-level and multi-criteria) [12] and relative (depending on the time, the way and the condition of measurement) in nature [13]. Therefore, an analytical framework for performance expression, which can meet these characteristics, is proposed in this section to increase the reliability of the final results of performance evaluation.

A. Analytical framework for performance expressions

From a generic level, it is assumed that the performance of an industrial project can be comprehensively identified from three necessary perspectives: stakeholders, evaluation periods and variables.

- *Stakeholders* represent the points of view to be taken into account.
- *Evaluation periods* indicate the time or intervals at which an evaluation must be performed. It can be an instant, a life cycle phase or the whole life-span of the project.
- *Performance variables* describe the components to be considered for performance evaluation with multiple levels and criteria for the problem at hand.

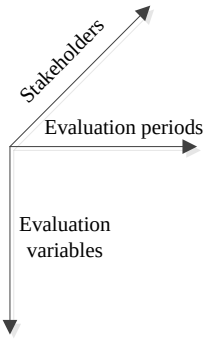


Fig. 1. Analytical framework for performance expression

Because it is supposed that the performance of an industrial system can be comprehensively expressed by its overall benefit, cost, value and risk, the evaluation variables can be allocated along these four dimensions. However, it may happen that in some particular cases, some elements, which normally fall within one of these four categories, must be considered as extremely important regarding the specification of decision making. In this context, evaluators can consider them as additional or alternate dimensions for overall performance expression.

For instance, in many projects the time dimension to monitor delays is crucial, it can be treated as a separate dimension in specific cases. In this paper, time was considered as a component of the value dimension because delays directly impact the satisfaction of stakeholders.

Fig. 1 presents the global structure of the proposed framework applied to the generic project management context.

The proposed framework includes a set of basic elements within each of the three independent perspectives. Each basic element represents the performance of one evaluation variable at one evaluation period along the project life history from one specific point of view. Based on the particular context of the decision problem, the evaluator selects the performance components (relevant stakeholders, specific evaluation periods and main evaluation variables) to define the scope of the performance evaluation problem.

The purpose of this analytical framework for performance expression is to help decision makers to organise their particular performance evaluation environment by identifying all relevant stakeholders, defining when to make evaluations and specifying all variables to be taken into account in the performance equations of each evaluation period for assessing cost, value and risk of a specific industrial project. To better explain how it works, each perspective needs to be detailed in the following sections.

B. Performance perspectives

Stakeholder identification is a process which aims to identify the individuals or organisations that can impact or be impacted by a decision, activity or outcome of the project. It also includes the analysis on their interests, interactions and potential impact on the project success. Fig. 2 provides an overview of the project stakeholder identification phase.

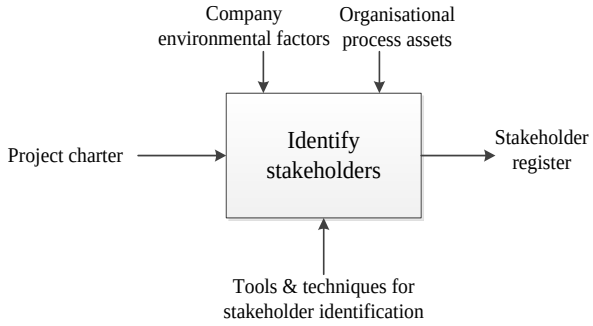


Fig. 2. Overview of project stakeholder identification

The input of the project stakeholder identification phase is the project charter, which is used to present the high-level description of the result of the project.

The enterprise environmental factors refer to conditions, not under the control of the project team, that positively or negatively influence or constrain the project and its outcome. The organisational process assets are the formal and informal plans, processes, policies, procedures and knowledge bases, specific to and used by the performing organisation.

The tools and techniques for stakeholder identification include stakeholder analysis, expert judgment and meetings with end users.

The stakeholder register is used to identify stakeholders who can provide information on the requirements [14]. This document contains all detailed information related to the identified stakeholders, including identification information (stakeholder contact information and its role in the project), assessment information (stakeholder major requirements, main expectations and potential influence for the project) and stakeholder classification (for example, internal/external, supporter/neutral/resistor).

In the proposed framework, the main project life cycle phases are assumed to be the usual evaluation periods over which performance of an industrial project needs to be measured and assessed. However, the definition of these evaluation periods is left to the responsibility of the analysts or the decision makers because they are the only ones who really know the relevant points in time when performance must be assessed and why.

The perspective of performance evaluation variables identifies the elements that are directly related to or will be involved in the performance measurement. It consists of three facets: performance dimensions, performance variables and evaluation criteria.

- Performance dimensions

The performance dimension facet forms the highest level in this hierarchical structure. It represents the basic axis on which the overall performance of an industrial project will be evaluated. For almost all cases, one performance dimension cannot always be directly evaluated. Each dimension contains a set of elements that are commonly utilised as elementary measures within organisations to evaluate the performance.

Therefore, performance variables are identified to represent the components of the performance dimensions.

- Performance variables

Each performance dimension is evaluated through one or several evaluation variables. Besides, these variables do not have the same level during the evaluation. It means that the relations among different variables can be described by a hierarchical structure. So, a breakdown analysis should be performed in order to express all the evaluation variables.

- Performance criteria

In many cases, a performance variable cannot be directly measured. Therefore, performance criteria should be identified in order to express each variable into a criterion that can be qualitatively or quantitatively measured. Although different performance variables are independent, they may have common criteria.

The proposed performance expression structure provides the evaluator with all the necessary elements to express the overall performance of a specific project. The framework has two main advantages: (1) comprehensive reflection of the multidimensional and relative characteristics of performance and (2) flexible structure to adapt particular decision context. Based on this framework, the evaluator can obtain the performance evaluation results through the proposed evaluation process, which contains a set of recommended methods and tools for performance management and decision support that s/he can freely use. The main operations will be described in more details in the following section.

III. Performance evaluation process

The objective of the proposed methodology is to comprehensively evaluate the overall performance of an industrial project based on qualitative and quantitative expressions of the selected four essential (but not necessarily sufficient) dimensions. To get a better understanding of the method, the idea is to be able to process the performance management problem by means of several activities. So, this section aims to present the global structure of the proposed methodological approach, which consists of four main phases: identify, quantify, aggregate and decide. They make up the performance evaluation process for decision making, as shown in Fig. 3.

Each phase provides deliverables that are further analysed in the following steps. Through this four-phase process, the overall performance expressions for all decision alternatives are generated. According to the final performance evaluations, the decision maker either chooses the preferred decision alternative or returns to the previous steps in the process to generate new evaluation results.

A. Identify

The first phase deals with the identification of all performance perspectives to express the overall performance regarding the particular decision context. More specifically, the goal is to identify all relevant stakeholders for the decision problem at hand, the different evaluation periods to be planned along the life history of the project and the evaluation variables to be considered to evaluate the performance for each evaluation period.

Large projects are often characterised by multi-dimensional objectives as well as complex processes, leading to complex and interacting relations among different evaluation variables In this context, it is necessary to develop top-bottom breakdown structures for each evaluation dimensions (as the four dimensions are independent at the highest level) to obtain an operational solution to guide the analysis. Based on the adapted definition of benefit (as described in Section I), it is assumed that this dimension can only be qualitatively analysed in the current research. So, the systemic breakdown analysis is applied to the other three axes: cost, value and risk.

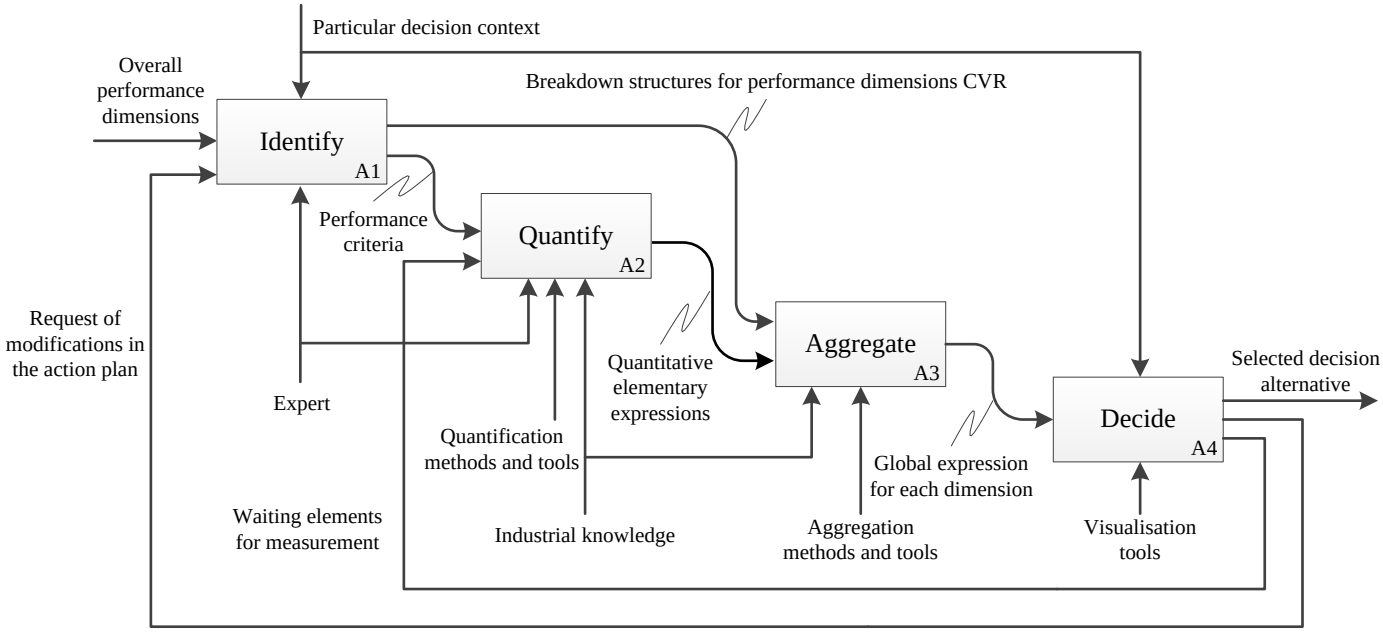


Fig. 3. Global structure of the proposed performance evaluation methodology

Depending on the particular decision context, it is recommended to use either the objective modelling techniques such as value focused thinking [15] or enterprise modelling techniques such as IDEF0 or SADT [16] and IDEF3 [17] as a common basis to integrate the breakdown development of each performance dimension into the same framework to make the analysis more efficient.

In many cases, a performance variable cannot be directly measured. Therefore, performance criteria should be identified in order to express each variable with qualitative or quantitative measurements.

B. Quantify

Based on the selection of performance measurement criteria, quantitative (expressions with numeric numbers) or qualitative (linguistic expressions) elementary performance measures are generated through a set of performance evaluation methods and tools.

Because qualitative performance expressions cannot be directly applied to performance evaluation (they may be subjective and are not additive), the aim of the second phase is to convert these elementary performance expressions into a quantifiable way to ease decision making.

Several methods, such as the Likert scaling [18] (for quantitative expressions) and the fuzzy handling [19] (for qualitative expressions), are used in the methodology to transform the qualitative expressions into quantitative values.

C. Aggregate

For the purpose of decision support, various performance expressions with different attributes should be unified into the same structure. The third phase allows the calculation with all relative performance elements in order to obtain the overall performance evaluation of each assessment dimension.

The elementary expressions with fuzzy numbers are normalised with a defuzzification procedure to find the best non-fuzzy performance value [20].

Based on the multi-attribute utility theory [21], the other elementary expressions with singles numeric numbers are normalized by using (1), (2) and (3).

$$C(x_i) = \begin{cases} 1, & x_i \le x_i^+ \\ \frac{B - x_i}{B - x_i^+}, & x_i^+ < x_i < B \\ 0, & x_i \ge B \end{cases} \quad (1)$$

$$V(x_i) = \begin{cases} 1, & x_i \ge x_i^+ \\ \frac{x_i - x_i^-}{x_i^+ - x_i^-}, & x_i^- < x_i < x_i^+ \\ 0, & x_i \le x_i^- \end{cases} \quad (2)$$

$$R(x_i) = \begin{cases} 1, & x_i \le x_i^+ \\ \frac{x_i - x_i^+}{x_i^- - x_i^+}, & x_i^+ < x_i < x_i^- \\ 0, & x_i \ge x_i^- \end{cases} \quad (3)$$

Where,

- $C(x_i)$ is the normalised elementary cost measure
- $V(x_i)$ is the normalised elementary value measure
- $R(x_i)$ is the normalised elementary risk measure
- x_i is the quantitative expression for the criterion i
- x_i^- is the worst result for the criterion i, it can be considered as the minimum performance that the stakeholder can accept
- x_i^+ is the best result for the criterion i, it is the ideal result that the decision maker expects
- B is the initial budget that is set by the project owner

Then, with the 1-9 linear scale, the evaluator gives the pair-wised judgments at each level of the individual breakdown structures to generate the weights for each criterion. To ensure the judgments are considered as not random, the consistency ratio is verified and it should be less than 10 % [5].

Then, through the selected aggregation operator, the normalised elementary performance expressions are transformed into global performance results for decision support. In this paper, the 2-additive Choquet integral [22], which considers only pairwise interactions between elementary expressions, is applied to generate overall performance expressions of cost ,value and risk.

D. Decide

Once the aggregation is finished for each scenario, the quantitative results of the overall performance for an industrial project can be evaluated. The objective of the decision support phase is to present the overall performance in a visual way (with a cost-value-risk 3D graph) to assist the evaluators in decision making.

IV. Experimental application

The application of the proposed methodology will be illustrated by using a case study dealing with building construction project implementation. The main results are presented in this section.

A. Background

A construction company simultaneously implements three different construction projects, namely P1, P2 and P3, which differ by complexity, duration, budget and variety of works. To make better use of shared material, financial and human resources to achieve higher profits, the company needs efficient performance evaluation for each of these projects during the implementation stage. This will ease the company decision making for the allocation of business resources based on the comparison of evaluation results.

B. Identification

The construction company is considered to be the main stakeholder in this case. In addition, the performance evaluation period focuses on the implementation phase of these construction projects.

The benefits in this decision problem are summarised as: (1) increasing efficiency of the construction company in projects implementation and (2) internal and external improvement of the enterprise image.

To generate elementary performance expressions, a set of common evaluation criteria are summarised according to practices in construction companies in [23]. Expert evaluation of factors affecting the overall performance is also taken into consideration in the selection of criteria. In this context, ten criteria are identified as shown in Table I.

In this case study, the cause and effect diagram is applied to develop the reference hierarchical structure. Then, the relations among evaluation criteria, basic elements in the hierarchical structure and selected performance dimensions can be identified by evaluators according to previous experiences. Based on these results, the individual breakdown structures of cost, value and risk can be developed. The breakdown structure of the project risk is shown in Fig. 4.

TABLE I. Selected performance criteria

Criteria	Descriptions	Units
C1: Project profitability	Project profit rate in the project income	%
C2: Estimated project expenditures	Percentage of project expenditures with the income	$ (million euros)
C3: Project team productivity	Ratio between income and the number of employees	$ (million euros)/person
C4: Number of accidents	Number of technical or security accidents	Units
C5: Nonconformities in project quality audits	Number of deviations from the specification or expectations of the project quality	Units
C6: Delay of project close	Additional time before closure	Month
C7: Budget compliance	Ratio between expenditures and budget	%
C8: Documentation management	Evaluation on the errors in documentation and mislead records	Linguistic expression
C9: Environmental appraisal	Evaluation on environmental impact	Linguistic expression
C10: Personal risk	Estimation of human errors	%

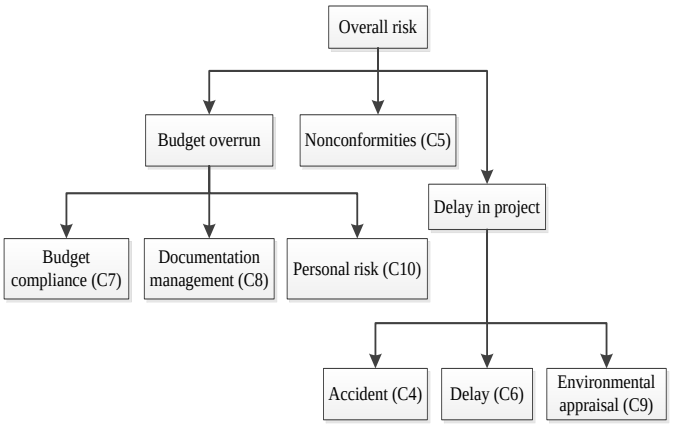


Fig. 4. Breakdown structure of project risk

C. Aggregation

The weights and consistency ratios (CR) for the elements in the breakdown structure of risk are shown in Table II.

Table III presents the elements for aggregation computation of elementary risk expressions with the 2-additive Choquet integral method (including consistent results of the Shapley parameters v_i and the interaction parameters I_{ij}). The similar operations can be applied to the dimensions of cost and value.

D. Decision support

To help the evaluator in the comparison of different candidate projects, the overall performance expressions for each project are displayed in the 3D (cost-value-risk) graph with the same operations in each dimension. Fig. 5 shows the graphical representation for all candidate projects.

This graph shows that projects 2 and 3 have much higher value than project 1, while their cost and risk performance are also better. Between projects 2 and 3, project 3 can generate higher value with less risk and cost. Therefore, it can be considered that the order of priority among these three projects is: project 3 > project 2 > project 1.

This case study deals with a decision problem when comparing projects at a certain evaluation period. Based on previous experiences in similar projects and objective modelling, the overall project performance can be broken down into individual hierarchical structures for each performance dimension with elementary performance expressions.

TABLE II. Consistency ratio and weights for risk elements

Elements (level 1)	Weights	Elements (level 2)	Weights
Budget overrun	0.52	Budget compliance (C7)	0.56
		Documentation management (C8)	0.12
		Personal risk (C10)	0.32
-		CR = 0.02	
Nonconformities (C5)	0.14	-	
Delay in project	0.34	Accident (C4)	0.23
		Delay (C6)	0.10
		Environmental appraisal (C9)	0.67
CR = 0.05		CR = 0.08	

TABLE III. Elements for aggregation with 2-additive Choquet integral

	Risk variables			Overall risk
	Budget overrun	Nonconformities	Delay in project	
P1	0.28	0.70	0.47	0.27
P2	0.41	0.50	0.41	0.40
P3	0.66	0.60	0.56	0.57
v_i	0.52	0.14	0.34	-
I_{ij}	I_{12}	I_{13}	I_{23}	-
	0.15	0.75	0.31	-

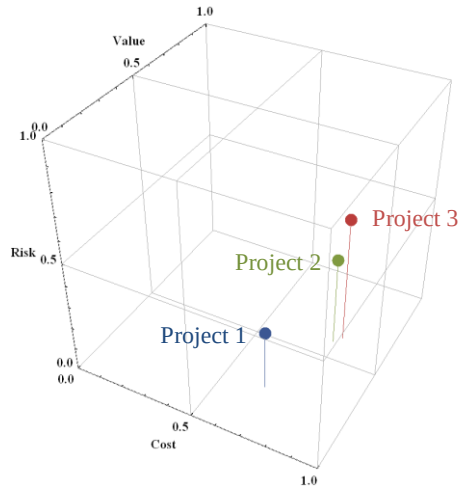


Fig. 5. Plot of overall performance expressions for all candidate projects

V. Conclusions and perspectives

The proposed methodology provides decision makers with an efficient methodological framework that is able to comprehensively evaluate the overall performance of an industrial project with limited complexity. The methodology is instrumented with mathematical tools to integrate both qualitative and quantitative elementary performance expressions into aggregated quantitative expressions to support performance evaluation. Finally, it includes graphical facilities to present the overall performance expressions to ease decision making process and support visual management. Although the methodology is developed and illustrated in the context of project management, it is considered that it can be extended to any type of industrial systems.

Concerning the research perspectives, the application of the proposed methodology should be enriched for performance evaluation in other decision making contexts. Another improvement is the development of an IT tool that allows the automated computation regarding various operations of the whole evaluation process.

References

[1] S. S. Nudurupati, U. S. Bititci, V. Kumar, and F. T. S. Chan, "State of the art literature review on performance measurement," Comput. Ind. Eng., vol. 60, no. 2, pp. 279–290, Mar. 2011.

[2] R. Cooper and R. S. Kaplan, "Measure costs right: make the right decisions," Harv. Bus. Rev., vol. 66, no. 5, pp. 96–103, 1988.

[3] R. S. Kaplan and D. P. Norton, "The balanced scorecard: Measures that drives performances," Harv. Bus. Rev., vol. 70, no. 1, pp. 71–79, 1992.

[4] M. Tweeddale, Managing risk and reliability of process plants. Gulf Professional Publishing, 2003.

[5] T. L. Saaty, Decision making for leaders: the analytic hierarchy process for decisions in complex world, Third Edition. RWS Publications, 2012.

[6] M. Bitton, "ECOGRAI : méthode de conception et d'implantation de systèmes de mesure de performances pour organisations industrielles," Université de Bordeaux 1, 1990.

[7] Y. Ducq and B. Vallespir, "Definition and aggregation of a performance measurement system in three aeronautical workshops using the ECOGRAI method," Prod. Plan. Control, vol. 16, no. 2, pp. 163–177, Mar. 2005.

[8] P. Suwignjo, U. S. Bititci, and A. S. Carrie, "Quantitative models for performance measurement system," Int. J. Prod. Econ., vol. 64, no. 1, pp. 231–241, 2000.

[9] A. Neely, C. Adams, and M. Kennerley, The Performance Prism: The ScoreCard for Measuring and Managing Business Success. FT Prentice Hall, 2002.

[10] F. Marmier, D. Gourc, and F. Laarz, "A risk oriented model to assess strategic decisions in new product development projects," Decis. Support Syst., vol. 56, pp. 74–82, Dec. 2013.

[11] F. Vernadat, L. Shah, A. Etienne, and A. Siadat, "VR-PMS: a new approach for performance measurement and management of industrial systems," Int. J. Prod. Res., vol. 51, no. 23–24, pp. 7420–7438, 2013.

[12] L. Shah, "Value-Risk based Performance Evaluation of Industrial Systems," Arts et Métiers ParisTech, 2012.

[13] F. Li, A. Etienne, A. Siadat, and F. Vernadat, "BCVR: A methodological framework for industrial performance management and decision-support," in Proceedings of 2015 IEEE International Conference on Industrial Engineering and Systems Management (IESM), 2015, pp. 1258–1267.

[14] PMBOK® Guide, "A guide to the project management body of knowledge (PMBOK® guide). -- Fifth edition." Project Management Institute, 2013.

[15] R. L. Keeney, Value-Focused Thinking: A Path to Creative Decision-making. Harvard University Press, 1992.

[16] D. T. Ross, "Structured Analysis (SA): A Language for Communicating Ideas," Ieee Trans. Softw. Eng., vol. SE-3, no. 1, pp. 16–34, 1977.

[17] R. J. Mayer, C. P. Menzel, M. K. Painter, P. S. deWitte, T. Blinn, and B. Perakath, "Information Integration for Concurrent Engineering (IICE), IDEF3 Process Description Capture Method Report." 1995.

[18] R. Likert, "A Technique for the Measurement of Attitudes," Arch. Psychol., vol. 140, pp. 1–55, 1932.

[19] V. Clivillé, L. Berrah, and L. Foulloy, "Fuzzy Symbolic Handling of Industrial Instantaneous and Trend Performance Expressions," in Advances in Production Management Systems: Innovative and Knowledge-Based Production Management in a Global-Local World: IFIP WG 5.7 International Conference, APMS 2014, vol. 440, 2014, pp. 68–75.

[20] S. Vinodh, M. Prasanna, and N. Hari Prakash, "Integrated Fuzzy AHP–TOPSIS for selecting the best plastic recycling method: A case study," Appl. Math. Model., vol. 38, no. 19–20, pp. 4662–4672, Oct. 2014.

[21] R. A. Ogle, S. J. Dee, and B. L. Cox, "Resolving inherently safer design conflicts with decision analysis and multi-attribute utility theory," Process Saf. Environ. Prot., vol. 97, pp. 61–69, Sep. 2015.

[22] C. Labreuche and M. Grabisch, "The Choquet integral for the aggregation of interval scales in multicriteria decision making," Fuzzy Sets Syst., vol. 137, no. 1, pp. 11–26, 2003.

[23] E. K. Zavadskas, T. Vilutienė, Z. Turskis, and J. Šaparauskas, "Multi-criteria analysis of Projects' performance in construction," Arch. Civ. Mech. Eng., vol. 14, no. 1, pp. 114–121, Jan. 2014.

A review of software capabilities for project planning and control

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Robert Pellerin
Department of mathematics and industrial engineering
École Polytechnique de Montréal
Montréal, Canada
robert.pellerin@polymtl.ca

Nathalie Perrier
Department of mathematics and industrial engineering
École Polytechnique de Montréal
Montréal, Canada
nathalie.perrier@polymtl.ca

Abstract—This paper reviews the main capabilities of project planning and control software developed during the last two decades. The review includes a classification scheme of the main characteristics of project management software in terms of project phases, management functions, and project management tools. Key research gaps are highlighted and promising future research trends are discussed. Our literature review demonstrates that most researchers have focused their studies on project planning and control software and ignored more integrated packages such as Enterprise Resource Planning and Project Management Information Systems.

Keywords—project management software; software capabilities; survey; review; planning and control.

I. Introduction

Project management generally consists of three phases: (1) planning, (2) scheduling, and (3) monitoring and control [1]. In the planning phase, the structure of the project is depicted at an aggregated level as a work-breakdown structure (WBS) and on a detailed level as a network diagram representing the technological dependencies between the activities (or tasks) of the project. The time required to perform each activity is also estimated, drawing on a stated resource level (manpower, machinery, etc.) for each activity. Based on the network and the duration of the activities, the earliest and latest start and finish times of the activities are then calculated and a resource profile showing the amount of resource capacity required over time is generated.

In the scheduling phase, the activities of the project are ordered subject to finite capacities of resources. Scheduling computations are commonly done using operations research methods, such as heuristics or exact methods. The time-cost trade-off of activity performance times may be considered in order to determine the cost of reducing the project completion time. Resource allocation, i.e., assigning the proper resource (staff) to each task, can also be done so that the feasibility of each schedule is checked with respect to resource requirements and availabilities.

In the last phase, the project is monitored and controlled by a comparison with actual status of the project against the planned baseline schedule. Project performance data are collected with respect to the baseline, performance measures are produced, and this information is reported. Based on these evaluations, predictions are made and revision of the project plan (re-planning) can be done when necessary.

Each phase is considered a leading indicator for assisting in the success of the delivery of the project. Each of these phases must thus be managed well for a project to be successful [2].

Over the years, a large number of project management software packages have been developed to assist the project manager in project management. The most common project management software packages included in experimental and comparative studies are Microsoft Project and Primavera [3, 4, 5, 1, 6, 7, 8, 9, 10, 11, 12, 13, 2, 14].

Despite these software are labelled as project management tools, they are in practice planning and control information systems and cannot be compared with integrated information systems. While not being able to cover the full range of project management functions, these packages can support the user in planning, scheduling, monitoring, and controlling the project. These packages also assist the user in managing project functions, such as time management, resource management, cost management, risk management, procurement management, and quality management. Each phase and each function can be described in terms of methods, tools, and techniques to support project managers in the decision-making process.

The aim of this paper is to provide a review of the studies on the evaluation and comparison of project management software capabilities for supporting planning and control functions. The emphasis is on the integration of project management phases, functions, methods, tools, and techniques into project management software. The review includes a detailed classification scheme of the main characteristics of project management software over the past 20 years, discussion of research areas demanding further effort, and identification of key opportunities and needs in future research.

The paper is organized as follows. The main characteristics of project management are first presented in Section II. An analysis of existing research on project management capabilities is then presented in Section III. Conclusions along with some promising future research opportunities are presented in the last section.

II. Characteristics of project management software

This section presents a comprehensive classification scheme of the main characteristics of project management software (Table I). The first level of the classification is composed of four categories: (1) Project management phases, (2) Project management functions, (3) Methods, tools, and techniques, and (4) Instance data.

The first category, Project management phases, classifies project management software features according to the project phase considered. The second category, Project management functions, provides the basic project management software functions. Typical basic function areas of this category include time management, resource management, cost management, and risk management. Since, time management is a key part in project management software, various software packages for project management include a procedure for resource-constrained project scheduling (RCPS), time-cost trade-off scheduling, resource allocation, and/or scheduling under uncertainty.

TABLE I. Characteristics of project management software

1. Project management phases
   - 1.1. Planning phase
   - 1.2. Scheduling phase
   - 1.3. Monitoring and control phases
2. Project management functions
   - 2.1. Time management
     - 2.1.1. RCPS
     - 2.1.2. Time-cost trade-off scheduling
     - 2.1.3. Resource allocation
     - 2.1.4. Scheduling under uncertainty
   - 2.2. Resource management
   - 2.3. Cost management
   - 2.4. Risk management
3. Methods, tools, and techniques
   - 3.1. Gantt charts
   - 3.2. Network diagrams
   - 3.3. Work-breakdown charts
   - 3.4. Task scheduler (or scheduling engine)
   - 3.5. PERT charts
   - 3.6. S-Curves
   - 3.7. Earned-value charting tools
   - 3.8. Resource-availability charting
   - 3.9. Resource-capacity tools
   - 3.10. Resource-leveling tools
   - 3.11. Cost-tracking tools
   - 3.12. Risk-simulation tools
   - 3.13. Issue-tracking system
4. Instance data
   - 4.1. Real world instances
   - 4.2. Randomly generated instances

The third category, Methods, tools, and techniques, comprises numerous project management software tools and calculation methods. Common tools in project management software packages include Gantt charts, WBS, scheduling engine, network diagrams, resource charting, tracking tools, cost-control mechanisms, project communication systems, and risk management features [15]. For the generation of resource feasible schedules, a variety of optimization methods are now embedded in project management software packages. High quality schedules (in terms of project duration) can be generated through exact branch-and-bound procedures [16] for small instances of projects, or classical heuristic solution procedures and metaheuristics, such as priority-rule heuristics [17, 18, 12], forward-backward improvement methods [19], tabu search [16], genetic algorithms [20], and evolutionary algorithms [21] for large-scale projects. The last category, Instance data, classifies the type of instance used for the evaluation and comparison of software packages.

In the next section, the categories presented in Table I are applied to summarize the main capabilities of project management software tools developed during the last two decades.

III. Analysis of existing research

In this section, we provide a classification of important studies on project management software capabilities over the past 20 years. To this end, the categories and characteristics presented in Table I are used as a basis to classify the research works in Table II. This table is arranged chronologically with the oldest works near the top. The categories from Table I are listed across the top of Table II. The remainder of this section follows the classification categories in Tables I-II to highlight key research contributions and gaps.

To compare the software packages, most authors concentrate on scheduling capabilities. This is not surprising, since resource-constrained project scheduling is definitively one of the principle keys to the quality of project management software. The problem of scheduling projects with resource constraints is especially difficult to solve, due to the combinatorial nature of the problem. Therefore, most of the task scheduler tools to solve the RCPS problem are heuristics.

The drawback of the RCPS problem is that it is assumed that a task can be processed in one specific way with a deterministic expenditure of time and a fixed resource requirement. Therefore, one interesting line of research would be the further implementation of more sophisticated task scheduling tools in software packages to solve more realistic scheduling problems. The multicriteria evolutionary metaheuristic presented by Tackenberg et al. [21] for multi-mode RCPS and multi-skill project scheduling is an illustrative example of this progress.

The dominance of the time management function is also clear in Table II, with studies including resource management, cost management, or risk management being quite rare [10, 11, 2]. This is surprising since, recently, risk management and adequate project planning and scheduling under uncertainty have been identified as critical conditions for successful project management [16].

In fact, the majority of the software packages reported in the studies perform the basic scheduling calculations assuming that activity durations are deterministic and known to the project manager. In practice, however, activity durations can

only be estimated and the actual activity durations may deviate from these estimates due to underestimation of the work content, errors during execution, unavailability of key staff or equipment. Also, in modeling stochastic project networks, the assumption that activity durations are independent might not be valid for many real-world cases [24]. Effective project management requires that the project manager takes these risks into account by performing a risk analysis on the activities, leading to activity duration distributions rather than deterministic activity durations. Using this information, proactive scheduling steps can be taken to make the project schedule more robust [16].

In addition to these tools that support scheduling and risk analysis, project managers require reliable warning systems for project monitoring and control. However, existing research on project management software reveals that these systems are rarely included. Analytical project control tools that support the managers to intervene effectively and stimulate them to undertake corrective measures at the right time are needed [24]. Monitoring and control tools to help decision makers in determining the timing and magnitude of corrective actions, such as rescheduling, effectively during project execution will contribute to fulfilling the needs of project managers.

Finally, most of the time, resource-constrained project scheduling capabilities of project management software are tested with resource constrained project scheduling problems taken from the PSPLIB library [3, 1, 7, 8, 9, 19, 12, 13]. Real-world projects are very rarely considered [18].

IV. Conclusion

In this paper, we reviewed the current studies on software capabilities for project planning and control. We discussed the main characteristics of project management software and analyzed existing research on project planning and control capabilities.

In this regard, we found that the quality of project management software packages is commonly evaluated and compared based on their scheduling capabilities [3, 4, 1, 7, 8, 9, 12]. In fact, the RCPS is still one of the principle keys to the quality of project management software today. We emphasized the importance of embedding monitoring and control tools as well as task scheduler tools to solve more realistic scheduling problems in project management software packages. For example, re-planning and dynamic scheduling methods are valuable tools as the project manager can observe the effects of modifications instantly by recalculating the schedule [5]. Robust scheduling tools and corrective action procedures are also required to generate a project schedule that anticipates on the occurrence of unforeseen events [16].

Another promising direction of research is the attention to the stochastic nature of projects [26]. In fact, most time management functions of project management software are restricted to deterministic approaches. Besides the classical probabilistic techniques of PERT, the uncertainty inherent to each project, such as activity duration uncertainty, can be taken into account by using critical chain/buffer management [16]. Random distribution functions can also be used to sample activity costs and durations and facilitates an understanding of the complex, stochastic, and dynamic behaviors of real project systems [24]. These stochastic techniques are able to provide early warning signals based on the expected behavior of the project [26].

With respect to enhanced validation, scheduling capabilities of project management software should be extensively tested using a large and diversified historical dataset based on real-world case studies or a well-designed simulation experiment [26]. Papers presenting only one example [14, 17] or a case study on one or two projects [18] merely provide an illustration of the scheduling capabilities of project management software.

Clearly, the integration of additional functions of project management, other than the classical planning and control functions, could be developed. For example, Pakgohar et al. [20] presented an integrated decision support system capable of greatly improving the bidding, subcontractor selection and project planning processes in a holistic integrated framework. The ultimate goal being the development of a fully integrated project management software that incorporates decision-aiding tools supporting the planning, scheduling, monitoring, and control phases of project management as well as the management of functions that are most often featured in studies as critical factors for successful project management.

From a methodological point of view, it is important to note that most researchers have focused on project planning software and disregard other more integrated tools such as ERP (Enterprise Resource Planning) and PMIS (Project Management Information System) which cover a larger scope of project management activities [27]. The non-consideration of these kinds of information systems and the emergence of new information systems such as Portfolio Management software clearly create a bias in our literature review.

In conclusion, it is surprising to note that many researchers in our field still associate project management with the only planning and control function. This may be explained in part by the marketing of traditional project planning software editors that brand their product as project management tools. In practice, the diversity and the gap between 'project management' information systems are becoming wider as their number continue to raise.

Acknowledgment

This work has been supported by the Natural Sciences and Engineering Research Council of Canada and the Jarislowsky/SNC-Lavalin Research Chair in the Management of International Projects. This support is gratefully acknowledged.

TABLE II. Classification of project management capabilities studies

Authors	1. Project phases			2. Management functions							3. Methods, tools, and techniques													4. Data	
				2.1. Time management																					
	1.1. Planning	1.2. Scheduling	1.3. Monitoring and control	2.1.1. RCPS	2.1.2. Time-cost trade-off	2.1.3. Resource allocation	2.1.4. Scheduling under uncertainty	2.2. Resource management	2.3. Cost management	2.4. Risk management	3.1. Gantt charts	3.2. Network diagrams	3.3. Work-breakdown charts	3.4. Task scheduler	3.5. PERT charts	3.6. S-Curves	3.7. Earned-value charts	3.8. Resource-availability charts	3.9. Resource-capacity tools	3.10. Resource-leveling tools	3.11. Cost-tracking tools	3.12. Risk-simulation tools	3.13. Issue-tracking tools	4.1. Real world instances	4.2. Randomly generated instances
Kolisch [3]		√		√										√											√
Maroto et al. [4]		√	√	√		√											√			√					√
Stilgenbauer et al. [5]		√	√											√					√						
Tormos and Lova [17]		√	√	√										√											
Bidanda and Hackworth [15]	√	√									√	√	√	√	√		√	√	√	√	√	√	√		
Hekimoğlu [1]	√	√		√	√							√	√	√											√
Bell et al. [6]		√												√											
Trautmann and Baumann [7]		√		√		√								√											√
Trautmann and Baumann [8, 9]		√		√										√											√
Cicibas et al. [10]		√	√					√		√				√									√		
Trautmann and Baumann [19]		√		√										√											√
Deblaere et al. [16]	√	√		√			√				√	√		√					√						
Trautmann and Brandinu [18]		√		√										√										√	
Chantrapornchai et al. [22]		√				√								√											
Duarte et al. [23]	√	√	√										√	√	√	√									
Pakgohar et al. [20]	√	√			√						√		√	√											
Gharaibeh [11]	√	√	√			√		√	√			√		√						√		√			
Damiani et al. [2]		√	√					√	√					√									√		
Hazir [24]		√	√		√									√			√								
Tiwari and Johari [14]		√		√	√									√											
Baumann and Trautmann [12, 13]		√		√										√											√
Tackenberg et al. [21]		√		√										√											√

References

[1] Ö. Hekimoğlu, "Comparison of the resource allocation capabilities of project management software packages in resource constrained project scheduling problems", Master thesis, Middle East Technical University, 2007.

[2] L. Damiani, R. Revetria, I. Svlilenova, P. Giribone, "Survey and comparison of the project management softwares used by engineering, procurement and construction companies", in Advances in Energy and Environmental Science and Engineering, A. Bulucea, Ed. WSEAS Press, 2015, pp.73-81.

[3] R. Kolisch, "Resource allocation capabilities of commercial project management software packages", Interfaces, vol. 29, 1999, pp.19-31.

[4] C. Maroto, P. Tormos and A. Lova, "The evolution of software quality in project scheduling", in Project Scheduling, J. Węglarz, Ed. US: Springer, 1999, pp.239-259.

[5] T.M. Stilgenbauer, T. Nicholas and A.L. Brizendine, "Scheduling transportation projects using Primavera Project Planner as part of the software series in civil engineering technology independent learning experiment at Fairmont State College", J. Eng. Technol., vol. 18, 2001, pp.24-30.

[6] M. Bell, T. Sulbaran, K. Williams and M. Blain, "Case study on scheduling software for prime contractors' efficient management of projects", 2008.

[7] N. Trautmann and P. Baumann, "Project scheduling with precedence constraints and scarce resources: an experimental analysis of commercial project management software", in Operations Research Proceedings 2008, B. Fleischmann, K.H. Borgwardt, R. Klein and A. Tuma, Eds., Berlin, Heidelberg: Springer-Verlag, 2009, pp.165-170.

[8] N. Trautmann and P. Baumann, "Resource-constrained scheduling of a real project from the construction industry: a comparison of software packages for project management", IEEE International Conference on Industrial Engineering and Engineering Management, pp.628-632, December 2009.

[9] N. Trautmann and P. Baumann, "Resource-allocation capabilities of commercial project management software: an experimental analysis", International Conference on Computers & Industrial Engineering, pp.1143-1148, July 2009.

[10] H. Cicibas, O. Unal and K.A. Demir, "A comparison of project management software tools (PMST)", Proceedings of the 9th Software Engineering Research and Practice, pp.560-565, July 2010.

[11] H. Gharaibeh, "Evaluating project management software packages using a scoring model—A comparison between MS Project and Primavera", J. Software Eng. Appl., vol. 7, 2014, pp.541-554.

[12] P. Baumann and N. Trautmann, "Resource-constrained project scheduling with project management information systems", in Handbook on Project Management and Scheduling, vol. 2, C. Schwindt and J. Zimmermann, Eds. Switzerland: Springer International Publishing, 2015, pp.1385-1400.

[13] P. Baumann and N. Trautmann, "A note on the selection of priority rules in software packages for project management", Flex. Serv. and Manuf. J., vol. 28, 2016, pp.694-702.

[14] S. Tiwari and S. Johari, "Project scheduling by integration of time cost trade-off and constrained resource scheduling, J. Ins. Eng. Ser. A, vol. 96, 2015, pp.37-46.

[15] B. Bidanda and D. Hackworth, "Project-management software: a guideline for system selection and use", in Field Guide to Project Management, D. I. Cleland, Ed. New Jersey: John Wiley & Sons, Inc., 2004, pp.426-447.

[16] F. Deblaere, E. Demeulemeester and W. Herroelen, "RESCON: Educational project scheduling software", Comput. Appl. Eng. Educ., vol. 19, 2011, pp.327-336.

[17] P. Tormos and A. Lova, "Tools for resource-constrained project scheduling and control: forward and backward slack analysis", J. Oper. Res. Soc., vol. 52, 2001, pp.779-788.

[18] N. Trautmann and G. Brandinu, "A serial scheme for minimizing the duration of resource-constrained projects within Microsoft Project", IEEE International Conference on Industrial Engineering and Engineering Management, pp.893-897, December 2011.

[19] N. Trautmann and P. Baumann, "An iterative backward/forward technique for the scheduling of resource-constrained projects within Microsoft Project", IEEE International Conference on Industrial Engineering and Engineering Management, pp.1558-1562, December 2010.

[20] A. Pakgohar, S.J. Childe and D.Z. Zhang, "A DSS solution for integrated automated bidding, subcontractor selection and project scheduling", in Decision Support Systems II - Recent Developments Applied to DSS Network Environments, J.E. Hernández, Liu S., Delibašić B., Zaraté P., Dargam F. and Ribeiro R., Eds. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 72-85.

[21] S. Tackenberg, S. Duckwitz, C. Schmalz and C.M. Schlick, "Resource-constrained project scheduling problem: investigation of the quality of project plans", in Advances in Ergonomic Design of Systems, Products and Processes, C.M. Schlick et al., Eds. Berlin, Heidelberg: Springer, 2017, pp.31-59.

[22] C. Chantrapornchai, D. Sawangkokrouk and T. Leadprathom, "Project management software: allocation and scheduling aspects", International Journal of u- and e- Service, Science and Technology, vol. 6, 2013, pp.35-42.

[23] D.L. Duarte, H.L. Miranda and J.R. Farias Filho, "A review of the literature about project management tools", 22nd International Conference on Production Research, July-August 2013.

[24] Ö. Hazir, "A review of analytical models, approaches and decision support tools in project monitoring and control", Int. J. Proj. Manag., vol. 33, 2015, pp.808-815.

[25] M.J. Liberatore, B. Pollack-Johnson and C.A. Smith, "Project management in construction: Software use and research directions", J. Constr. Eng. M., vol. 127, 2001, pp.101-107.

[26] L.L. Willems and M. Vanhoucke, "Classification of articles and journals on project control and earned value management", Int. J. Proj. Manag., vol. 33, 2015, pp.1610-1634.

[27] R. Pellerin, N. Perrier, X. Guillot and P.M. Leger. "Project characteristics, project management software utilization and project performance: an impact analysis based on real project data", Int. J. Inf. Sys. Proj. Manag., vol. 3, 2013, pp.5-28.

Supplier Selection and Evaluation

A Supplier Selection Model Development with Six Sigma Quality Indices for Products with Multiple Quality Characteristics

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Kuen-Suan Chen
Department of Industrial Engineering and Management,
National Chin-Yi University of Technology
Taichung 41170, Taiwan, R.O.C.
Email: kschen@ncut.edu.tw

Ching-Hsin Wang*
Department of Leisure Industry Management,
National Chin-Yi University of Technology
Taichung 41170, Taiwan, R.O.C.
Email: thomas_6701@yahoo.com.tw

Kim-Hua Tan
Business School,
University of Nottingham
NG8 1BB Nottingham, U.K.
Email: Kim.Tan@nottingham.ac.uk

Abstract—To satisfy reasonable costs and demands for high quality products, good supplier selection mechanisms have also come to play vital roles. A number of studies indicate that quality is an important criterion for a supplier selection. However, a product generally possesses several quality characteristics, all of which must attain certain quality levels in order for the product to be considered a good product that meets customers' needs. This study therefore employed the Six Sigma Quality Indices, which can directly reflect the yield and Six Sigma quality levels of processes, to develop a supplier selection model for products with multiple larger-the-better, smaller-the-better, and nominal-the-best quality characteristics. This model provides a selection criterion for enterprises choosing suppliers. We first proposed Six Sigma Quality Indices for products with multiple quality characteristics. Based on the definition of Six Sigma, we established the quality levels required of products and the quality levels required of each quality characteristic. Next, we derived the upper confidence limits for three types of Six Sigma Quality Indices and used the upper confidence limits to perform a systemic test on the Six Sigma Quality Indices. Based on the required quality levels of the quality characteristics and the upper confidence limits of the Six Sigma Quality Indices, we then constructed a supplier selection radar chart. Finally, we demonstrated the application of the proposed supplier selection model using high intensity discharge (HID) headlights manufactured in Taiwan.

Keywords—supplier selection, Six Sigma Quality Index, supplier selection radar chart, upper confidence limit

I. Introduction

As enterprises put more and more emphasis on their own core technologies as well as their professional manufacturing and services, it will become a future trend of business operation models to externally receive proper resources to participate into manufacturing and services. Enterprises realize that the outsourcing of non-nuclear activities not only can make the internal resource use more flexible but also can help them obtain suppliers' assistance in technology, profession, and efficiency. At present, the receipts of needs, such as raw materials, components, services, and equipment, in any form of business activities, all have to rely on purchasing. In the entire purchasing process, the most significant factor is the selection of suppliers. In addition, global enterprises have fiercely competed with each other in recent years. Therefore, the process of purchase plays an important role in business operation. Namely, if an enterprise wants to manufacture low-cost and high-quality products, it needs to rely on a good supplier.

The selection of suppliers should be made based on the suppliers' performances. The comparisons of the suppliers' performances have been discussed in numerous literatures. Dickson [1] did the earliest research on clearly defining the indices of the supplier selection, indicating that among 23 items of the supplier selection criteria, quality, delivery time, and the past historical performance, are the most important ones. On the basis of Dickson's 23 items of selection criteria, Weber et al. [2] studied the frequencies of these 23 criteria shown in different literatures and sorted out the most commonly used selection criteria in order - price, delivery time, quality, and production capacity. Patton [3] proposed 7 items of supplier evaluation criteria, including product quality, price, delivery time, sales support, equipment and technology, order status, and financial status. Verma and Pullman [4] stated that "quality", "price", and "delivery time" of a product were the critical factors for customers to select suppliers. Chen et al. [5] and Chen and Chen [6] all addressed that the present competitive environment had become the competition among the supply chains, so the quality of suppliers was a critical factor for determining the competition strength of a product., Wu et al. [7], Pearn et al. [8], and Kang et al. [9] all pointed out that when an enterprise selected a good supplier, the quality yields for raw materials, components, services, and equipment required by the supplier were crucial because they would influence the quality or function of the enterprise's end product. It can be perceived from the above-mention ed literatures that quality is one of the essentail criteria when selecting a supplier or a manufacturing process. Concerning the quality characteristics of a product or whether a supplier can reach the preset required level, the anaysis of Process Capability Indices (PCIs) canhelp manufacturers evaluate the process level of each product characteristic while clients can realize whether the manufacturer's or the supplier's process capability can meet their requirements and take this analysis as a reference for selecting a better supplier in the

future. As a result, PCIs have been applied to the supplier selection in many literatures. Chen and Chen [10] used PCI ratios as a basis for the supplier selection and conducted a simulation and comparison with the approximate distribution and three Bootstrap Resampling Methods to select a better supplier. Chen and Chen [6] developed TDD indices based on PCI C_{pp}, integrating various aspects, including quality, delivery time, and flexibility, for a supplier, presenting the supplier's overall performance, and leading the subsystem to do the right thing for the entire system to avoid the partial performance evaluation. Based on the probability density function of the estimate ratio for two indices induced from the process yield index S_{pk}, Lin and Pearn [11] conducted a two-staged one-tailed test to compare two suppliers and used the threshold value method as a selection policy. Liao et al. [12], Wu et al. [13] and Wu et al. [14] separately adopted PCIs C_{pk}, C_{pm} and C_{pmk} as a basis, used the fuzzy set in a way of α-cut to construct fuzzy membership functions of indices, and made use of the membership functions and the relationship among thresholds to establish a set of examination and evaluation criteria for the supplier selection.

In addition, because the Six Sigma quality management model can effectively enhance the process quality, many statistic scholars and quality engineers, such as Chen et al. [15], Huang et al. [16], and Chen et al. [17], individually discussed the correlations between the process capability indices and the Six Sigma quality levels and built evaluation improvement models. Although the methods proposed by the above-mentioned researchers can judge whether the process quality can meet the Six Sigma quality levels requested by customers, it can only be learned through conversion. Thus, it is relatively inconvenient in management. Concerning the larger-the-better (LTB), smaller-the-better (STB), and nominal-the-best (NTB) quality characteristics, Chen et al. [18] proposed the Six Sigma Quality Indices (SSQIs) for each separately. These three indices not only can directly reflect the Six Sigma quality level of the process but also can directly respond to the process yield. Furthermore, , Chen and Chen [19], Pearn and Cheng [20], and Chen et al. [17] all indicated that normally a product had several quality characteristics and could be seen as a good one which could satisfy customers' demands when the quality levels of all quality characteristics met the necessary requirements. Thus, this study applies the Six Sigma Quality Indices (SSQIs) proposed by Chen et al. [18] and develops a supplier selection model, a selection criterion for enterprises when they select suppliers, in the light of the larger-the-Better (LTB), smaller-the-better (STB), and nominal-the-best (NTB) quality characteristics. First, the Six Sigma evaluation indices corresponding to the products of multiple quality characteristics are presented in Chapter 2. Meanwhile, according to the definition of Six Sigma, the required values of product quality levels and the required values corresponding to the quality levels of quality characteristics are sorted out. Next, in Chapter 3, the upper confidence limits are induced based on three types of Six Sigma evaluation indices, and statistical tests of Six Sigma evaluation indices are conducted through the upper confidence limits. In Chapter 4, a supplier selection radar chart is drawn on the basis of all required values of quality levels of quality characteristics and the upper confidence limits of Six Sigma evaluation indices, and the high intensity discharge (HID) headlights manufactured in Taiwan is taken as an example to explain the application of the supplier selection model developed by this paper. Last, the conclusions of this paper are made in Chapter 5. The evaluation model presented in this paper can offer a company a reference when it carries out a supplier selection, and its criteria can enhance suppliers' self-evaluations and continual improvements, so that the entire product quality as well as business operation performance can be promoted.

II. Setting of Six Sigma Evaluation Indices and the Required Values for the Quality Levels of all Quality Characteristics

As noted by Chang et al. [21], if quality characteristic X is of the smaller-the-better (STB) type, then $0 < X \le USL$. Similarly, if quality characteristic X is of the larger-the-better (LTB) type, then $LSL \le X < \infty$. In addition, if quality characteristic X is of the nominal-the-better (NTB) type, then $LSL \le X < USL$, where USL is the upper specification limit and LSL is the lower specification limit. Chen et al. [17] stated that normally a product had several quality characteristics. If a product has n quality characteristics, including n_l quality characteristics of the LTB type, n_u quality characteristics of the STB type, and n_k quality characteristics of the NTB type, then $n = n_l + n_u + n_k$. Besides, because the Six Sigma quality management model can effectively enhance the process quality, many statistic scholars and quality engineers, such as Chen et al. [15], Huang et al. [16], and Wang et al. [17], individually discussed the correlations between the process capability indices and the Six Sigma quality levels and built evaluation improvement models. Although the methods proposed by the above-mentioned researchers can judge whether the process quality can meet the Six Sigma quality levels requested by customers, it can only be learned through conversion; thus, it is relatively inconvenient in management. Regarding the LTB, STB, and NTB quality characteristics, Chen et al. [18] separately proposed the Six Sigma Quality Indices (SSQIs). These three indices not only can directly reflect the Six Sigma quality level of the process but also can directly respond to the process yield. Considering the product which has multiple quality characteristics, this study applies Six Sigma quality indices to establish an evaluation model. First, concerning n_l quality characteristics of the LTB type, n_u quality characteristics of the STB type, and n_k quality characteristics of the NTB type, the corresponding Six Sigma Quality Indices are presented as follows:

$$Q_{phi} = \begin{cases} Q_{pli} = \dfrac{\mu_{li} - LSL_{li}}{\sigma_{lj}} + 1.5, i = 1,2,...n_l \ (LTB) \\ Q_{pui} = \dfrac{USL_{ui} - \mu_{ui}}{\sigma_{ui}} + 1.5, i = 1,2,...n_u \ (STB) \\ Q_{pki} = Min\left(Q_{pk_1 i},\ Q_{pk_2 i}\right) + 1.5, i = 1,2,...n_k \ (NTB) \end{cases} \quad (1)$$

where μ_{hj} is the process mean, σ_{hj} is the process standard deviation, $Q_{pk_1 i} = (\mu_{ki} - LSL_{ki})/\sigma_{ki}$ and $Q_{pk_2 i} = (USL_{ki} - \mu_{ki})/\sigma_{ki}$. If event E_{hj} represents the qualified quality characteristics under the assumed normality, then

$$(1)\ p(E_{li}) = p(X_{li} \geq LSL_{li}) = p\left(Z \geq -\left(Q_{pli} - 1.5\right)\right) = \Phi\left(Q_{pli} - 1.5\right), i = 1,2,...,n_l$$

$$(2)\ p(E_{ui}) = p(X_{ui} \leq USL_{ui}) = p\left(Z \leq Q_{pui} - 1.5\right) = \Phi\left(Q_{pui} - 1.5\right), i = 1,2,...,n_u$$

$$(3)\ p(E_{ki}) = p\left\{\left(X_{ki} \geq LSL_{ki}\right) \cap \left(X_{ki} \leq USL_{ki}\right)\right\} = 1 - p\left\{\left(X_{ki} > USL_{ki}\right) \cup \left(X_{ki} < LSL_{ki}\right)\right\}$$

$$\geq 1 - p\left(X_{ki} > USL_{ki}\right) - p\left(X_{ki} < LSL_{ki}\right) = 2\Phi\left(Q_{pki} - 1.5\right) - 1, i = 1,2,...,n_k,$$

where $\Phi(\bullet)$ is the Cumulative function of standard normal distributions $(i.e.\ N(0,1))$. By Eq. (1), Eq. (2) and Eq. (3), then $p(E_{hi}) \geq p_{hi}$, where p_{hi} can be shown as follows:

$$p_{hi} = \begin{cases} \Phi\left(Q_{phi} - 1.5\right), h = l\ or\ u \\ 2\Phi\left(Q_{phi} - 1.5\right) - 1, h = k \end{cases} \quad (2)$$

Let $1 - p_{hi} = q_{hi}$, then

$$q_{hi} = \begin{cases} 1 - \Phi\left(Q_{phi} - 1.5\right), h = l\ or\ u \\ 2 - 2\Phi\left(Q_{phi} - 1.5\right) - 1, h = k \end{cases} \quad (3)$$

As stated previously, when the process quality of each quality characteristic could meet the quality level which customers request, then it could be claimed as a qualified product. Therefore, if event E_T represents that the product is qualified, then $E_T = \bigcap E_{hi}$. Based on Boole's inequality and DeMorgan's rule, then

$$p(E_T) = p\left(\bigcap E_{hi}\right) = 1 - p\left(\bigcup E_{hi}^c\right) \geq p_T,$$

where $p_T = 1 - \sum_{h \in (l,u,k)} \sum_{i=1}^{n_h} q_{hi}$.

Chen et al. [22] and Wang et al. [17] all stated that when the requested product quality level is $k\sigma$, then the quality level of each quality characteristic must be as high as $k\sigma$. In other words, when the quality level of each quality characteristic is $k\sigma$ $\left(Q_{phj} = k'\right)$, then k' must be larger than or equal to k $(k' \geq k)$, so that the product quality level will have the chance to reach $k\sigma$. Meanwhile,

$$q_{hi} = \begin{cases} 1 - \Phi\left(k' - 1.5\right), h = l\ or\ u \\ 2 - 2\Phi\left(k' - 1.5\right) - 1, h = k \end{cases} \text{ and} \quad (4)$$

$$p_T = 1 - (n_l + n_u + 2n_k) \times \left(1 - \Phi(k' - 1.5)\right). \quad (5)$$

Breyfogle et al. [23] and Linderman et al. [24] indicated that the Six Sigma designed by Motorola was to allow the process mean to deviate from the process target 1.5σ below. Under such a circumstance, there was an average of merely 3.4 defects per million products (i.e. 3.4 Defects Per Million Opportunities, DPMO). Chen et al. [22] and Chen et al. [15] all addressed in their studies that when $|\mu - T| \leq 1.5\sigma$ and $\sigma = d/k$, the process quality of the quality characteristic reaches level $k\sigma$, and the yield $p(E_T)$ of the product reaches at least $p(k)$, that is,

$$p(E_T) \geq p(k) = \Phi\left(k - 1.5\right) + \Phi\left(k + 1.5\right) - 1 \quad (6)$$

If the product quality level is $k\sigma$ and each quality level of each quality characteristic is equal and at least $k'\sigma$ as requested, then based on Eq. (5) and Eq. (6), the relationship between k' and k received from $p_T = p(k)$ can be seen as follows:

$$1 - (n_l + n_u + 2n_k) \times \left(1 - \Phi(k' - 1.5)\right) = \Phi\left(k - 1.5\right) + \Phi\left(k + 1.5\right) - 1$$

$$\Rightarrow k' = \Phi^{-1}\left(1 - \frac{2 - \Phi\left(k - 1.5\right) - \Phi\left(k + 1.5\right)}{n_l + n_u + 2n_k}\right) + 1.5 \quad (7)$$

For example, suppose a product has 2 quality characteristics of the LTB type ($n_l = 2$), 1 quality characteristic of the STB type ($n_u = 1$), and 3 quality characteristics of the NTB type ($n_k = 3$). If the required product level reaches 6 Sigma, based on Eq. (7), the quality level for each corresponding important quality characteristic can be computed and reach 6.397 Sigma, so that it can be ensured that the product can reach the target of 6 Sigma quality level.

III. Upper Confidence Limits of Six Sigma Quality Indices

Consider X_{hi} a random variable with normal distribution $N\left(\mu_{hi}, \sigma_{hi}^2\right)$. Let $\left(X_{hi1}, X_{hi2}, \cdots, X_{hin}\right)$ be a random sample and let $\bar{X}_{hi}$ and S_{hi}^2 denote the sample mean and sample variance respectively, as follows:

$$\bar{X}_{hi} = \frac{1}{n}\sum_{j=1}^{n} X_{hij} \text{ and } S_{hi}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{hij} - \bar{X}_{hi}\right)^2.$$

The estimators of Q_{phi} can be show as follows:

$$\hat{Q}_{phi} = \begin{cases} \hat{Q}_{pli} = \dfrac{\bar{X}_{li} - LSL_{li}}{S_{li}} + 1.5, i = 1,2,...n_l\ (LTB) \\ \hat{Q}_{pui} = \dfrac{USL_{ui} - \bar{X}_{ui}}{S_{ui}} + 1.5, i = 1,2,...n_u\ (STB) \\ Min\left(\hat{Q}_{pk_1 i}, \hat{Q}_{pk_2 i}\right) + 1.5, i = 1,2,...n_k\ (NTB) \end{cases}, \quad (8)$$

where $\hat{Q}_{pk_1 i} = \dfrac{\bar{X}_{ki} - LSL_{ki}}{S_{ki}}$ and $\hat{Q}_{pk_2 i} = \dfrac{USL_{ki} - \bar{X}_{ki}}{S_{ki}}$.

Under the assumption of normality, considering the following random variables, $K = (n-1)S_{li}^2/\sigma_{li}^2$ and

$Z=\sqrt{n}\left(Q_{pli}-\hat{Q}_{pli}\times\left(\frac{S_{li}}{\sigma_{li}}\right)\right)$ are independently distributed as $N(0,1)$ and χ^2_{n-1}. Obviously, let $\alpha'=1-\sqrt{1-\alpha}$, then

$$P\{Z\le z_{\alpha'}\}=p\left\{Q_{pli}\le\left(\hat{Q}_{pli}-1.5\right)\left(\frac{S_{lj}}{\sigma_{lj}}\right)+\frac{z_{\alpha'}}{\sqrt{n}}+1.5\right\}=1-\alpha' \text{ and}$$

$$p\left\{K\ge\chi^2_{\alpha';n-1}\right\}=p\left\{\frac{S_{li}}{\sigma_{li}}\ge\sqrt{\frac{\chi^2_{\alpha';n-1}}{n-1}}\right\}=1-\alpha',$$

where $z_{\alpha'}$ is the upper α' quantile of $N(0,1)$ and $\chi^2_{\alpha';n-1}$ is the lower α' quantile of the chi-square distribution with $n-1$ degrees of freedom. From these relationships, we obtain

$$\begin{aligned}1-\alpha&=p\left\{Z\le z_{\alpha'},\ \chi^2_{\alpha';n-1}\le K\right\}\\&=p\left\{Q_{pli}\le\left(\hat{Q}_{pli}-1.5\right)\left(\frac{S_{lj}}{\sigma_{lj}}\right)+\frac{z_{\alpha'}}{\sqrt{n}}+1.5,\left(\frac{S_{li}}{\sigma_{li}}\right)\le\sqrt{\frac{\chi^2_{\alpha';n-1}}{n-1}}\right\}\\&\le p\left\{Q_{pli}\le\left(\hat{Q}_{pli}-1.5\right)\sqrt{\frac{\chi^2_{\alpha';n-1}}{n-1}}+\frac{z_{\alpha'}}{\sqrt{n}}+1.5\right\}\end{aligned}$$

Thus, the $(1-\alpha)\times100\%$ upper confidence limit of Q_{pli} is

$$UQ_{pli}=\left(\hat{Q}_{pli}-1.5\right)\sqrt{\frac{\chi^2_{\alpha';n-1}}{n-1}}+\frac{z_{\alpha'}}{\sqrt{n}}+1.5. \quad (9)$$

Similarly, the $(1-\alpha)\times100\%$ upper confidence limit of Q_{pui} is :

$$UQ_{pui}=\left(\hat{Q}_{pui}-1.5\right)\sqrt{\frac{\chi^2_{\alpha';n-1}}{n-1}}+\frac{z_{\alpha'}}{\sqrt{n}}+1.5. \quad (10)$$

For the same reason, the $(1-\alpha_j)\times100\%$ upper confidence limits of Q_{pk_ji} can be show as follows:

$$UQ_{pk_ji}=\left(\hat{Q}_{pk_ji}-1.5\right)\sqrt{\frac{\chi^2_{\alpha'_j;n-1}}{n-1}}+\frac{z_{\alpha'_j}}{\sqrt{n}}+1.5,\ j=1,2.$$

Let

$$A_j=\left\{Q_{pk_ji}\le\left(\hat{Q}_{pk_ji}-1.5\right)\sqrt{\frac{\chi^2_{\alpha'_j;n-1}}{n-1}}+\frac{z_{\alpha'_j}}{\sqrt{n}}+1.5\right\},\ j=1,2.,\ \text{then}$$

$p(A_j)=1-\alpha_j$, $p(A_j^c)=\alpha_j$ and $p(A_1\cap A_2)\ge1-p(A_1^c)-p(A_2^c)=1-(\alpha_1+\alpha_2)$.

Thus, we have

$$p\left\{Q_{pk_1i}\le\left(\hat{Q}_{pk_1i}-1.5\right)\sqrt{\frac{\chi^2_{\alpha'_1;n-1}}{(n-1)}}+\frac{z_{\alpha'_1}}{\sqrt{n}}+1.5,Q_{pk_2i}\le\left(\hat{Q}_{pk_2i}-1.5\right)\sqrt{\frac{\chi^2_{\alpha'_2;n-1}}{(n-1)}}+\frac{z_{\alpha'_2}}{\sqrt{n}}+1.5\right\}\ge1-(\alpha_1+\alpha_2)$$

Let $\alpha_i=\alpha/2$ and $\alpha''=1-\sqrt{1-\alpha/2}$, then $\alpha''_i=1-\sqrt{1-\alpha''_i}=\alpha''$, $i=1,2$, then

$$p\left\{Min\left(Q_{pk_1i},Q_{pk_2i}\right)\le\left(Min\left(\hat{Q}_{pk_1i},\hat{Q}_{pk_2i}\right)-1.5\right)\sqrt{\frac{\chi^2_{\alpha'';n-1}}{(n-1)}}+\frac{z_{\alpha''}}{\sqrt{n}}+1.5\right\}\ge1-\alpha.$$

Thus, the $(1-\alpha)\times100\%$ upper confidence limit of Q_{pki} is

$$UQ_{pki}=\left(\hat{Q}_{pki}-1.5\right)\sqrt{\frac{\chi^2_{\alpha'';n-1}}{(n-1)}}+\frac{z_{\alpha''}}{\sqrt{n}}+1.5. \quad (11)$$

By Eq. (9), Eq. (10) and Eq. (11), then the upper confidence limit of Q_{phi} can be shown as follows:

$$UQ_{phi}=\begin{cases}\left(\hat{Q}_{phi}-1.5\right)\sqrt{\frac{\chi^2_{1-\sqrt{1-\alpha};n-1}}{n-1}}+\frac{z_{1-\sqrt{1-\alpha}}}{\sqrt{n}}+1.5, h=l,u\\\left(\hat{Q}_{phi}-1.5\right)\sqrt{\frac{\chi^2_{1-\sqrt{1-\alpha/2};n-1}}{(n-1)}}+\frac{z_{1-\sqrt{1-\alpha/2}}}{\sqrt{n}}+1.5, h=k\end{cases}. \quad (12)$$

IV. The Construction and Case Study of the Supplier Evaluation Radar Chart

When a product has multiple quality characteristics, firstly, statistical tests are conducted based on the upper confidence limits of Six Sigma evaluation indices proposed by this paper. Nevertheless, observing from the test results of all quality characteristics to see whether each of them can meet the quality requirement, it is difficult to view the whole picture of the product. Hence, this paper subsequently presents all the test evaluation information of the quality characteristic process, so that the manager can obtain more complete information to carry out the supplier evaluation. According to Chang et al. [25], the radar chart is seen as a visualized 2D plane whose part is divided into multiple visual angles to show multiple objects. The same visual angle represents the same series and has one equiangular axis. The data points adjacent to the axis are linked to form a characteristic polygon. Observing the change of the relative distance axis in the polygon, the key point for each item is induced. In addition, the radar chart, applied to the field of engineering or education, able to implement quantative qualitative integrated data, and able to superimpose graphs of more than two systems for comparison, is regarded as a practical chart analysis, highly readable and widely applicable. For example, Tanimoto et al. [26] applied it to the pressure measurement for the spine-injured patient; Harding et al. [27] applied it to evaluate the performance of blended learning undertaken in the university courses; Chang et al. [25] adopted it to investigate the frequency of the multimedia communication system used by the internet users. All of the above are related studies assisted and analyzed by the radar chart employed in a wide range of fields with remarkable outcomes.

Consequently, this study proposes a supplier evaluation radar model. First, the upper confidence limits of the Six Sigma Quality Indices are used to examine whether each important quality characteristic process meet the requirement of quality level. Next, based on the product quality level (k) in Eq. (7), the corresponding required value k' for the

important quality characteristic index is discovered. To make the management convenient, this paper defines the judgment function of each quality characteristic for suppliers as follows:

$$J_{Shi}=\frac{U_S Q_{phi}}{k'} \tag{13}$$

When $U_S Q_{phi}<k'$, it can be concluded that $Q_{phi}<k'$ while $J_{shi}<1$, which means the quality level of this quality characteristic has not reached the requirement yet. Besides, when $U_S Q_{phi}\geq k'$, the minimum value of Q_{phi} is k', that is $Q_{phi}\geq k'$ while $J_{hi}\geq 1$, which shows that the quality level of this quality characteristic has already reached the requirement. Next, an evaluation can be conducted as long as the judgment function value J_{shi} for each quality characteristic is computed and drawn in the radar chart. This paper takes the Taiwanese manufacturer of the HID headlights as an example to explain its application. The product has 1 quality characteristic of the LTB type (the effective brightness must be larger than 37lm/w), 1 quality characteristic of the STB type (the electricity consumption must be smaller than40w), and 3 quality characteristics of the NTB type (light focus: 5 ± 1mm, color temperature: $300\pm$ 100k, and watt power: 40 $\pm$ 3mm, respectively). The subsequent evaluation steps are as follows:

1. If the required product quality level reaches 5-Sigma ($k=5$) and $(n=1+1+3\times2=8)$, then based on Eq. (7), it can be computed that the level of each important quality characteristic must reach 5.520-Sigma (k'=5.520).

2. According to the sampling data, the mean of samples, standard deviation, and the upper confidence limits of indices for each important quality characteristic of suppliers are as follows:

$$\bar{X}_{Shi}=\frac{1}{n}\sum_{j=1}^{n}X_{Shij},\ S^2_{Shi}=\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{Shij}-\bar{X}_{Shi}\right)^2,$$

$$U_S Q_{pli}=\left({}_S\hat{Q}_{pli}-1.5\right)\sqrt{\frac{\chi^2_{1-\sqrt{1-\alpha};n-1}}{n-1}}+\frac{z_{1-\sqrt{1-\alpha}}}{\sqrt{n}}+1.5,$$

$$U_S Q_{pui}=\left({}_S\hat{Q}_{pui}-1.5\right)\sqrt{\frac{\chi^2_{1-\sqrt{1-\alpha};n-1}}{n-1}}+\frac{z_{1-\sqrt{1-\alpha}}}{\sqrt{n}}+1.5,$$

$$U_S Q_{pki}=\left({}_S\hat{Q}_{pki}-1.5\right)\sqrt{\frac{\chi^2_{1-\sqrt{1-\alpha/2};n-1}}{(n-1)}}+\frac{z_{1-\sqrt{1-\alpha/2}}}{\sqrt{n}}+1.5.$$

3. Compute the judgment function for each important quality characteristic.

Let $J_{Shj}=\frac{U_S Q_{phi}}{k'}$, then

(1) When $U_S Q_{phi}<k'$, it can be concluded that $Q_{phi}<k'$ while $J_{hi}<1$.

(2) When $U_S Q_{phi}\geq k'$, the minimum value of Q_{phi} is k', that is $Q_{phi}\geq k'$ while $J_{hi}\geq 1$.

According to the previous statements, the quality level for each important quality characteristic of the HID headlight in this case is displayed in Table 1.

Table I. Quality Level for Each Important Quality Characteristic of the HID Headlight

	Effective brightness	Electricity consumption	Light focus	Color temperature	Watt power
Qualified standard	1.00	1.00	1.00	1.00	1.00
Real process quality	1.50	1.10	0.95	1.20	1.15

4. The judgment function for each important quality characteristic is drawn in the supplier radar evaluation chart, as shown in Fig. 1. When the values of J_{Shj} for all quality characteristics are larger than or equal to 1, then this supplier should be selected. If the value of J_{Shj} for one quality characteristic of a supplier is smaller than 1, the supplier should be informed to improve the quality characteristics before the follow-up re-evaluation.

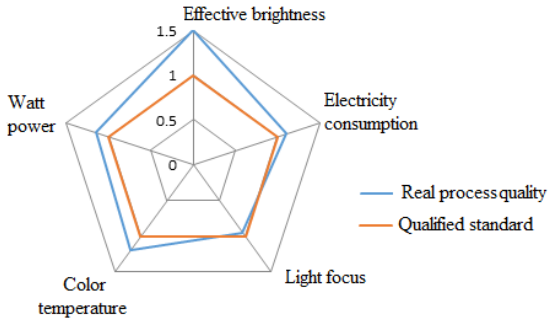


Fig. 1. HID Headlight Supplier Evaluation Radar Chart

V. Conclusions

A good supplier selection mechanism plays a critical role in contributing to the enterprise's mature development, so it is a future trend of business operation models, in which quality is one of the most important criteria. Therefore, we first proposed the Six Sigma Quality Indices which can reflect the process yields and quality levels for all the important quality characteristics of a product. We also proposed a relationship for the entire quality level of the product corresponding to the quality levels of important quality characteristics as an important standard for supplier selections.

First, the upper confidence limits for the Six Sigma Quality Indices of STB type, LTB type, and NTB type were concluded, and statistical inferences were strictly conducted based on these upper confidence limits, to examine whether the process level for each important quality characteristic can meet the requirement of quality level. In order to make management convenient, this paper constructed the judgment function based on the upper confidence limits as well as the required value (k') for the quality level of each quality characteristic and aggregated the judgment functions of all

quality characteristics on the supplier selection radar chart, so that the enterprise can clearly see the whole picture for all important quality characteristics of the supplier's product based on the supplier selection radar chart. By doing so, when the manufacture conducts a supplier selection, there will be better synthesized judgment effects. Meanwhile, the manufacture also provides the supplier with the processes of important quality characteristics against insufficient quality and feedback on the quality improvement, so that they can grow together and form a closer partnership with each other. Last, this study took the high intensity discharge (HID) headlight manufactured in Taiwan as an example to demonstrate the application of the supplier selection model proposed by this paper.

To sum up, the method proposed by this paper not only has strict statistical inferences but also presents the whole picture for all important quality characteristics of the product by means of the supplier selection radar chart. This method has good and convenient management efficacy. Meanwhile, it can also help suppliers grow together and maintain long-term partnerships with manufactures. In practice, this model can help the industrial chain develop and urge the promotion of quality.

References

[1] G. W. Dickson. "An analysis of vendor selection systems and decisions", Journal of Purchasing, vol. 2, no.1, pp.5-17, 1966.

[2] C. A. Weber, J. R. Current, and W.C. Benton. "Vendor selection criteria and methods", European Journal of Operational Research, vol. 50, no.1, pp. 2-18, January 1991.

[3] W. E. Patton. "Use of Human Judgment Models in Industrial Buyers' Vendor selection Decisions", Industrial Marketing Management, vol. 25, no.2, pp. 135-149, March 1996.

[4] R. Verma and M. E. Pullman. "An Analysis of the Supplier Selection Process", Omega-International Journal of Management Science, vol. 26, no. 6, pp. 739-750, December 1998.

[5] K. L. Chen, K. S. Chen, and R. K. Li. "Suppliers capability and price analysis chart", International Journal of Production Economics, vol. 98, no. 3, pp. 315-327, December 2005.

[6] K. S. Chen and K. L. Chen. "Supplier selection by testing the process incapability index", International Journal of Production Research, vol. 44, no. 3, pp. 589-600, 2006.

[7] C. W. Wu, M. Y. Liao, and T. T. Yang. "Efficient methods for comparing two process yields – strategies on supplier selection" International Journal of Production Research, vol. 51, no. 5, pp. 1587-1602, 2013.

[8] W. L. Pearn, C. H. Wu, C. C. Chuang. "An effective powerful test for one-sided supplier selection problem with multiple independent characteristics", Quality Technology and Quantitative Management, vol. 13, no. 2, pp. 182-196, 2016.

[9] H. Y. Kang, A. H. I. Lee, C. W. Wu, and C. H. Lee. "An efficient method for dynamic-demand joint replenishment problem with multiple suppliers and multiple vehicles", International Journal of Production Research, vol. 55, no. 4, pp.1065-1084, 2017.

[10] J. P. Chen and K.S. Chen. "Comparing the capability of two processes using Cpm", Journal of Quality Technology, vol. 36, no. 3, pp. 329-335, 2004.

[11] C. J. Lin, and W. L. Pearn. "Process selection for higher production yield based on capability index S_{pk}", Quality and Reliability Engineering International, vol. 26, no. 3, pp. 247-258, April 2010.

[12] M. Y. Liao, C. W. Wu, and J. W. Wu. "Fuzzy inference to supplier evaluation and selection based on quality index: A flexible approach", Neural Computing and Applications, vol. 23, pp. 117-127, December 2013.

[13] C. W. Wu, M. Y. Liao, and C. Y. Lin. "On ranking multiple touch-screen panel suppliers through the CTQ: Applied fuzzy techniques for inspection with unavoidable measurement errors", Neural Computing and Applications, vol. 25, no. 2, pp. 481-490, August 2014.

[14] C. W. Wu, M. Y. Liao, C. W. Lin, T. L. Lin. "Testing and ranking multiple wafer manufacturing processes with fuzzy quality data", Journal of Testing and Evaluation, vol. 44, no. 5, pp.1970-1977, September 2016.

[15] K. S. Chen, L.Y. Ouyang, C. H. Hsu, and C. C. Wu. "The communion bridge to Six Sigma and process capability indices", Quality and Quantity, vol. 43, no. 3, pp. 463-469, May 2009.

[16] C. T. Huang, K. S. Chen, and T. C. Chang. "An application of DMADV methodology for increasing the yield rate of surveillance cameras", Microelectronics Reliability, vol. 50, no. 2, pp. 266-272, February 2010.

[17] K. S. Chen, K. J. Wang, and T. C. Chang. "A novel approach to deriving the lower confidence limit of indices C_{pu}, C_{pl}, and C_{pk} in assessing process capability", International Journal of Production Research, pp. 1-19, 2017, in press.

[18] K. S. Chen, H. T. Chen, and T. C. Chang. "The construction and application of Six Sigma quality indices", International Journal of Production Research, vol. 55, No. 8, pp. 2365-2384, 2017.

[19] H. T. Chen and K. S. Chen. "Advanced multi-process performance analysis chart for an entire product with joint confidence regions", International Journal of Production Research, vol. 45, no. 9, pp. 2141-2159, 2007.

[20] W. L. Pearn, and Y. C. Cheng. "Measuring production yield for processes with multiple characteristics", International Journal of Production Research, vol. 48, no. 15, pp. 4519-4536, 2010.

[21] T. C. Chang, K. J. Wang, and K. S. Chen. "Sputtering process assessment of ITO film for multiple quality characteristics with one-sided and two-sided specifications", Journal of Testing and Evaluation, vol. 42, no. 1, pp. 196-203, January 2014.

[22] K. S. Chen, C. H. Wang, and H. T. Chen. "A MAIC approach to TFT-LCD panel quality improvement", Microelectronics Reliability, Vol. 46, No. 7, pp. 1189-1198, July 2006.

[23] F. W. Breyfogle, J. M. Cupello, and B. Meadows. Managing Six Sigma: A practical guide to understanding, assessing, and implementing the strategy that yields bottom-line success. New York: Wiley, 2001.

[24] K. Linderman, R. G. Schroeder, S. Zaheer, and A. S. Choo. "Six Sigma: a goal-theoretic perspective", Journal of Operations Management, vol. 21, no. 2, pp. 193-203, March 2003.

[25] Y. C. Chang, C. J. Chang, K. T. Chen, and C. L. Lei. "Radar Chart: Scanning for High QoE in QoS Dimensions", IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR) Canada, pp.1-6, 2010.

[26] Y. Tanimoto, H. Takechi, H. Nagahata, and H. Yamamoto. "Pressure measurement of air cushions for patients", IEEE Instrumentation and Measurement Technology Conference (IMTC) Italy, vol. 2, pp. 1245-1250, 1999.

[27] A. Harding, D. Kaczynski, and L. Wood. "Evaluation of blended learning: Analysis of qualitative data", Evaluation of Blended Learning: Analysis of Qualitative Data, pp. 56-62, January 2005.

Sustainable supplier monitoring problem: a fuzzy logic approach

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Abdelhak Elidrissi
Ecole Mohammadia d'Ingénieurs
Rabat, Morocco
abdelhakelidrissi@research.emi.ac.ma

Sophie Hennequin
Université de Lorraine / ENIM - LGIPM
Metz, France
sophie.hennequin@univ-lorraine.fr

Abstract—This paper considers a supplier selection problem which takes into consideration the sustainability trading scheme strategy. The considered problem is classic since a single-echelon supply chain, composed by several suppliers, a warehouse and a costumer, is studied. Our main contribution consists in evaluating the internalization of costs resulting from environmental and social impacts with the help of fuzzy logic. The objective is then the minimization of the total cost function which corresponds to the sum of purchasing, ordering, holding, transportation, environmental and social costs.

Keywords—multi-sourcing; supplier selection; sustainable development; nonlinear programming; fuzzy logic.

I. Introduction

The supplier selection problem is largely studied in the literature since it is one of the most competitive operational tasks and since it implies critical activities in the purchasing process [1]. The researches take two main directions. The first one concerns the definition of quantitative and / or qualitative criteria to select the best suppliers. Methods like analytical hierarchy process [2], grey relational analysis [3], data envelopment analysis [4], etc. could be developed. However, generally three main criteria dominate: price, quality and delivery [5]. The second direction of works is made for supporting decision-making and could be solved by: mathematical modelling, mathematical programing, multi-objective programing, mix-integer programing, multi-criteria decision-making methods, data envelopment analysis, etc. All these methods could be combined with artificial intelligence methods. A very good literature review could be found in [6]. In this paper, we only consider the assignment and monitoring of preselected suppliers based on economic costs.

In parallel, increasing public awareness, governmental regulations and market pressure have motivated corporates to integrate sustainability initiatives in the supply chain management to improve their performances and to retain their competitive edges [7]. Indeed, it is necessary to change the paradigm and to rethink our industries so, firstly to couple industrial efficiency with the social and environmental impacts and, secondly to integrate global legislative constraints and norms [8]. Many researches have been developed to define green and sustainable supply chains [9]. However, the sustainability debate mainly focuses on minimizing costs while reducing the environmental impacts. Good survey could be found in [10]. The social pillar of sustainability has been argued in terms of being a cause for, or possibly a solution to, environmental impacts, and not as something that deserves attention as a sustainability component in its own right [11]. The social pillar is now recognized as playing a major role in the overall performances of sustainable supply chains [12]. However, selecting a supplier from an environmental point of view (e.g. through avoiding hazardous materials, limiting greenhouse gases emission and considering alternative solutions that require less materials and/or energy) and from a social point of view (i.e. creates more better-paying job, improves the well-being and skills of employees to cite few examples) is somehow difficult. A firm's environmental and social efforts will not succeed without integrating the company's environmental and social goals with its purchasing activities and the willingness to pay (i.e. the firm consents to pay the price of various non-market benefits).

This work concerns this specific problem of internalizing the negative impacts of environmental and social pillars and their induced costs into the strategy of the firm. We will consider for each entity of the supply chain an environmental cost (carbon tax for greenhouse gases for example) and a social cost (which could be equal to the mean cost of work accident, or costs due to congestion, etc.). However, the integration of all negative impacts and induced costs is anywise difficult because of the complexity of involved phenomena. That is the reason why we also consider two penalties, one for each pillar, to limit as much as possible environmental and social impacts. These penalties, defined by the firm based on estimated impact of the suppliers, allow the firm to have a higher level of decision-making and also to integrate more uncertainties [13]. These penalties after calculation are added to the total cost. To solve this problem, we propose to consider the contribution of fuzzy logic [14] (since it is extremely difficult to precisely identify and quantify all factors impacting the environment and society) in the assignment and management of suppliers. Indeed fuzzy penalties allow enterprises to integrate additional constraints that may not be perfectly identified or identifiable vis-à-vis suppliers (from a social and environmental point of view).These fuzzy penalties also allow enterprises to have proactive comings in terms of regulation and social responsibility. It has to be noticed that in this work, the fuzzy logic will only be used to quantify the value of social (environmental) penalties in terms of the corresponding social

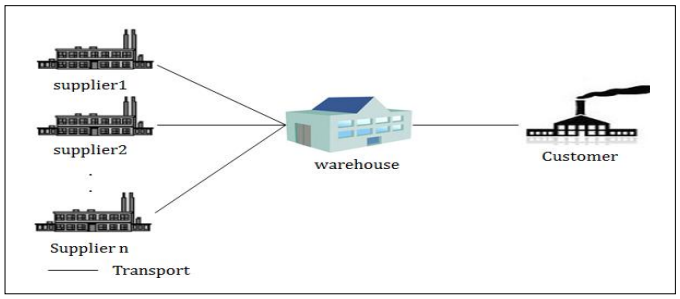
(environmental) impacts for each supplier and not to solve the supplier selection problem. Our basic idea comes from, for example, by exposure to a source of nuclear radiation: we are directly and seriously impacted by the radiation if it is very high, but also according to the duration of exposure if the radiation has a lower value. Thus, these two penalties are fixed costs once calculated. The calculation is done by fuzzy logic with three input variables: the importance of the call to a supplier (quantity and duration) and the environmental or social impacts according to the corresponding penalty.

This rest of paper is organized as follows. Section II presents the considered sustainable single-echelon supply chain. In section III, the mathematical model is detailed. The proposed algorithm is given and explained in section IV where a discussion is also proposed based on some numerical results. We conclude the paper and give some perspectives to our work in section V.

II. Single-echelon supply chain

We consider a supply chain network consisting of n suppliers, a warehouse, and a customer as illustrated in Fig. 1. The customer's request is assumed to be constant and known and its annual demand is denoted by D. It is assumed that the finite time horizon T, a year, is divided into several constant periods and the length of each period is denoted by T_i. The warehouse buyer must make a choice among the suppliers on the basis of a total cost criterion. The objective is then the minimization of this total cost function which includes ordering, purchasing, inventory and transportation costs but also environmental and social costs, subject to different constraints like the production capacity of each supplier and customer's delivery time (as described below).

Fig. 1. Supplier selection.



The ordered quantity for all suppliers at period T_i is equal to Q_i. If the supplier i supplies the customer at a period T_i then a binary variable is defined. It is equal to $Y_i = 1$ (0 otherwise). Y_i represents the fact that a supplier i is selected or not for the given period T_i. The percentage of the ordered quantity Q_i at T_i for the supplier i is equal to X_i. The ordering cost per order for the supplier i is given by Co_i, the unit purchasing cost for the supplier i is given by Cp_i. In this paper, we suppose that all costs are known and constant on the time horizon.

As the annual demand D is constant and X_i is the percent of Q_i at T_i for the supplier i, we obtain the following equations:

$$Q = \sum_{i=1}^{n} Q_i \tag{1}$$

$$Q_i = X_i . Q, \qquad i = 1, \dots, n \tag{2}$$

$$T_i = X_i . T, \qquad i = 1, \dots, n \tag{3}$$

$$0 \le X_i \le 1 \qquad i = 1, \dots, n \tag{4}$$

$$\sum_{i=1}^{n} X_i = 1 \tag{5}$$

$$Y_i = \begin{cases} 0 & \text{if } X_i = 0 \\ 1 & \text{if } X_i > 0 \end{cases} \qquad i = 1, \dots, n \tag{6}$$

The customer wants to have its ordered products with a maximal given delivery time denoted L. The replenishment delivery time between the supplier i and the warehouse is denoted Lr_i and the delivery time from the warehouse and the customer is given by Lw. The production capacity for each supplier i is denoted Lp_i. The holding rate by time unit in the warehouse is given by r_w and the holding rate by time unit for the customer is given by r_c. The distance between the ith supplier and the warehouse is given by d_{Siw} and the distance between the warehouse and the customer is given by d_{wc}. Different modes of transport, denoted m, could be used (road transport, air transport, etc.). The transportation costs are defined in function of the chosen transport mode, the transported tons (assimilated in this paper as transported quantities) and the distance. We denote by $Ct_i(m, Q_i, d_{Siw})$ the transportation cost between the ith supplier and the warehouse, and by $Ctw(m, Q, d_{wc})$, the transportation cost between the warehouse and the customer. Each mode of transport has a given limited capacity $Tc(m)$. In this work, we do not consider the loading / unloading process since we have only one warehouse. Generally for a known path, the transportation cost functions could be assimilated to linear functions with a fixed transportation cost, denoted Ctf_j^m, and a variable transportation cost, denoted Ctv_j^m, were j could be equal to i if we consider the transportation cost from the supplier i to the warehouse or equal to w if we have a transport between the warehouse and the customer. The fixed transportation cost corresponds, for example, to the depreciation of vehicles, wages, etc. and the variable cost to fuel consumed, maintenance activities, etc.

To take into account environmental impacts, we also define a pollution cost for the transport mode which includes greenhouse gases, fines particles, and other pollutants (like nitrogen oxides). It is given by $Ctp_i(m, Q_i, d_{Siw})$, a unitary pollution transportation cost, between the ith supplier and the warehouse, and by $Ctpw(m, Q_i, d_{wc})$, the transportation cost between the warehouse and the customer. For the suppliers, we define an environmental production cost (due to greenhouse gases –GHG- emission linked with energy consumption and different supplier's pollutant activities, waste, etc.), denoted by Cpp_i for the ith supplier. We suppose that this environmental production cost is a unitary cost per product.

To take into account social impacts, we also define accidents, traffic congestion and noise / disturbance costs for the transport mode. The sum of these costs is given by $Cts_i(m, Q_i, d_{siw})$, a unitary social transportation cost, between the ith supplier and the warehouse, and by $Ctsw(m, Q_i, d_{wc})$, the unitary social transportation cost between the warehouse and the customer. For the suppliers, we also define a social production cost (which may integrate the human development index of the supplier's production area – good salary, well-being at work, interesting vocational training, etc.), denoted by Csp_i for the ith supplier. We suppose that this social production

cost could be assimilated to a unitary cost per product (see [15]).

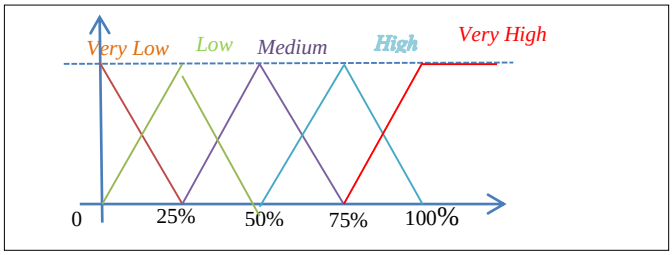
It has to be noticed that in this paper, since we only have one warehouse, we include the environmental and social impacts directly into the holding cost. For the warehouse, the holding cost per unit is given by *Cw*. The capacity of the warehouse is assumed to be unlimited (since the warehouse buyer may possibly call upon other warehouses if necessary but with different distances and holding costs, this case is not dealt in this work. Furthermore, we consider that the quantities Q_i of each supplier *i* are grouped in the warehouse and delivered in a single amount *Q* to the customer. The holding rate for the customer is given by r_c and the holding rate for the suppliers' products is given by r_w.

The warehouse buyer would like to choose the best suppliers from among the *n* ones taking into account their limited capacity but also a predefined environmental threshold, (for example a maximal GHG emission), and a predefined social threshold, which corresponds to a maximal social impact. These both predefined thresholds give rise to penalties if they are exceeded. The quantification of these penalties will be detailed in the following as well as the proposed nonlinear programming.

III. Model formulation

To define our both penalties, we chose to define two fuzzy systems based on the Mamdani fuzzy model [14] since it is well adapted to vague notions. These fuzzy models could be easily modified from real data. Each fuzzy model is composed by three input variables and an output variable. The input variables are defined as a percentage of predefined values (defined by the firm) which correspond to: an environmental impact value (named ie_i) (or respectively a social impact value named is_i), the sum of all requested products for the given supplier (named sum_i) and the duration (named $duration_i$) of the call to this supplier. For each fuzzy input variable, we define five fuzzy sets based on triangular and trapezoidal functions, denoted "*Very Low*", "*Low*", "*Medium*", "*High*" and "*Very High*" (see Fig 2). Since these fuzzy input variables (F_j with $j = 1$ to 3) are based on percentages, the corresponding universes of discourse are given by: $UF_j = \{U|0 < l < 100\%\}$ where F_j represents the considered fuzzy input variable.

Fig. 2. Fuzzy sets for input variables.



For the output variables, the environmental and the social penalties, we also define 5 fuzzy sets given by triangular and trapezoidal functions, denoted "*Very Low*", "*Low*", "*Medium*", "*High*" and "*Very High*" (same kind of fuzzy sets as the ones defined in Fig 2 to be sure to discover all the universe of discourse). Then, the fuzzy rules are given, for example for the environmental penalty denoted pe_i (same rules for the social penalty denoted ps_i) and for the supplier *i*, by the following rules denoted by R_k and R_{k+1}:

R^i_k: If sum_i of all requested products for the given supplier is "*Very Low*" and ie_i is "*Very Low*" and $duration_i$ is "*Very Low*" Then pe_i is "*Very Low*",

R^i_{k+1}: If sum_i of all requested products for the given supplier is "*Very High*" and ie_i is "*Very High*" and $duration_i$ is "*Very High*" Then pe_i is "*Very High*".

The well-known Min / Max method is adopted and the simple centroïd method applied for the defuzzification [14].

The objective is then to minimize the total cost function which is the sum of ordering, transportation, purchasing, holding, environmental and social costs. We have assumed that the quantities of each supplier are grouped into the warehouse and delivered in a single order to the customer. Then, the cost function is given by:

$$C_{total} = \frac{D}{Q}.\sum_{i=1}^{n} Co_i + \frac{D}{Q}.\left[\sum_{i=1}^{n} Cti(m,Qi,dsiw) + Ctw(m,Q,dwc)\right] + \sum_{i=1}^{n} X_i Cp_i D + Q.\sum_{i=1}^{n}\frac{X_i^2 Cp_i}{2}(r_w + r_c) + \frac{D}{Q}.\left[\sum_{i=1}^{n}(Cppi.Qi + Ctpi(m,Qi,dsiw) + pei) + Ctpw(m,Q,dwc)\right] + \frac{D}{Q}.\left[\sum_{i=1}^{n}(Cpsi.Qi + Ctsi(m,Qi,dsiw) + psi) + Ctsw(m,Q,dwc)\right] \quad (7)$$

This cost function is subject to several constraints such as the suppliers' capacity, the value of the demand, the delivery time expected by the customer and the capacity of each chosen vehicle. These constraints are given by the following equations:

$$X_i.D \le Lp_i \quad , i = 1,\dots,n \quad (8)$$

$$\sum_{i=1}^{n} X_i.D = D \text{ or } \sum_{i=1}^{n} X_i = 1 \quad (9)$$

$$\max_i Lri + l_w \le L \quad (10)$$

$$X_i.D \le Tc(m), \quad i = 1,\dots,n \quad (11)$$

Equation (8) allows to consider the limited production capacity of suppliers, equation (10) allows to respect the delivery time, equation (11) concerns the limited capacity of each mode of transport *m*.

To simplify and since we suppose that each environmental and social impact could be define in function of quantity of products, the transportation, environmental and social costs could be expressed as:

$$\sum_{i=1}^{n}[Cti(m,Qi,dsiw)] + Ctw(m,Q,dwc) = \sum_{i=1}^{n}[\,(Ct^{m}_{v_iw}.d_{S_iW}.Qi) + Ct^{m}_{f_iw}] + Ct^{m}_{vwc}.d_{wc}.Q + Ct^{m}_{fwc} \quad (12)$$

$$\sum_{i=1}^{n}[Cppi.Qi + Ctpi(m,Qi,dsiw) + pei] + Ctpw(m,Q,dwc) = Qi.\sum_{i=1}^{n}[\,Cenvi + Ct^{m}_{envS_iw}.d_{S_iW}] + Ct^{m}_{envwc}.d_{Wc}.Q + Cenv \quad (13)$$

$$\sum_{i=1}^{n}[Cspi.Qi + Ctsi(m,Qi,dsiw) + psi] + Ctsw(m,Q,dwc) = Qi.\sum_{i=1}^{n}[\,Csoci + Ct^{m}_{socS_iw}.d_{S_iW}] + Ct^{m}_{socwc}.d_{Wc}.Q + Csoc \quad (14)$$

where *Cenvi* represents the sum for a supplier *i* of all environmental costs for this supplier, Ct^{m}_{envSiw} represents the environmental cost between the supplier *i* and the warehouse and Ct^{m}_{envwc} represents the environmental cost between the warehouse and the customer and *Cenv* represents a fixed environmental cost (the sum of all penalties), *Csoci* represents the sum for a supplier *i* of all social costs for this supplier, Ct^{m}_{socSiw} represents the social cost between the supplier *i* and the warehouse and Ct^{m}_{socwc} represents the social cost between the warehouse and the customer and *Csoc* represents a fixed social cost (the sum of all penalties).

Then, after simplification (see equations (1) to (6) and equations (12) to (14)) and by substituting the values of Y_i in the integer programming, the final integer nonlinear programming model becomes a pure nonlinear programming. If the set *{Y}* is defined as the set of Y_i for which their values are equal to one, the pure nonlinear programming becomes:

$$Min\ C_{total} = \sum_{i=1}^{n} X_i.D.Cp_i + 2.\sqrt{D.\sum_{i=1}^{n}\frac{X_i^2.Cp_i}{2}(r_w + r_c)}.\sqrt{\left[\sum_{i=1}^{n} Co_i + Ct^{m}_{v_iW}.d_{S_iW}\right] + \sum_{i=1}^{n} Ct^{m}_{f_iW} + Ct^{m}_{vwC}.d_{wc} + Ct^{m}_{fwC}} + \sum_{i=1}^{n}[C_{env_i} + Ct^{m}_{envS_iW}.d_{S_iW}].X_i.D + Ct^{m}_{env\ \ wc}.d_{wc} + C_{env} + \sum_{i=1}^{n}[C_{soc_i} + Ct^{m}_{soc_{s_iw}}.d_{S_iW}].X_i.D + Ct^{m}_{soc_{wc}}.d_{wc} + C_{soc} \quad (15)$$

Subject to:

$$\begin{cases} X_i.D \le Lp_i\ , \quad i = 1,\dots,n \\ X_i.D \le Tc(m)\ , \quad i = 1,\dots,n \\ X_i \ge 0 \quad i = 1,\dots,n \\ \sum_{i=1}^{n} X_i = 1 \\ \max_i l_{r_i} + l_w \le L \end{cases} \quad (16)$$

Theorem 1: the objective function given by (15) subject to (16) is convex.

Proof of Theorem 1: To show that the objective function is convex, it is necessary to show that the Hessian matrix of the total cost is positive semi-definite (i.e. all diagonal elements are non-negative and all the principal determinants are non-negative [16]).

Since the ordering, transportation, environmental and social costs are constant (indeed, the penalties after calculation by the fuzzy system are constant and could not exceed a given maximal value) and (r_c+r_w) are constant, we only have to prove that the following equation (17) is convex:

$$\sqrt{\sum_{i=1}^{n} X_i^2.Cp_i} = T \quad (17)$$

Then, it can be stated that f=√T. The Hessian matrix of this function could be given by:

$$H(X) = \frac{Cp_1Cp_2\dots Cp_nX_n^n}{T^{3/2}}\begin{pmatrix} \frac{T}{X_n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ -\frac{Cp_1X_1^2}{X_n} & \cdots & 0 \end{pmatrix} \quad (18)$$

The determinant of this matrix and its *k*th order principal matrix are non-negative, thus the objective function is convex.

Q.E.D

This result will be used in an optimization algorithm to highlight our work in what follows.

IV. Proposed algorithm

To highlight our study, we use the data addressed in [16] and compare with our obtained results. We also consider three suppliers and the data are as follows (see also Tables 1 and 2).

Table 1. Suppliers data.

	Purchasing cost (MU)	Ordering cost (MU)	delivery time (week)	Capacity (parts)
Supplier 1	9	9	2	600
Supplier 2	16	4	1	700
Supplier 3	32	8	3	500

The holding rate in the warehouse is equal to 40%/year. The distances between suppliers and the warehouse and other suppliers' costs are given in Table 2. We suppose, in this paper, that the transportation is made by trucks (road transport); their capacity is equal to 40 tons. Air pollution, noise emissions, accidents, etc. during the transport generate social and environmental costs. They are chosen to be equal to: 0.9

MU.ton.km for the environmental impacts and 1.1 MU.ton.km for the social impacts (see [17] and [18]). The maximal values for the penalties are given by: 15MU (environmental) and 27MU (social). We decide in this work to define a priority to social impact since it is generally less considered. The distance between the warehouse and the customer is equal to 200 km.

Our proposed algorithm is detailed in Fig. 3.

Fig. 3. Algorithm.

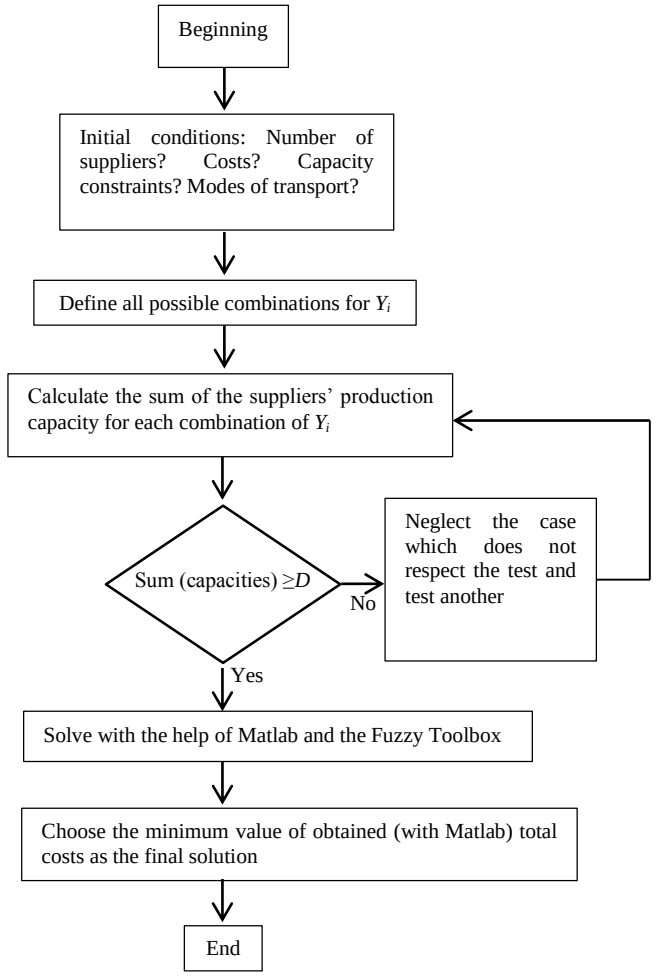


Table 2. Suppliers transportation, env. and social data.

	Fixed trans. cost (MU)	Var. trans. cost(MU)	Dist. (km)	Env. cost (MU)	Soc. cost (MU)
Supplier 1	0.480	0.157	200	6.62	2.3
Supplier 2	0.662	0.098	100	3.45	4.7
Supplier 3	0.586	0.234	150	4.5	3.9

The nonlinear programming is solved with the help of Matlab and the fuzzy toolbox (to calculate the penalties), and the results are given in Table 3. The presented cases correspond to all feasible cases. The best one is case 2 with an optimal quantity to be ordered equals to 486 parts and a minimum total cost given by 16103 MU.

Table 3. Results for 3 suppliers and transport by trucks.

Case	X_1	X_2	X_3	Q	Eco (MU)	Env (MU)	Soc (MU)
1	0.05	0.5	0.45	456	12450	2157	6623
2	0.2	0.8	0	487	9651	1863	4589
3	0.3	0	0.7	436	15230	2334	7210
4	0	0.6	0.4	464	11985	1969	5320

We can remark that if the firm wants to be economically efficient it should work with 20% of ordered products to supplier 1 and 80% of ordered products to supplier 2. These results are due, for an important part, to environmental and social costs. Indeed, if we compare with results obtained in [16], we can remark that the inscription of a supply chain into a sustainable framework strongly impacts the decision making.

Furthermore, with our chosen simulation data, it is more interesting for the firm to have two suppliers than one since if this supplier has negative impacts from a social and environmental point of view, it would be interesting for the firm to reduce this resulting impact by ordering parts to other suppliers. This basic idea is based on the policies currently being developed by governments to improve sustainable performances. In our case, if we only consider one supplier (with data given for supplier 1), the optimal quantity to be ordered is equal to 256 parts with a corresponding total cost equals to 11547 MU and 23% of this total cost are due to environmental and social costs. It has to be noticed that many important criteria are not considered in this work (such as quality, customer satisfaction, etc.). These criteria could also integrate environmental (like GHG emissions) and social (like well-being of inhabitants of the country where the suppliers' production sites are located) impacts.

V. Conclusion

In this paper, we propose a nonlinear programming to solve a supplier monitoring problem taking into account, in addition with traditional costs, the environmental and social pillars of sustainable development. In our work, these both pillars are considered also as economical costs and integrated to our model. The objective of this work is not to clearly identify and quantify all impacts that interact on sustainable performances but the internalization of these impacts as supplementary costs. To consider these two supplementary costs, we define firstly, based on other research works, environmental and social costs for each element of the single-echelon supply chain and secondly, we define with the help of fuzzy logic a penalty for each of these two pillars. These supplementary penalties permit to limit the use of suppliers that are not responsible corporate citizen. After defining the total cost function to minimize which corresponds to the sum of ordering, purchasing, holding, transportation, environmental and social costs (which are for examples due to accidents, pollution, layoffs of employees, job creation, emission of CO2 gas from transport and production

activities etc.) subject to limited capacity of transportation modes, limited production capacity of suppliers and the expected delivery time, we prove the objective function is convex. Based on this result, we propose an algorithm and implement it with Matlab to define by simulation the quantities to be ordered to suppliers in order to achieve sustainable performance and minimize costs. The obtained results are compared with other works and the impacts of environmental and social pillars are highlighted. However, the obtained results clearly depend on simulation parameters so an identification of all impacts and costs is necessary to help decisions made by managers.

In future works, we will consider other environmental and social impacts such as visual pollutions, ecological rucksack, wastes, use of space and ecological footprint, the impacts on biodiversity, other criteria of societal development in the neighborhood of suppliers, etc. For this, as the main difficulty is to know and quantify all impacted elements, fuzzy logic could also be of great interest by allowing to integrate more uncertainties (and also to consider more simply the precautionary principle). Furthermore, it would be interesting to place our study in a circular economy framework and trying to integrate all returns loops when it is possible. It will be also necessary to generalize our problem to variables demands and different customers and to consider more complex supply chains with much more other objectives than only economical cost to minimize to obtain more efficient and effective decisions for the managers.

References

[1] Seuring S. and Müller M. (2008). "From a litterature review to a conceptual framework to sustainable supplier management", Journal of Cleaner Production, 16(15), p. 1699-1710.

[2] W. Ho, X. Xu, P.K. Dey (2010). "Multi-criteria decision making approaches for supplier evaluation and selection: A literature review", European Journal of Operational Research, 202 (1), p. 16–24.

[3] Li G., Yamaguchi D., Nagai M. (2007). "A grey-based decision-making approach to the supplier selection problem", Math Comput Model 46(3–4), p. 573–581.

[4] Ping Shi, Bo Yan, Song Shi and Chenxu Ke, (2015). "A decision support system to select suppliers for a sustainable supply chain based on a systematic DEA approach", Information Technology and Management, 16(1), p. 39-49.

[5] Dowlatshahi, S., (2000). "Designer – buyer – supplier interface: theory versus practice". International Journal of Production Economics, 63, p. 111-130.

[6] Genovesea, A., Koha, S.C. L., Brunob, G. and E. Espositob, (2013). "Greener supplier selection: state of the art and some empirical evidence", International Journal of Production Research, 51(10), p. 2868–2886.

[7] Kannan D., Khodaverdi R., Olfat L., Jafarian A., Diabat A. (2013). "Integrated fuzzy multi criteria decision making method and multi-objective programming approach for supplier selection and order allocation in a green supply chain", Journal of Cleaner Production, 47, p. 355-367.

[8] Brown, A.S. (2009). "The Many Shades of Green", Mechanical Engineering, 131, p. 22-28.

[9] Ahi P., Searcy C. (2013). "A comparative literature analysis of definitions for green and sustainable supply chain management", Journal of Cleaner Production, 52, p. 329-341.

[10] McKinnon, A., Browne, M., and Whiteing, A. (2012). "Green logistics", 2nd Edition, Improving the environmental sustainability of logistics, Kogan Page Publishers, Ireland, 2012.

[11] Littig, B. and Grießler, E. (2005). "Social sustainability: a catch word between political pragmatism and social theory", International Journal on Sustainable Development, 8(1-2), p. 65-79.

[12] Cheraghi, S. H., Dadashzadeh, M. and M. Subramanian, (2011). "Critical Success Factors For Supplier Selection: An Update", Journal of Applied Business Research, 20(2), p. 91-108.

[13] Arıkan, E., and Jammernegg, W., (2014). "The single period inventory model under dual sourcing and product carbon footprint constraint", International Journal of Production Economics, 157, p. 15-23.

[14] Piegat, A. (2001). "Fuzzy Modeling and Control", Studies in fuzziness and soft computing, vol 69, Physica-Verlag Heidelberg New-York, published in Germany.

[15] Hennequin, S. and Ramirez Restrepo, L. M., (2016). "Fuzzy model of a jointly maintenance and production control under sustainability constraints", proceedings of the 8 th IFAC Conference on Manufacturing Modelling, Management and control (MIM 2016), IFAC-Papers On Line, 49(12), p. 1216–1221.

[16] Ghodsypour S. H., and C. O'Brien (2001). "The total cost of logistics in supplier selection, under conditions of multipe soursing, multiple criteria and capacity constraint", International Journal of Production Economics, 73, p. 15-27.

[17] Sawadogo, M., (2011), « Intégration de l'impact environnemental, sociétal et économique du transport intermodal au sein des chaînes logistiques vertes : Optimisation multi-objectif par les colonies de fourmis », PhD thesis, Lorraine University, under the Direction of D. Anciaux, in French.

[18] Kreutzberger, E., Macharis, C., Vereecken, L. and J. Woxenius, (2003). "Is intermodal freight transport more environmentally friendly than all raod freight transport? A review", proceedings of NECTAR conference n°7, Unea, Sweden, June 13-15.

Evaluating green supplier development programs by a new decision interval-valued fuzzy model-based possibilistic statistical concept

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

R. Tavakkoli-Moghaddam, N. Foroozesh
School of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran,
(tavakoli@ut.ac.ir; n.foroozesh@ut.ac.ir)

S.M. Mousavi
Department of Industrial Engineering, Faculty of Engineering, Shahed University, Tehran, Iran,
(sm.mousavi@shahed.ac.ir)

Abstract—Enhancing environmental performance of suppliers is important for green supply chains. Numerous companies have assigned resources to different green supplier development programs to improve their supplier performances. The decision to select the appropriate program for green supplier improvement is a challenging decision because of prior experience, limited quantitative information, and different supplier backgrounds. In this paper, a new decision interval-valued fuzzy model based on possibilistic statistical concepts is provided for appraising development programs of green suppliers. Two new versions of technique for order preference by similarity to ideal solution (TOPSIS) and Entropy method are introduced to solve the development programs evaluation problem. Then, an application example in the manufacturing industry is presented and solved by proposed decision interval-valued fuzzy model under uncertainty.

Keywords—Green supplier development program; interval-valued fuzzy sets; TOPSIS method; entropy method.

I. Introduction

Green supply chain management has observed significant advancement in the last two decades (Seuring and Muller, 2008). It is the combination of common ecological worries into supply chain management. There are numerous exercises and elements that might be taken within this organizational practice, for instance, incorporate selecting green suppliers and incorporating fusing supplier contribution to greening authoritative practices [1-4].

Regarding the evaluation of green supplier development programs, Chiou et al. [5] considered analytic hierarchy process (AHP) and provided fuzzy AHP for appraising green suppliers using six factors with 24 sub-factors. The case could decide the relative significance of appraising green suppliers over a multicultural setting. Yan [6] presented a combined methodology receiving a genetic algorithm and AHP to obtain and optimized appraisement of green suppliers. Buyukozkan and Cifci [7, 8] focused on a methodology with a fuzzy feedback approach with preference relationships under uncertainty. Fu et al. [9] presented an organized administrative methodology for companies to assess the impact connections amongst development programs of green suppliers by grey modeling. Blome et al. [10] regarded the contradicting hypothetical perspectives of authenticity in assessing firm performance along with administration responsibility. Akman [11] assessed suppliers to include green supplier development programs via fuzzy c-means and compromise ranking (VlseKriterijumska Optimizacija I Kompromisno Resenje, namely VIKOR) techniques. Awasthi and Kannan [12] developed and considered the issue of assessing green supplier advancement programs by a fuzzy NGT and VIKOR-based approach.

The above-related literature on the green supplier development programs denotes that an assessment of selection problem is a multi-criteria group decision-making (MCGDM) framework for the supply chain management (SCM), and is regarded as a new research area. In practice, several evaluation factors or criteria can influence this selection issue under uncertain conditions.

In this paper, an evaluation approach for development programs of green supplier is presented with a new decision-making model by interval-valued fuzzy sets and possibilistic statistical concepts. For this purpose, two new versions of technique for order preference by similarity to ideal solution (TOPSIS) and Entropy methods are introduced to solve the development programs evaluation problem. Two possibilistic mean and possibilistic standard deviation matrices are presented. Then, an application example in the manufacturing industry is presented and solved by proposed decision interval-valued fuzzy model under uncertainty.

The rest of the paper is organized as follows. Section 2 presents the research background followed by the proposed approach in Section 3. Section 4 contains the illustrative example. Finally, conclusions are given in Section 5.

II. Preliminaries

Interval-Valued Fuzzy Sets

The interval-valued fuzzy numbers have been considered as a special form of generalized fuzzy numbers. These fuzzy numbers can contain interval-valued trapezoidal fuzzy numbers, triangular shape, and interval-valued triangular fuzzy numbers. A graphical representation of an interval-valued triangular fuzzy number is depicted in Fig. 1. According to Yao

and Lin [13], an interval-valued triangular fuzzy number is represented as follows:

$$\tilde{A} = \left[\underline{\tilde{A}}, \underline{\underline{\tilde{A}}}\right] = \left[\left(\underline{a}_1, \underline{a}_2, \underline{a}_3; \underline{\hat{h}}_{\tilde{A}}\right), \left(\overline{a}_1, \overline{a}_2, \overline{a}_3; \hat{\overline{h}}_{\tilde{A}}\right)\right] \quad (1)$$

Suppose $\underline{\tilde{A}}$ and $\underline{\underline{\tilde{A}}}$ be two generalized triangular fuzzy numbers (GTFN) as depicted in Fig. 2; hence, $\underline{\hat{h}}_{\tilde{A}}$ and $\hat{\overline{h}}_{\tilde{A}}$ define the heights of $\underline{\tilde{A}}$ and $\underline{\underline{\tilde{A}}}$, and $\underline{a}_1, \underline{a}_2, \underline{a}_3$, $\overline{a}_1, \overline{a}_2, \overline{a}_3$ define the real values. $GTFN\ \tilde{A}$ denoted in the universe of discourse X is described by: $0 \le \underline{a}_1 \le \underline{a}_2 \le \underline{a}_3 \le 1$, $0 \le \overline{a}_1 \le \overline{a}_2 \le \overline{a}_3 \le 1$, $\overline{a}_1 \le \underline{a}_1$ and $\underline{a}_3 \le \overline{a}_3$. In addition, $\underline{\tilde{A}} = \left(\underline{a}_1, \underline{a}_2, \underline{a}_3; \underline{\hat{h}}_{\tilde{A}}\right)$, $\underline{\underline{\tilde{A}}} = \left(\overline{a}_1, \overline{a}_2, \overline{a}_3; \hat{\overline{h}}_{\tilde{A}}\right)$ and $\underline{\tilde{A}} \subset \underline{\underline{\tilde{A}}}$ are regarded.

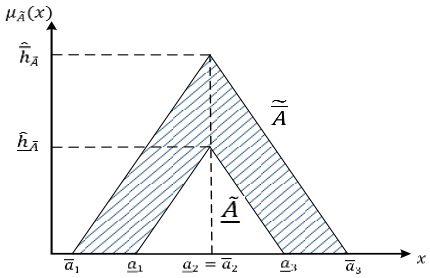


Fig. 1. Interval-Valued Fuzzy Number

A. Possibility Theory

A fuzzy number $\tilde{A}$ will be a fuzzy arrangement of the real line x with a normal, fuzzy convex and continuous membership function of limited support [14-16]. Let $\tilde{A} \in x$ be fuzzy number with $[\tilde{A}]^{\gamma} = [a_1(\gamma), a_2(\gamma)], \gamma \in [0,1]$. The possibilistic mean estimations of fuzzy number $\tilde{A}$ is denoted as $M(\tilde{A}) = \int_0^1 \gamma\left(a_1(\gamma) + a_2(\gamma)\right)$, and the possibilistic difference of $\tilde{A}$ as $Var(\tilde{A}) = \frac{1}{2}\int_0^1 \gamma(a_1(\gamma) - a_2(\gamma))^2 d\gamma$.

Let $\tilde{A}$ be a triangular fuzzy number with focus a, left-width $\tau > 0$, right-width $\sigma > 0$, i.e. $\tilde{A} = (a - \tau, a, a + \sigma)$. *A* γ-level arrangement of fuzzy number $\tilde{A}$ can be calculated by:

$$[\tilde{A}]^{\gamma} = [a - (1 - \gamma)\tau, a + (1 - \gamma)\sigma], \forall \gamma \in [0,1] \quad (2)$$

Based on above-mentioned explanations, the possibilistic mean estimation of triangular fuzzy number $\tilde{\xi} = (a, b, c)$ is presented by:

$$\begin{aligned} M(\tilde{\xi}) = e = \int_0^1 \gamma\big((a - (1 - \gamma)\tau) \\ + (a + (1 - \gamma)\sigma)\big)d\gamma \\ = (a + 2b + c)/4 \end{aligned} \quad (3)$$

Possibilistic variance of triangular fuzzy number $\tilde{\xi}$ can be given as follows:

$$\begin{aligned} &Var(\tilde{\xi}) \\ &= \int_0^1 \gamma\big((a - (1 - \gamma)\tau) + (a + (1 - \gamma)\sigma)\big)^2 d\gamma \\ &= (33\alpha_1^3 + 21\alpha_1^2\gamma + 11\alpha_1\gamma^2 - \gamma^3)/384\alpha_1 \end{aligned} \quad (4)$$

where $\alpha_1 = \max\{b - a, c - b\}$ and $\gamma = \min\{b - a, c - b\}$.

III. Proposed New Evaluation Model For Green Supplier Development Programs

In this section, a new procedure based on an interval-valued fuzzy group approach for green supplier development programs is presented based on possibility theory and statistical concepts. First, it is assumed that:

$DM = \{DM^k | k = 1, \dots, p\}$ as a set of supply chain team members, $X = \{X_i | i = 1, \dots, m\}$ as a set of green supplier development programs candidates, $C = \{C_j | j = 1, \dots, n\}$ as a set of criteria for green supplier development problem.

Since the information of green supplier development programs candidates is uncertain during group decision making in SCM, the supply chain decision makers or experts can consider an interval-valued fuzzy (IVF) $\tilde{A}_{ij}^k$ to estimate the judgment and opinion on sustainable supplier candidate X_i with respect to supplier criterion C_i. The problem of green supplier development programs with interval-valued fuzzy sets and statistical concepts can be expressed in matrix form:

$$\tilde{A}^k = \left[[(\underline{a}_{ij_1}, \underline{a}_{ij_2}, \underline{a}_{ij_3}), (\overline{a}_{ij_1}, \overline{a}_{ij_2}, \overline{a}_{ij_3})]^k\right]_{m \times n} \quad (5)$$

For convenience, we denote:
$\left[\left(\underline{a}_{ij_1}, \underline{a}_{ij_2}, \underline{a}_{ij_3}\right), \left(\overline{a}_{ij_1}, \overline{a}_{ij_2}, \overline{a}_{ij_3}\right)\right]^k = [(\underline{a}_{ij} - \underline{\tau}_{ij}, \underline{a}_{ij}, \underline{a}_{ij} + \underline{\sigma}_{ij}), (\overline{a}_{ij} - \overline{\tau}_{ij}, \overline{a}_{ij}, \overline{a}_{ij} + \overline{\sigma}_{ij})]^k$ for kth TM.

According to the above-mentioned descriptions, steps of the proposed interval-valued fuzzy procedure are presented for the evaluation and selection problem of the green supplier development programs as follows:

Step 1. Proper criteria are identified for the evaluation problem of the green supplier development programs.

Step 2. Provide the IVF-decision matrices of green supplier development programs candidates, and aggregate them. The aggregated IVF-decision matrices of the green supplier development programs candidates are obtained by the arithmetic average with the following relation, respectively:

$$\begin{aligned} &\underline{a}_{ijl} = \frac{1}{p}\sum_{k=1}^{p}\left(\underline{a}_{ij_l}\right)^k \ ; l = 1,2,3 \\ &\text{and } \overline{a}_{ijl} = \frac{1}{p}\sum_{k=1}^{p}(\overline{a}_{ij_l})^k \ ; l = 1,2,3 \end{aligned} \quad (6)$$

Step 3. Transform the aggregated IVF-matrix into the normalized matrix of the green supplier development programs

candidates. To transform different criteria scales into a comparable scale, the linear scale transformation method is used and presented by:

$$\tilde{a}'_{ij} = \left[\left(\underline{a}'_{ij_1}, \underline{a}'_{ij_2}, \underline{a}'_{ij_3}\right), \left(\overline{a}'_{ij_1}, \overline{a}'_{ij_2}, \overline{a}'_{ij_3}\right)\right] = \left[\left(\frac{\underline{a}_{ij} - \underline{\tau}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}, \frac{\underline{a}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}, \frac{\underline{a}_{ij} + \underline{\sigma}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}\right), \left(\frac{\overline{a}_{ij} - \overline{\tau}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}, \frac{\overline{a}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}, \frac{\overline{a}_{ij} + \overline{\sigma}_{ij}}{\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+}\right)\right], j \in \Omega_b \quad (7)$$

and

$$\tilde{a}'_{ij} = \left[\left(\underline{a}'_{ij_1}, \underline{a}'_{ij_2}, \underline{a}'_{ij_3}\right), \left(\overline{a}'_{ij_1}, \overline{a}'_{ij_2}, \overline{a}'_{ij_3}\right)\right] = \left[\left(\frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\underline{a}_{ij} + \underline{\sigma}_{ij}}, \frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\underline{a}_{ij}}, \frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\underline{a}_{ij} - \underline{\tau}_{ij}}\right), \left(\frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\overline{a}_{ij} + \overline{\sigma}_{ij}}, \frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\overline{a}_{ij}}, \frac{\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^-}{\overline{a}_{ij} - \overline{\tau}_{ij}}\right)\right], j \in \Omega_c \quad (8)$$

where Ω_b and Ω_c are the sets of benefit and cost attributes for the green supplier development programs evaluation problem, respectively.

$\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right)^+ = \max_i\left(\overline{a}_{ij} + \overline{\sigma}_{ij}\right), i = 1,..,m$, $\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right)^- = \min_i\left(\overline{a}_{ij} - \overline{\tau}_{ij}\right), i = 1,..,m$, and $k = 1, \dots, p$.

The maximum rating of each green supplier development programs candidate against each criterion and the minimum rating using the normalization process can be obtained. For convenience, it is denoted as follows:

$$\left[\left(\underline{a}'_{ij_1}, \underline{a}'_{ij_2}, \underline{a}'_{ij_3}\right), \left(\overline{a}'_{ij_1}, \overline{a}'_{ij_2}, \overline{a}'_{ij_3}\right)\right]^k = \left[\left(\underline{a}'_{ij} - \underline{\tau}'_{ij}, \underline{a}'_{ij}, \underline{a}'_{ij} - \underline{\sigma}'_{ij}\right), \left(\overline{a}'_{ij} - \overline{\tau}'_{ij}, \overline{a}'_{ij}, \overline{a}'_{ij} - \overline{\sigma}'_{ij}\right)\right]^k$$

Step 4. Build possibilistic mean interval value matrix for the assessment issue of the programs alternatives. The possibilistic mean interval values ($\breve{m}_{ij}$) of IVF $\widetilde{a'}_{ij} = \left[\left(\underline{a}'_{ij} - \underline{\tau}'_{ij}, \underline{a}'_{ij}, \underline{a}'_{ij} - \underline{\sigma}'_{ij}\right), \left(\overline{a}'_{ij} - \overline{\tau}'_{ij}, \overline{a}'_{ij}, \overline{a}'_{ij} - \overline{\sigma}'_{ij}\right)\right]$ are defined according to Eq. (3):

$$\breve{m}_{ij} = \left[\underline{m}_{ij}, \overline{m}_{ij}\right] = \left[\frac{\left(\underline{a}'_{ij_1} + 2 \times \underline{a}'_{ij_2} + \underline{a}'_{ij_3}\right)}{4}, \frac{\left(\overline{a}'_{ij_1} + 2 \times \overline{a}'_{ij_2} + \overline{a}'_{ij_3}\right)}{4}\right] \quad (9)$$

Then, the possibilistic mean interval value matrix is constructed for the evaluation problem of green supplier development programs candidates as follows:

$$\breve{M} = \left[\breve{m}_{ij}\right]_{m \times n} = \begin{bmatrix} \breve{m}_{11} & \breve{m}_{12} & \cdots & \breve{m}_{3n} \\ \breve{m}_{21} & \breve{m}_{22} & \cdots & \breve{m}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \breve{m}_{m1} & \breve{m}_{m2} & \cdots & \breve{m}_{mn} \end{bmatrix} \quad (10)$$

Step 5. Construct the possibilistic standard deviation interval value matrix for the evaluation problem of green supplier development programs candidates. The interval possibilistic standard deviation ($\widetilde{sd}_{ij}$) of IVF $\widetilde{a'}_{ij} = \left[\left(\underline{a}'_{ij} - \underline{\tau}'_{ij}, \underline{a}'_{ij}, \underline{a}'_{ij} - \underline{\sigma}'_{ij}\right), \left(\overline{a}'_{ij} - \overline{\tau}'_{ij}, \overline{a}'_{ij}, \overline{a}'_{ij} - \overline{\sigma}'_{ij}\right)\right]$ are determined according to Eq. (4):

$$\widetilde{sd}_{ij} = \left[\underline{sd}_{ij}, \overline{sd}_{ij}\right] = \left[\sqrt[2]{\frac{33\left(\underline{\alpha}_{ij}\right)^3 + 21\left(\underline{\alpha}_{ij}\right)^2\left(\underline{\gamma}_{ij}\right) + 11\left(\underline{\alpha}_{ij}\right)\left(\underline{\gamma}_{ij}\right)^2 - \left(\underline{\gamma}_{ij}\right)^3}{384\left(\underline{\alpha}_{ij}\right)}}, \sqrt[2]{\frac{33\left(\overline{\alpha}_{ij}\right)^3 + 21\left(\overline{\alpha}_{ij}\right)^2\left(\overline{\gamma}_{ij}\right) + 11\left(\overline{\alpha}_{ij}\right)\left(\overline{\gamma}_{ij}\right)^2 - \left(\overline{\gamma}_{ij}\right)^3}{384\left(\overline{\alpha}_{ij}\right)}}\right] \quad (11)$$

where $\underline{\alpha} = \max\left\{\underline{\tau}'_{ij}, \underline{\sigma}'_{ij}\right\}$ and $\underline{\gamma} = \min\left\{\underline{\tau}'_{ij}, \underline{\sigma}'_{ij}\right\}$ and $\overline{\alpha} = \max\left\{\overline{\tau}'_{ij}, \overline{\sigma}'_{ij}\right\}$ and $\overline{\gamma} = \min\left\{\overline{\tau}'_{ij}, \overline{\sigma}'_{ij}\right\}$.

Then, the possibilistic standard deviation interval value matrix is constructed for the evaluation problem of green supplier development programs candidates as follows:

$$\widetilde{SD} = \left[\widetilde{sd}_{ij}\right]_{m \times n} = \begin{bmatrix} \widetilde{sd}_{11} & \widetilde{sd}_{12} & \cdots & \widetilde{sd}_{3n} \\ \widetilde{sd}_{21} & \widetilde{sd}_{22} & \cdots & \widetilde{sd}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{sd}_{m1} & \widetilde{sd}_{m2} & \cdots & \widetilde{sd}_{mn} \end{bmatrix} \quad (12)$$

Step 6. Compute entropy weighting method based on mean-variance concepts.

Sub-steps 6.1. Calculate possibilistic mean entropy measure with interval values for each green supplier development evaluation criterion.

$$\breve{E}(m)_j = \left[\underline{e}(m)_j, \overline{e}(m)_j\right] = \left[-\frac{1}{Ln(m)}\sum_{i=1}^{m}\overline{m}'_{ij}Ln\left(\overline{m}'_{ij}\right), -\frac{1}{Ln(m)}\sum_{i=1}^{m}\underline{m}'_{ij}Ln\left(\underline{m}'_{ij}\right)\right] \quad (13)$$

where $\overline{m}'_{ij} = \left[\underline{m}'_{ij}, \overline{m}'_{ij}\right] = \left[\frac{\underline{m}_{ij}}{max_i\overline{m}_{ij}}, \frac{\overline{m}_{ij}}{max_i\overline{m}_{ij}}\right]$.

Sub-steps 6.2. Calculate possibilistic standard deviation entropy measure with interval values for each green supplier development evaluation criterion.

$$\check{E}(sd)_j = \left[\underline{e}(sd)_j, \overline{e}(sd)_j\right] = \left[-\frac{1}{Ln(m)}\sum_{i=1}^{m}\overline{sd}'_{ij}Ln\left(\overline{sd}'_{ij}\right), -\frac{1}{Ln(m)}\sum_{i=1}^{m}\underline{sd}'_{ij}Ln\left(\underline{sd}'_{ij}\right)\right] \quad (14)$$

where $\overline{sd}'_{ij} = \left[\underline{sd}'_{ij}, \overline{sd}'_{ij}\right] = \left[\frac{\underline{sd}_{ij}}{max_i \overline{sd}_{ij}}, \frac{\overline{sd}_{ij}}{max_i \overline{sd}_{ij}}\right]$.

Step 7. Calculate final weights of green supplier development evaluation criteria based on mean-variance concepts.

Sub-step 7.1. Calculate proposed entropy weight based on the possibilistic mean interval values.

$$\check{W}(m)_j = \left[\underline{w}(m)_j, \overline{w}(m)_j\right] = \left[1-\overline{e}(m)_j, 1-\underline{e}(m)_j\right] \quad (15)$$

Sub-step 7.2. Calculate proposed entropy weight based on the possibilistic standard deviation interval values.

$$\check{W}(sd)_j = \left[\underline{w}(sd)_j, \overline{w}(sd)_j\right] = \left[1-\overline{e}(sd)_j, 1-\underline{e}(sd)_j\right] \quad (16)$$

Step 8. Define positive-ideal (PIV) of possibilistic mean interval values for the evaluation problem of green supplier development programs candidates. The PIV ($\check{M}^*$) are calculated by:

$$\check{M}^* = \left\{\check{M}_1^*, \check{M}_2^*, \dots, \check{M}_n^*\right\} = \left\{\left[\underline{m}_j^*, \overline{m}_j^*\right]\right\} = \left\{\max_i \check{m}_{ij} \middle| i = 1,2,\dots,m\right\} \quad (17)$$

Step 9. Define positive-ideal and negative-ideal vector (PIV and NIV) of possibilistic standard deviation of interval values. The PIV ($\widetilde{SD}^*$) are determined by:

$$\widetilde{SD}^* = \left\{\widetilde{SD}_1^*, \widetilde{SD}_2^*, \dots, \widetilde{SD}_n^*\right\} = \left\{\left[\underline{sd}_j^*, \overline{sd}_j^*\right]\right\} = \left\{\min_i \widetilde{sd}_{ij} \middle| i = 1,2,\dots,m\right\} \quad (18)$$

Step 10. Calculate the separation measures matrix according to mean and standard deviation of interval-number from the PIV ($\check{M}^*$, $\widetilde{SD}^*$), respectively.

The separation vectors of mean, standard deviation of interval value matrix from the PIV are obtained for the selection problem of sustainable suppliers as below:

$$\Gamma_i = m_{D_{ij}^*}\left(\check{m}_{ij}, \check{M}_j^*\right) = \sqrt{\sum_{j=1}^{n}\left(\underline{w}(m)_j\left(\underline{m}_j^* - \underline{m}_{ij}\right)^2 + \overline{w}(m)_j\left(\overline{m}_j^* - \overline{m}_{ij}\right)^2\right)} \quad (19)$$

$$K_i = sd_{D_{ij}^*}\left(\widetilde{sd}_{ij}, \widetilde{SD}_j^*\right) = \sqrt{\sum_{j=1}^{n}\left(\underline{w}(sd)_j\left(\underline{sd}_j^* - \underline{sd}_{ij}\right)^2 + \overline{w}(sd)_j\left(\overline{sd}_j^* - \overline{sd}_{ij}\right)^2\right)} \quad (20)$$

Step 11. The final score is calculated using a linear combination closeness coefficient of alternative X_i.

$$Sc_i = \alpha\Gamma_i + \beta K_i \quad (21)$$

where $\sum \alpha + \beta = 1$, $0 \le \alpha \le 1$, $0 \le \beta \le 1$.

Step 12. Rank the preference order of all alternatives as per the incorporated relative closeness coefficient, and select the best one.

IV. Illustrative Example

An automobile manufacturing company is considered from the recent study in the literature [12]. To accomplish the objective, the organization is advancing the utilization of environment friendly practices in all procedures of the association, green provider advancement being one of them. To make best utilization of time and assets, the organization considered on surveying three green supplier improvement programs, including mandatory ISO 14000 certification (P_1), supplier training (P_2), and employee deployment with environment expertise at supplier locations (P_3) for the implementation a board of three expert specialists (DM_1, DM_2, and DM_3) toward surveying and regarding the best green supplier development program(s) for usage.

The committee applies the criteria for green supplier development program assessment as below: Time (C_1), Cost (C_2), Labor (C_3), Resources (C_4), Energy usage (C_5), Water (C_6), Emissions (C_7), Noise (C_8), Waste (C_9), Green packaging (C_{10}), Green manufacturing (C_{11}), Green product design (C_{12}), Green transportation (C_{13}), Green warehousing (C_{14}), Green procurement (C_{15}), and Reverse logistics (C_{16}).

Linguistic ratings regarding to the appraisement criteria and the development programs of green suppliers as candidates are given using Tables I & II.

TABLE I. Linguistic variables

Linguistic variables	Interval-valued fuzzy numbers
Very Poor (VP)	[(0.00,0.00,2.00), (0.00,0.00,3.50)]
Poor (P)	[(1.00,2.50,4.00), (0.00,2.50,6.00)]
Fair (F)	[(3.50,5.00,6.50), (2.00,5.00,8.00)]
Good (G)	[(6.00,7.50,9.00), (4.00,7.50,10.00)]
Very Good (VG)	[(8.00,10.00,10.00), (6.50,10.00,10.00)]

Following are the final weights of green supplier development evaluation criteria based on mean-variance concepts calculated for the evaluation problem of the green supplier development programs candidates. Table III shows the computational results of the proposed new evaluation model.

$\tilde{W}(m)_j = \left[\underline{w}(m)_j, \overline{w}(m)_j\right] =$
[(0.55,0.58), (0.65,0.71), (0.64,0.69), (0.48,0.52), (0.46,0.49), (0.66,0.68), (0.54,0.58), (0.47,0.48), (0.46,0.49), (0.70,0.73), (0.62,0.66), (0.68,0.62), (0.60,0.64), (0.33,0.33), (0.72,0.76), (0.35,0.39)].
and
$\tilde{W}(m)_j = \left[\underline{w}(m)_j, \overline{w}(m)_j\right] =$
[(0.03,0.64), (0.07,0.92), (0.04,0.74), (0.07,0.85), (0.04,0.85), (0.05,0.79), (0.04,0.91), (0.05,0.77), (0.08,0.92), (0.04,0.91), (0.03,0.67), (0.06,0.76), (0.04,0.78), (0.05,0.87), (0.07,0.98), (0.03,0.86)].

TABLE II. Linguistic assessments for three alternatives

Criteria	P1			P2			P3		
	DM1	DM2	DM3	DM1	DM2	DM3	DM1	DM2	DM3
C1	P	P	VG	VP	VG	VP	G	G	P
C2	VP	P	VP	VP	VP	F	G	P	VP
C3	G	VG	VP	VG	VG	F	VG	G	VP
C4	P	VP	VG	VP	VP	F	G	G	VP
C5	F	G	P	G	F	F	P	VG	P
C6	VG	P	VG	F	F	F	P	F	P
C7	VP	F	F	G	P	G	F	F	F
C8	F	VG	VG	F	F	VP	F	VP	P
C9	VP	G	G	G	VP	VP	F	VG	P
C10	G	F	F	G	P	VG	G	F	P
C11	G	VP	VG	VG	G	P	P	VG	VP
C12	VP	F	VG	VP	VP	VP	G	VP	F
C13	P	G	F	VG	VG	F	P	VG	VG
C14	VP	P	P	VP	VP	F	F	VG	F
C15	VG	P	G	VP	G	F	P	G	VG
C16	G	G	G	P	F	F	F	VP	F

TABLE III. Computational results of the model

Green supplier development programs	$\Gamma_i = m_{D_{ij}^*}(\tilde{m}_{ij}, \tilde{M}_j^*)$	$K_i = sd_{D_{ij}^*}(\widetilde{sd}_{ij}, \widetilde{SD}_j^*)$	Sc_i	Ranking order
P_1	0.945061	0.067936	0.506498	2
P_2	1.071516	0.072928	0.572222	3
P_3	0.844923	0.074872	0.459898	1

The resulting outcomes of the proposed model would help managers prioritize their green supplier development programs concentrating on programs by delivering greater supplier performance. The assessment and selection can assist resource and investment allocations.

V. Conclusion

Environmental performance of an organization in the supply chain could be evaluated by the organization's own natural endeavors and the execution of its suppliers. Helpful association with suppliers respected ecological issues permits organizations to enhance their ecological execution. Organizations request their suppliers to regard the needed natural accreditations and to characterize green applications. The end goal to build up their ecological capabilities, organizations ought to make agreeable, close and coordinated associations with their suppliers, assess green execution of their suppliers, and help suppliers about green programs' issues. This paper presents a new evaluation model for green supplier development programs by integrating interval-valued fuzzy sets and possibilistic statistical concepts. For this purpose, a new decision-making process under an interval-valued fuzzy environment is introduced based on the technique for order of preference by similarity to ideal solution (TOPSIS) method, two possibilistic mean and possibilistic standard deviation. In addition, a new version of entropy method is developed for computing the weight of each evaluation criterion with possibilistic statistical concepts. New closeness coefficient of alternative indexes are also presented to provide a final order of green supplier development programs. Finally, the validity of the proposed model was illustrated by an application example in the manufacturing industry from the recent literature. The decision modelling effort is able to regard uncertainties, and users of the proposed model should understand that the uncertainties additionally available in what elements ought to be incorporated and managerial perceptions related to the factors.

References

[1] Sarkis J, editor. Greening the supply chain. Berlin: Springer; 2006.

[2] Fahimnia B, Sarkis J, Davarzani H. Green supply chain management: A review and bibliometric analysis.

International Journal of Production Economics. 2015; 162:101-14.

[3] Dubey R, Gunasekaran A, Ali SS. Exploring the relationship between leadership, operational practices, institutional pressures and environmental performance: A framework for green supply chain. International Journal of Production Economics. 2015; 160:120-32.

[4] Kusi-Sarpong S, Bai C, Sarkis J, Wang X. Green supply chain practices evaluation in the mining industry using a joint rough sets and fuzzy TOPSIS methodology. Resources Policy. 2015; 46:86-100.

[5] Chiou CY, Hsu CW, Hwang WY. Comparative investigation on green supplier selection of the American, Japanese and Taiwanese electronics industry in China. InIndustrial Engineering and Engineering Management, 2008. IEEM 2008. IEEE International Conference on 2008 Dec 8 (pp. 1909-1914).

[6] Yan GE. Research on green suppliers' evaluation based on AHP & genetic algorithm. In2009 IEEE International Conference on Signal Processing Systems 2009 (pp. 615-619).

[7] Büyüközkan G, Çifçi G. A novel fuzzy multi-criteria decision framework for sustainable supplier selection with incomplete information. Computers in Industry. 2011;62(2):164-74.

[8] Büyüközkan G, Çifçi G. A novel hybrid MCDM approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy TOPSIS to evaluate green suppliers. Expert Systems with Applications. 2012;39(3):3000-11.

[9] Fu X, Zhu Q, Sarkis J. Evaluating green supplier development programs at a telecommunications systems provider. International Journal of Production Economics. 2012;140(1):357-67.

[10] Blome C, Hollos D, Paulraj A. Green procurement and green supplier development: antecedents and effects on supplier performance. International Journal of Production Research. 2014;52(1):32-49.

[11] Akman G. Evaluating suppliers to include green supplier development programs via fuzzy c-means and VIKOR methods. Computers & industrial engineering. 2015; 86:69-82.

[12] Awasthi A, Kannan G. Green supplier development program selection using NGT and VIKOR under fuzzy environment. Computers & Industrial Engineering. 2016; 91:100-8.

[13] Yao JS, Lin FT. Constructing a fuzzy flow-shop sequencing model based on statistical data. International journal of approximate reasoning. 2002;29(3):215-34.

[14] Zhang WG, Wang YL, Chen ZP, Nie ZK. Possibilistic mean–variance models and efficient frontiers for portfolio selection problem. Information Sciences. 2007;177(13):2787-801.

[15] Deng X, Li R. Gradually tolerant constraint method for fuzzy portfolio based on possibility theory. Information Sciences. 2014; 259:16-24.

[16] Li X, Qin Z, Kar S. Mean-variance-skewness model for portfolio selection with fuzzy returns. European Journal of Operational Research. 2010;202(1):239-4.

An exact approach for solving the supplier selection problem in multi-service outsourcing

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Pham Thi Hoai
School of Applied Mathematics and Informatics,
Hanoi University of Science and Technology,
No1 Dai Co Viet, Hanoi, Vietnam
Email: hoai.phamthi@hust.edu.vn

Le Thi Hoai An
Theoretical and Applied Computer Science Lab,
University de Lorraine,
Ile du Saulcy, 57045 Metz, France
Email: hoai-an.le-thi@univ-lorraine.fr

Nguyen Canh Nam
School of Applied Mathematics and Informatics,
Hanoi University of Science and Technology,
No1 Dai Co Viet, Hanoi, Vietnam
Email: nam.nguyencanh@hust.edu.vn

***Abstract*—We study the supplier selection problem in multi-service outsourcing (MSO) that has been modeled by Feng et al. [11] (2011). This model is indeed in the class of multiobjective combinatorial optimization (MOCO) problems. An heuristic decision method was proposed by Feng et al. [11] as well. However, to the best of our knowledge, no exact approach was explored for this problem so far. Hence, inspired by the recent developments on solving MOCO problems, we propose in this paper an exact algorithm based on polyblock approach for generating all nondominated points of MSO problems. Stack policies are also analyzed, pointing out their performance effectiveness by means of the results of computational experience.**

I. Introduction

Supplier selection (SS) is one of the most important problems in optimal sourcing area and has attracted many researchers in purchasing with methodologies ranging from conceptual to empirical and modelling streams. SS is the process of finding capable suppliers to provide the buyer with the right quality products and/or services at the right prices, at the right quantities, and at the right time [17], [21], [25]. SS involves two main tasks that are central to any decision-making problem [4]. One is determining the criteria to be considered, and the other is the method to be used to compare suppliers [17]. In fact, there are two kinds of SS problems: single SS and multiple SS. For the first case, each supplier meets all buyer's needs. The management only has to decide who the best supplier is. In the second case of SS, no single one satisfies all the demands of the buyer so that he must purchase the identical part from multiple suppliers. And obviously, the multiple SS is more difficult than the single one.

To our best knowledge, there are at least seven papers that have reviewed the literature of SS problems (see e.g., [1], [7], [8], [13], [14], [18], [30]). Anton et al. 2016 [1] categorizes the operational streams researched from 1990 while the remaining researchs focus on the mathematical approaches for SS problems by different periods: before 2000 [7], [8], [30]; 2000-2008 [18]; 2008-2012 [14]; 2007- the submitted time of the paper [13]. Historically, various decision-making methods have been proposed for handling SS problems but it can be divided into two groups: the group of individual methods and the group of hybrid methods. The individual methods are such as DEA (data envelopment analysis), MP (mathematical programming), AHP (analytic hierarchy process), CBR (the case-based reasoning), ANP (analytic network process), FANP (fuzzy ANP), GA (genetic algorithm), SMART (simple multi-attribute rating technique), GRA (grey ralational analysis), NN (neural network), etc. The hybrid methods are established by combining several individual methods; for instances, integrated AHP approaches, integrated ANP approaches, integrated DEA approaches, integrated MP approaches, etc. Among them MP is a widely used method. A numerous research papers choose MP as a main approach or a part of hybrid methods (by combining with AHP, ANP, TOPSIS+ANP, FAHP, FANP). MP methods for SS problems include: linear programming, integer linear programming, integer nonlinear programming, goal programming, multiobjective programming, stochatic programming and nonlinear programming. For more details, see [1], [3], [7]–[9], [13], [14], [18], [30], [31] and the references therein.

Some recent papers that use MP methods for SS problems from 2011 are summarized in the table below.

TABLE I: Research on MP models for SS problems

Model (solution method)	Chosen criteria	Constraint	Ref.
Multi-objective integer nonlinear programming (heuristic algorithm)	Min (oursourcing cost, wating time), Max collaborative utility	Buyer's demand	[11] (2011)
Integer programming (heuristic algorithm)	Min the sum of purchase and transportation cost	Amount of product, number of visited suppliers, discount inteval, buyer's demand	[20] (2012)
Mixed integer nonlinear programming (linearized and using commercial software)	Min the sum of anual ordering, inventory holding, and purchasing costs	Quality, capacity	[22] (2012)
Multi-choice goal programming	Min (purchasing cost, rejects and late deliveries)	The buyer's demand and the supplier's capacity	[12] (2014)
Multi-objective mixed integer NLP (weighted sum scalarization)	Min (TAPC, rejects and late deliveries)	The buyer's demand, capacity, quantity discount, discount level, number of suppliers	[26] (2015)

It is worth noting that most of the mathematical models in TABLE I tackle the supplier selection in manufacturing industries but the model of Feng et al. [11] (2011) is a problem in service industries (multi-service outsourcing). SS in multi-service outsourcing belongs to the class of multiple SS. It has become an important issue due to the rapid developments recently (see e.g., [2], [11], [19]). However, all researches concerning this problem is quite limited. In particular, only an heuristic approach was proposed for solving the model of SS in MSO by Feng et al. [11] (2011). One drawback of this method is that it does not guarantee the globality of

the obtained solutions. Therefore, an exact algorithm for this model is motivated and will be studied in this paper.

The model proposed by Feng et al. [11] (2011) is a multiobjective 0-1 programming involving three objectives which are collaborative utility, service outsourcing cost and service waiting time. It is indeed in the class of multiobjective combinatorial programming which is an interesting and challenging topic, receiving a lot of attentions recently, see e.g., [5], [6], [15], [16], [23]. Most solution methods for this kind of problems aim at outputting a pool of solutions that are maintained and updated whenever new solutions are found. The information of known solutions is used to decide whether a new solution should be detected. It can be used to guide the search process within the objective space. The part of the objective space where nondominated points could lie therefore plays an important role in the generation of the whole set of nondominated points of a multiobjective programming (or, equivalent to, for solving multiobjective optimization problems). It is defined by *the search area* in [23] or *the search region* in [16]. The generic method based on the representation of the search region is generalized by Klamroth et al. (see Algorithm 1 in [16]). This method proceeds by iteratively producing new solutions in the updated search region whenever a new point is found. Consequently, a key issue is to determine and efficiently update the search region during the exploration.

In recent works [5], [6], [16], [23], the search region is defined by a set of *corner points* or *upper bounds* or *local upper bounds* and some algorithms for updating the search region are also proposed. In particular, Klamroth et al. [16] (2015) developed the updating algorithm of Przybylski et al. [23] (2010) and proposed two efficient algorithms that are RE (redundancy elimination) and RA (redundancy avoidance). A specific neighborhood structure among local upper bounds has been elaborated in [5] for tricriteria problems and be further refined for any criteria cases in [6] (2016). A new algorithm based on the new insight into the neighborhood structure is then proposed in [16] for updating the search region. However, all those algorithms are designed without the mention of the stack mangement of the stored local upper bounds.

Hence, inspired by recent developments on multiobjective combinatorial optimization, in this paper we propose an exact algorithm based on polyblock approach for solving MSO problems by generating all nondominated points of it. To the best of our knowledge, it is the first time MSO problems are solved by an exact solution method. Stack policies are then also analyzed, pointing out their performance effectiveness by means of the results of computational experience. As a byproduct, a concise interpretation for the search region of MOCO problems is proposed by using the variant concepts of "polyblock" (half-open polyblock, in the minimization case) and "copolyblock" (half-open copolyblock, in the maximization case) which have been known as the basic concepts in theory of monotonic optimization [28]. A visualization of the search region is then obtained. It will be seen that there is a one to one correspondence between half-open polyblocks and the search region defined by local upper bounds as well.

The rest of the paper is organized as follows. In the upcoming section, we describe the mathematical model of the supplier selection in multi-service outsourcing constructed by Feng et al. [11]. We recall the preliminaries on the polyblock and copolyblock and MOCO problems in Section 3. Section 4 is dedicated to present the search region in the literature and in our paper. The application of the search region for finding the whole set of nondominated points of MOCO problems will be stated in Section 5. The computational results are shown in Section 6. The paper will be closed with some conclusions in Section 7.

II. Problem description

This section is dedicated to describe the mathematical model of MSO problem proposed by Feng et al. [11].

Suppose that a decision maker is going to select K suppliers from M candidate suppliers for providing K service process/product elements (SPE) ($2 \leq K \leq M$). It is assumed that each SPE is outsourced to one supplier. This assumption aims to avoid the risk of service delay and bad performance, see [11] for more details. Denote $I = \{1, \ldots, M\}$ be the set of suppliers and $S = \{1, \ldots, K\}$ the set of SPEs. For $i \in I$ and $s \in S$, the variable x_{is} is defined as follows

$$x_{is} = \begin{cases} 1 & \text{supplier } i \text{ is selected for SPE } s \\ 0 & \text{otherwise.} \end{cases}$$

The criteria for selecting a pool of desired suppliers include the outsourcing cost, the waiting time and the collaborative utility. The followings are the required data of MSO problem.

p_{is} : the unit price of SPE s that supplier$i$ charges the outsourcing firm,
t_{is} : the average waiting time of SPE s if it is supplied by supplier i,
d_s : the demand quantity of SPE s,
u_{ij} : the collaborative utility between partner firm i and j,
p_s^A : the highest acceptable price of SPE s,
t_s^A : the longest acceptable waiting time of SPE s.

We omit the computation of the collaborative utility $u_{ij}, (i, j \in I)$ which is not relevant for our purpose. Interested readers are refered to [11] for details. The supplier selection problem in MSO proposed by Feng et al. [11] is followed

$$\text{Min } Z_1 = \sum_{i=1}^{M}\sum_{s=1}^{K} p_{is} d_s x_{is} \tag{1}$$

$$\text{Min } Z_2 = \sum_{i=1}^{M}\sum_{s=1}^{K} t_{is} x_{is} \tag{2}$$

$$\text{Max } Z_3 = \sum_{i=1}^{M}\sum_{j=1}^{M} u_{ij} \sum_{s=1}^{K} x_{is} \sum_{s=1}^{K} x_{js} \tag{3}$$

subject to

$$\sum_{s=1}^{K} x_{is} \leq 1, i = 1, \ldots, M, \tag{4}$$

$$\sum_{i=1}^{M} x_{is} = 1, s = 1, \ldots, K, \tag{5}$$

$$\sum_{i=1}^{M} p_{is} x_{is} \leq p_s^A, s = 1, \ldots, K, \tag{6}$$

$$\sum_{i=1}^{M} t_{is} x_{is} \leq t_s^A, s = 1, \ldots, K, \tag{7}$$

$$x_{is} \in \{0, 1\}, i = 1, \ldots, M; s = 1, \ldots, K. \tag{8}$$

The objective function (1) aims to minimize the total service outsourcing cost; the criterion (2) presents the minimization

of service waiting time and Z_3 shows the maximization of collaborative utility between partner firms. Constraint (4) restricts that one supplier can provide at most one service. Equation (5) is for assigning each supplier to exactly one supplier. Constraints (6) and (7) ensure the service cost and service waiting time promised by the suppliers are not larger than the acceptable level.

It is seen that the difficulties in this model is not only in the discrete of multiobjective programming but also in the nonlinearity of the objective funtion, particularly, the third criterion Z_3.

III. Background

Throughout this paper, for any two vectors $x, y \in \mathbb{R}^m$ we write $x \leqq y$ ($x < y$, respectively (resp.)) if $x_i \leq y_i$ ($x_i < y_i$, resp.) for every $i = 1, \ldots, m$. We denote $x \leq y$ if $x \leqq y$ and $x \neq y$. If $a \leqq b$ then the rectangle (the box) $[a, b]$ is the set of all $x \in \mathbb{R}^m$ satisfying $a \leqq x \leqq b$, the half-open rectangle $[a, b)$ is the set of all $x \in \mathbb{R}^m$ such that $a \leqq x < b$. Note that, if $a \leqq b$ and there exists an index $i \in \{1, \ldots, m\}$ such that $a_i = b_i$ then $[a, b)$ is empty set.

For $i = 1, \ldots, m$; e^i denotes the ith unit vector of $\mathbb{R}^m$, i.e. $e^i_i = 1, e^i_j = 0\ \forall j \neq i$. We also denote $e \in \mathbb{R}^m$ be a vector of all ones, i.e. $e = \sum_{i=1}^{m} e^i$.

A. Polyblock and copolyblock

In this section, we recall the background of polyblock and copolyblock in the literature (see e.g., [24], [27]–[29]) and coressponding results to the variant concepts "half-open polyblock" and "half-open copolyblock". Let a, b be two vectors in $\mathbb{R}^m$ satisfying $a \leqq b$ and T a finite subset of $[a, b]$. The set $P = \bigcup_{z \in T} [a, z]$ ($Q = \bigcup_{z \in T} [z, b]$, resp.) is called a *polyblock* (*copolyblock*, resp.) with the vertex set T. A vertex z is called *proper* if there is no vertex $z' \neq z$ dominating z, i.e., such that $z' \geq z$ (dominated by z, i.e., such that $z' \leq z$). A vertex which is not proper is called an *improper* vertex. Obviously, a polyblock (copolyblock, resp.) is fully determined by its proper vertex set.

The set $P' = \bigcup_{z \in T} [a, z)$ ($Q' = \bigcup_{z \in T} (z, b]$, resp.) is called a *half-open polyblock* (*half-open copolyblock*, resp.) with the vertex set T. Similarly, we have the definitions of *proper* and *improper* vertex of a half-open polyblock (half-open copolyblock, resp.), and a half-open polyblock (half-open copolyblock, resp.) is fully determined by its proper vertex set.

Remark 1. The union of finitely many polyblocks (copolyblocks, half-open polyblocks or half-open copolyblocks, resp.) is also a polyblock (copolyblock, half-open polyblock or half-open copolyblock, resp.), see e.g. [28].

The following result is similar to Lemma 4 in Tuy et al. [29].

Proposition 2. (i) *If $a \leqq x < b$ then the set $[a, b) \setminus [x, b)$ is a half-open polyblock with vertices*

$$u^i = b + (x_i - b_i)e^i, i = 1, \ldots, m. \tag{9}$$

In other words, $[a, b) \setminus [x, b) = \bigcup_{i=1}^{m} [a, u^i)$.

(ii) *If $a < x \leqq b$ then the set $(a, b] \setminus (a, x]$ is a half-open copolyblock with vertices*

$$v^i = a + (x_i - a_i)e^i, i = 1, \ldots, m. \tag{10}$$

In other words, $(a, b] \setminus (a, x] = \bigcup_{i=1}^{m} (v^i, b]$.

B. Multiobjective combinatorial optimization problems

Consider a general MOCO problem with m objectives ($m \geq 2$):

$$\begin{aligned} \min \quad & f(x) = (f_1(x), f_2(x), \ldots, f_m(x)) \\ \text{s.t.} \quad & x \in \mathcal{X}, \end{aligned} \tag{11}$$

where, the feasible set $\mathcal{X}$ is nonempty bounded and discrete in the space $\mathbb{R}^m$, $f_i(x) : \mathcal{X} \to \mathbb{R}$ for $i = 1, 2, \ldots, m$ are m objectives. Denote $\mathcal{Y} = f(\mathcal{X})$ as the image of the feasible set $\mathcal{X}$. A feasible solution $\hat{x} \in \mathcal{X}$ is called *efficient or Pareto optimal* if there is no other $x \in \mathcal{X}$ such that $f(x) \leq f(\hat{x})$. If $\hat{x}$ is efficient then $f(\hat{x})$ is called a *nondominated point*. The set of all efficient solutions $\hat{x} \in \mathcal{X}$ is denoted by $\mathcal{X}_E$ and is called *the efficient set*. The set of nondominated points $\hat{y} = f(\hat{x}) \in \mathcal{Y}$, where $\hat{x} \in \mathcal{X}_E$, is denoted by $\mathcal{Y}_{ND}$ and is called *the nondominated set*.

The ideal point $y^I = (y^I_1, \ldots, y^I_m)$ defined by

$$y^I_i := \min_{x \in \mathcal{X}} f_i(x)\ \forall i = 1, 2, \ldots, m$$

is the lower bound of $\mathcal{Y}_{ND}$. The point $y^U = (y^U_1, \ldots, y^U_m)$ with

$$y^U_i := \max_{x \in \mathcal{X}} f_i(x)\ \forall i = 1, 2, \ldots, m$$

is an upper bound of $\mathcal{Y}_{ND}$ (see Ehrgott [10]). Let $b = y^U + e$ we then have $\mathcal{Y}_{ND} \subset [y^I, b)$.

IV. The search region

In this section, we first give the new interpretation for the search region of MOCO problems explicitly as half-open polyblocks and then relate to the corresponding concepts and results in the literature.

A. Explicit representation of the search region by half-open polyblocks

Let N be the finite set of known nondominated points of problem (11). Then the set $\cup_{y \in N} [y, b)$ is the part of objective space which dominates N and therefore not to be considered for further exploration of new nondominated points. Consequently, *the search region* $S(N)$ defined as the part of objective space where remaining nondominated points must lie (see e.g. [6], [16]), can be computed as follows.

$$S(N) = \begin{cases} [y^I, b) & \text{if } N = \emptyset \\ [y^I, b) \setminus (\bigcup_{y \in N} [y, b)) & \text{otherwise.} \end{cases}$$

$$\Leftrightarrow S(N) = \begin{cases} [y^I, b) & \text{if } N = \emptyset \\ \bigcup_{y \in N} ([y^I, b) \setminus [y, b)) & \text{otherwise.} \end{cases}$$

By Proposition 2 and Remark 1, $S(N)$ has the structure of a half-open polyblock which is explicitly defined by its proper vertex set $T(N)$, i.e.,

$$S(N) = \bigcup_{z \in T(N)} [y^I, z).$$

Note that, in the case $N = \emptyset$ we have the initial search region $S(\emptyset) = [y^I, b)$, and $T(\emptyset) = \{b\}$.

The key issue has been concerned with in literature and in this paper is updating the search region whenever a new nondominated point y^* is detected, i.e., determining the set $S(N \cup \{y^*\}) = S(N) \setminus [y^*, b)$ by computing $T(N \cup \{y^*\})$ from $T(N)$ and y^*.

B. Related concepts and results in the literature

The study about the search region was considered in Przybylski et al. [23] (2010) in their two-phase method to MOCO problems with the concept of *the search area*. However, it was really constructed systematically in Klamroth et al. [16] (2015). We now recall some important concepts in their work. Setting $Z = (y^I, b), \hat{Z} = [y^I, b]$ and denoting N be the *stable set* of Z i.e., for any $z^1, z^2 \in N$, $z^1 \not\leq z^2$ (remember that, in the setting of MOCO problems, if N is the finite set of nondominated points then it is obvious stable). The search region $S(N)$ in [16] are characterized by the set of local upper bounds $U(N)$

$$S(N) = \cup_{u \in U(N)} C(u),$$

where $C(u) = \{z \in Z : z < u\}$ is called *a search zone*, and $U(N) \subset \hat{Z}$ are determined as those points that (i) are not strictly dominated by any points in N and (ii) are maximal with this property (see Proposition 2.6 in [16] for more details). It is clear that there is one to one correspondence between the proper vertex set $T(N)$ (for defining our search region) and the set of local upper bounds $U(N)$. And every corresponding search zone is actually a half-open box $[y^I, u)$ which is a part of our half-open polyblock. Hence, from now on, in order to be consistent in presentation we will use $T(N)$ instead of $U(N)$ when describing the related results in [6], [16], [23].

As mentioned above, the issue of updating the search region of a general MOCO problem plays an important role for generating the whole set of nondominated points. It has been received a lot of attentions recently with some main approaches as follows.

(i) The first one constructs $T(N \cup \{y^*\})$ from $T(N)$ and a new nondominated point y^* by eliminating the improper vertex set of the half-open polyblock $S(N \cup \{y^*\})$. This method was considered by Przybylski et al. [23] (2010) and then was improved by Klamroth et al. [16] (2015) with enhanced filtering step (see Algorithm 3 in [16] for more details), says RE (redundancy elimination) algorithm. This algorithm updates $T(N \cup \{y^*\})$ from $T(N)$ and y^* based on redundancy elimination, i.e., eliminating the redundant local upper bounds by using filtering steps.

(ii) The second approach is based on theoretical properties of local upper bounds which is called RA (redundancy avoidance) algorithm (Algorithm 5 in [16]). According to this, for any $j = 1, \ldots m$, $u + (y^*_j - u_j)e^j \in T(N \cup \{y^*\})$ is characterized by a neccessary and sufficient condition, see Theorem 4.4 in [16] for more details. The filtering step can be avoided then.

(iii) The third one is based on the neighborhood structure among local upper bounds. It was proposed in Däechert and Klamroth [5] (2015) in the version of tricriteria problems. This algorithm is then further improved for general case of multiobjective optimization problems by Däechert et al. [6] (2016).

V. Solution method for solving MOCO problems

A. A generic method based on the representation of the search region

In this section, we present a generic method based on the solution of budget constrained programs for finding the whole set of nondominated points of MOCO problems. Based on the representation of the search region as a half-open polyblock (or a set of search zones as the interpretation of Klamroth et al. [16]), all nondominated points can be iteratively generated by solving the following mathematical program associated to each proper vertex u of the updated search region $S(N)$ ($u \in T(N)$, noting that $S(N)$ is shown to have the structure of a half-open polyblock with proper vertex set $T(N)$).

$$\min\{g(f(x)); x \in \mathcal{X}, f(x) < u\}, \qquad (P(u))$$

where g is any strongly increasing aggregation function of $f_j, j = 1, \ldots, m$. In this paper, we choose $g(f(x)) = \sum_{j=1}^{m} f_j(x)$. Interested readers can see [16] for more details. Below is the pseudocode of the algorithm generating nondominated set of an MOCO problem. It should be noticed that the important step in Algorithm 1 is Step 9 where the search region (here, the proper vertex set of the search region) is updated after detecting a new nondominated point. And we can use any method as mentioned in previous sections for updating the search region. Although the algorithm based on the neighborhood structure [6] is verified as the best performance compared to RA and RE algorithm under the assumption GP (general position)[1] of the stable set N but such a case is not realistic for many instances of MOCO problems. For NGP (non GP) case which appears in most of MOCO problems, Algorithm RA is better than the other ones for small number of points in nearly all the cases of p. Hence, in this paper, we use the RA procedure for Step 9 of Algorithm 1.

B. The choice of u and stack management

Regarding the choice of parameter u for each iteration we can although pick an arbitrary vertex from the set of proper vertices T but, in order to improve the efficiency of the algorithm, it is necessary to study to find an appropriate stack policy for managing the stored subproblems. As it is well known that, a stack can be managed with various policies such as FIFO (First in First out), LIFO (Last in First out), Priority and so on. In particular, for Algorithm 1 we will investigate four strategies of picking u from T (Step 3)

- FIFO: the first appended element is first considered;
- LIFO: the last appended element is first considered;
- Priority 1: u satisfies

$$u = \text{argmax}_{z \in T}\{W_z : W_z = \sum_{i=1}^{m} z_i\};$$

[1] N is a nonempty and stable set of points in *general position* i.e., for all $z, z' \in N, z \neq z'$, we have $z_i \neq z'_i$ for all $i = 1, ..., m$, see [6] for more details.

Algorithm 1: For finding the set $\mathcal{Y}_{ND}$

Input : Objective function $f(x) = (f_1(x), f_2(x), ..., f_m(x))$, feasible set $\mathcal{X}$.

Output: The whole set of nondominated points $\mathcal{Y}_{ND}$.

/* Initialize the set of proper vertices T of search region and the nondominated set $\mathcal{Y}_{ND}$. */

$T \leftarrow \{b\}$, $N \leftarrow \emptyset$.

while $T \neq \emptyset$ **do**

  /* Choose a proper vertex $u \in T$. */

  Select $u \in T$

  /* Solve problem $(P(u))$ to get an efficient solution $\tilde{x}$. */

  $\tilde{x} \leftarrow$ solving$(P(u))$

  if $\tilde{x}$ *is NaN* **then**

    /* There is no nondominated point in $[y^I, u)$, hence delete u from T. */

    $T \leftarrow T \setminus \{u\}$

  else

    /* Add a new nondominated point $f(\tilde{x})$ to N */

    $N \leftarrow N \bigcup \{f(\tilde{x})\}$

    /* Update the proper vertex set T by using the RA procedure */

    $T \leftarrow T(N)$

  end

end

Return $\mathcal{Y}_{ND} = N$

- Priority 2: u satisfies

$$u = \text{argmin}_{z \in T}\{W_z : W_z = \sum_{i=1}^{m} z_i\}.$$

Note that, the stack mangement for the generic method for solving MOCO problems has not been considered so far.

VI. Computational results

We implemented Algorithm 1 with four stack policies: FIFO, LIFO, Priority 1 and Priority 2 for finding the whole set of nondominated points of MSO problems.

All algorithms were coded in Matlab R2012a and ran on an PC with RAM 8G, Intel core i7 2.26 GHz and Win7-64bit operating system.

For solving each subproblem $(P(u))$, we used CVX 2.1 containing Gurobi Solver 6.0 (academic version at http://cvxr.com/cvx/ and http://www.gurobi.com/). We limited time for each instance to 7200s.

The first instance is the example of the supplier selection in MSO of CSA company given by Feng et al. [11]. The numerical results of Algorithm 1 for the example in [11] are shown in TABLE II. Note that, the heuristic algorithm proposed in [11] outputted five nondominated points while Algorithm 1 generated six nondominated points. As our approach is exact, these six points are all nondominated points of the example.

In addtion, we also randomly generated test problems according to the similar rules in [11] which can be divided into two groups.

- Group 1: the parameters p_{is}, t_{is}, u_{ij} were generated using random numbers from the uniform distribution $N(1, 10)$; and $p_s^A = 8, t_s^A = 8, d_{is} = 10$, i.e., all suppliers can provide any SPE.

Efficient solution	Z_1	Z_2	Z_3
$P_1 \to S_1, P_5 \to S_2, P_8 \to S_3, P_{12} \to S_4, P_{15} \to S_5$	1914.3	6.8	199.7
$P_1 \to S_1, P_5 \to S_2, P_8 \to S_3, P_{13} \to S_4, P_{15} \to S_5$	1917.2	6.6	189.0
$P_1 \to S_1, P_6 \to S_2, P_8 \to S_3, P_{12} \to S_4, P_{15} \to S_5$	1914.4	6.6	181.7
$P_1 \to S_1, P_6 \to S_2, P_8 \to S_3, P_{13} \to S_4, P_{15} \to S_5$	1917.4	6.4	173.7
$P_1 \to S_1, P_5 \to S_2, P_8 \to S_3, P_{12} \to S_4, P_{15} \to S_5$	1916.4	6.5	171.7
$P_1 \to S_1, P_5 \to S_2, P_8 \to S_3, P_{13} \to S_4, P_{15} \to S_5$	1919.4	6.3	165.7

TABLE II: Computational results of Algorithm 1 for the CSA example from Feng et al. [11]

- Group 2: Analogous to Group 1, however p_{is}, t_{is} were both set to be $L = 10000$ for some randomly generated places of the matrices $\{p_{is}\}_{i=1,...,M}^{s=1,...,K}$, $\{t_{is}\}_{i=1,...,M}^{s=1,...,K}$, i.e., the candidate supplier i cannot provide SPE s.

The computational CPU time on the set of randomly generated instances are reported in TABLE III. For each problem size we reported the average of the running time of each algorithm over 5 instances. These results show that the performance of Algorithm 1 for MSO problems depends on stack management. The Priority 2 stack results to be efficient for Algorithm 1 in all the cases of Group 1 (we emphasized such these cases by bold numbers). For Group 2, Priority 1 and Priority 2 take turns in the best performance for Algorithm 1.

	K	M	$\lvert\mathcal{Y}_{ND}\rvert$	CPU time (s)			
				Prio. 1	Prio. 2	LIFO	FIFO
Group 1	4	10	25.6	59.5	**59.4**	66.4	62.4
		11	25	144.2	**81.4**	164.5	114.4
		12	43	264	**214.4**	260.8	268.6
		13	34.2	213.9	**210.3**	210.6	211.5
		14	36.6	581.2	**455.9**	557.8	565.5
		15	44.6	695.6	**444.5**	722.1	614.6
	5	10	39.8	1584.6	**1398.5**	1592.8	1446.9
		11	53.4	1717	**1317.3**	1866.8	1657.7
		12	50	3504.2	**3239.3**	3693.8	3259.5
Group 2	5	14	13.6	**23.8**	24.6	27.5	26.2
		15	37	421.8	**326.9**	404.7	367.1
		16	25.2	**224.9**	262.6	303.5	246.1
	6	10	29.6	217.7	**187.4**	197.3	223.1
		11	14.6	49.2	**48.8**	52.6	50.4
		12	17.6	**102.5**	136.2	153.2	106.4
		13	15.2	**103.6**	106.6	110.4	117.6
	7	10	36.2	**492.3**	494.6	515.3	525.7
		11	35.2	1084.6	**994.1**	1100.5	1072.3
		12	21.8	**345.2**	363.6	410.6	398.4

TABLE III: Computational results of Algorithm 1 with different stack policies

VII. Conclusion

We have proposed an exact approach for solving the supplier selection problem in multi-service outsourcing which was modelled by Feng et al. [11] (2011). The considered problem is in the form of binary nonlinear multiobjective

programming which is hard to be solved. Unlike heuristic approach proposed in previous work [11], the globality of our algorithm can generate the whole set of nondominated points of MSO problem.

The second contribution of the paper is to the solution method for general multiobjective combinatorial optimization problems. We give a new concise description of the search region of multiobjective combinatorial optimization (MOCO) problems by half-open polyblocks for the minimization problems and half-open copolyblocks for the case of maximization problems. The generic method based on the representation of the search region for generating all nondominated points of MOCO problems has been considered with different stack management policies in this paper too. By the obtained numerical results of the investigation of stack policies we have pointed out the suitable stack management for solving MSO problems.

ACKNOWLEDGMENT

The authors wish to thank the anonymous referees for their useful comments which helped them to improve the paper. The first author would like to thank National Foundation for Science and Technology Development (Nafosted) fund in Vietnam for supporting her to study this research.

REFERENCES

[1] W. Anton, H. Evi, B. W. C, H. Nils-Ole, *A Systematic Assessment of Supplier Selection Literature - State-of-the-Art and Future Scope. Intern. Journal of Production Economics.* DOI: http://dx.doi.org/10.1016/j.ijpe.2016.06.022, 2016.

[2] M. Antelo, L. Bru, *Outsourcing or restructuring: the dynamic choice.* International Journal of Production Economics 123, 1-7, 2010.

[3] K.S Bhutta, F. Huq, *Supplier selection problem: a comparison of the total cost of ownership and analytic hierarchy process approaches.* Supply Chain Management: An International Journal 7(3), 126-135, 2002.

[4] M. Braglia, A. Petroni, *A quality assurance-oriented methodology for handling trade-offs in supplier selection.* Int. J. Phys. Distrib. Logist. Manag. 30(2), 96-111, 2000.

[5] K. Därchert, K. Klamroth, *A linear bound on the number of scalarizations needed to solve discrete tricriteria optimization problems.* J. Glob. Optim. 61(4), 643-676, 2015.

[6] K. Dächert, K. Klamroth, R. Lacour and D. Vanderpotten, *Efficient computation of the search region in multi-objective optimization.* Eur. J. Oper. Res., available online, 2016.

[7] L. De Boer, E. Labro, P. Morlacchi, *A review of methods supporting supplier selection.* European Journal of Purchasing and Supply Management 7 (2), 75-89, 2001.

[8] Z. Degraeve, E. Labro, F. Roodhooft, *An evaluation of supplier selection methods from a total cost of ownership perspective.* European Journal of Operational Research 125 (1), 34-58, 2000.

[9] E.A. Demirtas, Ö. Üst ün, *An integrated multiobjective decision making process for supplier selection and order allocation.* Omega 36, 76-90, 2008.

[10] M. Ehrgott, *Multicriteria Optimization.* Springer, Berlin, 2005.

[11] B. Feng, ZZ.P. Fan, Y. Li, *A decision method for supplier selection in multi-service outsourcing.* Int. J. Production Economics 132, 240-250, 2011.

[12] O. Jadidi, S. Cavalieri, S. Zolfaghari S, *An Improved Multi-Choice Goal Programming Approach for Supplier Selection Problems.* Appl. Math. Modelling , $doi : http : //dx.doi.org/10.1016/j.apm.2014.12.022$, 2014.

[13] I.M Joan and M.C. Ana, *Supplier Evaluation and Selection: A Review of the literature since 2007.* Managing Complexity, Lecture Notes in Management and Industrial Engineering, DOI $10.1007/978-3-319-04705-8_25$, Springer International Publishing Switzerland, 2014.

[14] C. Junyi, N.K.L. James, W.T.N. Eric, *Application of decision-making techniques in supplier selection: A systematic review of literature.* Expert Systems with Applications 40, 3872-3885, 2013.

[15] G. Kirlik, S. Sayın, *A new algorithm for generating all nondominated points of multiobjective discrete optimization problems.* Eur. J. Oper. Res. 232: 479-488, 2014.

[16] K. Klamroth, R. Lacour and D. Vanderpooten, *On the representation of the search region in multi-objective optimization.* Eur. J. Oper. Res. 245(3): 767-778, 2015.

[17] F. Hamed, M. Iraj, T.A. Minoo, K. Somayeh and M-A. Nezam, *A multi-objective decision-making process of supplier selection and order allocation for multi-period scheduling in an electronic market.* Int. J. Adv. Manuf. Technol. 52, 1039-1052, 2011.

[18] W. Ho, X. Xu and P. Dey, *Multi-criteria decision making approaches for supplier evaluation and selection: A literature review.* Eur. Jour. of Oper., 2010.

[19] I. McCarthy, A. Anagnostou, *The impact of outsourcing on the transaction costs and boundaries of manufacturing.* International Journal of Production Economics 88, 61-71, 2004.

[20] R. Mansini, M. W. P. Savelsbergh, B. Tocchella, *The supplier selection problem with quantity discounts and truckload shipping.* Omega Int. J. Manage. Sci. 40(4), 445-455, 2012.

[21] A. Mandal, S.G. Deshmukh, *Vendor selection using interpretive structural modelling (ISM)* Int. J. Oper. Prod. Manage14(6), 52-59, 1994.

[22] A. Mendoza, J. A. Ventura, *Analytical models for supplier selection and order quantity allocation.* Appl. Math. Modell. 36(8), 3826-3835, 2012.

[23] A. Przybylski, X. Gandibleuc and M. Ehrgott, *A two phase method for multi-objective integer programming and its application to the assignment problem with three objectives.* Discrete Optimization 7(3), 149-165, 2010.

[24] N.T.H. Phuong, H. Tuy, *A Unified Monotonic Approach to Generalized Linear Fractional Programming.* J. Glob. Optim. 26(3), 226-259, 2003.

[25] J. Sarkis, S. Talluri, *A model for strategic supplier selection* J. Supply Chain Manag 38(1), 18-32, 2002.

[26] W. C. Tsai, *Order allocation for multi-item sourcing with supply disruptions in shipment quality and delivery.* International Journal of Logistics Research and Applications: A Leading Journal of Supply Chain Management, DOI: 10.1080/13675567.2015.1012153, 2015.

[27] H. Tuy, *Normal Sets, Polyblocks, and Monotonic Optimization.* Vietnam Journal of Mathematics 27(4), 277-300, 1999.

[28] H. Tuy, *Monotonic optimization: Problems and solution approaches.* SIAM J. Optim. 11, 464-494, 2000.

[29] H. Tuy, M. Minoux, N.T.H. Phuong, *Discrete monotonic optimization with application to a discrete location problem.* SIAM J. Optim. 17(1), 78-97, 2006.

[30] C.A. Weber, J.R. Current, W.C. Benton, *Vendor selection criteria and methods.* European Journal of Operational Research 50 (1), 2-18, 1991.

[31] C.A. Weber, L.M. Ellram, *Supplier selection using multiobjective programming: a decision support system approach.* International Journal of Physical Distribution & Logistics Management, 23 (2), 3-14,1993 MCB University Press, 0960-0035.

Supply Chain Coordination and Risk Management

A fuzzy approach for supply chain vulnerabilities assessement: Application to an agrifood supply chain

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Elleuch Hatem
Unit of Logistic, Industrial and Quality Management
Faculty of Economic and Management Sciences,
University of Sfax
Sfax, Tunisia
elleuch_hatem@hotmail.com

Dafaoui El Mouloudi,
El Mhamedi Abderrahman
QUARTZ Laboratory EA 7393
University Institute of Technology
University of Paris 8
Paris, France
{e.dafaoui, a.elmhamedi}@iut.univ-paris8.fr

Chabchoub Habib
Unit of Modeling and decision-making in organizations
Institute of Higher Business Studies
University of Sfax
Sfax, Tunisia
Habib.chabchoub@fsegs.rnu.tn

***Abstract*—Supply chain managers always look for invulnerable supply chains with minimum risks and more resilience. This aspiration is all the more insistent when it comes to an agrifood supply chain with a complex and large volume flow in addition to the constraints related to processes, safety and business. Likewise, any successful strategy for resilience improvement is conditioned by an effective analysis of supply chain vulnerabilities. In this paper, a methodology is provided for dealing with vulnerabilities in an agrifood supply chain. The scope of this work is to incorporate fuzzy linguistic modeling in FMECA (Failure mode, effects and criticality analysis) method to handle imprecise judgments in the supply chain vulnerability evaluation matrix. The proposed approach offer a framework for identifying supply chain vulnerability factors using Ishikawa diagram and fuzzy FMECA for assessing their criticalities. The proposed approach is illustrated through a real agrifood supply chain case study in one of SMEs in Tunisia.**

Keywords—supply chain; agrifood midstream, vulnerability; fuzzy failure mode, effects and criticality analysis

I. Introduction

As a result of changes in the economic, business and ecological environments, modern supply chains seem to be more vulnerable for several reasons (increased disasters in number and in intensity, supply chain complexity, competitive pressures, etc.). Supply chains are becoming increasingly complex, leading to increased interdependencies among companies or individuals, and can be disrupted in myriad ways in today's rapidly changing global business environment [1]. According to [2], the focus on efficiency and the complexity of global supply chains makes them vulnerable. In today's uncertain environment, these complex networks of suppliers, manufacturers, and customers are vulnerable to internal and external disruptions and may be disrupted in many ways. Many disruptions and risk factors have threatened production and retail distribution systems. Moreover, the supply chain disruptions may occur in a variety of forms. The threats to the supply chain have been indeed sharpened by long and global supply chain, by products with increasingly reduced life cycles and by volatile and unpredictable markets. Unfortunately, there is no possibility to avoid these risks, thus the supply chain resilience and vulnerability have become a particularly important issues for many companies [3]. So, the challenge for current companies is to manage and mitigate these risks through the creation of more resilient supply chains, able to opportunely and efficaciously face unexpected events. Predicting and mitigating system vulnerabilities and designing appropriate interventions can lead to the development of more resilient systems, capable of delivering a sustained level of value. Thus, supply chain managers need to have the ability to measure and quantify the vulnerability of their supply chains [4]. With respect to complexity and the type of considered management problem and tacking account for the vagueness in decision-making, an integrated fuzzy FMECA approach for supply chain vulnerability assessment is proposed in this paper.

The remaining sections of this paper are organized as follow. In section 2, literature review covering the concept of supply chain vulnerability and existing methods for assessing supply chain vulnerability is presented. In section 3, a description of methodology is provided and theoretical background is presented. In section 4, application of the proposed approach is presented in the agrifood industry case study. Finally, section 5 provides a conclusion and potential extensions for future researches.

II. Literature Review

In general, vulnerability is the degree to which a system is susceptible to, or unable to cope with the adverse effects of environmental changes. The vulnerability of any system at any scale reflects the exposure and sensitivity of that system to hazardous conditions and the ability, capacity, or resilience of the system to cope, adapt, or recover from the effects of those conditions. It is characterized by the capacity of the system and its preparation to face the hazard or anticipated consequences. As in [5], vulnerability is defined as "The potential for a system to be affected by internal and / or external hazards. Vulnerability is determined from three characteristics: the predisposition to risk, elasticity to withstand shock and strength-building [6].

In the case of supply chains, there are many causes of increased vulnerability. Supply chain vulnerability is an inherent trait of the supply chain, which is determined by the structure and characteristics of the supply chain itself. The supply chains were previously local and integrals, they have

become global and modular [7]. Supply chain vulnerability has been defined broadly by [8] as ‘‘an exposure to serious disturbance, deriving by risks inside the supply chain, as well as to risk outside the supply chain’’ and in [9] as ‘‘the propensity of risk sources and risk drivers to outweigh risk mitigating strategies, thus causing adverse supply chain consequences’’. This concept has been defined by [10] and [11] as “a sensitivity to external or internal events, caused by lack of robustness or resilience”. According to these authors, the vulnerability could be seen as a set of system (i.e. supply chain) characteristics that could be used to evaluate the level of exposure to risks and the severity of its impacts by measuring the system sensitivity and ability to deal with disruptive events".

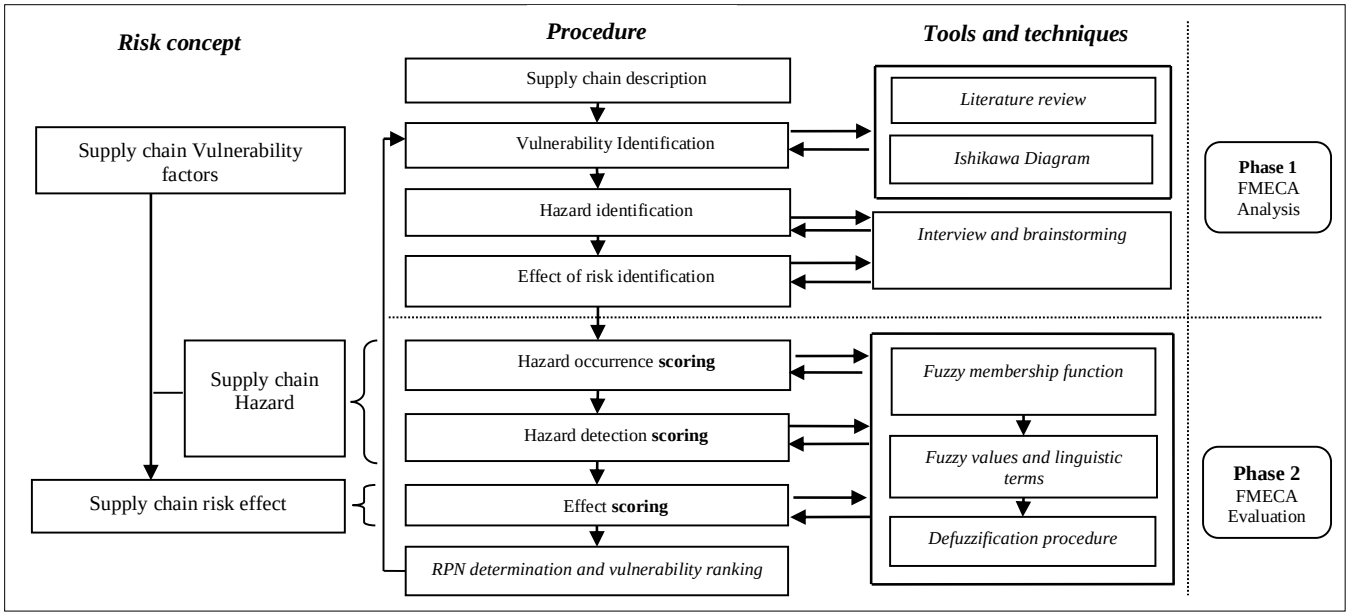
Analyzing the vulnerability in industrial supply chain system is presented as a means to reduce the risk, to become better prepared to manage the system‘s vulnerabilities and to improve the system’s resilience [12]. Measurement of supply chain vulnerability is by and large regarded as difficult because it is multi-dimensional and there are no well-developed metrics for evaluating the factors on which vulnerability depends [13]. Qualitatively, a vulnerability assessment is considered the one-step of a general risk assessment [14]. The vulnerability assessment typically occurs after hazard and threat assessment but before risk assessment and involves the identification of potential vulnerabilities. A vulnerability assessment should address whom or what is vulnerable, along with how it is vulnerable [15]. The two main outputs of a critical infrastructure vulnerability assessment are the identification of critical elements and the quantification of system vulnerability indicators [14].

The problematic of supply chain vulnerability assessment has been given little attention in the literature as reviewed in [16]. Reference [17] proposed an analytical approach for assessing and quantifying vulnerability index for control and monitoring using graph theory. Reference [18] proposed a conceptual study to focus on assessing vulnerability from different perspectives to measure vulnerability index of different supply chain structures. The same, there are few studies that have investigated vulnerability assessment on real case studies. Reference [19] proposed an approach for evaluating the vulnerability by measuring the impact of a disturbance on the system performance. This evaluation is based on two performance indicators namely the time related performance indicator and magnitude related performance indicator. Reference [20] have used pairwise comparison of AHP method to allow the decision maker to evaluate vulnerability factors in terms of criticality. To conclude, integrated qualitative and quantitative approaches for assessing vulnerabilities in supply chain are limited especially those focusing on real case studies. Also there is a lack of vulnerability assessment approaches that consider all the dimensions of risk, namely vulnerability, disruption and the effects. Considering the above conditions, this research present an integrated approach for supply chain vulnerability assessment using Ishikawa diagram as a structural method for identifying the vulnerability in supply chain and fuzzy FMECA method for identifying, evaluating and analyzing related concepts of vulnerability and impacted risks. Also fuzzy approach is used to take into consideration the vagueness of decision maker’s judgments. The proposed approach is applied to a real case study in an agrifood supply chain in Tunisia.

III. Proposed Approach

The problematic of this research consists of determining a new approach for supply chain vulnerabilities identification and assessment. The proposed approach illustrated in (fig. 1) is based on the conceptual framework of the risk process described in [16]. The approach framework show the concept of risk, the procedure steps and used tools and techniques.

Fig. 1.Conceptual framework for supply chain vulnerability assessment



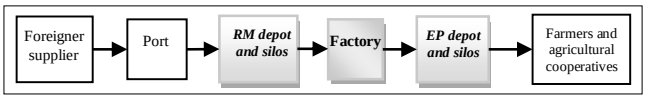
As illustrated in fig. 1, Vulnerability is the drivers of risk, consecutively, supply chain hazard or disruption represent the trigger that lead to the supply chain risk occurrence and subsequently the risk effects. Following this concept, the proposed approach consist of identifying the vulnerability factors and then, assessing their criticality by evaluating their related concepts namely the potential hazard and the potential risk effects. These latter determines the overall risk severity and give insight on the criticality of each corresponding vulnerability factors. The proposed FMECA approach is based on two main phases. First phase named FMECA analysis, include supply chain description, identification of vulnerability factors based on literature review and Ishikawa diagram and finally the identification of potential related hazard and related risk effects based on interviews and brainstorming technique with the technical staff of the agrifood company. The task of vulnerability analysis is to identify the factors that have the biggest impact on the performance state of supply chain in the future. Ishikawa diagram is used for this purpose as suggested in reference [21]. The second phase named FMECA evaluation consists of evaluating the hazard occurrence, the hazard detection and the risk effects. This evaluation is achieved by the decision maker judgments. The product of the variables values determine the risk severity and thus reflecting the criticality of the corresponding vulnerabilities. The vulnerability assessment is apprehended through the severity of the overall risk effects. This second phase employs fuzzy theory to deal with subjective decision maker judgments. The vector of critical weights of supply chain vulnerability factors with respect to each business category according to the five M of Ishikawa diagram is given by using fuzzy FMECA. Trapezoidal fuzzy numbers models the decision maker's judgments of supply chain vulnerability factor values. First, decision maker is solicited to give their judgment about the FMECA variables for each vulnerability factor using linguistic terms. Then, decision matrix will be converted into defuzzified values. Then, the risk priority number correspondent to each vulnerability factor is calculated as the multiplication of the fuzzy rating values of hazard occurrence, hazard detection and risk effects.

IV. Supply Chain Vulnerability Assessment: Case study

A. Supply chain description

The methodology will be illustrated by a case study conducted in an agrifood company in Tunisia. The supply chain under study can be modeled as subsystems of the same company (fig. 2). To simplify the illustration, we will only consider the factory and the immediate connection namely in upstream, deposits and storage silos for raw materials and in downstream, finished products storage depots and silos. Only physical and informational flows are considered. The operational policy of the company is under the push paradigm. Hence, the process begins with importing large cargo of grain for livestock feed. These cereals arrived at the port are transported to the depots and silos respectively for medium and short term storage. The cereals are then processed to make compound feed. The latter are then transferred to storage warehouses of finished products before being delivered to customers, namely, farmers and agricultural cooperatives.

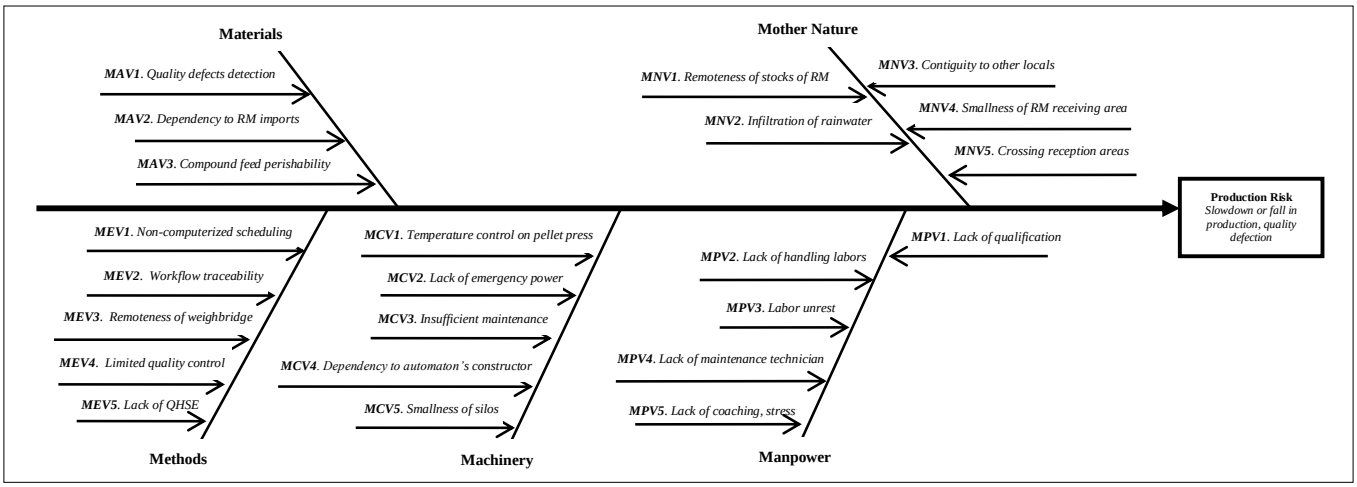
Fig. 2. Agrifood supply chain



B. Vulnerability identification

This step is achieved by exploring literature review on supply chain vulnerability factors and brainstorming technique with technical staff of the company. Fig. 3 shows the vulnerabilities related production risks namely the quality defection and the slowdown or fall in production. Identified vulnerabilities are classified into the five classes of Ishikawa diagram method. The result show three material vulnerability factors, five mother-nature vulnerability factors, five methods vulnerability factors, five machinery vulnerability factors and five manpower vulnerability factors. The identified vulnerability factors are examined through the fuzzy FMECA method in the next section.

Fig. 3. Ishikawa diagram for vulnerability factors identification [16]

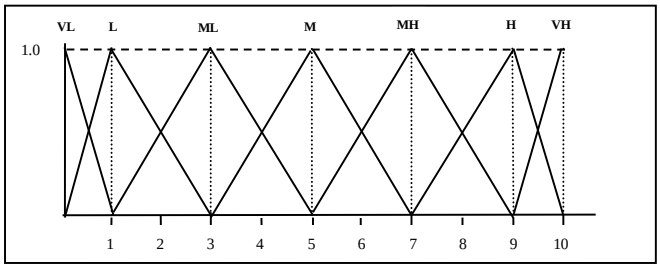


C. Fuzzy FMECA Approach

FMECA is a widely used engineering technique typically used for providing information for risk management decision-making [22]. A typical FMECA is consisted of the following components; the identification and listing of failure modes and the consequent faults, assessment of the chances of the occurrence of faults, then assessment of the chances of the detection of faults, assessment of the occurrence of the hazard or risk and assessment of the severity of the consequences of the failure mode. Each failure mode can be evaluated by three factors as severity, likelihood of occurrence, and the difficulty of detection of the failure mode. In a typical FMEA evaluation, a number between 1 and 10 (with 1 being the best and 10 being the worst case) is given for each of the three factors. By multiplying the values for severity (S), occurrence (O), and detectability (D), a risk priority number (RPN) is obtained, which is RPN = S*O*D [23]. Then the RPN value for each failure mode is ranked to find out the failures with higher risks.

The crisp values of traditional FMECA method have been criticized extensively for different reasons, mainly its subjectivity. For that reason, we consider fuzzy FMECA using fuzzy number for the calculation of different set of (O), (S) and (D). We use linguistic variables for rating the O, S and D variables. The linguistic variable is a convenient way for decision makers to express their assessments, which is very useful in dealing with ill-defined situations. These linguistic variables can be expressed in triangular numbers as shown in fig. 4. The following linguistic variables describe seven evaluation levels namely very low (VL), low (L), medium low (ML), medium (M), medium high (MH), high (H) and very high (VH).

Fig. 4. Membership function for O, S and D variables



Linguistic terms have to be quantified through defuzzification procedure. This latter is an important step in fuzzy approach, which converts fuzzy numbers into crisp scores. Many different techniques for such defuzzification are available, such as the mean of maximal (MOM), center of area (COA) and converting fuzzy data into crisp scores (CFCS) [24]. In this paper, the canonical representation of operation on triangular fuzzy numbers [25], which is based on the graded mean integration representation method, is used for the defuzzification procedure. The graded mean integration representation of triangular number $\tilde{a} = (a_1, a_2, a_3)$ can be expressed by:

$$\tilde{x}_0(\tilde{a}) = \frac{1}{6}(a_1 + 4a_2 + a_3) \quad (1)$$

, where $\tilde{x}_0(\tilde{a})$ is the defuzzified value.

Fuzzy numbers and their correspondent crisp value are given in table 1 according to each linguistic term. Crisp values are calculated using equation (1).

TABLE I. Linguistic Variables [25]

Linguistic variables	**Fuzzy numbers**	**Crisp value** $\tilde{x}_0(\tilde{a})$
Very Low (VL)	(0,0,1)	0.166
Low (L)	(0,1,3)	1.166
Medium low (ML)	(1,3,5)	3
Medium (M)	(3,5,7)	5
Medium high (MH)	(5,7,9)	7
High (H)	(7,9,10)	8.833
Very high (VH)	(9,10,10)	9.833

After providing evaluation for each FMECA variables (O), (S) and (D) using linguistic terms, this latter are converted into correspondent crisp values. Then, RPN for each vulnerability factors are calculated by multiplying said variables. Results are presented in table 2. The result show that dependence to the automaton's constructor, dependence on imports of large cargo of raw materials, the trade union strength and labor unrest, the remoteness of raw material plant depots and the non-computerized production scheduling are respectively the most critical vulnerability factors that lead to the most severe risks. While the least critical vulnerable factors are lack of qualification of staff, rainwater infiltration vulnerability at the plant, lack of handling workers, lack of coaching stress at work and lots of noise and remoteness of factory weighbridge. It is worth mentioning that the decision-makers were very satisfied with the results of the research and with the general ranking of vulnerability factors

As a continuity of this research, managers should consider the most critical supply chain vulnerability factors. The manager regards vulnerability factors that lead to the most severe risks that have the highest RPN in the FMECA matrix to then, design the appropriate strategy of supply chain resilience. The number of considered vulnerability factors falls within the preferences of decision maker.

V. Conclusion

As the dynamic and complexity of global business increases, risks and vulnerabilities tends to grow throughout the overall supply chains. This context involve implementing an effective supply chain risk management strategy. For this purpose and in particular, vulnerability assessment is the basis of any supply chain risk management approach. To deal with this context, an integrated approach for identifying and evaluating vulnerability factors in a fuzzy decision making environment is developed. The proposed approach consists of using Ishikawa diagram for vulnerability factors identification and fuzzy FMECA for assessing their criticalities. It admit that the most critical vulnerabilities are those leading to the most severe risks. The integrated proposed framework combining Ishikawa diagram and fuzzy FMECA provide a novel approach for supply chain vulnerability assessment and management.

TABLE II. Fuzzy FMECA for Supply Chain Vulnerability Assessment

	Cause (Vulnerability)	Hazard Occurrence	Hazard	Hazard detection	Effect (Risk consequence)	Risk severity	RPN
Material	MAV1: Detection Problem of raw material quality defects	MH	Arrival of raw material with quality defects	ML	Slower production and degradation of end product quality	ML	63.00
	MAV2: Dependence on imports of large cargo of RM	M	Damage to the ship, strike at the port, delay in loading with the supplier	H	Slowing down or stopping production	VH	433.99
	MAV3: Perishability of compound feed	MH	Increase in demand for End product	L	Restructuring constraint and delay in delivery	VH	80.23
Mother-nature	MNV1: Remoteness of raw material plant depots	ML	Hazard in the transport of dump trucks (e.g. strike, etc.)	H	Slowdown and potential shutdown of production	VH	260.40
	MNV2: Rainwater infiltration vulnerability at the plant	L	Torrential rains	ML	Wiring, raw material stocks in silos damage and slowing production	ML	10.49
	MNV3: Contiguity of the plant to other storage buildings	L	Fire in adjacent spaces	M	Factory fire and production stoppage	H	51.48
	MNV4: Smallness of raw material reception area and inability to receive big truck	H	Lack of raw materials and forced to bring it to distant suppliers	L	Late arrival of raw material, slowdown in production	MH	72.07
	MNV5: Connecting spaces of raw material reception and end product located in the factory	ML	Receipt of contaminated raw material	ML	Contamination of end products by insect pests	M	45.00
	MNV6: Remoteness of factory weighbridge	M	Congestion in contiguous space of the factory	L	Slowdown in production	MH	40.81
Methods	MEV1: Non-computerized production scheduling	MH	Human errors during planning or execution	M	Planning errors, poor quality PF, loss of time and delay in delivery	MH	210.00
	MEV2: Absence of workflow Traceability	M	Information transmission errors	ML	Qualitative or quantitative production errors	MH	105.00
	MEV3: Low quality control of batches of end product	ML	End product batches with quality defects	M	Delivery of degraded quality of end product to customers and impact on brand image	MH	105.00
	MEV4: Absence of a quality system , hygiene, safety and environment	ML	QHSE Failure	ML	Slower production and degradation of quality	M	45.00
Machinery	MVC1: Absence of computerized control on the temperature of pellet press	ML	Temperature rise or failure at the pellet press	M	Fire on the pellet press	VH	147.45
	MCV2: Absence of current generator and insufficient maintenance of the transformer	ML	Power failure from state electricity company or due to transformer failure	MH	Production shutdown	VH	206.43
	MCV3: Insufficient maintenance material of dosage and routing equipment	M	Wear of equipment	M	Slowdown and stoppage of production	MH	175.00
	MCV4: Dependence to the automaton's constructor	M	Failure of the programmable logic control	VH	Production shutdown	VH	483.14
	MCV5: Smallness of end product storage silos	H	Increase in end product demand or change in production chain	ML	Production slowing down	M	132.45
Manpower	MPV1: Lack of qualification of staff	MH	Overwork or lack of staff	VL	Qualitative or quantitative production errors	M	5.81
	MPV2: Lack of handling workers	M	Production program loaded	L	Slowdown of production	M	29.15
	MPV3: Trade union strength and labor unrest	MH	Claim and conflict between union and employer	M	Slowing down or stopping production	VH	344.05
	MPV4: Lack of maintenance technician	H	Electrical and mechanical failure	L	Slowing down or stopping production	MH	72.07
	MPV5: Lack of coaching, stress at work and lots of noise	M	High volume of work	L	Slowdown and production errors	M	29.15

This proposed framework enables continuous assessment of supply chain vulnerabilities level to allow to implement the most relevant actions resilience in a second time. The proposal is illustrated through a real agrifood supply chain case study in one of SMEs in Tunisia that make animal feed products. This study can be spread to other domains of decision making for evaluating and ranking vulnerability factors. As further research and improvements to the present approach, well-known fuzzy numbers such as Gaussian fuzzy numbers could be used to better refine the results. Also, it seems interesting to consider the use of multi-criteria decision methods and simulation to compare the performance of supply chain systems with different supply chain vulnerability scenarios. Other ranking methods could be used in the proposed approach like Shannon's entropy method. In the other hand and as an extension to this research, a development of a supply chain vulnerability management could be developed by simultaneously reducing vulnerability factors and increasing resiliency capacities while respecting performance and efficiency constraints.

Acknowledgment

The authors would like to express their thanks to all the personal of the factory ALCO and especially the factory manager for their support during this research and their help in the data collecting stage.

References

[1] T. Nowakowski, S. Werbińska-Wojciechowska & M. Chlebus "Supply chain vulnerability assessment methods—possibilities and limitations," Safety and Reliability of Complex Engineered Systems, 2015–ISBN 978-1-138-02879-1

[2] C.W. Craighead, J. Blackhurst, M.J Rungtusanatham, R. B Handfield, "The Severity of Supply Chain Disruptions: Design Characteristics and Mitigation Capabilities,". Decision Sciences, 2007, 38(1), pp. 131-156.

[3] T. Murino, E Romano, L.C Santillo "Supply chain performance sustainability through resilience function,". Proceedings of the Winter Simulation Conference, 2011, pp. 1605–1616.

[4] S.M. Wagner and N. Neshat, "A comparison of supply chain vulnerability indices for different categories of firms,". International Journal of Production Research, 2012, vol. 50. pp. 2877-2891.

[5] F. Petit, "Concepts d'analyse de la vulnérabilité des infrastructures essentielles- prise en compte de la cybernétique," Montreal Polytechnic Thesis. 2009

[6] C. Gondard-Delcroix, and S. Rousseau "Vulnérabilité et stratégies durables de gestion des risques : une étude appliquée aux ménages ruraux de madagascar". DOI: 10.400/développement durable, 2004, 1143.

[7] J.J. Laville "Comment sécuriser sa supply chain?" Logistique & Management 2006, Vol. 14 – N°1.

[8] M. Christopher, and H. Peck, "Building the resilient supply chain," International Journal of LogisticsManagement, 2004, vol. 2, no. 15, pp. 1–13.

[9] U. Jüttner, H. Peck & M. Christopher, "Supply chain risk management: outlining an agenda for future research," International Journal of Logistics Research and Applications: A Leading Journal of Supply Chain Management, 2003, vol 6:4, pp 197-210.

[10] V. Albino and A.C Garavelli, "A methodology for the vulnerability analysis of just-in -time production systems,". In: International journal of production economics, 1995, Vol. 41, No.1 , pp. 71-80.

[11] B.E. Asbjornslett, "Assess the vulnerability of your production system,". In: Production Planning & Control, 1999, vol.10, No.3 , pp.219-229.

[12] H. Elleuch, E.M. Dafaoui, A. El Mhamedi and H. Chabchoub, "A Quality Function Deployment approach for Production Resilience improvement in Supply Chain: Case of Agrifood Industry,", 12th IFAC Conference on Intelligent Manufacturing Systems, (IMS 2016), 05-07 December, 2016, Austin, USA.

[13] S.M. Wagner, and C. Bode, "An empirical investigation into supply chain vulnerability,". Journal of purchasing and supply management, 2006, pp. 301-3012.

[14] W. Kröger, E. Zio, "Vulnerable systems" ISBN: 978-0-85729-654-2 (Print) 978-0-85729-655-9 (Online), 2011.

[15] C.M. Schnaubelt, E.V Larson & M.E. Boyer, "Vulnerability Assessment Method Pocket Guide: A Tool for Center of Gravity Analysis". Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/tools/TL100/TL129/RAND_TL129.pdf, 2014.

[16] H. Elleuch, E.M. Dafaoui, A. El Mhamedi and H. Chabchoub, " Resilience and Vulnerability in Supply Chain: Literature review,", 8th IFAC Conference on Manufacturing Modelling, Management and Control (MIM 2016), 28-30 Juin 2016, University of Technology of Troyes, France.

[17] S.M. Wagner and N. Neshat, "Assessing the vulnerability of supply chains using graph theory,". International journal of production economics, 2010, 126. pp. 121-129.

[18] B.D. Neureuther and G. Kenyon, "Mitigating supply chain vulnerability,". Journal of marketing channels, 2009, 16. pp. 245-263.

[19] J.V. Vlajic, S.W.M. Lokven, R. Haijema and G.A.J. Vandervorst "Using vulnerability performance indicators to attain food supply chain robustness". Production Planning & Control, 2012

[20] M.M..H. Chowdhury, M.A. Quaddus, "A multiple objective optimization based QFD approach for efficient resilient strategies to mitigate supply chai vulnerabilities: The case of garment industry of bangladesh" Omega, 2015, vol 57, pp 5-21.

[21] B.E. Asbjornslett, "Assessing the vulnerability of supply chains. In: Supply chain risk". A Handbook of Assessment, Management and Performance. Zsidisin G.A., Ritchie B. (eds.). Springer Science+Business Media, LLC, NY, USA, 2009.

[22] A. Pillay and J. Wang, "Modified failure mode and effects analysis using approximate reasoning," Reliability Engineering and System Safety, 2003, 79, pp 69–85.

[23] K.S Chin, A. Chan & J.B. Yang, "Development of a fuzzy FMEA based product design system". The International Journal of Advanced Manufacturing Technology, 2008, 36, 633–649.

[24] H.C. Liu, J.X. You, Q.L. Lin & H. Li, "Risk assessment in system FMEA combining fuzzy weighted average with fuzzy decision-making trial and evaluation laboratory," International Journal of Computer Integrated Manufacturing, 2014.

[25] C. Chou, "The Canonical Representation of Multiplication Operation on Triangular Fuzzy Numbers," Computers & Mathematics with Applications, 2003, 45 (10–11): 1601–1610.

Identifying and ranking of clearance chain management (CCM) strategies

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Rihab Mechmech, Slim Harbi, Atidel Hadj-Alouane
OASIS Laboratory - ENIT
University of Tunis El Manar
Tunis, Tunisia
slim.harbi@enicarthage.rnu.tn
atidel.hadj@enit.rnu.tn
ryhabmechmech@gmail.com

Sami Sboui
SQLI services
Manouba, Tunisia
JASSP SAS.
Champs sur Marne, France
sbouis@yahoo.fr

Abstract —A critical problem in many manufacturing, distribution, and retailing companies is the selection of the best solution to liquidate the unsold inventory using alternate distribution channels. In this paper, preliminary research concepts and findings concerning the disposal of the unsold inventory are presented. Different attributes and strategies are identified based on academic literature review, business practices and expert's opinion. A taxonomy for the clearance chain management (CCM) is proposed. Then, a structured AHP approach is developed to rank and prioritize clearance strategy and/or channel. The paper concludes with an illustration and comments on opportunities and direction for further research on the area of clearance chain management.

Keywords— Clearance chain; liquidators; unsold inventory, AHP, Risk-performance criteria.

I. Introduction

The problem of unsold inventory is one of the most common problems faced by supply chain managers. The potential causes of unsold inventory are multiple. Economic crisis, price increases, customer cancellations, product life cycle shortening, the introduction of a new product, changing business conditions may be the basis for the unsold products. Similarly, an abrupt decrease in demand, forecasting errors, poor quality in final product or over-supply may lead to an excess inventory situation. Whatever the causes, unsold inventory is a major concern for every business. Therefore, the rate of unsold inventory can be used as a key performance indicator for supply chain performance [1]. The accumulation of unsold products includes everything from apparels and jewelry to home appliances, multimedia products, alcoholic beverages, drugs and toys. According to [2], the industry average of discounted apparel range between 30% to 50% and unsold items range between 15% to 20%. Other studies show that the rate of unsold products in the textiles sector reaches 20%; the cultural products (CDs, DVDs, books) are about 40% [3]. In the food and agribusiness industries, the unsold inventory represents more than 40% and in the publishing business, it exceeds 60% [4].

The disposal of unsold inventory has its challenges since it is a major inventory cost. Consequently, companies look to "secondary channels" or "secondary market". Indeed, the secondary markets for electronics components were estimated at $15 B in 2008 [5], and those for pharmaceutical supply chains were about $20 B in 2007 [6].

The management of the unsold inventory has led to the participation of new players in the supply chain such as closeout sellers, liquidators and discounters to unclog their inventory pipeline [7]. In this paper, the concept of "clearance chain" (CC) refers to this new kind of supply chain.

The concept of CCM was introduced by Dr. Sami Sboui in 2003 [8,9] and it has become an important and new research topic [10].

The disposal of unsold inventory through a well-designed CC has the potential of being a financial and sustainability key performance indicator. Current literature lacks to capture the development of CC strategies to deal with the unsold inventory.

Nowadays, no business or company can escape from the unsold inventory situation. Consequently, in a strategic-decision level, the main problem for a company (manufacturers, retailers and wholesalers) is to select the appropriate clearance strategy and channel to manage unsold inventory. However, unsold inventory liquidation can be a complex process. It is important for a business to realize the best value for its unsold inventory to minimize risks (direct and indirect losses) and maximize performance (profits). Therefore, the research questions are: (a) what are the clearance strategies being used in practice? (b) What are the main objectives and criteria to choose a clearance strategy? Moreover, (c) Which approach a company should adopt to select the most suitable clearance strategy and channel.

To better understand this phenomenon and answer the research questions, a multidisciplinary approach including qualitative and quantitative methods has been adopted. The used approach is based on deductive reasoning, through the review of the relevant literature to the topic, cases studies and experts' opinions.

The paper is organized as follows: the second surveys the academic and professional literature on "clearance chain management" and proposes a taxonomy of "clearance chain". Section 3 describes the novel proposed approach to select the appropriate clearance strategies and channel. In section 4, an illustration is given and finally, a summary of the main findings of the study, discuss limitations and outline future research directions is propounded.

II. Towards a taxonomy of CCM

This section attempts to extend the understanding of the unsold inventory phenomena and presents the basis for a taxonomy of CCM. Whilst there is a growing recognition of the need to manage unsold inventory, there is still limited research into clearance strategies decision frameworks and support tools.

A. Academic literature review

The academic literature was exanimated through two perspectives: (1) strategies of managing unsold inventory and (2) selection of clearance strategy.

1) Strategies of managing unsold inventory

During the last two decades, supply chain uncertainties and excess inventory has been key issues of concern to researchers and practitioners involved in inventory management. Reference [11] considers the excess inventory as a "deadweight"; "among other adverse effects, it uses valuable storage space, inflates assets, diminishes working capital, and causes a reduction in return on investment". Reference [12] was the first to study the problem of excess inventory disposal, given stochastic demand. According to [13], disposal creates benefits in at least two ways; namely, the salvage

revenue obtained from surplus unit disposal, and the savings in inventory carrying charges. One of the consequence of the excess inventory was the emergence of secondary markets. In the marketing field, according to [14], this market refers to as AREI "After-Retail Excess Inventory" channel. This made a great transformation in the consumer trends and buying behavior through smart buying based on discount products [14]. In the economic and financial fields, the operation of "selling off or disposing off obsolete, scrap, surplus, or waste goods or material in a manner that maximizes the return while minimizing the costs and liabilities", is called "asset recovery" [15].

As pointed out in [16], "secondary markets can thus benefit, in most cases, every member of the supply chain". An option to dispose of excess inventory in the supply chain by selling it in the secondary market can increase a company's profits, reduce its financial risks, and improve the mismatch between supply and demand. During the last years, secondary markets have growing rapidly in size and have becoming also available to supply chains in various industries [17]. According to [18] secondary market sales complicate the structure of the supply chain. He introduces a novel class of inventory policies, referred to as the disposal saturation policies.

Moreover, many studies on reverse supply chains' have stressed the importance of the secondary market in the management of the returned products and overstock [19] [20] [21]. According to [22] the inventory is in fact a liability if it costs more than it earns. They suggest a variety of means (strategies) of disposing of excess inventory: return to supplier, third-party sale, and even charitable donation. Reference [23] discusses the channels inventory disposal and describe why companies should integrate product donations within their overall corporate sustainability strategy. They highlight four main options, Return and Reuse, Liquidation, Disposal, and Donation. Recent study on the business benefits of product donation shows that in most cases the low-cost recovery from liquidation and high costs of return processing make donating surplus inventories to registered charities the most attractive option [24]. The impact of clearance sales on consumer behavior was outlined in [25]. If a firm uses clearances sales, its customers face the trade-off between buying the good at a high price with certainty and waiting for a lower price but then facing the risk of being rationed. According to [10] the main goal of clearance sales for any retailer is to maximize the revenue derived from unsold merchandise when a new collection is about to be introduced. Their research describes the development, implementation, and evaluation in the field of the markdown optimization system that the Spanish clothing brand Zara has.

2) Selection of clearance strategies.

A number of classifications schemes have been proposed in the literature to guide the choice of supply chain strategy [26]. It is evident that the choice of supply chain strategy should be based upon a careful analysis of the demand/supply characteristics of the various product/markets served by a company. Many studies evaluated the supply chain strategies using AHP methodology [27] [28]. The question of the choice of clearance strategy to manage unsold inventory is not clearly studied in the academic literature [21]. Therefore, in this section, several similar researches topics will be presented.

It can be noted that the distribution network design has been for a long time an interesting research topic within the supply chain management [29]. Researchers aim to shape the structure of the distribution network [30]. The clearance channel is quite similar to forward and reverse channel. Reverse channel is different from the traditional distribution channel. Indeed, it has high degree of uncertainty in supply both in terms of quantity and quality of returned products [31]. A comprehensive review of literature of the design of networks for the reverse logistics as well as for the reverse logistics coupled with forward logistics is given in [32]. Some researchers have addressed reverse channel selection issues. For example, [33] compared reverse channels from a profit performance perspective and [34] examined CLSC channel selection issue under two competing retailers setting, while [35] investigated pricing strategy and examined the performance of CLSCs under different channel leaderships. In practice, companies have adopted various reverse channels.

Many studies provide a great insight of the decision-making process about disposition issues in the reverse logistics [36]. Hybrid models, based on AHP and TOPSIS under fuzzy environment was proposed for the selection and evaluation of alternatives including manufacturer's operation, third party operation or joint operation [37]. Reference [38] also developed a fuzzy rule-based decision-making approach using fuzzy AHP for ranking the recovery/disposal strategies for brown products. Many other studies evaluated green supply chain management strategies based on identified criteria using AHP, Fuzzy AHP and TOPSIS [39] [40] [41]. According to a recent survey on Fuzzy MCDM techniques, FAHP is the second most widely used technique in a stand-alone mode (just after AHP) [42]. Therefore, this study was focusing on works using AHP and FAHP methods to select supply chain strategy and design the chain network.

B. Clearance strategies through business practices analysis

To understand the strategies adopted by companies to dispose of unsold inventory, a number of studies are reviewed in many specialized magazines and websites. Two main subjects have been revealed: (1) Clearance strategies practices in various business, and (2) Attributes for the selection of clearance strategies.

1) Classification of clearance strategies

The examination of business practices in various sectors (Textile sector (T), IT sector (I), the household appliances sector (A) and the cosmetics sector (C), has revealed eight main strategies to cope with unsold inventory (Table I and Table II). These strategies are classified into two types:

- Cash recovery strategies (Type I): (S1): Clearance sales; (S2): Factory outlet (and its derivatives: outlets and brand villages); (S3): Online clearance (auction marketplace); (S4): Discounters (liquidators; closeouts stores); (S5): Clearance wholesalers;
- Disposal strategies (Type II): (S6): Recycling; (S7): Destruction and (S8): Donation.

TABLE I. Clearance strategies (type I)

S	Comments and references	T	I	A	C
S1	• Liquidate a high quantity of unsold items. • Represent 40% of "T" sector surplus (in France) [43], and in "A" sector, purchasing in sales period has been progressed between 2015 and 2016 with 3.9 points [44].	X	X	X	X
S2	• Start in 1970, largely adopted by manufacturers. It is a B to C process "to sell only their goods" [45]. • "Factory outlet" has a legal definition, whereas retail "outlet" has not any legal definition [46]. • Factory outlets and "outlets shopping centers" [47] are called also "brand villages". • Zara sells its leftover articles, after the sales period, under the trademark LEFTIES [48].	X	X	X	X
S3	• Clearance websites such as ventesprivees.com, destockplus.com, overstock.com, [43]. There are two types of websites: B-to-B and B-to-C.	X	X	X	X
S4	• "Brick and mortar" liquidators that buy the unsold products and resale it to consumers (B to C). There are two types of liquidators: international and national ones [47] [48].	X	X		
S5	• They are like discounters; but they resell unsold products in bulk (B to B).	X	X		

There are no real guidelines that help company to conclude which strategy is most often used in the disposal of unsold inventory, but generally, the strategies S1, S2 and S3 are mainly used in all business sectors to convert unsold inventory to cash immediately. Moreover, for

many raisons, Type 2 strategies are chosen by some companies to disposal all unsold inventory or "residual unsold products" that remains in stock after using type one strategies.

TABLE II. Clearance strategies (Type II)

S	Comments and references	T	I	A	C
S6	• Generally used in electronics components.		X	X	
S7	• To preserve the brand image and used generally for prestigious and luxury brands.	X			X
S8	• Represents the last solution for small firms after the sales period, and for others one, it is a way to enhance their social image [49].	X	X	X	X

In the following, some examples of clearance strategies adopted by manufacturers, wholesalers and retailers are presented. The primary method for sale-through unsold merchandise is through rigorous markdowns and clearance sales. If still unsold, it may either be given away, sold to a liquidator, or to a discount store, which only buys overstocks. In the textile sector, there are various strategies: some companies sell unsold products to outlets or discount stores (TJ Maxx, Ross, NOZ, …) or discount websites (overstock.com, destockplus.com, …), donate them to charities or foundations (Goodwill, Salvation Army, "Association Don en Nature", …), hire recyclers to re-purpose the clothes, and some places outright destroy or dump their products (shredding, burning).

Companies, with high-end products and middle-range products, generally have their integrated clearance strategies like the factory outlets [49]. These companies use also other alternatives like discounters, clearance wholesalers and the online strategy. However, for most small companies or stores, which do not have the possibilities to invest in costly clearance strategies, the most suitable solution is the clearance sales even if it has a minimum "profit margin" and they donate residual unsold goods afterwards [49]. As mentioned also in [48], "big retailers use their own built-in liquidation channels to handle their vast quantity of returns, and products end up being resold through a secondary retail market. For smaller retailers, they have several options for selling excess inventory: Third-party online auctions or selling to a BtoB liquidation marketplace; selling goods to a liquidation company or donating goods to charity for a nice tax deduction".

2) Attributes for the selection of clearance strategies

Table III presents the cost and benefits of each disposal option [23].

TABLE III. Benefits and risks of disposal methods [23]

Options	Costs & Benefits	Risks
Return or reuse	Significant cost recovery Reverse logistics cost Return processing cost	Product obsolescence Quality issues
Liquidation	Limited cost recovery Easy to do	Risk of brand reputation Market cannibalization
Disposal	Waste processing cost Easy to do	Environmental regulations Reputational affects
Donation	Waste reduction Operational cost savings Tax benefits Potential social impact	Risk to brand reputation Product liability Market cannibalization

According to Xerfi's study [50] two major challenges are facing companies that adopt the clearance solutions: how to preserve the brand image and how to limit the risk of cannibalization with traditional distribution channels. Prestigious brands companies have always chosen to destroy their unsold products to protect their brand image (Brand protection) especially when it comes to high-end luxury goods such as Louis Vuitton or Gucci, these companies do not want the market to be flooded with their merchandise as it depreciates the value of their goods. Another important risk, related directly to the consumer and should be considered, is the change in the consumer buying behavior. According to Michel-Edouard Leclerc, purchasing in the sales period has become a common instinct; consumers generally are waiting for markdown prices [43].

It may be noted that companies adjust the discount rates based on how fast they want to sell the stock. Some other donate the unsold remaining inventory to charity. This will give them a tax write off that can save more money than they would have made by making a deal with a liquidator. Consequently, the short-term goal for a company holding a high volume of unsold products, is to liquidate the greatest quantity during the shortest delay and to maximize benefits. However, in a long-term horizon, the company must take into account the negative effects that may be caused. Therefore, to decide on clearance strategies, two main criteria and six sub-criteria are proposed:

- Performance: Drained volume (P_1); Flow speed (P_2): Benefit/Price (P_3).
- Risks: Deterioration of the brand image (R_1); Cannibalization (R_2); Change in the consumer buying behavior (R_3).

In the next section, a performance-risk based approach is developed to assist companies selecting their clearance strategies and channels. Based on the interaction between risk and performance attributes, a decision support tool linking clearance strategies (scenario) (S) to performance (P) and risk (R) is proposed.

III. Clearance strategies and channels selection: A Performance-Risk based approach

A. Ranking of clearance strategies

The ranking process of CC strategies starts with the business goal of the company, based on a risk-performance trade-off and shows how the clearance strategy can contribute to achieving company goals. As mentioned in previous sections, strategy ranking and selection issues are one of the largest problem that have been studied in the field of multi-criteria decisions models [51]. The AHP method, introduced by Saaty [52], directs how to determine the priority of a set of alternatives and the relative importance of attributes in a multi-criteria decision-making problem. To help companies to choose the "most suitable clearance strategy", and based on the clearance strategies' evaluation criteria outlined previously, a compromise strategy evaluation between two criteria is proposed: the strategy performance and the risks that generates. Therefore, a score was assigned for each strategy performance and strategy risk.

A clearance strategy performance represents the achievement of company's goals in terms of benefit (price), drained volume and flow speed. With regard to the risks, a strategy is evaluated against negative effects that can be caused. These risks are the changing in the buying consumer behavior, the cannibalization of the traditional distribution channels and the deterioration of the brand image (brand reputation). To choose one strategy amongst others decision maker needs to rank all alternative strategies regarding their "Performance Index" and "Risk Index". The AHP framework of ranking CC strategies is structured as a hierarchy which includes three levels: goal, criteria (2 dimensions), and sub-criteria (6 items) for determining the priority strategy. The AHP hierarchy scheme is constructed and shown in Fig.1.

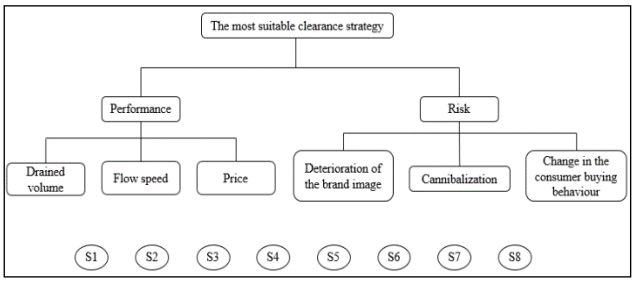


Fig. 1. Hierarchical Structure of clearance strategies

Therefore, each strategy (Si) will be scored regarding two main criteria (P_i, R_i) and six sub-criterions (P_{i1}, P_{i2}, P_{i3}) and (R_{i1}, R_{i2}, R_{i3}). In the conventional AHP, the pairwise comparisons for each level with respect to the goal are conducted using a nine-point scale proposed by Saaty [52].

B. The Choice of clearance channels

The clearance channel (C_j) is a combination of two or more clearance strategies (S_i). Then:

$$Cj = \{Si\} \qquad (1)$$

The main idea of the proposed approach is to calculate a "Channel performance index" CP_j and a "Channel risk index" CR_j for each clearance channel C_j. CP_j was estimated as an additive aggregation of the strategies' performances. Consequently, it will be equal to the sum of the performance of each strategy weighted by the percentage of the drained volume. Then:

$$CPj = \sum_{i=1}^{nj} (\mu i * Pi) \qquad (2)$$

Where:

- μ_i : the % of the drained volume by the strategy S_i,
- n_j : the number of strategies in the clearance channel C_j.

A risk R_i of a strategy S_i, situated even in the end of the channel, can obviously has an impact on all the clearance chain. Therefore, an addition of the different risks R_i of the strategies Si will be a poor estimation for the GR index. It is necessary thus to consider the GR as a multiplicative factor or as a max bounded value. Therefore, the channel GR will be considered as the maximum of the strategies' risks in the channel. Then:

$$CR_j = MAX\ (R_i) \qquad (3)$$

Example: Considering the clearance channel C = S_1(μ_1, P_1, R_1), S_2 (μ_2, P_2, R_2), S_3 (μ_3, P_3, R_3) and S_4 (μ_3, P_4, R_4). The CP and the CR of that contains are:

$$CP = \sum_{i=1}^{4} (\mu i * Pi) \qquad (4)$$

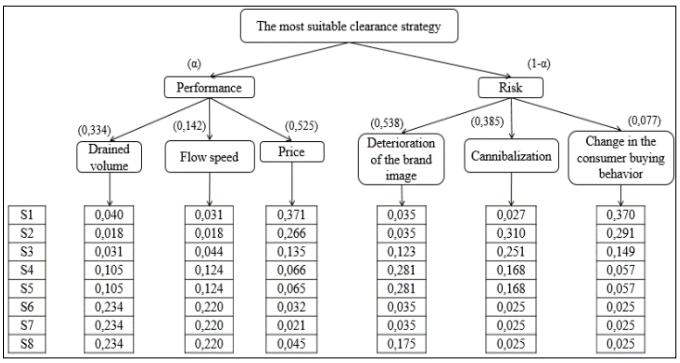
$$CR = MAX\ (Ri) = R_1 \qquad (5)$$

In the next section, an illustration of the proposed approach based on a case study is presented.

IV. ILLUSTRATION

A. Ranking of clearance strategies'

To illustrate the proposed approach, a numerical example based on judgments has been used to extract a strategies' ranking for a company operating in the textile sector. This section presents these results. The first step was to build the pairwise comparison matrices (two for the sub-criteria, six comparison matrices for the eight alternatives with respect to sub-criteria). Then, the ranking of each strategy regarding each sub-criterion is obtained by calculating the weights of each sub-criterion. Moreover, the consistency index for each step was verified. Fig.2 and 3 show the different weights and scores.



S1	0,040	0,031	0,371	0,035	0,027	0,370
S2	0,018	0,018	0,266	0,035	0,310	0,291
S3	0,031	0,044	0,135	0,123	0,251	0,149
S4	0,105	0,124	0,066	0,281	0,168	0,057
S5	0,105	0,124	0,065	0,281	0,168	0,057
S6	0,234	0,220	0,032	0,035	0,025	0,025
S7	0,234	0,220	0,021	0,035	0,025	0,025
S8	0,234	0,220	0,045	0,175	0,025	0,025

Fig. 2. Sub-criterion's weights and scores

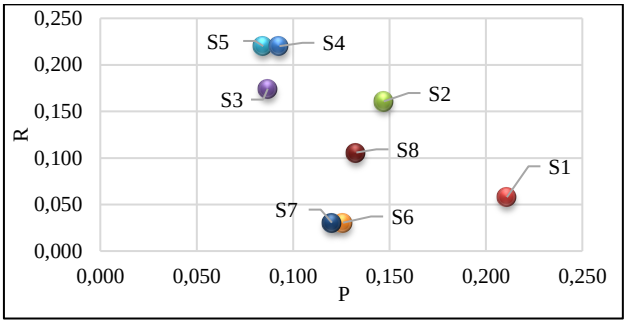


Fig. 3. Risk-performance ranking of CS

Fig.3 represents the strategies scores regarding performance and risk. The build-in strategies such clearance sales and factory outlets represent the most performant ones. This is because they have the highest score regarding (benefit) price that represents the most important sub-criteria in term of performance. Their main risk is the change in the consumer buying behavior. However, the factory outlet generates also the cannibalization risk. It may be noted that the discounters and the clearance wholesalers have a highly risks against their performance values. Moreover, they can easily lead to the risks of cannibalization and deterioration of the brand image. Online clearance's performance is quite similar than discounters and clearance wholesalers but its risks is lower because the company keep a control on the distribution channel.

Destruction and recycling are the less risky strategies, but their performances are limited compared to other solutions. The donation has a higher risk because it presents a threat of deterioration of the brand image. To calculate the global score for each strategy, a variable α is introduced in order to represent the trade-off between the performance and the risk. $\alpha \in [0,1]$ and should be fixed by the decision-maker as a risk aversion parameter. Equation 6, which scores the strategies risk value, will be used to choose the less risky strategy.

$$R_i^{*=} \frac{1}{Ri * \sum_i {}^1/_{Ri}} \qquad (6)$$

The global score for each strategy will depend on the risk-performance trade-off (α) of the decision maker. The strategy score (SS_i) is calculated using the equation 7.

$$SSi = \alpha \times Pi + (1-\alpha) \times R_i^* \qquad (7)$$

To rank the different strategies, α has been varied, and the obtained results are shown in the Fig.4.

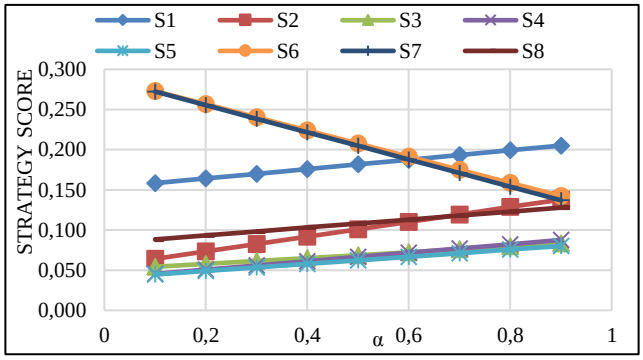


Fig. 4. SS_i evolution

- For 0.1 ≤α≤0.5, the score is dominated by the risk and from a value α>0.5, the score will be dominated by the performance.
- In this study, for α≤0.6 the company adopt a risk based strategy. Then, the recycling and the destruction strategies have the highest score since they have the highest value of R*.

- For α>0.6, the company adopt a performance based strategy, clearance sales have the highest score.
- S3, S4 and S5 were the riskiest strategies, and in terms of performance, they are not well ranked too. Therefore, these strategies are less sensitive to α variation. However, S2 and S8 are more sensitive to α variation.

B. The choice of clearance channel

In this illustration example, based on the professional literature review, six clearance channels have been chosen. Based on [42] that is interested in the clearance strategies of non-food products and the views of the experts, it was possible to extract representative percentages of the drained volumes by each strategy in each channel. Using equation 2 and equation 3, the values of channel performance (CP) and channel risk (CR) for each clearance channel were calculated (Table IV, Fig. 5).

TABLE IV. Channels Risk-Performance

Clearance channels	CP	CR	NCP	NCR	NCR*
C1 : S1➔S7	0,331	0,058	0,128	0,062	0,361
C2 : S1 ➔S8	0,344	0,106	0,133	0,113	0,198
C3 : S1 ➔ S2 ➔ S3 ➔S4	0,538	0,220	0,208	0,234	0,095
C4 : S1 ➔ S2 ➔ S3 ➔S8	0,578	0,174	0,224	0,185	0,120
C5 : S1 ➔S2 ➔S8	0,491	0,161	0,190	0,171	0,130
C6 : S1 ➔ S4	0,304	0,220	0,118	0,234	0,095

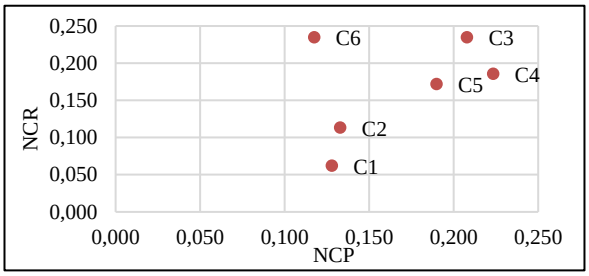


Fig. 5. Channels comparison

Table IV shows also normalized channel performance NCR, normalized channel risk (NCR) and Normalized channel risk score NCR* using the same idea in the equation 6. Fig.5 shows that the channels belong to three categories: A high risk with a low or medium performance (C6); a high risk and performance (C3, C4, C5), and a medium risk and performance (C1, C2).

For more analysis, the α value has been varied to extract the global score for each channel according to the equation 8 and the results are shown in Fig.6.

$$CSi = \alpha \times NCPi + (1-\alpha) \times NCR_i^* \quad (8)$$

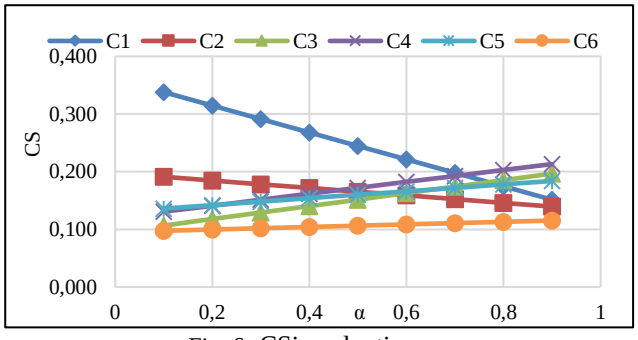


Fig. 6. CSi evaluation

Channels' analysis according to α value is similar to the previous analysis regarding clearance strategies. Fig.9 shows that for low values of α (α≤0.4), channels which have the highest risk score are dominant (C1 and C2). Gradually as the value of α is increasing, those that have a significant performance values overcome the others (C4, C5 and C3).

V. CONCLUSION

Based on the literature, a new taxonomy for CCM was presented. Facing various clearance strategies to liquidate their surplus, the company must choose the most suitable strategies that will better match its objectives, business value and global strategy. Therefore, a structured approach was proposed to assist the clearance strategy's selection process. To demonstrate the feasibility of the proposal, a case illustration was made, which has been analyzed later. The findings of the study will provide useful insight to the supply chain managers and researchers for clearance chain decision-making.

Some assumptions were taken in this study that represent some limits in the proposed approach. Indeed, the unsold quantities and drained volume by each strategy are considered known; however, in practice they are uncertain. Thus, further analysis can be suggested by considering the uncertainty of the quantities of unsold inventory. Moreover, the fuzziness of the selection criteria also can be considered [50]. In this research, the clearance strategies have been prioritized and the existing clearance channels have been evaluated. Therefore, it is important in the next steps of this research to optimize the clearance channel considering the company goals, constraints and the clearance strategies scores. The use of multi-objective optimization and simulation tools could be an interesting research direction.

References

[1] B. Chae, "Developing key performance indicators for supply chain: an industry perspective", Supply Chain Management: An International Journal, vol.14, no.6, 2009, pp. 422-428.

[2] A. Desai, N. Nassar, and M. Chertow, "American Seams: An Exploration of Hybrid Fast Fashion and Domestic Manufacturing Models in Relocalised Apparel Production", The Journal of Corporate Citizenship, vol.45, 2012, pp.53-78.

[3] L'expresse entrerpise, "Comment écouler ses stocks sans casser sa marque ?, [online]. Available: http://lentreprise.lexpress.fr/ [Accessed 20 03 2017].

[4] N. Monier, "Vendre :Comment les invendus peuvent se transformer en or". Available: http://www.lejournaldesentreprises.com/ [Accessed 16 03 2017].

[5] B. Judge, "Best Buy enters secondary market", Retailer Daily, 2009, http://www.retailerdaily.com/.

[6] K. Leininger, "The U.S. pharmaceutical secondary market". White paper, IntegriChain, Princeton, NJ, 2007.

[7] Dale S. Rogers, Ron Lembke, and John Bernardino, Reverse Logistics: A New Core Competency, Supply Chain Management Review, vol.13, no.3, 2013, pp.40-47.

[8] Sami SBOUI, Unsold and Excess Inventory: between optimization and management: A New challenge for the Supply Chain Management, IEEE, 2006.

[9] Sami SBOUI, Clearance Chain Management. : Définition, Concept et Modélisation de la gestion, Ministère de l'Economie (France), 13ème appel à projets du Fonds unique interministériel. Paris, 1 mars 2012.

[10] Caro F. and Gallien J, Clearance Pricing Optimization for a Fast-Fashion Retailer, Journal of Operations Research, vol.60, no.6, 2012, pp.1404-1422.

[11] Tersine, R.J. and R.A. Toelle, Optimum stock levels for excess inventory items, Journal of Operations Management, Vol.4, 1984, pp. 245-258.

[12] Donald B. Rosenfield, Disposal of Excess Inventory. Operations Research, vol. 37, no.3, 1989, pp.404-409.

[13] Rosenfield, D.B., Optimality of myopic policies in disposing excess inventory, Operations Research, vol. 40,1992, pp.800-803.

[14] C. G. Brown and A. Narayanan, After Retail Excess Inventory Channel and Supply Chain Economics in US, Proceedings of the Annual Meeting of the Association of Collegiate Marketing Educators, 2010, pp. 25-31.

[15] N. C. Craig and A. Raman, "Improving Store Liquidation", Manufacturing & Service Operations Management, vol. 18, no. 1, 2015.

[16] Lee, H., S. Whang., The impact of the secondary market on the supply chain. Management Sci, vol. 48, no. 6, 2002, pp.719-731.

[17] Petruzzi, N. S., G. E. Monahan., Managing fashion goods inventories: Dynamic recourse for retailers with outlet stores.IIE Trans, vol.35, no.10, 2003, pp.1033–1047.

[18] A. Angelus, "A Multiechelon Inventory Problem with Secondary Market Sales". Management Science, vol. 57, no. 12, 2011, pp.2145-2162.

[19] Bennekrouf M, Benyoucef L, Sari Z., Problèmes de conception et pilotage des chaînes logistiques inverses et globales : Etat de l'art, MOSIM'10, mai 2010.

[20] Saman Hassanzadeh Amin & Guoqing Zhang, A proposed mathematical model for closed-loop network configuration based on product life cycle, Int J Adv Manuf Technol, vol. 58, 2012, pp.791–801.

[21] Dale S. Rogers1, Benjamin Melamed, and Ronald S. Lembke, Modeling and Analysis of Reverse Logistics, Journal of Business Logistics, vol. 33, no. 2, 2012, pp.107–117.

[22] Toelle, R.A. and R.J. Tersine, Excess inventory financial asset or operational liability, Production and Inventory Management Journal, vol. 30,1989, pp. 32-35.

[23] Md. Islam, John Vate., Vittal Prabhu, Marco Taisch; Dimitris Kiritsis., Surplus Product Donation and Sustainability Strategy: Channels and Challenges for Corporate Product Donations, 20th Advances in Production Management Systems (APMS), Part I, IFIP AICT 414, 2013, pp.54-61.

[24] Ross, J., McGiverin-Bohan, K.L, The Business Case for Product Philanthropy, School of Public and Environmental Affairs, Indiana University, Indiana, 2012.

[25] Volker Nocke and Martin Peitz, A THEORY OF CLEARANCE SALES, The Economic Journal, vol 117, July 2007, pp.964–990.

[26] Martin Christopher, Helen Peck and Denis Towill, A taxonomy for selecting global supply chain strategies, The International Journal of Logistics Management vol. 17, no. 2, 2006, pp. 277-287.

[27] Miah Md. Rubel , Hemendra Nath Roy, Sudipta Saha, Md. Shohel Parvez, Md. Jubaer Alom and Nikhil Ranjan Dhar, Is Leagile Supply Chain Suitable for Apparel Manufacturing Organizations? A Multicriteria Decision Making Perspective, International Journal of Scientific & Engineering Research, vol. 4, no. 8, 2013, pp. 933-938.

[28] Samman Thaeir Ahmed Saadoon Al, Modelling Lean, Agile, Leagile Manufacturing Strategies: An Fuzzy Analytical Hierarchy Process Approach For Ready Made Ware (Clothing) Industry In Mosul, Iraq, International Journal of Advances in Engineering & Technology, vol.7, no.3, 2014, pp.1091-1108.

[29] Riccardo Mangiaracina, Guang Song and Alessandro Perego , Distribution network design: a literature review and a research agenda, International Journal of Physical Distribution & Logistics Management, vol. 45, no. 5, 2015, pp. 506 – 531.

[30] Sharma M.J., Moon I., Bae H., Analytic hierarchy process to assess and optimize distribution network, Applied Mathematics and Computation, 2008, vol.202, no.1, pp. 256–265.

[31] Xiaohua Han, Haiyan Wu, Qianxia Yang and Jennifer Shang, Reverse Channel Selection under Remanufacturing Risks: Balancing Profitability and Robustness, Intern. Journal of Production Economics, vol.182, no.C, 2016, pp.63-72.

[32] Aravendan, M. and Panneerselvam, R., Literature Review on Network Design Problems in Closed Loop and Reverse Supply Chains, Intelligent Information Management, Vol 6, 2014, pp.104-117.

[33] Savaskan, R. C., Bhattacharya, S., & Van Wassenhove, L. N., Closed-loop supply chain models with product remanufacturing. Management Science, vol.50, no.2, 2004, pp. 239-253.

[34] Savaskan, R. C., & Van Wassenhove, L. N., Reverse channel design: the case of competing retailers. Management Science, vol. 52, no. 5, 2006, pp.1-14.

[35] Choi, T. M., Li, Y., & Xu. L., Channel leadership, performance and coordination in closed loop supply chains, International Journal of Production Economics, vol 146, 2013, pp.371-380.

[36] S Agrawal, R. K. Singh, Q. Murtaza, Disposition decisions in reverse logistics: Graph theory and matrix approach, Journal of Cleaner Production, vol 137, 2016, pp. 93-104.

[37] Senthil, S., Srirangacharyulu, B., Ramesh, A., A decision making methodology for the selection of reverse logistics operating channels, Proce. Eng, vol 38,2012, pp. 418-428.

[38] Dehghanbaghi, M., Hosseininasab, H., Sadeghieh, A., A hybrid approach to support recovery strategies (A case study), J. Clean. Prod, vol. 113,2016, pp. 717-729.

[39] Muralidhar. P, K Ravindranath and V Srihari, Evaluation of Green Supply Chain Management Strategies Using Fuzzy AHP and TOPSIS, IOSR Journal of Engineering, vol. 2, no. 4, 2012, pp. 824-830.

[40] Luthra, S., Garg, D., & Haleem, A., Identifying and ranking of strategies to implement green supply chain management in Indian manufacturing industry using Analytical Hierarchy Process. Journal of Industrial Engineering and Management, vol.6, no.4,2013, pp. 930-962.

[41] Odeyale, S.O., Oguntola, A.J., & Odeyale, E.O., Evaluation and selection of an effective green supply chain management strategy: A case study. International Journal of Research Studies in Management, vol.3, no.1,2014, pp. 27-39.

[42] A. Mardani, A. Jusoh, E. K. Zavadskas, Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014, Expert Systems with Applications, vol.42, no. 8, 2015.pp. 4126–4148.

[43] Journal du net, « Après les soldes, encore 15% d'invendus à écouler », [online], available : http://www.journaldunet.com/economie/distribution/enquete/soldes-invendus/1-soldes.shtml [Accessed 20 03 2017].

[44] LSA commerce & consommation, 19 07 2016, Soldes d'été 2016 : les ventes d'électroménager grimpent, quelle enseigne est gagnante ? [Sondage], [online], Available : http://www.lsa-conso.fr/soldes-d-ete-2016-les-ventes-d-electromenager-grimpent-quelle-enseigne-est-gagnante-sondage,242634 [Accessed 31 03 2017]

[45] Jean-Christophe Courjault, Laurent Pouquet, Le développement des magasins d'usine : impact sur les structures de marche dans le commerce de détail, CREDOC, 2000.

[46] La presse.ca, « les outlets selon les détaillants, [online], Available : http://affaires.lapresse.ca/ [Accessed 28 03 2017].

[47] J. Dennis Lord, The Outlet/Off-Price Shopping Centre as a Retailing Innovation, The service Industries Journal, vol.4, no.1, 1984, pp.9-18.

[48] Libération, 08 08 2001, Après les soldes, les soldeurs finissent le rest, [online]. Available : http://www.liberation.fr/futurs/2001/08/08/apres-les-soldes-les-soldeurs-finissent-les-restes_373693 [Accessed 01 03 2017]

[49] La tribune, 26 06 2014, Soldes : comment les magasins se débarrassent des invendus, [online]. Available : http://www.latribune.fr/actualites/economie/20140625trib000836945/soldes-comment-les-magasins-se-debarrassent-des-invendus-.html [Accessed 28 03 2017]

[50] Xerfi, 19 01 2015, Le déstockage dans la mode, [online]. Available : http://www.xerfi.com/presentationetude/Le-destockage-dans-la-mode_4DIS63 [Accessed 20 03 2017].

[51] B.ROY & D. BOUYSSOU, méthodes multicritères d'aide à la décision, Economica, Paris, 1993.

[52] R. W. Saaty, The Analytic Hierarchy Process - what it is and how it is used, U.S.A, vol.9, no.3-5, 1987, pp.161-176.

[53] Sylvain Kubler, Jérémy Robert, William Derigent, Alexandre Voisin, Yves Le Traon, A state-of the-art survey & testbed of Fuzzy AHP (FAHP) applications, Expert Systems With Applications, vol.65, no.6, 2016, pp. 398-422.

Investigating the interrelationships of risk factors affecting outsourcing logistics: An Analysis with DEMATEL Method in the pharmaceutical supply chain context

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Nadine KAFA,
Dafaoui El Mouloudi
*Laboratoire QUARTZ EA 7393,
IUT de Montreuil - Université Paris8,
140, rue de la nouvelle France, 93100
Montreuil, France
{n.kafa, e.dafaoui}@iut.univ-paris8.fr

Asmae El Mokrini*/**,
Abderrahman El Mhamedi
*Laboratoire QUARTZ EA 7393,
IUT de Montreuil - Université Paris8,
140, rue de la nouvelle France, 93100
Montreuil, France
{a.elmokrini, a.elmhamedi}@iut.univ-paris8.fr

Abdelaziz Berrado
**AMIPS-Ecole Mohammadia d'Ingénieurs, Université Mohamed 5
Avenue Ibn Sina B.P. 765 Agdal,
Morocco
berrado@emi.ma

Abstract— Outsourcing logistics has nowadays become an attractive option for firms because it allows them to focus on their core competencies and efficiently manage their functions through an external expert supplier. However, alongside the numerous benefits of outsourcing logistics functions, various risks arise with the implementation of this option. Evaluating the risks of outsourcing can be regarded as a Multiple Criteria Decision Making (MCDM) problem, which requires considering many quantitative and qualitative factors as evaluative criteria simultaneously. This evaluation becomes more complex considering the interactions between various criteria of disruption risks. Indeed, managing supply chain risk is a challenging task because individual risks can be interconnected and actions that mitigate one risk can end up exacerbating another. In this paper, we propose an approach to evaluating interactions between criteria of outsourcing logistics using the DEMATEL method. This method allows the conversion of the cause and effect relations between criteria into a visual structural model, and handling the inner dependences within a set of criteria. The proposed approach is illustrated and validated by a real-life industry case of a pharmaceutical supply chain. The results provide managerial insights as to which of the factors are to be given priority for supply chain risk mitigation.

Keywords—DEMATEL, Pharmaceutical supply chain, Outsourcing, Logistics, Criteria interactions, Risks.

I. Introduction

In today's highly competitive environment, the pharmaceutical supply chain, one typical industry of products with a high benefit per mass unit, is facing an era of transformation, from the supply to the distribution channel. It is challenged to reduce costs to counter recurrent fluctuations in the business environment. Outsourcing logistics has become as a critical strategic factor for many firms due to the competitive market and the need for cooperation amongst firms. It allows them to focus on their core competencies and counter challenges related to the constant evolution of medicinal development and manufacturing in the pharmaceutical industry.

Outsourcing logistics for the pharmaceutical supply chain in particular is a critical process since the supply chain entails specific complexity and high personalized characteristics [2]. According to [3], the increase in outsourcing activities due to globalization has grown the complexity of the supply networks. They have directly contributed to the complexity of the relationships of the supply chain risk factors. This complexity is difficult to manage[4] and systems with a large number of elements can be vulnerable to nonlinear interactions. Since an effective mitigation strategy is influenced by the degree of understanding of the variety and interconnected supply chain risks [5], the interaction and interdependence relationships among the evaluation criteria are critical and should be taken into account in the evaluation process. Risks can be influenced by these interrelationships through factors such as cost, quality, flexibility, economic environment and supplier confidence level[6]. Thus, actions that mitigate one risk can end up exacerbating another due to the influential factor amongst multiple criteria. Supply chain risk management can be regarded as a multi-criteria problem containing both quantitative and qualitative criteria with complex relationships which may have trade-offs and interaction with each other. These criteria are related to each other directly and indirectly. Thus, it can be difficult for a decision maker to analyze them separately [7].

Supply chain risk management has been tackled by several studies [8]. Different MCDM methods like AHP, TOPSIS, VIKOR, and PROMETHEE have been widely applied in risk management. However, these methods have usually been used for risk evaluation solely. Moreover, the causal relationships among the enablers for supply chain risk mitigation has not been demonstrated in literature [9]. Despite the large number of articles that propose the use of MCDM methods, none of these studies have investigated the interdependent relationships amongst the risk factors in the specific case of pharmaceutical supply chains. Criteria interactions are usually neglected in the context of supply chain risk management. In this sense, this research work aims to investigate the evaluation criteria for outsourcing risks and highlight their interactive relationships

using the decision-making trial and evaluation laboratory (DEMATEL) method.

This paper is organized as follows. In the next section, the review of literature related to outsourcing risk management and DEMATEL approach are mentioned. In section III, the proposed approach is presented. In section IV, a numerical illustration from a real-life industry case and results are discussed for validating the proposed approach. Finally, in section V, conclusions and future research directions are presented.

II. Literature review

A. Outsourcing risk management in supply chain

Outsourcing in pharmaceutical supply chains has been tackled by several studies whether on the strategic, tactical, or operational levels ([1], [10], [11]). To focus on their core competencies, many organizations are investigating the feasibility of outsourcing logistics to improve the performance of their supply chain. Therefore, many organizations adopt a strategy to outsource all or part of their logistics activities in order to reduce costs, improve delivery time and satisfy customers. On one hand, organizations may expect to achieve many different benefits through successful outsourcing. On the other hand, there are various risks that are encountered by firms if outsourcing is not successful. It is therefore necessary for managers to strengthen their understanding of supply chain risks and implement corresponding risk management strategies [12]. To do that, risks need to be identified and assessed during the feasibility study of the outsourcing decision process. An important phase in the decision process of outsourcing deals with the identification and assessment of potential risks [8]. In this way, many researchers have studied the problems of risk identification in the supply chain, strategies to avoid these risks, and the bullwhip effect related to them[12]. Interested readers are referred to [13] for a closer discussion of risks identification and evaluation related to outsourcing logistics in the pharmaceutical supply chain. Despite the large number of studies proposing the use of MCDM methods; there were no studies taking into account complex interactions among the multiple criteria in the context of supply chain risk. Therefore, a suitable MCDM method could be useful in this area, considering the interactions between quantitative and qualitative criteria. In this paper, the interactions between supply chain risk management criteria is demonstrated using DEMATEL method. The proposed approach is used to prioritize the important factors influencing outsourcing logistics in pharmaceutical supply chain.

B. DEMATEL Approach

The decision making trial and evaluation laboratory (DEMATEL) method [14] is developed by the Science and Human Affairs Program of the Battelle Memorial Institute of Geneva between 1972 and 1976. The DEMATEL was developed with the belief that pioneering scientific research methods and their appropriate use could improve the understanding of a specific cluster of intertwined problems, thus contributing to the identification of workable solutions using a hierarchical structure[15]. DEMATEL approach is considered as one of the best tools for dealing with the importance and casual relationships among the evaluation criteria [14],[16],[18]. This method can confirm interdependence among variables/attributes and restrict the relationship reflecting their characteristics using an essential system and a development trend [18]. This method puts all factors into two distinct categories called "cause" and "effect" by applying impact values between factors. This categorization results in a more thorough realization of system's components and correspondingly finding solutions to resolve complex system's problems [19]. DEMATEL is a helpful approach that visualizes the structure of complex causal relationships among evaluation criteria using matrices and digraphs [20]. Hence, DEMATEL can convert the relationship between the causes and effects of criteria into an intelligible structural model of the system. One of the major advantages of DEMATEL method is the development of a directed graph that reflects the interrelationships between variables. The final output of the DEMATEL method is the 'impact versus relation' map (IMP), which represents the experts' mind-sets in the real world for the intertwined problem groups [21].

In the literature, numerous examples of the application of DEMATEL method were discussed, either alone or in a combination with other methods (AHP, ANP, VIKOR, ELECTRE, TOPSIS…etc), in a conventional form or in the fuzzy environment. This approach has been successfully applied in many fields such as sustainable supply chain management [22], [23], [19], supplier selection [15], [20], outsourcing provider selection [7], [24], e-learning programs [17],[24], SaaS adoption [6], air transport management [16], project management [25], customer behavior [18], and business process management [26].

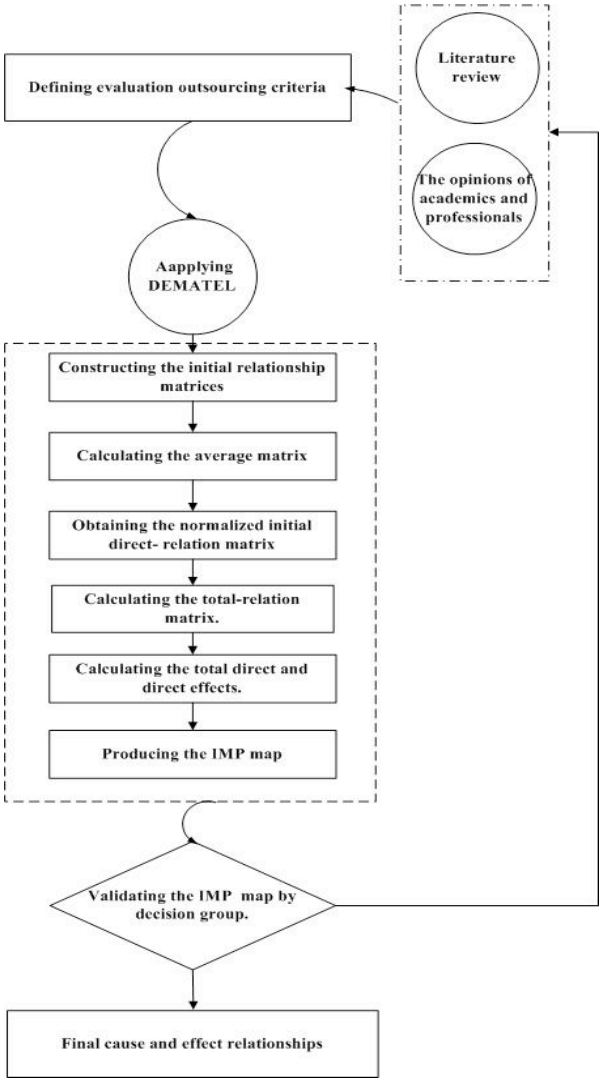
III. Proposed Approach

This section explains the details and various steps involved in the DEMATEL approach for evaluating outsourcing criteria in the context of supply chain risks. The DEMATEL method is selected in our study in order to provide more valuable information for decision making in this problem with complex interdependencies. To achieve this aim, firstly, the main criteria are investigated by a comprehensive literature review and experts' opinion [1]. Then, the DEMATEL procedure and matrices are structured based on decision makers' judgments. The various steps involved in the proposed approach are explained in Fig.1, and can be organized in 4 main steps.

- Determining decision group.
- Selecting evaluation criteria based on literature reviews and validation of decision makers, including experts.
- Applying DEMATEL method to analyze the importance of criteria and the casual relations among them. The judgement of nine experts was obtained using a questionnaire in a real case study.
- Validating the network relationship map by decision group and analyzing results.

The validity and efficacy of this approach is demonstrated through an application dealing with outsourcing logistics in the context of a Moroccan Pharmaceutical Supply chain.

Fig.1 Proposed approach using DEMATEL method



The detailed DEMATEL method procedure is presented in the following sub-steps.

a) Constructing the initial relationship matrices

Let the number of identified evaluation criteria for outsourcing risks be 'n' and the experts selected to be "m". Each expert is asked for evaluating the direct influence of criterion "c_i" over criterion "c_j" according to an integer linguistic scale from 0 (no influence), 1 (low influence), 2 (medium influence), 3 (high influence), and 4 (very high influence). Thus "$X_1, X_2, X_3,\ldots,X_m$" initial relation matrices as $X_k = [\, x_{ij}^k \,]$ were constructed based on the influence ratings from each expert, where "k" is the expert number of participating in evaluation process with $1 \le k \le m$.

b) Calculating the average matrix

To aggregate all judgments from m experts, the average matrix A= $[a_{ij}]$ is calculated as below:

$$A = [a_{ij}] = 1/m \sum_{i=1}^{m} x_{ij}^k \quad (1)$$

c) Obtaining the normalized initial direct- relation matrix

The normalized initial direct-relation matrix R= $[r_{ij}]$, where r_{ij} [0, 1] can be obtained as below.

$$R = \lambda * A \quad (2)$$

$$\lambda = Min\left[\frac{1}{\max\limits_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|}, \frac{1}{\max\limits_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|}\right] \quad (3)$$

The amount of $\max\limits_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$ indicates the total direct effects that a factor with maximum direct effect has on other factors. The amount of $\max\limits_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|$ indicates the total direct effects that the factor which is most affected by other factors, receives. Positive scalar matrix λ as the upper bound allocates the lowest value of these two to itself, and matrix R is obtained through dividing each of matrix A elements by scalar matrix λ. It is significant that each r_{ij} element from matrix R is a value between zero and one.

d) Calculating the total-relation matrix

A continuous decrease of the indirect effects of problems along the powers of matrix R, e.g. $R_2, R_3 \ldots, R_\infty$, guarantees convergent solutions to the matrix inversion, similar to an absorbing Markov chain matrix. Note that $\lim\limits_{m\to\infty} R^h = [0]_{n\times n}$ where $[0]_{n\times n}$ is the $n \times n$ null matrix. The full influence matrix $F=[f_{ij}]$, where f_{ij} represents full direct and indirect influence exerted from criterion c_i to criterion c_j can be acquired by using the formulation (4) as follows:

$$F = R\,(I-R)^{-1} \quad (4)$$

Where I is the identity matrix.

e) Calculating the total given and received both directly and indirectly effects

Each row sum and column sum of matrix T can be calculated as follows:

$$q = [q_i]_{n\times 1} = \left[\sum_{j=1}^{n} f_{ij}\right]_{n\times 1} \quad (5)$$

$$p = [p_j]'_{1\times n} = \left[\sum_{i=1}^{n} f_{ij}\right]'_{1\times n} \quad (6)$$

The value of "q_i" indicates the total given both directly and indirectly effects, that criterion i has on the other criteria. The value of "p_j" indicates the total received both directly and indirectly effects, that all other criteria have on that criterion j.

Thus, when $j = i$, the sum $(q_i + p_i)$ gives us an index representing the total effects both given and received by criterion i. In other words, $(q_i + p_i)$ represents the degree of importance that criterion i plays in the system. As well, the difference $(q_i - p_i)$ indicates the net effect that criterion c_i contributes to the system. When $(q_i - p_i)$ is positive, criterion c_i is a net causer, and when $(q_i - p_i)$ is negative, criterion i is a net receiver [14], [24].

f) Producing the network relationship map

The threshold value "μ" is computed by the average of the elements in matrix F in order to eliminate some minor effects elements in matrix F using (7), [28]:

$$\mu = \frac{\sum_{i}^{n}\sum_{j}^{n}[f_{ij}]}{L} \qquad (7)$$

Where L is the total number of elements in the matrix F. The values of elements in matrix F are zero if their values less than μ. That is, there are lower influences with other criteria when their values are less than μ. Finally, the network relationship map can be acquired by mapping the dataset of $(q + p, q - p)$, where the horizontal axis $(q + p)$ is made by adding q to p, and the vertical axis $(q - p)$ is made by subtracting q from p.

IV. Case Study

We present a numerical example for application of the DEMATEL method. The objective is to examine the interrelationships among criteria that affect the level of risk for a pharmaceutical organization. This organization is situated in Morocco and wishes to outsource logistics functions to a 3PL.

To evaluate the impact of the risks related to logistics outsourcing in the pharmaceutical supply chain, several criteria were selected based on the literature review. Six impact areas were considered as the most relevant criteria that were suggested by decision makers and have been incorporated in this paper. The definition of criteria represents a continuation to [1]. The following list presents all the criteria taken into consideration and their description:

- Financial impact (C_1): represents the severity of the loss in budget allocated for the implementation of the outsourcing project, if the risk R was to occur.
- Impact on the pharmaceutical supply chain reactivity (C_2): the severity of impact on the supply chain of pharmaceuticals that is related to the efficient and effective responsiveness of the supply chain if the risk R was to occur.
- Impact on patients' safety (C_3): represents the severity of the effect on patients' health, security, and wellbeing if risk R was to occur.
- Impact on public regulation and legitimacy (C_4): represents the impact on the conformity to the rules and laws that are designed to govern and control the pharmaceutical logistics functions in Morocco, if risk R was to occur.
- Impact on organizational culture (C_5): represents the impact on the social and psychological environment in the Moroccan context if risk R was to occur.
- Risk mitigation effort(C_6): represents the amount of financial and timely effort required

In this study, we have applied DEMATEL approach to define the cause and effect relationship among the evaluation criteria for outsourcing risks in the pharmaceutical supply chain. X_k represents the data gathered in terms of the evaluation criteria of expert k=nine (Logistics managers, Pharmacists …), where $X_k = [\, x_{ij}^k \,]$. The average matrix A was established by using (1).

$$A = \begin{bmatrix} 0 & 1.7778 & 0.6667 & 1.0000 & 2.2222 & 1.2222 \\ 1.1111 & 0 & 2.8889 & 0.5556 & 0.8889 & 0.1111 \\ 2.3333 & 0.5556 & 0 & 2.1111 & 1.6667 & 0.3333 \\ 0.8889 & 0.7778 & 1.0000 & 0.2222 & 2.2222 & 1.2222 \\ 0.4444 & 1.4444 & 0.3333 & 1.0000 & 0 & 1.1111 \\ 1.6667 & 0.6667 & 0.1111 & 0 & 1.0000 & 0.1111 \end{bmatrix}$$

The normalized initial direct-relation matrix R was calculated by using (2).

$$R = \begin{bmatrix} 0 & 0.2222 & 0.0833 & 0.1250 & 0.2778 & 0.1528 \\ 0.1389 & 0 & 0.3611 & 0.0694 & 0.1111 & 0.0139 \\ 0.2917 & 0.0694 & 0 & 0.2639 & 0.2083 & 0.0417 \\ 0.1111 & 0.0972 & 0.1250 & 0.0278 & 0.2778 & 0.1528 \\ 0.0556 & 0.1806 & 0.0417 & 0.1250 & 0 & 0.1389 \\ 0.2083 & 0.0833 & 0.0139 & 0 & 0.1250 & 0.0139 \end{bmatrix}$$

The total relation matrix F was calculated by using (4) as shown below.

$$F = \begin{bmatrix} 0.3463 & 0.5182 & 0.3827 & 0.4003 & 0.6710 & 0.3886 \\ 0.4748 & 0.3005 & 0.5826 & 0.3808 & 0.5349 & 0.2508 \\ 0.6126 & 0.4333 & 0.3101 & 0.5537 & 0.6874 & 0.3390 \\ 0.4211 & 0.3931 & 0.3709 & 0.2944 & 0.6447 & 0.3778 \\ 0.2901 & 0.3686 & 0.2525 & 0.2984 & 0.2932 & 0.2892 \\ 0.3700 & 0.2722 & 0.1805 & 0.1624 & 0.3606 & 0.1588 \end{bmatrix}$$

The sums of rows and columns of matrix F were calculated by using (5) to (6) as shown in Table 1.

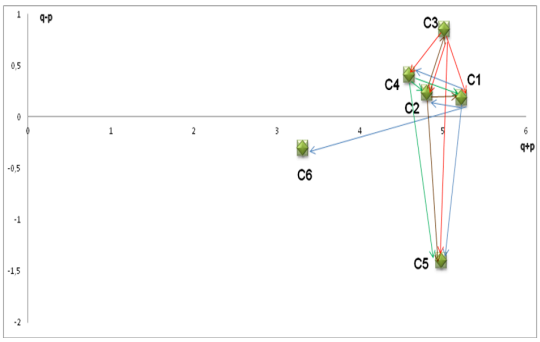
The threshold value μ = 0.3879 was calculated using (7). The values of fij in Table (1) were greater than μ = 0.3879, were represented using different colors. These values presented the interaction between criteria. For example, the value of f12 (0.5182) > μ (0.3879), the arrow in the network relationship map was drawn from C_1 to C_2. It was found also that pharmaceutical supply chain reactivity (C_2) and patients' safety (C_3) had a mutual interaction as both the value of f_{23} (0.5826) and f_{32} (0.4333) were greater than μ = 0.3879.

TABLE I. The sums of total effects and the net affect among criteria

Criteria	C_1	C_2	C_3	C_4	C_5	C_6	q	p	q+p	q-p
C_1	0.3463	0.5182	0.3827	0.4003	0.6710	0.3886	2.7071	2.5149	5.2221	0.1922
C_2	0.4748	0.3005	0.5826	0.3808	0.5349	0.2508	2.5245	2.2859	4.8103	0.2386
C_3	0.6126	0.4333	0.3101	0.5537	0.6874	0.3390	2.9361	2.0793	5.0155	0.8568
C_4	0.4211	0.3931	0.3709	0.2944	0.6447	0.3778	2.5020	2.0900	4.5920	0.4120
C_5	0.2901	0.3686	0.2525	0.2984	0.2932	0.2892	1.7919	3.1918	4.9838	-1.3999
C_6	0.3700	0.2722	0.1805	0.1624	0.3606	0.1588	1.5045	1.8042	3.3086	-0.2997

Then, the cause and effect diagram was built, Fig.2.

Fig.2 Network relationship map among criteria



V. Discussion and Results

The importance of evaluation criteria was determined by *(q+p)* values as index representing the total effects both given and received by each criterion. Based on Table 1, Financial loss (C_1) was the most important evaluation criterion with the largest *(q+p)* value = 5.2221, whereas Risk mitigation effort (C_6) was the least important criterion with the smallest *(q+p)* value = 3.3086. Regarding to *(q+p)* values, the prioritization of the importance of criteria is $C_1 > C_3 > C_5 > C_2 > C_4 > C_6$.

Based on the net effect *(q-p)* values, the criteria are divided into (i) cause group when *(q- p)* is positive and (ii) effect group when *(q - p)* is negative.

(i) In this study, the criteria financial loss (C_1), pharmaceutical supply chain reactivity (C_2), patients' safety (C_3), and public regulation and legitimacy (C_4) are considered as net cause, having the *(q-p)* values of 0.1922, 0.2386, 0.8568, and 0.4120 respectively. It is also pointed out patients' safety (C_3) is the most critical impact criterion on the others in the decision of outsourcing.

(ii) In the effect group, organizational culture (C_5) and risk mitigation effort (C_6) are classified as net receiver, with the *(q-p)* values of -1.3999 and -0.2997, respectively. Then, organizational culture (C_5) is the most affected by the other criteria (C_1), (C_2), (C_3), and (C_4).

VI. Conclusion

Outsourcing logistics in the pharmaceutical industry requires a thoughtful evaluation of the risks related to this decision. The enablers of these risks are usually interconnected. Identifying decision criteria and the relationships among them can help decision makers identify primary causal enablers and therefore monitor the vulnerability issues of their supply chain. This study proposed a novel decision-making approach based on DEMATEL method to identify and analyze the causal relations among the evaluation criteria that affect the level of risk for a Moroccan Pharmaceutical Supply chain. In this context, six enablers for supply chain risk mitigation were identified. The cause and effect enablers of supply chain risk mitigation were categorized. Furthermore, a network relationship map was designed. The results of this approach can help managers in taking proactive steps to address primary causal enables of risk mitigation up in the strategic level to reduce total risk impacts of supply chains.

In future research, the proposed approach will be validated by other case studies for comparison. Also, DEMATEL technique in its original form is incapable to capture the uncertainty during value judgment elicitation. Given ambiguous or uncertain criteria, it would be interesting to incorporate the fuzzy approach for handling uncertainty and achieving results that are more accurate. We suggest grey-DEMATEL to determine interrelationships among the risk criteria in an uncertain environment in terms of perspective. It will be also interesting to reduce the model complexity in the case of a large number of elements (criteria and factors) using other MCDM methods such as TOPSIS, VIKOR or hybrid methods. The idea is to define a critical set of criteria or factors in order to propose a model with a smaller number of relations. Another perspective is to investigate the interactions between factors and attributes as well as risks using Cauchy's theorem.

REFERENCES

[1] A. El Mokrini, E. Dafaoui, A. Berrado, et A. El Mhamedi, « An approach to risk Assessment for Outsourcing Logistics: Case of Pharmaceutical Industry », IFAC-Pap., vol. 49, no 12, p. 1239‑1244, 2016.

[2] C. W. Craighead, J. Blackhurst, M. J. Rungtusanatham, et R. B. Handfield, « The Severity of Supply Chain Disruptions: Design Characteristics and Mitigation Capabilities », Decis. Sci., vol. 38, no 1, p. 131‑156, févr. 2007.

[3] C. Harland, R. Brenchley, et H. Walker, « Risk in supply networks », J. Purch. Supply Manag., vol. 9, no 2, p. 51‑62, mars 2003.

[4] R. B. Mason, « COPING WITH COMPLEXITY AND TURBULENCE - AN ENTREPRENEURIAL SOLUTION », J. Enterprising Cult., vol. 14, no 04, p. 241‑266, déc. 2006.

[5] S. Chopra et M. S. Sodhi, « Managing risk to avoid: Supply-chain breakdown », MIT Sloan Manag. Rev., vol. 46, no 1, sept. 2004.

[6] W.-W. Wu, L. W. Lan, et Y.-T. Lee, « Exploring decisive factors affecting an organization's SaaS adoption: A case study », Int. J. Inf. Manag., vol. 31, no 6, p. 556‑563, déc. 2011.

[7] J. J. H. Liou et Y.-T. Chuang, « Developing a hybrid multi-criteria model for selection of outsourcing providers », Expert Syst. Appl., vol. 37, no 5, p. 3755‑3761, mai 2010.

[8] A. El Mokrini, N. Kafa, E. Dafaoui, A. El Mhamedi, et A. Berrado, « Evaluating outsourcing risks in the pharmaceutical supply chain: Case of a multi-criteria combined fuzzy AHP-PROMETHEE approach », IFAC-Pap., vol. 49, no 28, p. 114‑119, 2016.

[9] R. Rajesh et V. Ravi, « Modeling enablers of supply chain risk mitigation in electronic supply chains: A Grey–DEMATEL approach », Comput. Ind. Eng., vol. 87, p. 126‑139, sept. 2015.

[10] Y. Boulaksil et J. C. Fransoo, « Implications of outsourcing on operations planning: findings from the pharmaceutical industry », Int. J. Oper. Prod. Manag., vol. 30, no 10, p. 1059‑1079, sept. 2010.

[11] J. Momme, « Framework for outsourcing manufacturing: strategic and operational implications », Comput. Ind., vol. 49, no 1, p. 59‑75, sept. 2002.

[12] C. SUN, Y. XIANG, et Q. CHE, « A Supply Chain Risk Evaluation Method Based On Fuzzy Topsis », Int. J. Saf. Secur. Eng., vol. 5, no 2, p. 150‑161, 2015.

[13] A. El Mokrini, E. M. Dafaoui, A. El Mhamedi, et A. Berrado, « A decision framework for outsourcing logistics in the pharmaceutical supply chain », 2015, p. 748‑756.

[14] A. Gabus et E. Fontela, « World Problems, An Invitation to Further Thought within the Framework of DEMATEL », janv. 1972.

[15] T. Kuo, C.-W. Hsu, et J.-Y. Li, « Developing a Green Supplier Selection Model by Using the DANP with VIKOR », Sustainability, vol. 7, no 2, p. 1661‑1689, févr. 2015.

[16] J. J. H. Liou, G.-H. Tzeng, et H.-C. Chang, « Airline safety measurement using a hybrid model », J. Air Transp. Manag., vol. 13, no 4, p. 243‑249, juill. 2007.

[17] G. Tzeng, C. Chiang, et C. Li, « Evaluating intertwined effects in e-learning programs: A novel hybrid MCDM model based on factor analysis and DEMATEL », Expert Syst. Appl., vol. 32, no 4, p. 1028‑1044, mai 2007.

[18] Y.-J. Chiu, H.-C. Chen, G.-H. Tzeng, et J. Z. Shyu, « Marketing strategy based on customer behaviour for the LCD-TV », Int. J. Manag. Decis. Mak., vol. 7, no 2/3, p. 143, 2006.

[19] K. Govindan, R. Khodaverdi, et A. Vafadarnikjoo, « Intuitionistic fuzzy based DEMATEL method for developing green practices and performances in a green supply chain », Expert Syst. Appl., vol. 42, no 20, p. 7207‑7220, nov. 2015.

[20] J. Chai, J. N. K. Liu, et E. W. T. Ngai, « Application of decision-making techniques in supplier selection: A systematic review of literature », Expert Syst. Appl., vol. 40, no 10, p. 3872‑3885, août 2013.

[21] C.-W. Lin, S.-H. Chen, et G.-H. Tzeng, « Constructing a cognition map of alternative fuel vehicles using the DEMATEL method », J. Multi-Criteria Decis. Anal., vol. 16, no 1‑2, p. 5‑19, janv. 2009.

[22] S. Rahman et N. Subramanian, « Factors for implementing end-of-life computer recycling operations in reverse supply chains », Int. J. Prod. Econ., vol. 140, no 1, p. 239‑248, nov. 2012.

[23] H.-H. Wu et S.-Y. Chang, « A case study of using DEMATEL method to identify critical factors in green supply chain management », Appl. Math. Comput., vol. 256, p. 394‑403, avr. 2015.

[24] Ö. Uygun, H. Kaçamak, et Ü. A. Kahraman, « An integrated DEMATEL and Fuzzy ANP techniques for evaluation and selection of outsourcing provider for a telecommunication company », Comput. Ind. Eng., vol. 86, p. 137‑146, août 2015.

[25] C.-K. Lin, Y.-S. Chen, et H.-M. Chuang, « Improving Project Risk Management by a Hybrid MCDM Model Combining DEMATEL with DANP and VIKOR Methods—An Example of Cloud CRM », in Frontier Computing, vol. 375, J. C. Hung, N. Y. Yen, et K.-C. Li, Éd. Singapore: Springer Singapore, 2016, p. 1033‑1040.

[26] C. Bai et J. Sarkis, « A grey-based DEMATEL model for evaluating business process management critical success factors », Int. J. Prod. Econ., vol. 146, no 1, p. 281‑292, nov. 2013.

[27] M. Tamura, H. Nagata, et K. Akazawa, « Extraction and systems analysis of factors that prevent safety and security by structural models », 2002, vol. 3, p. 1752‑1759.

[28] J. Yang, « Innovation capability and corporate growth: An empirical investigation in China », J. Eng. Technol. Manag., vol. 29, no 1, p. 34‑46, janv. 2012.

Coordinating contracts for a capital-constrained supply chain with retailer effort
(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Dinh Anh PHAN(*)
dinhanhdhkt@gmail.com
Thi Le Hoa VO(*)
thi-le-hoa.vo@univ-rennes1.fr
Anh Ngoc LAI(*)
anh-ngoc-lai@univ-rennes1.fr

(*) IGR-IAE - CREM UMR CNRS 6211
University of Rennes 1
11 Rue Jean Macé, CS 70803, 35708 Rennes Cedex 7, France.

Abstract—In this paper, we study the issue of coordinating a capital constrained supply chain in the presence of retailer effort (CCSCE). We consider a CCSCE consists of a supplier (leader) and a retailer (follower) where the retailer may exert promotional effort and has no access to bank financing due to low credit rating. Conversely, the supplier has adequate funds to offer trade credit to the retailer without borrowing from external channels. We then examine whether the existing coordination contracts in a supply chain (SC) with retailer effort but without capital constraints (SCE) can still coordinate the CCSCE. Our result shows that these contracts can achieve coordination of the SC (SC) when the interest rate of trade credit is competitively priced. In addition, we propose a non-linear trade credit schedule and design a generalized contract based on risk compensation for SC coordination (SCC) when both the capital constraints and the retailer's promotional effort exist. We also demonstrate that the CCSCE coordination can be achieved only when risk compensation depends on the order quantity. Finally, we perform a numerical study to illustrate the findings.

Keywords— SC coordination; Capital constraint; Trade credit; Promotional effort; Risk compensation.

I. Introduction

In the decentralized SC, the SC members make decisions independently, or make their decisions in order to maximize their own profit. Both of these strategies can lead to local optimum solutions which may provide lower total profit for the SC [1-2]. The SCC plays a remarkable role because it improves the profit of the SC. In the coordination state, participants must work as a centralized system (centralized SC) in which the retailer pursues the profit of the entire SC, i.e., the retailer tries to maximize the sum of the profit of the retailer and the supplier. The SCC could be achieved via several approaches including SC contracts, information technology, information sharing and joint decision making [3]. In this paper, we focus on the SC contract as a critical component in an SCC since it provides incentives to all the parties so that the decentralized SC behaves as nearly, or exactly the same as the centralized SC [4]. In other words, the SC contracts indicate the parameters to the retailer so that the retailer's self-profit-maximizing objective is aligned with the objective of the entire SC. For example, Cachon [5] proposes some contracts to coordinate the SC such as buyback contracts, quantity-flexibility contracts, quantity-discount contracts, sales-rebate contracts, and revenue-sharing contracts. These traditional contracts are then applied to coordinate the SC under various conditions. The two main streams are capital constraints [6-9] and retailer effort - dependent demand [10-13]. However, the majority of the current literature on SCC isolates capital constraints from retailer's promotional efforts in a firm's operational decisions.

In reality, promotional efforts have become increasingly more common in business practices. A retailer can increase demand for the goods by applying a promotional effort such as advertising, corporate social responsibility, sales price, etc. However, all of those activities are costly. As a result, if the supplier does not provide sufficient incentives, then the retailer will have no motivation to enhance effort level [13]. Additionally, capital constraints are also a great challenge for many firms (especially for small and medium-sized start-ups,). In such cases, it is often the case that the supplier offers trade credit to the retailer by allowing them to delay their payment until the products are sold [7]. Consequently, the contracts that coordinate the SC without capital constraints may be ineffective in coordinating the capital-constrained SC.

As both capital constraints and the retailer's promotional effort are important factors affecting the success of the SCC, this study investigates the issue of SCC and designs coordinating contracts in a CCSCE in which the retailer has capital constraints and has to exert costly effort to increase market demand for the products.

Our two main research questions are as follows:

- Can the existing coordination contracts in an SCE still coordinate a CCSCE under trade credit?
- If these contracts cannot coordinate such a SC, can we design a new contract combining the advantages of these existing contracts and trade credit in order to maintain the coordination?

To address these questions, we will consider a SC consisting of a supplier and a capital constrained retailer who acts as the follower. The supplier, as a leader in the Stackelberg game, has adequate funds to provide trade credit

to the retailer without borrowing from external channels (such as banks or other financial institutions). The retailer, on the other hand, faces a shortage of working capital and cannot obtain a bank loan when needed because of low credit rating.

We then develop a model in order to consider the effect of trade credit and a coordinating contract on the retailer's profit when both capital constraints and promotional effort exist. Moreover, we transfer the retailer's decisions in a decentralized CCSCE (DCCSCE) into the optimization problem. We next examine and compare the retailer's optimal decisions in a DCCSCE with the optimal solutions in a centralized SCE (CSCE). Therefore, we propose a trade credit schedule and design a new generalized contract based on risk compensation for coordinating the DCCSCE.

This paper is organized as follows: we present the CCSCE model with a description and assumptions in Section II. In Section III, we present the model in the case of a DCCSCE with a coordinating contract in order to analyze the retailer's optimal decisions. In Section IV, we first describe the CCSE decision making problem and then design the coordination contracts for a DCCSCE. Section V illustrates a numerical example. The last Section concludes the paper by showing our significant results and research perspective.

II. Model description and assumptions

We consider a two-echelon SC consisting of a supplier as a Stackelberg leader and a capital constrained retailer who faces a non-negative random demand that is sensitive to retailer promotional effort. We assume the capital constrained retailer has an initial capital, B while the supplier has no capital constraints. The supplier produces a single product which the retailer sells to customers at a constant retail price, p and he needs to decide a supply contract to achieve the best performance for the entire SC. The retailer then chooses the quantity ordered q from the supplier and the level of promotional effort e to increase demand.

For simplicity, we assume that a stockout does not cause any goodwill loss and the unsold products have zero salvage value. Let c be the unit production cost of the supplier. The supplier charges the retailer a unit fixed wholesale price w, an interest rate r under trade credit and may request an additional payment A, that may depend on the retailer's decision variables q and/or e. We use a single effort level e to summarize the retailer's activities in promoting efforts. Let $g(e)$ be the retailer's cost of effort which we assume to be increasing and convex in e with $g(0)=0$. Let x be a random variable representing customer demand. We assume that demand satisfies the multiplication from $x = e\xi$, where ξ is a random variable which follows a probability density function (PDF) $f(\xi)$ and the cumulative distribution function (CDF) is $F(\xi)$. This type of effort–demand model has been used in many studies [10, 14, 15].

Let $f(x|e)$ be the PDF and let $F(x|e)$ be the CDF of demand given the effort level e. The complementary CDF is $\bar{F}(x|e) = 1 - F(x|e)$. We assume that the demand is stochastically increasing with the retailer's effort; that is $\frac{\partial F(x|e)}{\partial e} < 0$. The distribution is the common knowledge of the supplier and retailer. According to the properties of conditional probability, the PDF and the CDF of the market demand is: $f(x|e) = \frac{1}{e} f\left(\frac{x}{e}\right)$ and $F(x|e) = F\left(\frac{x}{e}\right)$. (1)

III. Decentralized SC

A. Retailer's Problem in the DCCSCE without trade credit

Without consideration of the trade credit, we use a general supply contract (RE hereafter) which is constructed as RE$\{w, A\}$. It includes a fixed wholesale price w and an additional payment realized at time 0, denoted by A, from the retailer to the supplier ($A<0$ would indicate a payment from the supplier to the retailer). This additional payment may depend on the retailer's decision variables q and/or e. Yan and Zaric [11] show that the coordinating contract in the SC with the retailer's effort but without capital constraint, in previous literature, can be seen as special cases of contract RE$\{w, A\}$. Based on their work, the additional payment is a function of the retailer's decision variables q and/or e, i.e., $A=A(q,e)$. We assume that the additional payment function is continuously differentiable in q and e. However, we do not assume a specific form of the additional payment.

In a DCCSCE, let $\widehat{\Pi}_r(q, e, RE)$ be the expected profit functions of the retailer under RE. Based on the basic retailer-effort model, the retailer's optimization problem without trade credit being considered, can be written as follows:

$$\max_{q\geq 0, e\geq 0} \widehat{\Pi}_r(q, e, RE) = R(q, e) - \big(wq + A + g(e)\big)\big(1 + r_f\big) \quad (2)$$

$$\text{subject to : } 0 \leq wq + A + g(e) \leq B \quad (3)$$

The first term $R(q, e)$ is the expectation of sale revenue at the end of the selling season (referred to as time 1) when the order quantity is q and retailer's effort is e.

$$R(q, e) = pE[min(x, q)] = pe\int_0^{q/e} \bar{F}(\xi)d\xi.$$

The second term in (2) is the ordering cost for q products at fixed wholesale price and the additional payment plus the cost of exerting effort level. Here the term $(1 + r_f)$ reflects an adjustment for the time value for money at the risk-free interest rate (r_f).

If the retailer does not have capital constraints, (3) is redundant and has no impact on the retailer's decisions. Hence, (2) can be reduced to the unconstrained retailer-effort model. Let $(\tilde{q}, \tilde{e})$ be the optimal quantity - effort level pair of retailer for this model. By solving the following first-order equation : $\partial\widehat{\Pi}_r(q, e, RE)/\partial q\big|_{\{\tilde{q},\tilde{e}\}} = \partial\widehat{\Pi}_r(q, e, RE)/\partial e\big|_{\{\tilde{q},\tilde{e}\}} = 0$,

Therefore, the retailer's optimal effort level $\tilde{e}$ satisfies $(\partial g(e)/\partial e|_{\{\tilde{q},\tilde{e}\}} + \partial A/\partial e|_{\{\tilde{q},\tilde{e}\}})\big(1 + r_f\big) = p\Omega(\dot{q})$, and the optimal order quantity is: $\tilde{q} = \tilde{e}\dot{q}$, where, $\dot{q} = \bar{F}^{-1}\big[(w + \partial A/\partial q|_{\{\tilde{q},\tilde{e}\}})(1 + r_f)/p\big]$ and $\Omega(q) = \int_0^q \xi dF(\xi) = qF(q) - \int_0^q F(\xi)d\xi = \int_0^q \bar{F}(\xi)d\xi - q\bar{F}(q)$.

However, if the retailer's initial capital is not sufficient to support their optimal decisions $\{\tilde{q}, \tilde{e}\}$, i.e, $B < w\tilde{q} + A(\tilde{q}, \tilde{e}) + g(\tilde{e})$, the capital constraint (3) is active and the retailer cannot achieve their optimal quantity - effort level pair $\{\tilde{q}, \tilde{e}\}$. Equations (2)-(3) indicate that the retailer's capital constraints may influence their optimal decisions. For the retailer with capital constraints, when no trade credit is

considered, the optimal quantity - effort level pair, $\{q_0, e_0\}$, is obtained by solving the following program :

$\max_{q\geq 0, e\geq 0} \widehat{\Pi}_r(q, e, RE)$, subject to: $wq+A+g(e)=B$.

We also refer to it as the capital constrained retailer-effort model.

B. Retailer's Problem in the DCCSCE with trade credit

In a DCCSCE with trade credit, the supplier, as a leader, offers a general contract TC (TC hereafter). It is constructed as TC$\{r, w, A\}$ where r, w, and A denote the interest rate of trade credit, the fixed wholesale price and an additional payment A realized at time 0 respectively. The additional payment A is defined as in the RE(w, A). The supplier then announces the TC contract with r, w, A to the retailer. After observing the supplier's offer, the retailer determines the level of effort as well as the quantity ordered from the supplier and asks for trade credit when necessary to maximize their expected profit. With trade credit financing, the retailer will transfer all of their remaining cash after paying effort cost when delivering the goods and the remaining payment will be paid at time 1. In the case of the retailer's initial capital not being sufficient to support their cost of effort, i.e., $g(e)>B$, the supplier must transfer funds to the retailer amount, $g(e) - B$, which is paid back to the supplier after the demand is realized.

We first formulate the retailer's expected profit under TC$\{r, w, A\}$. In a DCCSCE, under TC$\{r, w, A\}$, the retailer needs to borrow a trade credit amount of $L= wq + A + g(e) - B$ from the supplier with an interest of r. (L=0 refers to the case when the retailer spends their own funds entirely but does not use trade credit from the supplier). The repayment amount at time 1 is $L(1+r)$. Let $z=z(q, e, r, w, A)$ be the demanded amount for which the revenue at the end of period is just sufficient to cover the promised payment to the supplier.

$$z = L(1+r)/p = (wq + A + g(e) - B)(1+r)/p. \quad (4)$$

After the demand is realized, the retailer needs to pay back $L(1+r) = pz$, which represents the full repayment of the loan plus interest to the supplier. However, if the retailer's total revenue is not sufficient such that $pMin\{x, q\} < L(1+r)$, i.e., $x < z$, the demand falls below z which means that the sum of the value from sold products is less than that which should be paid by the retailer. The retailer is then forced to declare bankruptcy. In this case, the supplier acquires the total revenue realized, ie., px. Otherwise, the retailer pays the debt in full and their total cost is: $L(1+r) + B(1+r_f) = pz + B(1+r_f)$. As a result, the expected cost of the retailer under TC$\{r, w, A\}$ is $C(q, e, TC) = pe\int_0^{z/e} \bar{F}(\xi)d\xi + B(1+r_f)$.

Thus, the retailer's expected profit under TC$\{r, w, A\}$ is:

$$\Pi_r(q, e, TC) = pe\int_y^{q/e} \bar{F}(\xi)\, d\xi - B(1+r_f). \quad (5)$$

Where, $y=z/e$ is the number of units that must be sold to fully pay off the trade credit undertaken by the retailer for a given retailer's effort level. We refer to this as the retailer's default threshold. Hence, the retailer defaults with probability $F(y)$. The objectives of the retailer are to maximize their expected profits. Therefore, their problem in a DCCSCE under TC$\{r, w, A\}$ can be formulated as:

$$\max_{q\geq 0, e\geq 0} \Pi_r(q, e, TC) = pe\int_y^{q/e} \bar{F}(\xi)\, d\xi - B(1+r_f). \quad (6)$$

For the supplier, at the end of the period, he receives a repayment from the retailer that includes two parts: the repayment when the retailer defaults which is the retailer's total realized sales revenue and the repayment when the retailer fully repays the loan $L(1+r) = pz$ when $x > z$. Then, the expected total repayment the supplier gets from the retailer is:

$$\phi(s) = E_{x<z}[px] + E_{x\geq z}[pz] = pE[min\{x, z\}],$$
$$= p\int_0^z xdF(x|e) + pz\bar{F}(x|e) = pe\int_0^y \bar{F}(\xi)d\xi.$$

Therefore, the supplier's expected profit in a DCCSCE under TC$\{r, w, A\}$ can be expressed as:

$$\Pi_s(q, e, TC) = pe\int_0^y \bar{F}(\xi)d\xi + (B - g(e))(1+r_f) - qc(1+r_f). \quad (7)$$

If the supplier is risk-neutral and indifferent to the risk, the trade credit should be priced so that the supplier's discounted expected payoff of trade credit financing is equal to the borrowed amount. Hence, the supplier sets the interest rate of trade credit according to the following equation:

$$pe\int_0^y \bar{F}(\xi)d\xi = L(1+r_f). \quad (8a)$$

Let $\bar{\bar{y}} = (wq + A + g(e) - B)(1+r_f)/pe$, then the above equation can be transformed into:

$$\int_0^{\bar{\bar{y}}(1+r(\bar{\bar{y}}))/(1+r_f)} \bar{F}(\xi)d\xi = \bar{\bar{y}}. \quad (8b)$$

(8b) determines the interest rate of trade credit $r(\bar{\bar{y}})$ with respect to the retailer's default threshold at the risk-free interest rate, $\bar{\bar{y}}$.

We next consider the case where the interest rate of trade credit is not competitively priced. When the retailer's initial capital is not enough to make the optimal decision $(\tilde{q}, \tilde{e})$, they have two options : either to choose trade credit or cash transactions. In a cash transaction, the retailer uses up their initial capital and as mentioned in Section III.A, the retailer's optimal quantity - effort level pair are $\{q_0, e_0\}$, and their expected profit is $\widehat{\Pi}_r(q_0, e_0, RE)$.

If the capital constrained retailer decides to accept the trade credit provided by the supplier. The objective of the retailer is to maximize their expected profit by choosing the optimal quantity - effort level pair (q^*, e^*). For a given supply contract TC$\{w, r, A\}$, it is straightforward to show that if $\partial^2 A/\partial q^2 \geq 0$, the retailer's expected profit function in a DCCSCE, $\Pi_r(q, e, TC)$, is concave in q for any effort level e. If it is assumed that the cost of effort function and the additional payment are chosen such that the existence of an optimal solution (q^*, e^*), is assured, the optimal solution, if it exists, can be obtained through the first order conditions. By solving the following first-order equation: $\partial\widehat{\Pi}_r(q, e, TC)/\partial q\big|_{\{q^*, e^*\}} = \partial\widehat{\Pi}_r(q, e, TC)/\partial e\big|_{\{q^*, e^*\}} = 0$, we can get the retailer's optimal effort level, e^* and the retailer's optimal order quantity, q^* satisfying : $(1+r)\bar{F}(y^*)(\partial g(e)/\partial e|_{\{e^*\}} + \partial A/\partial e|_{\{q^*, e^*\}}) = p\Omega(\ddot{q}) - p\Omega(y^*)$ and $q^* = e^*\ddot{q}$, where $\ddot{q} = \bar{F}^{-1}\left[\left(w + \partial A/\partial q|_{\{q^*, e^*\}}\right)(1+r)\bar{F}(y^*)/p\right]$, and $y^* = (wq^* + A(q^*, e^*) + g(e^*) - B)(1+r)/pe^*$.

IV. Coordinating contracts for a DCCSCE with trade credit

A. Conditions for coordination

In a CSCE, both the supplier's and the retailer's decisions are fully coordinated under central control for the benefit of the entire SC. This is a benchmark case with which we compare the case of a coordination scheme in a DCCSCE. The total profit of a CSCE at time 1 can be expressed as:

$$\Pi_c(q,e) = R(q,e) - \big(qc + g(e)\big)\big(1+r_f\big). \quad (9)$$

The objective of a CSCE is to maximize $\Pi_c(q,e)$. Based on Taylor [10], the optimal order quantity is $q_c = e_c\hat{q}$ and the optimal effort level e_c satisfies $\partial g(e)/\partial e|_{\{e_c\}}\big(1+r_f\big) = p\Omega(\hat{q})$ where $\hat{q} = \bar{F}^{-1}\big[c(1+r_f)/p\big]$. The first-best CSCE's profit is $\Pi_c = pe_c\Omega(\hat{q}) - g(e_c)(1+r_f)$.

We then derive the supplier's coordination scheme that induces the retailer to replicate the optimal outcomes of the CSCE in making decisions. Let (q^*, e^*) be the retailer's optimal decision in a DCCSCE. Similar to the results of Cachon [5], the DCCSCE is coordinated if the general contract TC$\{r, w, A\}$ satisfies the following two conditions:

C1: Under the TC$\{r, w, A\}$, the retailer's optimal decisions in the DCCSCE are the same as in the CSCE, i.e., $(q^*, e^*) = (q_c, e_c)$.

C2: Under the TC$\{r, w, A\}$, each party obtains an arbitrary pre-negotiated portion of the total system profit, i.e., $\Pi_r(q^*, e^*, TC) = \lambda\Pi_c$, for any $\lambda \in (0,1)$ denotes the percentage of the coordinated SC profit that the retailer can share.

B. Coordinating contract in the DCCSCE when the interest rate of trade credit is competitively priced

The existing research shows that the general contract RE$\{w, A\}$ can coordinate DSCE [14]. However, in a DCCSCE with trade credit, because the suppliers bear the default risk, can RE$\{w, A\}$ still coordinate DCCSCE?.

Theorem 1: If a contract RE$\{w, A\}$ coordinates the DSCE, the TC$\{r, w, A\}$ in which the interest rate of trade credit is competitively priced, achieves coordination in the DCCSCE.

Proof. If the supplier is risk-neutral, he sets the interest rate of trade credit based on (8a). If (8a) is substituted into (5), the retailer's profit function in the DCCSCE becomes:

$\Pi_r(q,e,TC) = R(q,e) - (wq + A + g(e) - B)\big(1+r_f\big) - B\big(1+r_f\big) = \widehat{\Pi}_r(q,e,RE)$.

If RE$\{w, A\}$ can coordinate DSCE then

$\partial\widehat{\Pi}_r(q,e,RE)/\partial q|_{\{q_c,e_c\}} = \partial\widehat{\Pi}_r(q,e,RE)/\partial e|_{\{q_c,e_c\}} = 0$

Therefore,

$\partial\Pi_r(q,e,TC)/\partial q|_{\{q_c,e_c\}} = \partial\Pi_r(q,e,TC)/\partial e|_{\{q_c,e_c\}} = 0$

In addition, $\widehat{\Pi}_r(q_c,e_c,RE) = \lambda\Pi_c$ thus $\Pi_r(q_c,e_c,TC) = \lambda\Pi_c$. Hence, C1 and C2 are satisfied. Thus, the contract presented in Theorem 1 achieves coordination in the DCCSCE. □

Theorem 1 indicates that when the interest rate of trade credit is competitively priced, the coordinating contracts in the DCCSCE have similar structural properties to those in the DSCE. This result implies that SC coordinating contracts for the noncapital constrained SC in the presence of retailer effort continue to be coordinated for the capital constrained SC under trade credit as long as the needed financing is competitively priced.

C. Coordination of a DCCSCE with wholesale price and trade credit schedules

In this section, we propose a trade credit schedule to coordinate the DCCSCE. Under such trade credit schedule, the supplier first chooses a fixed wholesale price w and then proposes an appropriate trade credit rate that depends on the retailer's decisions in order to induce the retailer to replicate the optimal outcomes in the CSCE. We refer to it as the contract with (w, r). Let $\Pi_r(q,e,w,r)$ and $\Pi_s(q,e,w,r)$ be the expected profit functions of the retailer and the supplier under the trade credit schedule respectively. Let λ be a value that determines the profit allocation, defined over the continuous interval [0, 1]. We then obtain the following result.

Theorem 2*:* For given a fixed wholesale price, if the supplier sets a trade credit schedule *r(q, e)* defined in (10). The contract with *(w, r)* coordinates the DCCSCE.

$$(1-\lambda)R(q,e) - pe\int_0^y \bar{F}(\xi)d\xi + (\lambda cq + \lambda g(e) - B)\big(1+r_f\big) = 0 \quad (10)$$

Where $y = (wq + g(e) - B)(1 + r(q,e))/pe$ and $\lambda \in (0,1)$

Proof. If (10) is substituted into (5), (7) and simplified, we have $\Pi_r(q,e,w,r) = \lambda\Pi_c(q,e)$ and $\Pi_s(q,e,w,r) = (1-\lambda)\Pi_c(q,e)$. That means that under the contract with (w, r), the retailer's and the supplier's expected profit functions are linear functions of the expected total SC profit, thus ensuring that the retailer's optimal decision is the same as the first-best solution of the CSCE. Moreover, under the contract with (w, r), the SC members can arbitrarily split the SC profit by setting λ. Therefore, C1 and C2 are satisfied □

D. Design of a coordinating contract in a DCCSCE based on risk compensation

Yan, Wang, Cheng and Huang [8] consider the credit risk compensation and regulate that the retailers will compensate the credit risk to the suppliers and achieve SC coordination with trade credit by designing a contract using quantity - discount contract. This work allows us to design the general contract TC$\{r, w, A\}$ based on risk compensation and the coordinating contracts for the noncapital constrained SC in the presence of retailer effort to achieve coordination in the DCCSCE. Under this TC, the supplier sets the interest rate of trade credit (r) equal to the risk-free interest rate (r_f). He charges a fixed wholesale price (w) the same as in the situation in which the retailer is not capital constrained. However, the additional payments (A) need to be added to an amount equal to risk compensation (T).

Theorem 3: In a DCCSCE where the supplier offers trade credit to the capital constrained retailer and the interest rate of trade credit set by the supplier equals the risk-free interest rate. If RE$\{w, A\}$ achieves coordination in a DSCE with an arbitrary profit allocation between the SC members, then, the contracts proposed TC$\{r_f, w, A+T(q, e)\}$ achieves coordination in the DCCSCE, where $T(q, e)$ depends on both q and e. The specific expression of $T(q, e)$ is:

$$T(q,e) = \frac{pe\int_0^y F(\xi)d\xi}{(1+r_f)}, \quad (11)$$

where $y = (wq + A + T(q,e) + g(e) - B)\big(1+r_f\big)/pe$.

Proof. Under a TC{r_f, w, $A + T(q, e)$}, the expected profit function of the retailer in a DCCSCE is

$\Pi_r(q,e,TC) = R(q,e) - (wq + A + T(q,e) + g(e))(1 + r_f) + pe\int_0^y F(\xi)d\xi$.

Using (5), we get $\Pi_r(q,e,TC) = \widehat{\Pi}_r(q,e,RE) - T(q,e)(1+r_f) + pe\int_0^y F(\xi)d\xi = \widehat{\Pi}_r(q,e,RE)$.

Similar to the proof of Theorem 1, we know that a TC{r_f, w, $A + T(q, e)$} achieves coordination in a DCCSCE □

Proposition 2: If the interest rate of trade credit $r(\bar{\bar{y}})$ is competitively priced, the risk compensation $T(q,e)$ is related to the trade credit rate $r(\bar{\bar{y}})$ as follows:

$$(1 + r_f)T(q,e) = L(r(\bar{\bar{y}}) - r_f),$$

where $\bar{\bar{y}} = L(1 + r_f)/pe$ is the retailer's default threshold at the risk-free interest rate before adding the risk compensation to the trade credit amount.

Theorem 4: If the RE{w, A} achieves coordination in a DSCE, then the contracts proposed TC{r_f, w, $A+ T$}, where T depends only on the level of effort e cannot coordinate a DCCSCE.

Proof. If the risk compensation T depends only on e. We suppose that $T(e)$ is satisfied $(1 + r_f)\partial T(e)/\partial e|_{\{q_c,e_c\}} = \partial pe\int_0^y F(\xi)d\xi\,/\partial e|_{\{q_c,e_c\}}$. Therefore, $\partial\Pi_r(q,e,TC)/\partial e|_{\{q_c,e_c\}} = 0$. Hence, the first order conditions for e are satisfied. However, T does not depend on q so $\partial T(e)/\partial q|_{\{q_c,e_c\}} = 0$. After taking the first partial derivatives of $\Pi_r(q,e,TC)$ with respect to q, we get $\partial\Pi_r(q,e,TC)/\partial q|_{\{q_c,e_c\}} = (w + \partial A/\partial q|_{\{q_c,e_c\}})F(y_c)(1 + r_f)$. In addition, for y>0 we have $F(y)$ >0. Therefore, $\partial\Pi_r(q,e,TC)/\partial q|_{\{q_c,e_c\}} = cF(y_c)(1 + r_f) > 0$, hence C1 is not completely satisfied. That means that the TC{r_f, w, A +T} in which the risk compensation depends only on e fails to coordinate the DCCSCE □

E. A general coordinating contract based on risk compensation

Theorem 5 indicates the necessary and sufficient conditions for the existence of the coordinating contract in the DCCSCE based on risk compensation.

Theorem 5: If the following conditions are satisfied, then there exists a contract TC {r_f, w, A} which achieves coordination in the DCCSCE when A, *ie.*, $A=A(q,e)$ depend on q and/or e.

- $R(q,e) - A + pe\int_0^y F(\xi)d\xi$ is concave in q and e (12a)
- $\partial A/\partial q|_{\{q_c,e_c\}} = \frac{c - w\bar{F}(y_c)}{\bar{F}(y_c)}$, (12b)
- $(1 + r_f)\bar{F}(y_c)\frac{\partial A}{\partial e}\Big|_{\{q_c,e_c\}} = (1 + r_f)F(y_c)\frac{\partial g(e)}{\partial e}\Big|_{\{q_c,e_c\}} - p\Omega(y_c)$, (12c)
- $A(q_c,e_c) = (1-\lambda)\frac{\Pi_c}{1+r_f} - q_c(w - c) + \frac{pe_c\int_0^{y_c}F(\xi)d\xi}{1+r_f}$. (12d)

Where $y_c = (wq_c + A(q_c,e_c) + g(e_c) - B)(1 + r_f)/pe_c$ and, we recall, $\Omega(y_c) = \int_0^{y_c}\xi dF(\xi)$.

Condition (12a) guarantees that the optimal solution (q^*,e^*) exists and is unique. Condition (12b) incentivizes the retailer to order the quantity q_c. Condition (12c) incentivizes the retailer to decide on the effort level e_c, and (12d) guarantees that each party obtains an arbitrary pre-negotiated portion of the total system profit. From Theorems 5, if we let $y_c = 0$, conditions 12 a, b, c, and d can be reduced to the conditions under which a coordinating contract exists in a noncapital constrained SC in the presence of the retailer's effort.

Proof. We omit the details of the proof for brevity.

V. Numerical example

In this section, we present some examples to obtain further managerial insights based on the theoretical results above. We assume that the random variables ξ follows an Exponential distribution with a mean of 2.5; the cumulative distribution function is $F(\xi) = 1 - \text{Exp}(-0.4\,\xi)$ for $\xi \geq 0$. We also assume that the retailer can influence the demand by exerting promotional effort with cost of $g(e) = 0.05e^2$. We set: the market price at p=\$6, the unit production cost at c = \$3 and the risk free interest rate at r_f = 0.

Using the above assumptions, in the CCSE, the optimal order quantity is q_c = 39.88, the optimal effort level is e_c = 23.014 and the total expected SC profit will be Π_c = \$ 26.482.

Yan and Zaric [11] show that the quantity-discount contracts, which consist of a basic constant price component, $w = c + \frac{(1-\lambda)}{2(1+r_f)q_c}\Pi_c$, and a variable price component (the additional payment), $A(q) = (1-\lambda)\frac{\Pi_c}{2(1+r_f)}\frac{q_c}{q}$ can be used to coordinate the DSCE and arbitrarily allocate the total profit between the SC members. For our numerical studies, we will show that this contract continues to be coordinating for the DCCSCE under trade credit by adding the supplier's decision variables (i.e. the interest rate of trade credit and the risk compensation).

We suppose that the retailer and supplier together agree on the portion of the profit allocation λ, then the quantity-discount contract has specific parameters depending on the portion of profit allocation and the order quantity q as follows: $w = 3 + 0.33201 * (1 - \lambda)$ and $A(q) = 528.0396 * (1 - \lambda)/q$. Next, we examine the effectiveness of this contract on the SC coordination under different conditions of the retailer's initial capital.

A. Coordination analysis when the interest rate of trade credit is competitively priced

With Theorem 1, the supplier can determine the interest rate of trade credit by using (8b) so that the retailer makes the optimal decisions equal to the optimal solutions in the centralized case and arbitrarily allocates the profit. . In this case, the interest rate of trade credit will depend on the retailer's default threshold at the risk-free interest rate, $\bar{\bar{y}}$. By solving (8b), we have $(\bar{\bar{y}}) = -1 - \frac{2.5\text{Log}[1-0.4\bar{\bar{y}}]}{\bar{\bar{y}}}$. Column 4 in Table 1 represents the value of the interest rate that the supplier applies to the trade credit when the retailer replicates the optimal solutions in the centralized case. For example, if B=\$90, λ=0.4, $q^* = q_c = 39.88$ and $e^* = e_c = 23.014$, then $\bar{\bar{y}} = 0.522$, the interest rate of trade credit will be $r(\bar{\bar{y}}) = 0.122$ and the retailer's default threshold is $y_c = 0.522 * (1 + 0.122) = 0.585$ respectively (column 9 of Table 1 represents the value of y_c). At these values, the retailer's expected profit is \$10.593 (equivalent to 40% of Π_c) and the supplier's expected profit is \$15.889 (equivalent to 60% of Π_c).

Table 1: The supplier's decision variables for each capital level and profit allocations under trade credit

λ	$ *B*	*w* ($ per unit)	$r(\bar{\bar{y}})$	$r(q_c, e_c)$	$\$T(q_c, e_c)$	$\$A(q_c, e_c)$	$\bar{\bar{y}}$	y_c	$\$\Pi_r(q_c, e_c)$	$\$\Pi_s(q_c, e_c)$	$\$\Pi_c$
	40	3.232	0.241	0.340	30.004	39.272	0.903	1.120	7.944	18.538	26.482
	70	3.232	0.169	0.296	15.978	25.246	0.686	0.801	7.944	18.538	26.482
0.3	90	3.232	0.127	0.287	9.466	18.735	0.541	0.609	7.944	18.538	26.482
	110	3.232	0.089	0.311	4.846	14.115	0.396	0.431	7.944	18.538	26.482
	164.66	3.232	0.000	0.000	0.000	9.269	0.000	0.000	7.944	18.538	26.482
	40	3.199	0.234	0.320	28.532	36.476	0.884	1.090	10.593	15.889	26.482
	70	3.199	0.163	0.273	14.996	22.941	0.666	0.775	10.593	15.889	26.482
0.4	90	3.199	0.122	0.261	8.752	16.697	0.522	0.585	10.593	15.889	26.482
	110	3.199	0.084	0.279	4.362	12.307	0.377	0.408	10.593	15.889	26.482
	162.012	3.199	0.000	0.000	0.000	7.945	0.000	0.000	10.593	15.889	26.482
	40	3.166	0.227	0.299	27.108	33.729	0.864	1.061	13.241	13.241	26.482
	70	3.166	0.157	0.250	14.053	20.673	0.647	0.749	13.241	13.241	26.482
0.5	90	3.166	0.116	0.234	8.070	14.691	0.502	0.561	13.241	13.241	26.482
	110	3.166	0.079	0.246	3.907	10.527	0.357	0.386	13.241	13.241	26.482
	159.363	3.166	0.000	0.000	0.000	6.621	0.000	0.000	13.241	13.241	26.482

The total profit of the SC will be $26.482. Thus, the quantity-discount contract achieves SC coordination in the case where the interest of trade credit is competitively priced.

B. Coordination analysis based on trade credit schedule

With Theorem 2, the supplier combines the fixed wholesale price and the interest rate of trade credit so that the retailer makes optimal decisions equal to the optimal solutions of the centralized case, assuming that the supplier still retains the fixed wholesale price as in the quantity-discount contract (without variable price component). In column 3 of Table 1, we show the value of the fixed wholesale price corresponding to each predetermined profit allocation. In this case, the supplier will set the trade credit rate depending on the order quantity, the effort level and the retailer's initial capital by using (10). For example, if *B*=$110, λ=0.3, then the trade credit rate function $r(q, e)$ is:

$$\frac{2200-64.65q-e^2-300e\text{Log}[-\frac{0.066(0.899q-4.5e-10.49e\text{Exp}(-0.4q/e)+0.015e^2-110)}{e}]}{64.65q+e^2-2200}$$

This is applied when the retailer's default threshold y is greater than zero, ie., $3.232q$+0.05e^2-110>0. In the coordination state, the retailer's quantity - effort level pair $\{q^*, e^*\}$ are $q^* = 39.88$ and $e^* = 23.014$, the interest rate of trade credit is shown in column 5 of Table 1. For example, if *B*=$110 and λ=0.3, the trade credit rate set by the supplier will be 0.311. Then $y_c = (3.232 * 39.88 + 0.05 * 23.014^2 - 110)(1 + 0.311)/(6 * 23.014) = 0.431$. At these values, the retailer's expected profit and the supplier's expected profit are $7.944 (equivalent to 30% of Π_c) and $18.538 (equivalent to 70% of Π_c). The profit of the SC will be $26.482. Therefore, the combination of the wholesale price and the trade credit interest rate can be used to achieve SC coordination and arbitrarily allocate the SC profit.

C. Coordination analysis based on risk compensation

With Theorem 3, the supplier coordinates the SC by adding risk compensation $T(q,e)$ to the additional payment $A(q)$ (the variable price component). The risk compensation is calculated by solving (11). For example, if B = $90, λ= 0.5, the risk compensation depends on the retailer's decision variables q and e as follows:

$$T(q,e) = -\frac{264.028}{q} - 3.166q - 0.05e^2 + 15e\text{Log}\left[-\frac{300qe\text{Exp}(6/e)}{5280.57 + 63.320q^2 + q(e^2 - 1800 - 300e)}\right]$$

which is applied when the retailer's default threshold is greater than zero, ie., 3.166q+264.019/q + 0.05e^2-90>0. The total additional payment respectively is $A(q,e) = A(q) + T(q,e)$

$$= 90 - 3.166q - 0.05e^2 - 15e\text{Log}[\frac{6-\frac{17.60}{q}-0.211q+e-0.0033e^2}{e}].$$

In the coordination state, the value of risk compensation and the total additional payment are $8.070 and $14.691, respectively. At these values, the retailer's expected profit and the supplier's expected profit will be $13.241 (equivalent to 50% of Π_c) and $13.241 (equivalent to 50% of Π_c) as calculated by $6 * 23.014 \int_{y_c}^{39.88/23.014} \text{Exp}[-0.4\xi]\, d\xi - 90$ and $6 * 23.014 \int_0^{y_c} \text{Exp}[-0.4\xi]\, d\xi + 90 - 0.05 * 23.014\text{^}2 - 39.88 * 3$ respectively, where $y_c = (3.166 * 39.88 + 14.691 + 0.05 * 23.014^2 - 90)/(6 * 23.014) = 0.561$. The profit of the SC will be $26.482. In columns 6 and 7 of Table 1, we present the value of the risk compensation and the total additional payment with respect to B and λ when the retailer replicates the optimal decisions of the CSCE. We can also use Proposition 2 to show that the value of risk compensation in the state of coordination is satisfied

$$T(q_c, e_c) = L_c r(\bar{\bar{y}}) = (172.603 - 26.481\lambda - B) r(\bar{\bar{y}}).$$

VI. Conclusions

In this paper, we have studied the SC coordination under a complex situation: the retailer needs to exert promotional effort but has no access to bank financing and the supplier has adequate funds to offer trade credit to the retailer without borrowing from external sources. We then develop models to analyze the retailer's optimal decisions in a decentralized capital constrained SC in the presence of retailer effort (DCCSCE) with trade credit and coordinating contracts. Our results show that when the interest of trade credit is

competitively priced, the contracts that coordinate the SC without capital constraints but with retailers' effort (SCE) continue to coordinate the DCCSCE. However, when the interest of the trade credit is not competitively priced, these contracts need adjustments to achieve coordination in the DCCSCE.

In fact, our research indicates that if the supplier is properly compensated for the amount of credit risk in the form of risk premium, coordinating contracts in the DCCSCE have similar structural properties to those in decentralized SCE (DSCE). Otherwise, if the interest rate of trade credit is not enough to compensate for credit risk, the coordinating contracts in the DSCE need adjustments to achieve coordination in the DCCSCE.

We then propose a non-linear trade credit schedule and design a generalized contract based on risk compensation to coordinate the DCCSCE. We have shown that if risk compensation depends only on level of effort, this contract fails to coordinate the DCCSCE. Finally, we point out that the coordination of the DCCSCE under contracts can exist under some necessary and sufficient conditions of additional payment and propose some specific contracts that achieve coordination in the DCCSCE based on risk compensation.

For future study, our model may be used to examine this coordination during bankruptcy and in the case where the supplier is also capital constrained. Indeed, we are interested in understanding whether the coordination can exist under these situations and how to modify the existing coordinating contracts to reach this coordination.

References

[1] T.F. Bresnahan and P.C. Reiss, "Dealer and manufacturer margins", Rand J. Econ, 1985, 16, pp. 253–268.

[2] C.J. Corbett, D. Zhou and C.S. Tang, "Designing supply contracts : contract type and information asymmetry", Management Science, 2004, 50(4), pp. 550–559.

[3] A. Kanda and S.G. Deshmukh, "Supply chain coordination : perspectives, empirical studies and research directions", Int. J. Prod. Econ, 2008, 115(2), pp. 316–335.

.

[4] A. Tsay, "The quantity flexibility contract and supplier–customer incentives", Management Science, 1999, 45 (10), pp. 1339–1358.

[5] G.P. Cachon, "Supply chain coordination with contracts", Handbooks in Operations Research and Management Science, vol. 11, Elsevier, North Holland, 2003, pp. 227–339

[6] M. Dada and Q. Hu, "Financing newsvendor inventory", Operations Research Letters, 2008, 36(5), pp. 569-573.

[7] C.H. Lee, B.D. Rhee, "Trade credit for SC coordination", European Journal of Operational Research 2011; 214(1), pp.136–146.

[8] J.Yan, X. Wang, H. Cheng and L. Huang, "Study on the coordination contract in Supply chain under trade credit based on risk compensation", Chaos, Solitons & Fractals, 2016, 89, pp. 533-538.

[9] P. Kouvelis and W. Zhao, "Supply chain contract design under financial constraints and bankruptcy costs", Management Science, 2015, 62(8), pp. 2341–2357

[10] T. Taylor, "Supply chain coordination under channel rebates with sales effort effects", Management Science, 2002, 48(8), pp.992–1007.

[11] X. Yan and G. S Zaric, "Families of supply chain coordinating contracts in the presence of retailer effort", International Journal of Production Economics, 2016, 175, pp. 213-225.

[12] Li, Liying, and W. Yong, "Coordinating a supply chain with a loss-averse retailer and effort dependent demand", The Scientific World Journal, 2014.

[13] H. Krishnan, R. Kapuscinski and D.A. Butz, "Coordinating contracts for decentralized Supply chain with retailer promotional effort", Management Science, 2004, 50(1), pp. 48–63.

[14] Y. Gerchak and M. Parlar, "A single period inventory problem with partially controllable demand", Computers & operations research, 1987, 14(1), pp. 1–9.

[15] R.C. Rao, "Compensating heterogeneous salesforces: Some explicit solutions", Marketing Science, 1990, 9(4), pp. 319–341.

Supply Chain Management

MPC Approach for Synchronized Supply Chains of Perishable Goods

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Tomás Hipólito*, João Lemos Nabais*†, Rafael Carmona-Benítez‡, Miguel Ayala Botto* and Rudy R. Negenborn§
*IDMEC, Instituto Superior Técnico
Universidade de Lisboa, Lisboa, Portugal
Email: tom.hipolito@gmail.com, ayalabotto@tecnico.ulisboa.pt
†School of Business Administration
Instituto Politécnico de Setúbal, Setúbal, Portugal
Email: joao.nabais@esce.ips.pt
‡School of Business and Economics
Universidad Anáhuac México, México
Email: rafael.carmona@anahuac.mx
§Department of Maritime and Transport Technology
Delft University of Technology, The Netherlands
Email: R.R.Negenborn@tudelft.nl

***Abstract*—The movement of perishable goods is growing worldwide. Perishable goods need to be available to the market before the expiration date. With the decrease in inventory levels the components of a supply chain become even more integrated and dependent on coordinated decisions. Information regarding perishable goods must be visible throughout the supply chain for avoiding goods loss. A Model Predictive Control (MPC) heuristic to address operations management at supply chains of perishable goods is proposed in this paper. The approach is capable to follow the remaining time until expiration date which is critical to avoid losses. The supply chain is modeled using a state-space representation and operations management at the supply chain is formulated as an MPC Problem. In order to cope with operational decisions, the problem is solved on a periodic basis. The proposed approach is capable to deal with production decisions, monitor work-in-progress (WIP), and make transport assignments while monitoring the remaining time until the expiration date. Flows over the supply chain can be synchronized and therefore we named this type of supply chain a Synchronized Supply Chain (SSC). The approach is modular and easily scalable for large-scale supply chains. Numerical results illustrate these statements.**

I. Introduction

In supply chains (SC) multiple partners (suppliers, manufacturers, retailers) are contributing to move commodities from a source node to a destination node [1]. The challenges posed to operations management at supply chains are increasing in complexity with the spatial distribution of the network. These challenge increase when the value of goods is limited on time, such is the case when dealing with perishable. Moving these goods wisely is critical to assure they are delivered at the right location on the agreed time and quantity in order to be able to collect a revenue.

Supply chains can be found in different domains such as food [2], [3], distribution and retail. A supply chain can move a single commodity or multiple commodities. Commodities can be categorized in two different classes:

- Time-Constant Classes: such as the type of goods (fruit, milk, meat, fish,...), the volume (type of container, individual box,...), the final client location, the client priority, the weight, the temperature of transport;
- Time-Varying Classes: related to the available time to deliver the goods such that a revenue can be received, for instance the expiration date.

A distinction between raw materials (products without an expiration date) and perishable goods (products with an expiration date associated with) is made in this paper. Examples of perishable goods are: different types of fruits, yoghurt, fresh milk, fresh meat and fresh fish. In the case of perishable goods, the economic value of the product is limited in time which means that it has to be transported and delivered where needed respecting its expiration date. Also in some domains, the delivery of cargo should respect a pre-defined time window, which has to be fulfilled or contractual penalties will be triggered. In either case, operations management should track the remaining time until due time to proceed with decision or update decisions in case of need. It is important that information exchange happens throughout the supply chain and that the remaining time to deliver perishable goods is transparent for the different partners. In current supply chain management approaches the available time to deliver cargo is not taken into account [4], [5].

In this paper an MPC Heuristic is proposed to cope with operations management for synchronized supply chains of perishable goods making use of integrated modeling and control techniques, starting in the design phase. In this paper, market demand is interpreted as an exogenous input that disturbs the inventory levels across the supply chain state. Operations management is required to assign flows between nodes such that demand is satisfied without loosing the good value. Model Predictive Control (MPC) has shown to be successfully applied in process industry applications [6], and is now gaining increasing attention in fields like supply chains [7], power

networks [8], and water distribution networks [9]. By using mathematical models to describe the flows inside supply chains it is possible to make predictions about their future behavior. In supply chains, costs can be associated to flows and quantities of stored commodities. The model predictive controller can determine which actions have to be chosen in order to obtain the best performance taking into account the desired goals, existing constraints, disturbances and prediction information if available.

This paper is organized as follows. In Section II the modeling of supply chains is addressed using a state-space approach. Operations management is formulated and addressed using an MPC Heuristic in Section III. The performance of the proposed approach is tested through numerical simulations in Section IV. In Section V conclusions are drawn and future research topics are indicated.

II. MODELING

A. Conceptual Approach

1) Supply Chain: At a macroscopic level a supply chain exhibits two major phenomena [10]:

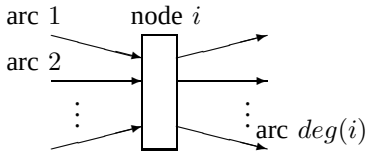
- Potential: related to the storage capability in well-defined areas, where commodities can be produced, manufactured or simply stored. These locations are modelled as *center nodes*, and are used to capture source nodes, intermediate nodes or downstream nodes (see Fig. 1(a));
- Flow: related to the transport or production delay. In this paper, this is modelled using a connection, which is composed of a succession of nodes with an in-degree and out-degree equal to one. Connection j is composed of n_{c_j} nodes and $n_{\mathrm{c}_j}+1$ flows (see Figure 1(b)).

In this paper, the production is not restricted to industrial manufacturing, meaning that if some features of a good have been changed that creates a new good. All supply chains are generally composed of center nodes and connections.

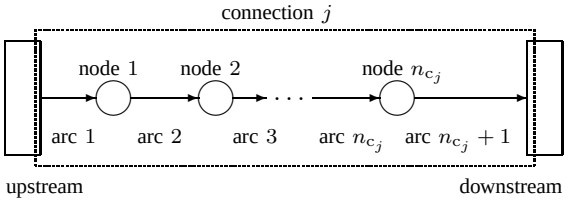
2) Commodities: This paper assumes that commodities in supply chains are categorized in two categories:

- raw materials: n_{m} is the amount of raw materials in the supply chain (e.g., clothes, package, bottles) which can be measured in units, volume or weight;
- perishable goods: n_{g} is the number of perishable goods in the supply chain (e.g., pears, pineapple, bananas). Moreover for each perishable good the remaining time until due time is considered via, n_{dt_p}, the number of time steps until due time of perishable good p, typically measured in days. This property is used to consider either the expiration date of perishable cargo or the agreed delivery time. For the sake of readability, from now on this property will be mentioned as due time.

The number of commodities in the supply chain is given by $n_{\mathrm{c}} = n_{\mathrm{m}} + n_{\mathrm{g}}$. The number of properties to track over time



(a) Center node of store type.



(b) Connection between two center nodes.

Fig. 1. Elementary components in a supply chain ($deg(i)$ stands for node degree).

is given by the contributions of raw materials and perishable goods due time, i.e.,

$$n_{\mathrm{p}} = n_{\mathrm{m}} + \sum_{i=1}^{n_{\mathrm{g}}} n_{\mathrm{dt}_i}. \tag{1}$$

B. Supply Chain Model

The total number of nodes inside the supply chain is associated with the topology and is given by

$$n_{\mathrm{y}} = n_{\mathrm{n}} + \sum_{j=1}^{n_{\mathrm{c}}} n_{\mathrm{c}_j}, \tag{2}$$

where n_{n} is the number of center nodes and n_{c_j} is the number of nodes belonging exclusively to connection j. For each node in the supply chain a state-space vector $\bar{\mathbf{x}}_i(k)$, $i = 1, \ldots, n_{\mathrm{y}}$ is defined, and these are merged to form the state-space vector $\mathbf{x}$ of the complete supply chain,

$$\bar{\mathbf{x}}_i(k) = \begin{bmatrix} x_i^1(k) \\ x_i^2(k) \\ \vdots \\ x_i^{n_{\mathrm{m}}}(k) \\ x_i^{1,1}(k) \\ \vdots \\ x_i^{1,n_{\mathrm{dt}_1}}(k) \\ \vdots \\ x_i^{n_{\mathrm{g}},1}(k) \\ \vdots \\ x_i^{n_{\mathrm{g}},n_{\mathrm{dt,g}}}(k) \end{bmatrix}, \mathbf{x}(k) = \begin{bmatrix} \bar{\mathbf{x}}_1(k) \\ \bar{\mathbf{x}}_2(k) \\ \vdots \\ \bar{\mathbf{x}}_{n_{\mathrm{y}}}(k) \end{bmatrix}, \tag{3}$$

where $x_i^{p,\mathrm{dt}_p}(k)$ is the amount of perishable good p with the maximum due time dt_p at node i at time step k, and $n_{\mathrm{dt,g}} = n_{\mathrm{dt}_{n_{\mathrm{g}}}}$ is the number of due times for time a given good n_{g}. The state-space dimension is given by $n_{\mathrm{x}} = n_{\mathrm{y}} n_{\mathrm{p}}$ corresponding to the number of properties handled and the

number of nodes existing in the supply chain. For the case of considering only raw materials, $n_{\mathrm{dt}_p} = 1$, what reduces to $n_{\mathrm{x}} = n_{\mathrm{y}} n_{\mathrm{c}}$. The state-space vector contains information about the amount per commodity not only at the center nodes, with significant storage capacity, but also at flow nodes. The total amount per commodity inside the supply chain is always accessible through the state-space vector: either in storage, transport or work-in-progress (WIP).

The demand is seen as an exogenous input $\mathbf{d}$ with length $n_{\mathrm{d}} = n_{\mathrm{c}}$. This means that the remaining due time is not tracked at the retail shop, and it is assumed that a FIFO policy is used at the retail shop shelves. Consider $u_j^{p,\mathrm{dt}_p}(k)$ as the amount of commodity p and due time dt_p to be pulled from node j at time step k. For all admissible flows inside the Supply Chain a control action vector is defined $\bar{\mathbf{u}}_j$ with length n_{p}. All $\bar{\mathbf{u}}_j$ ($j = 1, \ldots, n_{\mathrm{f}}$), where n_{f} is the number of available flows, are merged to form the overall control action vector $\mathbf{u}(k)$ with length $n_{\mathrm{u}} = n_{\mathrm{f}} n_{\mathrm{p}}$:

$$\bar{\mathbf{u}}_j(k) = \begin{bmatrix} u_j^1(k) \\ u_j^2(k) \\ \vdots \\ u_j^{n_{\mathrm{m}}}(k) \\ u_j^{1,1}(k) \\ \vdots \\ u_j^{1,n_{\mathrm{dt},1}}(k) \\ \vdots \\ u_j^{n_{\mathrm{g}},1}(k) \\ \vdots \\ u_j^{n_{\mathrm{g}},n_{\mathrm{dt,g}}}(k) \end{bmatrix}, \mathbf{u}(k) = \begin{bmatrix} \bar{\mathbf{u}}_1(k) \\ \bar{\mathbf{u}}_2(k) \\ \vdots \\ \bar{\mathbf{u}}_{n_{\mathrm{f}}}(k) \end{bmatrix}. \quad (4)$$

The model for the supply chain can then be represented in a state-space form as

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}_{\mathrm{u}}\mathbf{u}(k) + \mathbf{B}_{\mathrm{d}}\mathbf{d}(k), & (5) \\ \mathbf{y}(k) &= \mathbf{x}(k), & (6) \\ \mathbf{x}(k) &\geq \mathbf{0}, & (7) \end{aligned}$$

where $\mathbf{y}$ is the current amount per commodity at all nodes with dimension $n_{\mathrm{y}} = n_{\mathrm{x}}$, $\mathbf{A}$, $\mathbf{B}_{\mathrm{u}}$ and $\mathbf{B}_{\mathrm{d}}$ are the state-space matrices. The supply chain state $\mathbf{x}$ at the next time step, $k+1$, is determined using (5) as a function of the current supply chain state $\mathbf{x}$ plus the contribution due to the control action $\mathbf{u}$ and the demand. The control action $\mathbf{u}$ is the flow of commodities between nodes and is constrained by the available infrastructure resources.

C. Model Insight

In order to take advantage of some structural features, supply chain nodes should be ordered using the following policy: i) connections related to production lines are addressed first (flow nodes are numbered from upstream to downstream) after, ii) connections related to transport are addressed (flow nodes are numbered from upstream to downstream) then finally, iii) center nodes related to source, intermediate and destination nodes are numbered (the center nodes are numbered from upstream to downstream).

III. Operations Management

The use of Model Predictive Control (MPC) is justified by the ability to include constraints, predictions about the system behavior and exogenous inputs [6]. At each time step the controller first obtains the current state of the system. Then it formulates an optimization problem, using the desired goals, existing constraints, disturbances, and prediction information if available.

A. Problem Formulation

The cost function is generally a function of the supply chain states, control actions and desired states over the prediction horizon N_{p},

$$J(\tilde{\mathbf{x}}_k, \tilde{\mathbf{u}}_k, \tilde{\mathbf{x}}_{\mathrm{ref}}) = \sum_{l=0}^{N_{\mathrm{p}}-1} f(\mathbf{x}(k+1+l), \ldots \mathbf{u}(k+l), \mathbf{x}_{\mathrm{ref}}(k+1+l)), \quad (8)$$

where $\tilde{\mathbf{x}}_k$ is the vector composed of the state-space vectors for each time step over the prediction horizon $\begin{bmatrix} \mathbf{x}^{\mathrm{T}}(k+1) & \ldots, & \mathbf{x}^{\mathrm{T}}(k+N_{\mathrm{p}}) \end{bmatrix}^{\mathrm{T}}$, $\tilde{\mathbf{u}}_k$ is the vector composed of the control action vectors for each time step over the prediction horizon $\begin{bmatrix} \mathbf{u}^{\mathrm{T}}(k) & \ldots, & \mathbf{u}^{\mathrm{T}}(k+N_{\mathrm{p}}-1) \end{bmatrix}^{\mathrm{T}}$, $\mathbf{x}_{\mathrm{ref}}$ is the state-space reference vector and $\tilde{\mathbf{x}}_{ref}$ is the vector composed of the state-space reference vectors for each time step over the prediction horizon $\begin{bmatrix} \mathbf{x}_{ref}^{\mathrm{T}}(k+1) & \ldots, & \mathbf{x}_{ref}^{\mathrm{T}}(k+N_{\mathrm{p}}) \end{bmatrix}^{\mathrm{T}}$. The weights to be used in the objective function (8) are considered time-varying as a parameter that can change over time according to the variability of resources available. The MPC problem for the supply chain can now be formulated as:

$$\begin{aligned} \min_{\tilde{\mathbf{u}}_k} \quad & J(\tilde{\mathbf{x}}_k, \tilde{\mathbf{u}}_k, \tilde{\mathbf{x}}_{\mathrm{ref}}) & (9) \\ \text{s.t.} \quad & \mathbf{x}(k+1+l) = \mathbf{A}\mathbf{x}(k+l) + \mathbf{B}_{\mathrm{u}}\mathbf{u}(k+l) + \mathbf{B}_{\mathrm{d}}\mathbf{d}(k+l), & (10) \\ & \mathbf{y}(k+l) = \mathbf{x}(k+l), l = 0, \ldots, N_{\mathrm{p}}-1, & (11) \\ & \mathbf{x}(k+1+l) \geq \mathbf{0}, & (12) \\ & \mathbf{u}(k+l) \geq \mathbf{0}, & (13) \\ & \mathbf{y}(k+l) \leq \mathbf{y}_{\max}, & (14) \\ & \mathbf{P}_{\mathrm{uu}}\mathbf{u}(k+l) \leq \mathbf{u}_{\max}, & (15) \\ & \mathbf{x}(k+l) \geq \mathbf{P}_{\mathrm{xu}}\mathbf{u}(k+l), & (16) \\ & \mathbf{P}_{\mathrm{dx}}\mathbf{x}(k+1+l) \leq \mathbf{d}_{\mathrm{d}}(k+l), & (17) \end{aligned}$$

where $\mathbf{d}_{\mathrm{d}}$ is the vector introducing the market demand prediction, $\mathbf{y}_{\max}$ is the maximum storage capacities per node, $\mathbf{u}_{\max}$ the available infrastructure resources according to the supply chain layout, $\mathbf{P}_{\mathrm{xu}}$ is the projection from the control action set $\mathcal{U}$ into the state-space set $\mathcal{X}$ and $\mathbf{P}_{\mathrm{uu}}$ is the projection matrix from the control action set $\mathcal{U}$ into the infrastructure resource capacity set $\mathcal{U}_{\max}$ and $\mathbf{P}_{\mathrm{dx}}$ is the projection matrix from the state-space set into the disturbance set.

Constraints (12)–(17) are necessary in this framework as a guarantee that the supply chain behaves as expected:

- nonnegativity of states and control actions: negative storage and negative control actions (flows) are not physically possible, which is imposed by constraints (12)–(13);

- storage capacity: each supply chain node has to respect its storage capacity, this is captured in constraint (14);
- maximum control actions: the supply chain design in terms of available hardware used to implement desired flows is represented by constraint (15);
- feasible control actions: not all control actions that satisfy constraints (13) and (15) are feasible. The control action has to respect the existing amount per commodity in the each supply chain node at each time step. Constraint (16) imposes this relation;
- market demand: constraint (17) introduces the prediction if available.

B. Problem Insight

The MPC problem to solve has been formulated as a linear programming problem. In order to set the desired flow behavior for the supply chain through a pull-flow behavior, the following guidelines are given: i) the inventory reference at each center node is set as an upper-bound (by default the system will be producing to stock); ii) the weights at the flow nodes should be set as positive values in order to prevent the storage of commodities at these nodes, such that once a cargo starts a connection it should be delivered as soon as possible at the corresponding downstream node; and iii) weights at the center nodes need to be coordinated in order to promote the flow from upstream to downstream, that is to say, storing commodities at a downstream node has a higher benefit for 1 time period stay than the cost of transporting or producing the commodity.

IV. Numerical Experiments

In this section the MPC approach is applied for a supply chain composed of three tiers and handling raw materials and perishable goods.

A. Scenario Description

A time step of one day is considered for this supply chain. There are three tiers (supplier, distribution center and retail shop) in the supply chain (Fig. 2). Four commodities are supplied to the market: two raw materials (M_1 and M_2) and two perishable goods (G_1 and G_2) with a life time of 12 time steps (manufactured from the raw materials). The perishable goods are produced at the supplier tier according to a given proportion of raw materials (Table I). Once produced and made available at the supplier (node 5), the perishable good has only 12 time steps (days) to be delivered at the retail shop. Once delivered at the retail shop the due time of the perishable good is no longer tracked and it is assumed that the retail shop has into consideration a FIFO policy at the shelves.

		Raw materials		due time
		M_1	M_2	
perishable goods	G_1	1	1	12
	G_2	2	1	12

TABLE I. Perishable goods manufactured in the supply chain.

The supply chain is described by 7 nodes and 8 flows. These are used to capture the main features: i) Facilities –

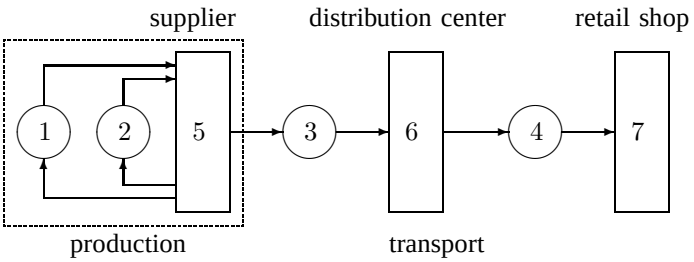


Fig. 2. Supply chain composed of three tiers, handling two raw materials and two perishable goods (boxes: center nodes; circles: flow nodes).

described by three center nodes (supplier, distribution center and retail shop); ii) Production lines – used to produce G_1 and G_2. Both have a lead-time of 2 time steps, composed of two flows and one flow node. They both start and end at the supplier (node 5); and iii) Transport connections – are used to transport all commodities from the supplier towards the retail shop. Both have a due time of 2 time steps, therefore are composed of two flows and one flow node. The supply chain capacities are discriminated per commodity and maximum capacity at each node and flow (Table II and Table III). The lead time between the supplier and the retail shop is 4 time steps for all commodities. The lead time between starting producing to availability at the retail shop is 6 time steps due to the production time for both perishable goods.

			total	commodities			
				M_1	M_2	G_1	G_2
flow nodes		node 1	2	0	0	2	0
		node 2	3	0	0	0	3
		node 3	14	4	5	2	3
		node 4	14	4	5	2	3
connections	connection 1	inflow	2	0	0	2	0
		outflow	2	0	0	2	0
	connection 2	inflow	3	0	0	0	3
		outflow	3	0	0	0	3
	connection 3	inflow	14	4	5	2	3
		outflow	14	4	5	2	3
	connection 4	inflow	14	4	5	2	3
		outflow	14	4	5	2	3

TABLE II. Connection capacities in the supply chain.

		total	commodities			
			M_1	M_2	G_1	G_2
Capacity	node 5	4000	1000	1000	1000	1000
	node 6	2000	500	500	500	500
	node 7	400	100	100	100	100
Inventory	node 5	2760	900	900	480	480
	node 6	68	18	20	14	16
	node 7	68	18	20	14	16

TABLE III. Node capacities in the supply chain and desired inventories at the retail shop.

According to the topology of the supply chain the incidence

matrix (relating nodes and flows) is given by,

$$\mathbf{D}(\mathcal{G}) = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{18}$$

B. Scenario Configuration

Consider that selling to public is starting in 20 days. Initially there are only inventory of raw materials at the supplier, in a sufficiently large amount to fulfill the expected demand. This means that all perishable goods must be produced. The market demand is assumed constant with value of 1 unit per time step for all commodities after day 20. The weights used for the construction of the optimization problem to be solved at each time step are indicated in Table IV, pulling commodities downstream but penalizing over-due time costs.

	commodity			
	M_1	M_2	G_1	G_2
Supplier	0	0	-1	-1
Distribution Center	-1	-1	-2	-2
Retail Shop	-2	-2	-20	-20
Over Due	–	–	200	250

TABLE IV. COST FUNCTION WEIGHTS.

C. Numerical Results

For different prediction horizons N_{p} and demand knowledge, different numerical experiments are evaluated in terms of:

- over-due cargo: capability to manage perishable goods without loosing value;
- warehouse capacity used at the distribution center;
- deviation to desired inventory at the retail shop;
- the amount of goods moved for each due time at the supply chain, as a measure of the risk of loosing goods value.

With a bigger prediction horizon the methodology approach performance increases, either with or without demand knowledge (Table V). Only for $N_p = 6$ perishable goods with a due time lower than de maximum lead-time (6 time steps) are being moved (Table VI). For N_p equal or higher to 8 no over-due happens for both scenarios. The performance in terms of warehouse capacity used at the distribution center and the deviation of inventory levels at the retail shop are equal for N_p equal or higher to 8. For this case scenario a prediction horizon higher then maximum due time available at the supply chain is not necessary. Recall that the maximum lead-time in the supply chain is lower than the maximum due time at the supply chain ($6 < 12$). In the case of knowledge of the demand, the inventory levels at the retail shop increase until the desired value is reached. Even when the demand starts, at day 20, the inventory level remains unaltered. For the case with no knowledge of the demand market, once the demand starts, inventory levels drops and remains for the rest of the simulation (see Fig 3).

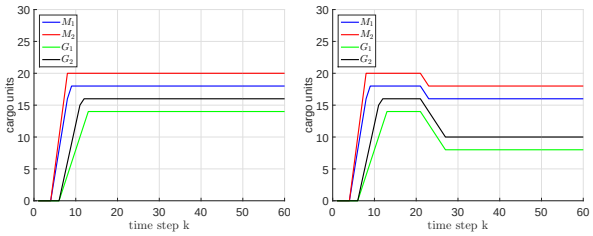


Fig. 3. Inventory levels at the retail shop for known demand (left) and unknown demand (right).

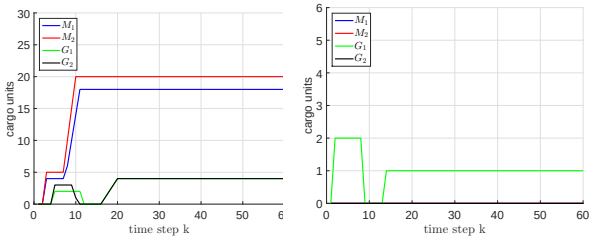


Fig. 4. Inventory per commodity at the distribution center (left) and Work In Progress (WIP) for connection 1 (right).

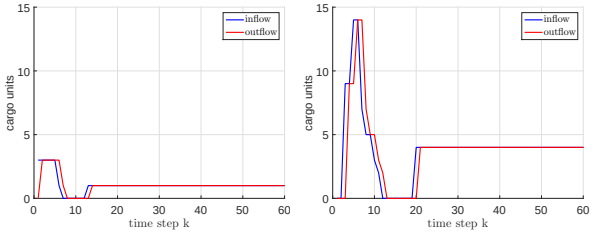


Fig. 5. Inflow and outflow for connections 2 (left) and 4 (right).

1) Insights in Prediction Horizon $N_{\mathrm{p}} = 14$: Due to a sufficiently large prediction horizon operations management is capable to fulfill the market demand for all commodities, while eliminating waste (no unnecessary production or unjustified transport). For detail on operations see Fig 4–5.

2) Insights in Prediction Horizon $N_{\mathrm{p}} = 6$: Due to a question of short vision a low performance is achieved for the supply chain (Fig. 6). The proposed approach is still able to pull the commodities (raw material and perishable goods) taking into account the market demand but at the cost of over-due cargo. It is possible to see that for the final time steps the inventory of G_1 and G_2 are dropping at the retail shop (in Fig. 7). Due to unnecessary production, too many resources of M_1 have been consumed, leading to the run out of stock at the supplier. Production of G_1 and G_2 is halted as well as transport.

V. CONCLUSION

In this paper an MPC heuristic was proposed to address operations management for perishable goods in supply chains. The approach is capable to track the remaining time until due time making it suitable to handle perishable goods. With this

Horizon	Over Due Goods		Inventory used at Distribution Center		Inventory error at Retail Shop	
	Known Demand	Unknown Demand	Known Demand	Unknown Demand	Known Demand	Unknown Demand
$N_{\mathrm{p}} = 6$	111	116	3710	3655	492	803
$N_{\mathrm{p}} = 8$	0	0	2436	2084	471	1009
$N_{\mathrm{p}} = 12$	0	0	2436	2082	471	1079
$N_{\mathrm{p}} = 14$	0	0	2436	2082	471	1079

TABLE V. PERFORMANCE EVALUATION FOR DIFFERENT PREDICTION HORIZONS.

Horizon		Commodities Due Times											
		dt_1	dt_2	dt_3	dt_4	dt_5	dt_6	dt_7	dt_8	dt_9	dt_{10}	dt_{11}	dt_{12}
Kown Demand	$N_{\mathrm{p}} = 6$	72	72	0	0	0	0	0	0	30	30	213	213
	$N_{\mathrm{p}} = 8$	0	0	0	0	0	80	82	0	30	30	120	122
	$N_{\mathrm{p}} = 12$	0	0	0	0	0	80	82	0	30	30	120	122
	$N_{\mathrm{p}} = 14$	0	0	0	0	0	80	82	0	30	30	120	122
Unkown Demand	$N_{\mathrm{p}} = 6$	69	69	0	0	0	0	0	0	30	30	215	215
	$N_{\mathrm{p}} = 8$	0	0	0	0	0	0	0	0	100	102	104	106
	$N_{\mathrm{p}} = 12$	0	0	0	0	0	0	0	0	98	100	102	104
	$N_{\mathrm{p}} = 14$	0	0	0	0	0	0	0	0	98	100	102	104

TABLE VI. EVALUATION OF DUE TIME FLOWS.

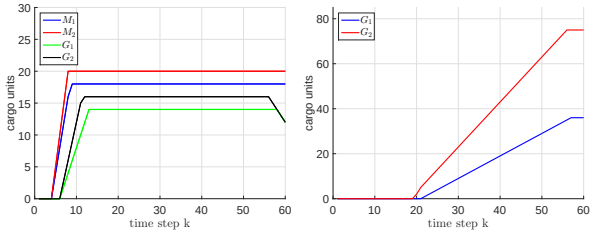


Fig. 6. Inventories at the retail for known demand (left) and unknown demand (right).

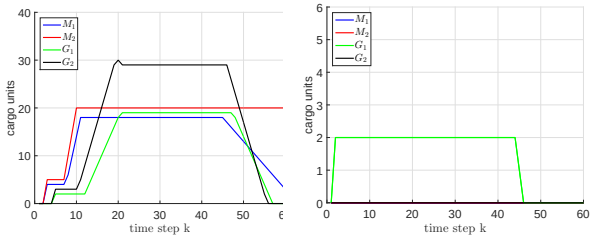


Fig. 7. Inventory per commodity at the distribution center (left) and Work In Progress (WIP) for G_1 (right).

capability it is possible to launch production orders according to the market demand, reducing the risk of losing value due to over production. Operations are synchronized throught the supply chain and intermediary nodes can be used soleley for X-docking activity. The approach is modular being easily scalable for large-scale systems. As future research the inclusion of a forecast module for market demand is worth to be considered and the analysis of the Forrester effect on perishable goods.

ACKNOWLEDGMENT

This work was supported by Fundação para a Ciência e Tecnologia (FCT), through IDMEC, under LAETA, project UID/EMS/50022/2013 and by the FCT, through IDMEC, under LAETA Pest-OE/EME/LA0022.

REFERENCES

[1] Ballou, R.H.: Business Logistics – Supply Chain Management: Planning, Organizing, and Controlling the Supply Chain. Prentice Hall (2004)

[2] Kuo, J-C., Chen, M-C.: Developing an advances Multi-Temperature Joint Distribution System for the food cold chain. Food Control **4** 559–566 (2010)

[3] Lin, X., Negenborn, R.R., Lodewijks, G..: Towards quality-aware control of perishable goods in synchromodal transport networks. In: Proceedings of the 5th IFAC Conference on Sensing, Control and Automation for Agriculture (IFAC AgriControl 2016), pp. 132–137, Seattle, Washington, August (2016)

[4] Wang, W., Rivera, D.E.: Model Predictive Control for Tactical Decision-Making in Semiconductor Manufacturing Supply Chain Management. IEEE Transactions on Control, Systems Technology, **16**, 5, 841–855 (2008)

[5] Alessandrini, A., Gaggero, M., Tonelli, F.: Min-Max and Predictive Control for the Managemetn of Distribution in Supply Chains. IEEE Transactions on Control, Systems Technology, **19**, 5, 1075–1089 (2011)

[6] Maciejowski, J.M.: Predictive control with constraints. Harlow, UK: Prentice Hall (2002)

[7] Maestre, J.M., Muñoz de la Pena, D., Camacho, E.F.: Distributed MPC: a supply chain case. In: Proceedings of the 48th IEEE CDC and 28th CCC, pp. 7099–7104, Shanghai, China, December (2009)

[8] Geyer, T., Larsson, M., Morari, M.: Hybrid emergency voltage control in power systems. In: Proceedings of the European Control Conference. Cambridge, UK. paper 322 (2003)

[9] Negenborn, R.R., Van Overloop, P.J., Keviczky, T., De Schutter, B.: Distributed model predictive control of irrigation canals. Networks and Heterogeneous Media **4** 359–380 (2009)

[10] Nabais, J.L., Negenborn, R.R., Carmona-Benítez, R., Mendonça, L., Botto, M.A.: Hierarchical Model Predictive Control for Multiple Commodity Transportation Networks, Distributed MPC Made Easy, Springer (2013)

[11] Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows. New Jersey, Prentice Hall (1993)

[12] Nabais, J.L., Negenborn, R.R., Carmona-Benítez, Mendonça, L.F., Lourenço, J., Botto, M.A.: Multi-Agent Architecture for Supply Chains using a Predictive Pull-Flow Perspective. Proceedings of the 11^{th} Int. Conf. on Practical Application of Agents and Multi-Agent Systems, pp.94–105, Salamanca, Spain, May (2013)

[13] Kvasnica, M., Grieder, P., Baotić, M.: Multi-parametric toolbox (MPT), (2004)

The selling mode in a supply chain with online retailing

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Wenchang Liu, Danqin Yang*
School of Economics and Management
Nanjing University of Science and Technology
Nanjing 210094, China
luwnca@163.com, yangdanqin@163.com

Abstract—In this paper, we consider a supply chain in which two competitive manufacturers sell products through an exclusive pure play e-tailer besides their traditional offline channels. Using a game-theoretical model, we analyze the e-tailer's optimal choice for the selling mode in the presence of upstream competition, cross-channel spillover and cost asymmetry between the online and offline channels. We reveal that the e-tailer always prefers the agency selling mode to the reselling mode, if the spillover effect of the e-channel on the offline channel is negative. Otherwise, the e-tailer prefers the reselling mode to the agency selling mode, if the cost advantage of the e-channel over the offline channel is small and the competition between the upstream manufacturers is intense. In addition, we find that the gap of the optimal retail price between the agency selling and reselling mode is increasing in the cross-channel spillover coefficient with symmetric selling cost between the e-channel and offline channel as well as the effect of the upstream competition on the gap is not monotonic.

Keywords—electronic retailing; supply chain management; upstream competition; cross-channel spillover; game theory

I. Introduction

This is a golden age of information, commercialization, and globalization. Both business modes and people's lifestyles have been turned upside down by the advent of platform-based e-commerce. It has been a global phenomenon that many customers choose to purchase products through online channel because of its great convenience, large product variety, low price and search cost, etc[1,2]. Driven by this noticeable trend, some manufacturers have begun to sell their products through the Internet channel, such as Lenovo and Sony[3]. In 2013, Amazon had over 2 million sellers whose sales accounted for roughly 40% of its total sales[4]. Therefore, competition between online sellers is becoming increasingly fierce.

Many prior researches on multi-channel emphasize the cannibalization effect of the e-channel on the offline channel[5]. However, recently some scholars claim that e-channel and offline channel may target different kinds of customers, leading to consumer segmentation[6,7]. Therefore, the introduction of the e-channel actually expands the potential market of the offline channel instead of cannibalizing the offline sales.

Currently, the e-tailer mainly uses two different selling mode: reselling mode and agency selling mode. For example, eBay operates as a pure platform and uses agency selling mode. However, Zappos, a celebrated online shoe retailer, has broken away from its initial agency selling mode and turned to reselling mode. One key difference between these two selling formats lies in who retains the control rights over the retail prices[8,9]. For reselling mode, the e-tailer acts as a conventional merchant who purchases products from the manufacturers and then resells them to customers. Thus, it is the e-tailer that sets the ultimate retail price. However, for agency selling mode, the e-tailer acts as a multi-side platform in which the manufacturers and the customers can directly interact with each other. The manufacturers decide their retail prices independently, but they have to share a certain portion of their revenues with the e-tailer as referral fees. For instance, Amazon charges its sellers a different commission fee for every sale through its marketplace depending on the category of the goods, e.g., 6% for PCs, and 15% for mobile phones[10].

The comparison between the reselling mode and the agency selling mode for online platform is first introduced by Hagiu[11] in 2007. Some researches focus on the price difference between these two selling formats. Reference [12] studies the pricing contract between one platform provider and one application provider, finding that the price in the agency selling mode is lower than that in the reselling mode due to the double marginalization effect. Different from [12], we consider both the spillover effect and the selling cost advatange of the e-channel over the offline channel in a supply chain with two competitive manufacturers and one e-tailer. Although we draw similar conclusion with [12] in a specific setting without selling cost advantage of the e-channel, but we find that an increase of the cross-channel spillover will enlarge the price gap between the agency selling and reselling mode and the effect of the upstream competition is not monotonic. Intriguingly, other scholars obtain an opposite result that the retail price in the agency selling mode is higher than that in the reselling mode, when consumers are locked in the downstream retailers[13], and the competition intensity upstream is less than downstream[14]. In addition, [15] studies the impact of third-party information on quality and fitness on the e-tailer's preference for different selling arrangements, demonstrating that its choice for selling format can function as a strategic tool to extract more profits.

This paper is closely related to [9] which models competition between downstream platforms. In contrast, we focus on competition between upstream manufacturers. We find that the effect of the competition on the e-tailer's

*Corresponding author

This work is supported by (i) National Natural Science Foundation of China under grant 71201084; (ii) Social Science Foundation of Jiangsu Province under grant 15GLC006.

preference for different selling modes in our model is opposite to that in [9]. Moreover, we discuss the impact from the cost advantage of the online channel over the offline channel on the e-tailer's optimal choice for selling mode, which is not considered in [9].

Overall, this paper contributes to the related literatures by addressing the e-tailer's optimal choice for selling format in a supply chain with competition between the upstream manufacturers, cross-channel spillover and cost asymmetry between online and offline channels. Our main conclusions reveal that if the spillover effect of the e-channel on the offline channel is negative, the e-tailer always prefers agency selling mode to reselling mode. Otherwise, the e-tailer's preference for agency selling mode is strengthened as the cost advantage of the e-channel over offline channel becomes more pronounced, but it is weakened as the upstream competition becomes fiercer.

The remainder of this paper proceeds as follows: section 2 describes the basic model. In section 3, we give the optimal decisions of the manufacturers and the e-tailer and analyze the impact of upstream competition, cross-channel spillover and cost asymmetry on the optimal decisions in reselling and agency selling mode, respectively. In section 4, we numerically give conditions for the e-tailer's preference for different selling format and compare the optimal retail prices in agency selling and reselling mode in the case with no cost advatange between online and offline channels. Section 5 concludes our main findings, proposes some managerial implications and offers avenues for further research.

II. Model Description

We consider a supply chain consisting of two symmetric manufacturers and a pure play e-tailer[1]. The monopoly e-tailer has the power to determine its retailing format which may be reselling or agency selling mode. Each manufacturer sells products through both the traditional retailer (i.e., offline channel) and the e-tailer (i.e., e-channel).

Assume that the demand of manufacturer i 's products sold through the e-channel is given by ($i, j = 1,2, i \neq j$)

$$q_i = \frac{1}{1+\beta} - \frac{1}{1-\beta^2} p_i + \frac{\beta}{1-\beta^2} p_j. \quad (1)$$

In Eq. (1), $0 \le \beta < 1$ captures the competition intensity and also the degree of differentiation between the two manufacturers[8,9], which might be resulted by brand loyalty, customer service, etc. Larger β implies that the degree of differentiation is lower and the competition is more intense. If $\beta = 0$, the two manufacturers are fully different and there is no competition; otherwise, the e-channel demand of manufacturer i depends not only on its own e-channel retail price (p_i), but also on its rival's e-channel retail price (p_j). This kind of demand function has desirable features, i.e., as β increases, the price sensitivity $1/(1-\beta^2)$ increases and the potential market size $1/(1+\beta)$ shrinks, which coincides with the accepted wisdom that consumers tend to be more price sensitive to the less differentiated (or more sustituable) products and less differentiated products appeal to less customers.

Assume that the offline demand of each manufacturer is $Q + \gamma(q_1 + q_2)$, where Q represents the base demand of the offline channel, and $\gamma \in [-1,1]$ is the cross-channel spillover coefficient. Note that we do not consider $\gamma < -1$, because for this case the negative influence of the online sales on the offline sales is so strong that the manufacturers have no motivation to enter the e-channel. One can see that γ can be either positive or negative, suggesting that the e-channel may bring additional revenues or generate channel conflicts for the offline channel. Without loss of generality, we normalize the offline price to 1, implying that γ does not alter the offline channel price. The underlying reasons may include established reference prices, high menu costs and so on[8]. Moreover, the unit selling cost of the e-channel is assumed to be 0, but that of the offline channel is c , $0 \le c \le 1$. This assumption is reasonable because in practice the e-channel is typically more cost effective. Larger c represents larger cost advantage of e-channel over offline channel. Based on these assumptions, we know that the offline profit of each manufacturer is $(1-c)[Q + \gamma(q_1 + q_2)]$.

III. Analysis

According to the selling format that the e-tailer chooses, we analyze the optimal decisions of the manufacturers and the e-tailer in two different settings.

3.1. Reselling Mode

First, we consider the setting where the e-tailer chooses the reselling mode (RR) . In this case, the timing of the game is as follows:

(i) *Stage* 1: two manufacturers decide their wholesale prices w_1 and w_2 simultaneously;

(ii) *Stage* 2: the e-tailer jointly decides the retail prices p_1 and p_2 for the two products.

The manufacturers try to maximize their profits as follows:

$$\underset{w_1}{Max}\ \pi_1^{RR} = (1-c)[Q + \gamma(q_1 + q_2)] + w_1 q_1, \quad (2)$$

$$\underset{w_2}{Max}\ \pi_2^{RR} = (1-c)[Q + \gamma(q_1 + q_2)] + w_2 q_2. \quad (3)$$

The first items of Eqs. (2)-(3) represent the manufacturers' profits from the offline channel, and the second items represent the manufacturers' e-channel profits.

And the e-tailer is trying to maximize its profit as follows:

$$\underset{p_1, p_2}{Max}\ \pi_E^{RR} = (p_1 - w_1)q_1 + (p_2 - w_2)q_2. \quad (4)$$

The first item of Eq. (4) represents the e-tailer's profit from selling manufacturer 1's product, and the second item represents its profit from selling manufacturer 2's product.

[1] A pure play e-tailer only has an online presence, so all its profits stem from e-channel, such as Amazon and eBay.

The subscript E represents the e-tailer. By using backward induction method, we can obtain the symmetric sub-game perfect equilibrium for RR in Proposition 1.

Proposition 1. With reselling mode, the optimal wholesale prices and retail prices are as follows:

$$w^{RR*} = w_1^{RR*} = w_2^{RR*} = (1-\beta)[1-\gamma(1-c)]/(2-\beta),$$
$$p^{RR*} = p_1^{RR*} = p_2^{RR*} = \{2-\beta+(1-\beta)[1-\gamma(1-c)]\}/[2(2-\beta)].$$

Proof. First, we analyze the e-tailer's optimal decisions given the manufacturers' wholesale prices.

From Eq. (1) and (4), we know that the Hessian matrix of π_E^{RR} with respect to p_1 and p_2 is $\begin{bmatrix} -2/(1-\beta^2) & 2\beta/(1-\beta^2) \\ 2\beta/(1-\beta^2) & -2/(1-\beta^2) \end{bmatrix}$, which is negatively definite. Thus, the optimal solution satisfies the first-order conditions:

$$\begin{cases} \partial\pi_E^{RR}/\partial p_1 = (1-\beta-2p_1+w_1+2\beta p_2-\beta w_2)/(1-\beta^2) = 0, \\ \partial\pi_E^{RR}/\partial p_2 = (1-\beta-2p_2+w_2+2\beta p_1-\beta w_1)/(1-\beta^2) = 0. \end{cases}$$

By solving the above equations, we obtain the e-tailer's best response function $p_i^{RR*}(w_i) = (1+w_i)/2, i \in \{1,2\}$. Inserting them into Eqs. (2)-(3), we know that manufacturer i chooses w_i to maximize $\pi_i^{RR*} = (1-c)[Q+\gamma(q_1(p_1^{RR*}, p_2^{RR*}) + q_2(p_1^{RR*}, p_2^{RR*}))] + w_i q_i(p_1^{RR*}, p_2^{RR*})$. Differentiating π_i^{RR*} twice with respect to w_i, we have $\partial^2\pi_i^{RR*}/\partial w_i^2 = -1/(1-\beta^2) < 0$. Thus, π_i^{RR*} is a concave function of w_i, so that the unique optimal solution satisfies the first-order conditions, i.e.,

$$\begin{cases} \partial\pi_1^{RR*}/\partial w_1 = \{(1-\beta)[1-\gamma(1-c)]-2w_1+\beta w_2\}/[2(1-\beta^2)] = 0 \\ \partial\pi_2^{RR*}/\partial w_2 = \{(1-\beta)[1-\gamma(1-c)]-2w_2+\beta w_1\}/[2(1-\beta^2)] = 0 \end{cases}.$$

By solving the above equations, we obtain the optimal wholesale prices $w^{RR*} = w_1^{RR*} = w_2^{RR*} = (1-\beta)[1-\gamma(1-c)]/(2-\beta)$.

Inserting them into $p_i^{RR*}(w_i)$, the optimal retail prices are:

$$p^{RR*} = p_1^{RR*} = p_2^{RR*} = \{2-\beta+(1-\beta)[1-\gamma(1-c)]\}/[2(2-\beta)]. \quad \square$$

From Proposition 1, we derive the following observations:

(i) When the cross-channel spillover coefficient γ is negative, i.e., the online sales cannibalize the offline sales, the two manufacturers set a high wholesale price to induce a high e-channel retail price for the purpose of decreasing the online sales. In contrast, when γ is positive, i.e., the online sales increases the offline sales, they set a low wholesale price to induce a low retail price in order to accelerate the online sales. In other words, as γ increases, the optimal wholesale prices and retail prices both decrease. Nevertheless, the negative effect of γ on the optimal wholesale price is larger than that on the optimal retail price. Therefore, the e-tailer's optimal profit margin increases in γ. The online sales are also fueled by low retail prices, so the e-tailer's optimal profit increases in γ.

(ii) The impact of the competition intensity β on the optimal wholesale prices, retail prices and profit margin is similar to that of γ. One important difference is that the e-tailer's optimal profit doesn't monotonically increase with β. This is because although larger β leads to a higher margin and spurs demand due to lower retail prices, the potential market size $1/(1+\beta)$ also shrinks, implying that the demand may decrease, which hurts the e-tailer. In other words, larger upstream competition is not always beneficial to the e-tailer. This is different from the convential wisdom which indicates that larger upstream competition restricts upstream suppliers' bargaining power and is beneficial to the downstream buyers[16,17].

(iii) The impact of the cost advantage of the e-channel over the offline channel c on the optimal wholesale prices, retail prices and the e-channel profit margin depends on γ. On one hand, for $\gamma < 0$, as c increases, the manufacturers' offline loss incurred by the e-channel is reduced. Thus they decrease the wholesale prices to secure a high profit by increasing the online sales, which induces the e-tailer to reduce the e-channel retail prices. On the other hand, for $\gamma > 0$, as c increases, the manufacturers' offline benefits from the e-channel is reduced. Thus they increase the wholesale prices to secure a high profit, which leads the e-tailer to raise the retail prices. Moreover, c has a larger impact on the wholesale prices than the retail prices. Thus, the e-channel profit margin is increasing in c for $\gamma < 0$, and decreasing in c for $\gamma > 0$.

3.2. Agency Selling Mode

In this section, we consider the setting where the e-tailer chooses the agency selling mode (AA). The fundamental difference of this mode compared with RR lies in that now it is the manufacturer that controls the e-channel price. In this case, the timing of the game is as follows:

(i) *Stage* 1: the e-tailer jointly decides the ratios of the revenue share (i.e., unit agency fees) α_1 and α_2 for the two manufacturers;

(ii) *Stage* 2: two manufacturers decide their e-channel retail prices p_1 and p_2 simultaneously.

The two manufacturers try to solve the following problem:

$$\underset{p_1}{Max}\ \pi_1^{AA} = (1-c)[Q+\gamma(q_1+q_2)]+(1-\alpha_1)p_1q_1, \quad (5)$$

$$\underset{p_2}{Max}\ \pi_2^{AA} = (1-c)[Q+\gamma(q_1+q_2)]+(1-\alpha_2)p_2q_2. \quad (6)$$

And the e-tailer is is trying to solve the following problem:

$$\underset{\alpha_1,\alpha_2}{Max}\ \pi_E^{AA} = \alpha_1 p_1 q_1 + \alpha_2 p_2 q_2. \quad (7)$$

By using similar solving method as section 3.1, we obtain the sub-game perfect equilibrium of AA in Proposition 2.

Proposition 2. With agency selling mode, the optimal ratios of revenue share and retail prices are as follows:

$$\alpha^{AA*}=\alpha_1^{AA*}=\alpha_2^{AA*}=1-\{M+\gamma(1-c)[\beta-\gamma(1-\beta)(1-c)]/(3M)\},$$

$$p^{AA*}=p_1^{AA*}=p_2^{AA*}=(1-\beta)[1-\alpha^{AA*}-\gamma(1-c)]/[(1-\alpha^{AA*})(2-\beta)],$$

where $M=\{54\gamma^2(1-c)^2(1-\beta)+[54^2\gamma^4(1-c)^4(1-\beta)^2 -108\gamma^3(1-c)^3(\beta-\gamma(1-\beta)(1-c))^3]^{1/2}\}^{1/3}/(3\times 2^{1/3})$.

Proof. First, we analyze the manufacturers' optimal decisions given the e-tailer's ratios of revenue share. The second-order derivatives of Eqs. (5)-(6) with respect to p_i ($i=1,2$) are as follows:

$$\partial^2\pi_i^{AA}/(\partial p_i)^2=-2(1-\alpha_i)/(1-\beta^2)<0,$$

which indicates that the second-order condition holds. Thus, the optimal solution satisfies the first-order conditions, i.e.,

$$\begin{cases}\dfrac{\partial\pi_1^{AA}}{\partial p_1}=\dfrac{(1-\alpha_1)[1-2p_1-\beta(1-p_2)]-\gamma(1-\beta)(1-c)}{1-\beta^2}=0,\\ \dfrac{\partial\pi_2^{AA}}{\partial p_2}=\dfrac{(1-\alpha_2)[1-2p_2-\beta(1-p_1)]-\gamma(1-\beta)(1-c)}{1-\beta^2}=0.\end{cases}$$

As the two manufacturers are symmetric, we let $\alpha_1=\alpha_2=\alpha$. By solving the above equations, we obtain the manufacturers' best response function $p_1^{AA*}=p_2^{AA*}=p^{AA*}(\alpha)$ $=(1-\beta)(1-\alpha-\gamma)/[(1-\alpha)(2-\beta)]$. Inserting them into Eq. (7), the e-tailer chooses α to maximize

$$\pi_E^{AA\prime}=\frac{2\alpha(1-\beta)[1-\alpha-\gamma(1-c)][1-\alpha+\gamma(1-\beta)(1-c)]}{(1-\alpha)^2(2-\beta)^2(1+\beta)}.$$

Differentiating $\pi_E^{AA\prime}$ twice with respect to α, we have

$$\partial^2\pi_E^{AA\prime}/\partial\alpha^2=-2(1-\beta)(1-c)\phi(\gamma)/[(1-\alpha)^4(2-\beta)^2(1+\beta)]$$

where $g(\gamma)=\gamma(1-\beta)(1-c)(2+\alpha)+\beta(1-\alpha)$, $\phi(\gamma)=\gamma g(\gamma)$. According to the sign of γ, we divide it into three settings:

(i) For $\gamma=0$, we have $\pi_E^{AA\prime}=2\alpha(1-\beta)/[(2-\beta)^2(1+\beta)]$ and $\phi(\gamma)=0$. $\pi_E^{AA\prime}$ is increasing with α. Thus, the e-tailer will set the optimal unit agency fee $\alpha^*\big|_{\gamma=0}=1$.

(ii) For $\gamma<0$, we have $\lim_{\gamma\to 0^-}g(\gamma)=\beta(1-\alpha^*\big|_{\gamma=0})=0$. From $g'(\gamma)=(1-\beta)(1-c)(2+\alpha)>0$, we have $g(\gamma)<0$ and $\phi(\gamma)>0$ for $\gamma<0$.

(iii) For $\gamma>0$, we have $\lim_{\gamma\to 0^+}g(\gamma)=\beta(1-\alpha^*\big|_{\gamma=0})=0$. Similarly, from $g'(\gamma)>0$, we have $g(\gamma)>0$ and $\phi(\gamma)>0$ for $\gamma>0$.

Taken together, we have $\phi(\gamma)\geq 0$, so that $\partial^2\pi_E^{AA\prime}/\partial\alpha^2\leq 0$. Thus, the optimal solution satisfies the first-order condition:

$$\partial\pi_E^{AA\prime}/\partial\alpha=(1-\beta)\{3\alpha^2-\alpha^3+[1-\gamma(1-c)][1+\gamma(1-c)(1-\beta)]+\alpha$$
$$[3+\gamma^2(1-c)^2-\beta\gamma(1-c)(1+\gamma(1-c))]\}/[(1-\alpha)^3(2-\beta)^2(1+\beta)]=0.$$

Therefore, we can obtain the optimal unit agency fee α_1^{AA*} and α_2^{AA*}. Inserting them into $p^{AA*}(\alpha)$, we get the optimal retail prices p^{AA*} in Proposition 2. □

In order to characterize the effects of β and γ on the equilibrium decisions of AA, we consider a symmetric case in which the e-channel does not have selling cost advatange over the offline channel, i.e., $c=0$. Based on Proposition 2, we numerically describe the effects of β and γ on the optimal retail price and unit agency fee of AA in Figs. 1-2. From Figs. 1-2, we derive the following observations:

(i) With AA, the optimal retail price decreases with γ. The optimal unit agency fee increases with γ if $\gamma<0$; otherwise, it decreases with γ. We can explain it as follows: for $\gamma<0$, the e-channel is harmful to the offline channel, so the manufacturers are inclined to reduce their online sales. Consequently, they set high retail prices to mitigate this negative impact. In order to increase online sales, the e-tailer charges lower unit agency fee (α^{AA*}) to induce the two manufacturers to set lower retail prices. As the negative effect of γ decreases (i.e., γ increases), the manufacturers tend to reduce their retail prices, whereas the e- tailer increases the unit agency fee. For $\gamma>0$, the effect of the e-channel on the offline channel is positive. To benefit from this demand stimulating effect, the manufacturers set lower retail prices, which reduces the e-tailer's profit margin. The larger the γ, the lower the retail prices. Therefore, as γ increases, the e-tailer decreases α^{AA*} to prevent retail prices from plummeting.

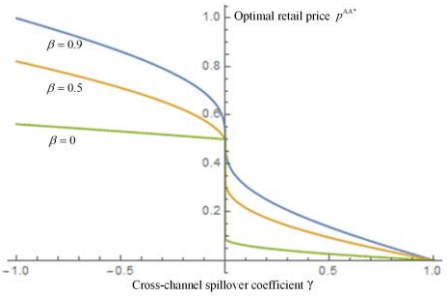


Fig. 1. The effect of β and γ on the optimal retail price for AA

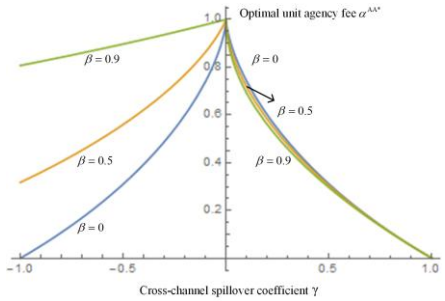


Fig. 2. The effect of β and γ on the optimal unit agency fee

(ii) If $\gamma < 0$, the optimal unit agency fee α^{AA*} increases with β; otherwise, α^{AA*} slightly decreases with β. The underlying reason is that the competition between the two manufacturers contributes to lowering the optimal retail prices, and this negative effect is weakened by an increasing γ. When $\gamma = 0$, the e-tailer extracts all profits of the e-channel irrespective of the magnitude of β, as there is no cross-channel spillover.

IV. Comparison between RR and AA

It is rather challenging to compare the equilibrium decisions for the two selling modes and give the e-tailer's the optimal choice for the selling mode theoretically, so we resort to numerical analysis.

4.1. The E-tailer's Optimal Choice for the Selling Mode

By comparing the e-tailer's optimal profits in RR and AA, we show the e-tailer's optimal choice for the selling format in Fig. 3. From Fig. 3, we observe that :

(i) For $\gamma < 0$, the e-tailer would rather choose AA than RR. This is because in RR, to weaken the negative effect of the e-channel on their offline channel, the manufacturers are inclined to charge high wholesale prices for the e-tailer, which incurs high retail prices and low e-channel demand as well. By comparison, in AA, the e-tailer is able to encourage the manufacturers to reduce their e-channel retail prices so that enhancing the online sales via adjusting the agency fee.

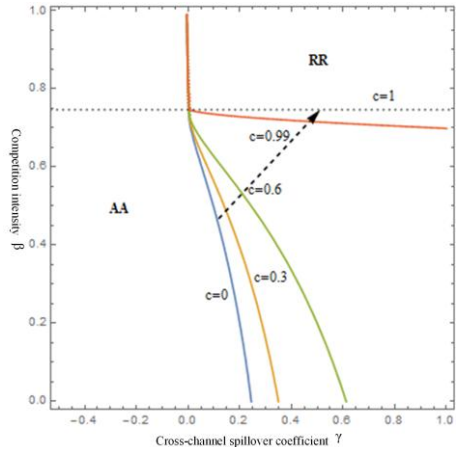


Fig. 3. The effect of β and γ on the e-tailer's preference for selling mode

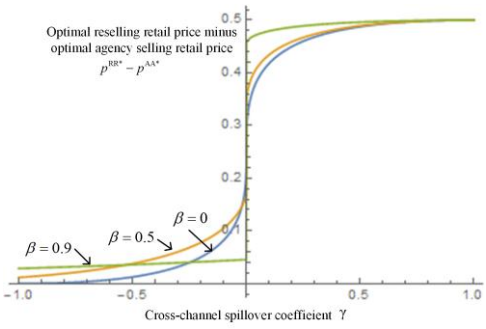


Fig. 4. The effect of β and γ on the gap of the optimal retail prices between RR and AA

(ii) For $\gamma > 0$, if γ is sufficiently large, the e-tailer prefers RR to AA, due to that the e-tailer benefits from low wholesale prices in RR, whereas the e-tailer is hurt by low retail price and unit agency fee in AA. Otherwise, if γ is sufficiently small, the e-tailer's optimal choice for selling format also depends on the magnitude of the competition intensity β and the cost advantage of the e-channel over the offline channel c. In RR, large β will induce a drastic decrease in wholesale prices, which weakens the double marginalization and benefits the e-tailer. Moreover, the e-tailer jointly decides the retail prices for the products of the two manufacturers to optimize its profit, which internalizes the competition. Thus, the e-tailer prefers RR to AA, if β is large; otherwise, the e-tailer sticks to AA.

(iii) When $\gamma > 0$, we have shown that as c increases, the manufacturers increase their wholesale prices in RR, so the double marginalization effect is enhanced. Thus, the e-tailer tends to switch from RR to AA as c becomes large. Particularly, $c = 1$ means that the manufacturers can not secure any profit from the offline channel, so that γ has no influence on the manufacturers' total profits. In this extreme case, the e-tailer's optimal selling format is only dependent on the competition between the manufacturers. As is indicated by the regions below or above the horizontal dotted line in Fig. 3, we know that if the competition between the two manufacturers is mild, the e-tailer chooses AA to leverage the upstream competition; otherwise it chooses RR to internalize the upstream competition. In general, the e-tailer's preference for AA is strengthened as c increases, but is weakened as β increases.

4.2. Comparison of the Optimal Retail Prices for RR and AA

In this part, we compare the optimal retail prices in RR and AA. To focus on the effect of β and γ, we assume $c = 0$. The results are demonstrated in Fig. 4. As is shown in Fig. 4, one can see that $p^{RR*} - p^{AA*}$ is always positive, implicating that the optimal retail price in AA is always lower than that in RR, due to the double marginalization effect. Similar result can be found in [12], which does not consider competition. In contrast, some other literatures draw an opposite conclusion because of the influence from the lock-in effect in the retailing terminal[13] and the downstream competition[14]. In addition, we can also see that the gap of the optimal retail price between RR and AA increases with γ, but is not monotonic with β. We can explain this phenomenon as follows: the optimal retail prices in both AA and RR decrease as γ increases, but γ has a larger impact on the retail prices in AA compared with RR. The underlying reason is that γ directly affects the retail prices in AA because it is the manufacturers that decide the retail prices. However, γ indirectly affects the retail prices in RR because it is the e-tailer that decides the retail prices. This indicates that the double marginalization effect is intensified by the cross-channel spillover in our model, which is not considered in [12].

V. Discussions and Conclusions

In this paper, we reveal that when the cross-channel spillover coefficient is negative, the e-tailer will unambiguously choose the agency selling mode, because it dampens the double marginalization, which is unavoidable in the reselling mode. When the cross-channel spillover coefficent is positive but small, the e-tailer prefers agency selling to reselling mode, if the competition intensity is small, otherwise the e-tailer prefers reselling to agency selling mode. Moreover, we find that the optimal retail price in the agency selling mode is always lower relative to that in the reselling mode. Therefore, the agency selling mode may be more popular with consumers. Finally, we show that the agency selling mode will become more prevalent, when the cost advantage of the e-channel over the offline channel is more pronounced. In summary, the e-tailer trades off between the channel efficiency in agency selling mode and the competition cushion effect in reselling mode, when choosing its optimal e-retailing format.

Most of the above conclusions are consistent with the observations in practice. Take Amazon for an example, for those goods that have negative cross-channel spillover such as electronic products[10], Amazon mainly adopts agency selling mode. By contrast, Amazon mainly uses reselling mode for those goods such as books and media products that may have possitive cross-channel spillover[9]. In addition, with more and more sellers flooding into Amazon's marketplce, the competition between upstream sellers becomes fiercer. As a result, Amazon also begins to apply the reselling mode (i.e., self-support retailing, logistics and services) to more product categories such as electronic products mentioned above that adopt the agency selling mode at the beginning.

These results shed some light on the operation management in the electronic retailing industry. We highlight the importance that the e-tailer should take into account the cross-channel spillover effect when deciding its optimal selling format. In our setting, we assume that the competition and the cross-channel spillover are exogenous. In fact, both the manufacturers and the e-tailer are able to endogenize these variables to some extent. On one hand, if the manufacturers achieve customer segmentation by implementing appropriate multi-channel scheme, the cannibalization effect can be reduced. If the manufacturers sell complementary products in different channels, the demand stimulating effect of the e-channel on the offline channel can be increased. On the other hand, the e-tailer can affect the competition intensity between the manufacturers via e.g., deliberately selecting its suppliers or frequently making part of its suppliers obsolete. In addition, with the rising of Industry 4.0, advanced production and management methods may change the operation cost of the manufacturers among multiple channels dramastically, which may influence the e-tailer's preference for different selling formats. Thus, it is necessary for the e-tailer to obtain, monitor and evaluate the detailed information about its supplier' operation cost before choosing the optimal selling mode in a e-retailing supply chain.

This paper remains to be explored further. Firstly, we assume a symmetric scenario and obtain symmetric equilibrium. However, power imbalance and asymmetric factors such as information and customer service exist widely in practice. Moreover, online price dispersion is a common phenomenon. Thus, we can extend our model by considering an asymmetric setting, e.g., the production costs of the two manufacturers may be different. Secondly, we don't consider demand uncertainty. Combining demand uncertainty can be a good outlet. Finally, pricing is usually a dynamic and evolving process. So it is a promising direction to extend this analysis to a multi-period setting.

References

[1] E. Brynjolfsson, Y. Hu, and M. D. Smith, "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers", Management Science, vol. 49, pp. 1580-1596, 2003.

[2] A. Dukes, and L. Liu, "Online Shopping Intermediaries: The Strategic Design of Search Environments", Management Science, vol. 62, pp. 1064-1077, 2015.

[3] R. Yan, "Managing channel coordination in a multi-channel manufacturer–retailer supply chain", Industrial Marketing Management, vol. 40, pp. 636-642, 2011.

[4] F. Zhu, and Q. Liu, "Competing with Complementors: An Empirical Look at Amazon.com", Harvard Business School Technology & Operations Mgt. Unit Working Paper, pp. 15-44, 2016.

[5] K. Cattani, W. Gilland, H. S. Heese, and J. Swaminathan, "Boiling Frogs: Pricing Strategies for a Manufacturer Adding a Direct Channel that Competes with the Traditional Channel", Production and Operations Management, vol. 15, pp. 40-56, 2006.

[6] B. Deleersnyder, I. Geyskens, K. Gielens, and M. Dekimpe, "How cannibalistic is the Internet channel? A study of the newspaper industry in the United Kingdom and the Netherlands", International Journal of Research in Marketing, vol. 19, pp. 337-348, 2002.

[7] F. J. Riggins, "A multichannel model of separating equilibrium in the face of the digital divide" , Journal of Management Information Systems, vol. 21, pp. 161-179, 2004.

[8] K. Jerath and Z. J. Zhang, "Store within a store", Journal of Marketing Research, vol. 47, pp. 748-763, 2010.

[9] V. Abhishek, and K. Jerath, and Z. J. Zhang, "Agency selling or reselling? Channel structures in electronic retailing". Management Science, vol. 62, pp. 2259-2280, 2015.

[10] J. Chen, J. E. Chen, K. Y. Goh, Y. C. Xu, and B. C. Y. Tan, "When do sellers bifurcate from Electronic Multisided Platforms? The effects of customer demand, competitive intensity, and service differentiation", Information & Management, vol. 51, pp. 972-983, 2014.

[11] H. Andrei, "Merchant or Two-Sided Platform?", Review of Network Economics, vol. 6, pp. 1-19, 2007.

[12] J. S. Gans, "Mobile application pricing", Information Economics and Policy, vol. 24, pp. 52-59, 2012.

[13] J. P. Johnson, "The Agency and Wholesale Models in Electronic Content Markets", Available at SSRN: https://ssrn.com/abstract=2126808, 2013.

[14] U. Foros, H. J. Kind, and G. Shaffer, "Turning the Page on Business Formats for Digital Platforms: Does the Agency Model Soften Competition?", Available at SSRN: https://ssrn.com/abstract=2317715, 2013.

[15] Y. Kwark, J. Chen, and S. Raghunathan, "Platform or Wholesale? A Strategic Tool for Online Retailers to Benefit from Third-Party Information", Working Paper. University of Texas at Dallas, 2015.

[16] Asker, John, and H. Bar-Isaac, "Raising retailers' profits: on vertical practices and the exclusion of rivals", The American Economic Review, vol. 2, pp. 672-686, 2014.

[17] Milliou, Chrysovalantou, and E. Petrakis, "Upstream horizontal mergers, vertical contracts, and bargaining", International Journal of Industrial Organization, vol.5, pp.963-987, 2007.

A Quality Model for a Multi-layered Sourcing Supply Chain

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

JrJung Lyu* and Chia-Wen Chen
Department of Industrial and Information Management
National Cheng Kung University
Tainan, Taiwan, R.O.C
jlyu@mail.ncku.edu.tw

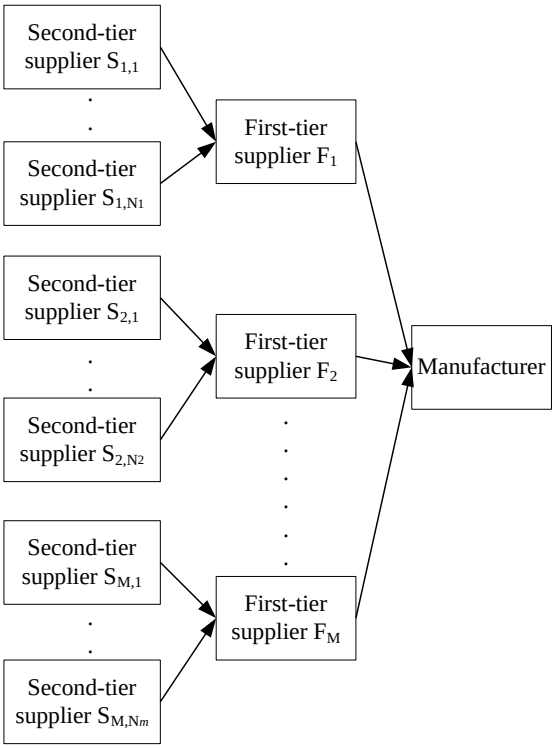
Abstract—Under globalization, many firms increasingly opt to blur their organizational boundaries to enhance the quality of both raw materials and final products. Manufacturers typically purchase components or parts from suppliers to manufacture products for customers. First-tier suppliers then purchase components from second-tier suppliers in a multi-layered sourcing supply chain. The quality of components or parts purchased from first- and second-tier suppliers affects the quality of finished products. In this study, a quality model is proposed to ensure the yield rate of each tier (second-tier suppliers, first-tier suppliers, and the manufacturer). This model can be applied to satisfy customers of manufacturers in a multi-layered sourcing supply chain.

Keywords—multi-layered sourcing supply chain, quality model

I. Introduction

As countless suppliers are available in the global marketplace, identifying appropriate suppliers via multiple sourcing to ensure product quality in a global supply chain is extremely important. Manufacturers typically purchase components or parts from suppliers to produce finished products for customers. First-tier suppliers then purchase components from second-tier suppliers in multi-layered sourcing supply chain. The quality of components and parts is central to supplier evaluation (Wu et al., 2008). Notably, the quality of components and parts purchased from first- and second-tier suppliers affects the quality of finished products. In a three-echelon supply chain, product defects are repaired prior to delivery to customers. Repair costs are based on the number of finished processes use to repair a defective product. Repair cost increases and has a significant effect on quality when a defect can be traced to the second-tier supplier. Therefore, only the considerations of processing yield rate of manufacturer and first-tier supplier are unsatisfied, and the second-tier supplier is also considered. As mentioned, the existence of multi-sourcing and multi-layered supply chains increase the difficulty of generating a quality model for a supply chain that ensures that product quality meets customer needs. In Fig. 1, Fm means the first-tier supplier providing the same part to a manufacturer; however, these parts have different quality. Sm,n purchases the required components from their second-tier suppliers, respectively. Therefore, how to select the best first- and second-tier suppliers to achieve product quality with minimal cost becomes increasingly complex and difficult.

Fig. 1. A typical multi-layered souring supply chain showing the relationships among the manufacturer, first-tier suppliers, and second-tier suppliers.



In this study, a multi-layered sourcing quality model (MSQM) is developed to address this problem in a supply chain. The proposed quality model can be applied to improve yield rate for each layer (second-tier supplier, first-tier supplier, and manufacturer) using inspection and repair to satisfy customers of manufacturers.

The remainder of paper is organized as follows. Section 2 reviews relevant literature. Section 3 describes the proposed quality model. Section 4 addresses the case study of a motorcycle manufacturer. Finally, Section 5 gives conclusions.

II. Literature Review

Quality cost estimation can be utilized to help manufacturers with different levels of risk aversion select suppliers to construct an appropriate quality model and satisfy customers in a multi-layered sourcing supply chain. In this era of global competition, suppliers are distributed worldwide; thus sourcing material is a significant task for global supply chains. Manufacturers typically pay close attention to supplier selection. Chan *et al.* (2008) proposed a fuzzy-based analytical hierarchy process to generate guidelines and directions for decision-makers when selecting appropriate global suppliers. An example from a manufacturing industry searching for a global supplier of a critical component was employed to demonstrate the effectiveness of the process. Abdel-Malek *et al.* (2005) developed a framework for assessing the performance of outsourcing strategies in a multi-layered supply chain. Cooper and Yoshikawa (1994) developed a three-phase supply chain framework that has an automobile assembly plant, first-tier supplier, and second-tier supplier. Inter-organizational systems were developed for sharing information to reduce the cost of products supplied by the chain. Ghodsypour and O'Brien (2001) generated a mixed-integer non-linear programming model to solve the multiple sourcing problem; this model accounts for the total cost of logistics. Tsai and Hsu (2010) proposed various cost of quality (CoQ) models to assist quality managers in administrating their requirements and resource constraints on selecting the optimal CoQ model developments. Plewa et al. (2016) proposed that total cost of quality and its failure cost component are significantly lower at higher levels of quality, while the prevention and appraisal cost components are not observed to be significantly higher at higher levels of quality.

Quality cost can be utilized to estimate the related costs between supplier and manufacturer. Freeman (1995) divided quality cost into prevention cost, appraisal cost, international failure cost, and external failure cost. Chang *et al.* (1996) developed a method for determining quality costs in multi-stage manufacturing systems by considering these four costs. Tagaras and Lee (1996) developed a quality cost model that considers quality cost, input quality, and imperfections during supplier selection.

A quality model can be constructed that considers quality cost. From the viewpoint of suppliers, Starbird (1994) applied numerical methods to examine how a risk-averse supplier reacts to acceptance of the sampling plan used by a customer. Starbird demonstrated that risk-averse suppliers deliver high-quality goods. Notably, manufacturers also have different levels of risk aversion. From the manufacturer perspective, Moskowitz and Plante (1984) presented an optimization algorithm that combines Bayesian sampling and a risk function to obtain an optimal sampling plan while maximizing expected utility. In considering the viewpoints of both suppliers and manufacturers, Reyniers (1992) indicated that quality control and a sampling plan are important to reducing cost for both suppliers and manufacturers.

III. Multi-layered Sourcing Quality Model (MSQM)

This study proposes a multi-layered sourcing quality model (MSQM) to improve the yield rate for each layer in a supply chain (second-tier supplier, first-tier supplier, and the manufacturer) using inspection and repair for manufacturers.

To meet the quality requirements of customers, the MSQM evaluates the total cost, including quality cost (cost of repairing defects) and purchase cost, for each supplier to help manufacturers select suppliers. A manufacturing operation can be protected against poor input quality by inspecting incoming materials.

Consider a manufacturer receiving one type of input material from a particular supplier. These materials are processed and transformed into finished products. Incoming materials and the manufacturing process may be imperfect. The simple module (SM) makes the following assumptions.

- Defective output is only identified in the inspection process.
- q_2, q_1, and q_0 are independent.
- No inspection of incoming material is performed.

Notably, q_2 and q_1 are the probabilities that a unit of incoming material from the second- and first-tier suppliers is defective, respectively. Additionally, q_0 is the probability that the manufacturing process is producing defective output.

The result of defective output can be rework, scrap, and external failure costs. The unit cost of defective output when defects are due to an imperfection is r_0. The unit cost of defective output is r_{01} and r_{02} when defects are produced from imperfect input from the first- and second-tier suppliers, respectively, and the manufacturing process is imperfect. The unit cost of defective output is r_{012} when the defect was produced using inputs from the first- and second-tier suppliers, and the manufacturing process is imperfect. Both r_1 and r_2 are the unit costs of a defective output when the defect is due to defective input from the first- and second-tier suppliers, respectively. Notably, r_{12} is the unit cost of defective output when the defect is due to defective input from the first- and second-tier suppliers. Thus, the probabilities for the incurrence of costs r_0, r_1, r_{01}, r_2, r_{02}, r_{12} and r_{012} are $(1-q_2)(1-q_1)q_0$, $(1-q_2)q_1(1-q_0)$, $(1-q_2)q_1q_0$, $q_2(1-q_1)(1-q_0)$, $q_2(1-q_1)q_0$, $q_2q_1(1-q_0)$, and $q_0q_1q_2$, respectively. Hence, for a given q_2 and q_1, the expected failure cost per unit processed is derived by Eq. (1).

$$\phi(q_2,q_1) = (1-q_2)(1-q_1)q_0r_0 + (1-q_2)q_1(1-q_0)r_1 + (1-q_2)q_1q_0r_{01} + q_2(1-q_1)(1-q_0)r_2 + q_2(1-q_1)q_0r_{02} + q_2q_1(1-q_0)r_{12} + q_0q_1q_2r_{012} \quad (1)$$

When the input material from the second- and first-tier suppliers is perfect, the expected failure cost per unit is

$\phi(q_2=0, q_1=0)=q_0 r_0$. The average quality cost of finished product with material from the second- and first-tier suppliers per unit, $Q(q_2, q_1)$ is derived by Eq. (2).

$$Q(q_2,q_1)=\phi(q_2,q_1)-\phi(q_2=0,q_1=0)=(1-q_2)q_1(1-q_0)r_1+(1-q_2)q_1q_0r_{01}+q_2(1-q_1)(1-q_0)r_2+q_2(1-q_1)q_0r_{02}+q_2q_1(1-q_0)r_{12}+q_0q_1q_2r_{012} \quad (2)$$

To evaluate a particular supplier, both purchase and quality costs must be considered. Generally, unit purchase cost depends on average product quality, *i.e.*, on probabilities q_2 and q_1 that a random unit purchased from a supplier is defective. Total acquisition and quality costs relevant to supplier evaluation are derived by Eq. (3). Let c be the unit cost of guaranteed conforming product (q_2=0, q_1=0), the unit purchase cost is then $C(q_2,q_1)=c(1-q_2)^2(1-q_1)^2$.

$$K(q_2,q_1)=C(q_2,q_1)+Q(q_2,q_1)=c(1-q_2)^2(1-q_1)^2+(1-q_2)q_1(1-q_0)r_1+(1-q_2)q_1q_0r_{01}+q_2(1-q_1)(1-q_0)r_2+q_2(1-q_1)q_0r_{02}+q_2q_1(1-q_0)r_{12}+q_0q_1q_2r_{012} \quad (3)$$

The relationship between q_2, q_1, and C(q_2, q_1) can be expressed in formulas. The optimal q_2 and q_1 (suppliers) satisfy the first derivatives $\frac{dK}{dq_2}=0$ and $\frac{dK}{dq_1}=0$, as well as the second derivatives $\frac{\partial K(q_2,q_1)}{\partial q_2^2}\geq 0$, $\frac{\partial K(q_2,q_1)}{\partial q_1^2}\geq 0$, and $\frac{\partial^2 K(q_2,q_1)}{\partial q_2 \partial q_1}=\frac{\partial^2 K(q_2,q_1)}{\partial q_1 \partial q_2}\geq 0$. The first derivatives of total cost with respect to q_2 and q_1 are derived by Eqs, (4) and (5), respectively.

$$\frac{\partial K(q_2,q_1)}{\partial q_2}=-2c(1-q_2)(1-q_1)^2-q_1(1-q_0)r_1-q_1q_0r_{01}+(1-q_1)(1-q_0)r_2+(1-q_1)q_0r_{02}+q_1(1-q_0)r_{12}+q_1q_0r_{012} \quad (4)$$

$$\frac{\partial K(q_2,q_1)}{\partial q_1}=-2c(1-q_2)^2(1-q_1)+(1-q_2)(1-q_0)r_1+(1-q_2)q_0r_{01}-q_2(1-q_0)r_2-q_2q_0r_{02}+q_2(1-q_0)r_{12}+q_0q_2r_{012} \quad (5)$$

The second derivatives of total cost with respect to q_2^2, q_1^2, and q_1q_2 (q_2q_1) are derived by Eqs. (6)–(8), respectively.

$$\frac{\partial^2 K}{\partial q_2^{\,2}}=2c(1-q_1)^2>0 \quad (6)$$

$$\frac{\partial^2 K}{\partial q_1^{\,2}}=2c(1-q_2)^2>0 \quad (7)$$

$$\frac{\partial^2 K}{\partial q_2 \partial q_1}=\frac{\partial^2 K}{\partial q_1 \partial q_2}=4c(1-q_2)(1-q_1)-(1-q_0)r_1-q_0r_{01}-(1-q_0)r_2-q_0r_{02}+(1-q_0)r_{12}+q_0r_{012} \quad (8)$$

To search for the minimum total cost, $\frac{\partial^2 K}{\partial q_2^2}*\frac{\partial^2 K}{\partial q_1^2}-\left(\frac{\partial^2 K}{\partial q_2 \partial q_1}\right)^2$ must be larger than zero. When the output of Eq. (9) exceeds zero, it means that $\frac{\partial K}{\partial q_2}=0$ and $\frac{\partial K}{\partial q_1}=0$. It indicates that the optimal solution (q_2^*, q_1^*) are exist resulting the minimum K(q_2, q_1).

$$4c^2(1-q_2)^2(1-q_1)^2-\{4c(1-q_2)(1-q_1)+[(r_{12}-r_1-r_2)+q_0(r_1+r_2+r_{012}-r_{01}-r_{12}-r_{02})]\}^2 \quad (9)$$

Let $W=(1-q_0)r_1+q_0r_{01}+(1-q_0)r_2+q_0r_{02}-(1-q_0)r_{12}-q_0r_{012}$, $\frac{\partial K}{\partial q_2}=0$ and $\frac{\partial K}{\partial q_1}=0$ be represented by Eqs. (10) and (11), respectively. The solutions of q_1 and q_2 are acquired using Mathematica.

$$-2c(1-q_2)(1-q_1)^2-q_1W+(1-q_0)r_2+q_0r_{02}=0 \quad (10)$$

$$-2c(1-q_2)^2(1-q_1)-q_2W+(1-q_0)r_1+q_0r_{01}=0 \quad (11)$$

IV. Case Study

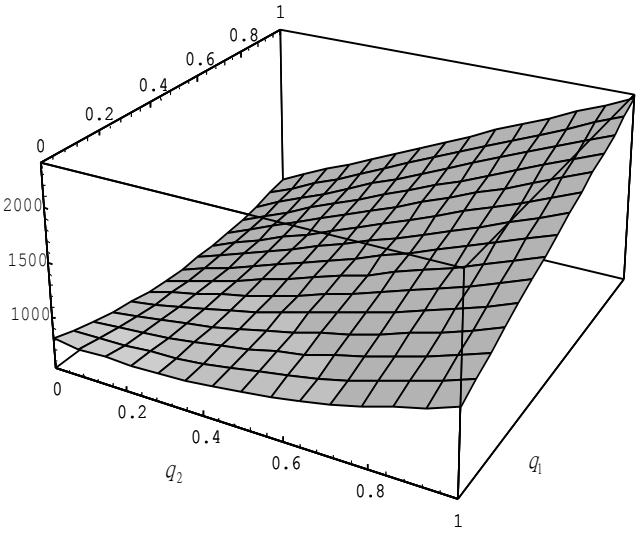
The case in this study is the supply chain of a motorcycle manufacturer, which includes the manufacturer (assembly plant), first-tier supplier (electrical machinery plant), and second-tier supplier (component plant). The case company's assembly plant purchases parts from the electrical machinery plant and sells motorcycles to customers. The first-tier supplier, the electrical machinery plant, purchases raw materials from the component plant and sells components to the assembly plant. The second-tier supplier, the component plant, provides raw materials to the electrical machinery plant for part production.

The defect rate of the assembly plant (q_0) is 0.04. The defect rates (q_1) of the electrical machinery plants, plants A and B (q_1^A and q_1^B), are 0.06 and 0.05, and defect rates (q_2) of the corresponding component plants (q_2^A and q_2^B), are 0.08, and 0.10. The unit cost of a defective output from the assembly plant (r_0), electrical machinery plants A and B (r_1), and component plants, second-tier supplier for first-tier supplier A and B (r_2) are 100, 800, and 1300. The unit costs of a defective output from the assembly plant and electrical machinery plant (r_{01}), assembly plant and component plant (r_{02}), and electrical

machinery plant and component plant (r_{12}) are 1500, 2000, and 2500, respectively. The unit cost of defective output from the assembly plant, electrical machinery plant, and component plant (denoted by r_{012}) is 5000. The unit purchase cost of guaranteed conforming parts from the electrical machinery plant (c) is 800.

Figure 2 illustrates feasible solutions of the total cost function when the manufacturer selects supplier A. The X-axis represents the defect rate of the second-tier supplier, Y-axis represents the processing defective rate of the first-tier supplier, and Z-axis represents the total cost. The SM can be used to determine the optimal yield rates for the first- and second-tier suppliers by considering quality cost and purchase cost to minimize total cost.

Fig. 2. Numerical illustration



After the defective rate and cost are read in, the total cost of the assembly plant purchasing parts from the electrical machinery plant via Eq. (3). For machinery plant A, total cost is 756.35. For machinery plant B, total cost is 761.24.

$$K(q_2^A, q_1^A) = C(q_2^A, q_1^A) + Q(q_2^A, q_1^A) = 598.30 + 158.05 = 756.35$$
$$K(q_2^B, q_1^B) = C(q_2^B, q_1^B) + Q(q_2^B, q_1^B) = 584.82 + 176.42 = 761.24$$

V. Conclusions

This study proposes a multi-layered sourcing quality model (MSQM) to help the manufacturer select the appropriate first- and second-tier suppliers while considering quality cost. In MSQM, assume that all parameters are fixed, except for defect rate of the second-tier supplier. The corresponding defect rate of the lowest total cost yields the optimal quality level of the supplier, resulting in the lowest cost for the manufacturer when considering both cost and manufacturer defect rate.

Under globalization, an appropriate quality model that considers quality cost is important for assuring product quality in a multi-layered sourcing supply chain. In addition to improving manufacturer capabilities, quality improvement in supplier product is required.

This study proposes MSQM to improve the yield rate for each layer (second-tier supplier, first-tier supplier, and manufacturer) in multi-layered sourcing supply chain. A motorcycle supply chain, which has a manufacturer (assembly plant), first-tier supplier (electrical machinery plant), and second-tier supplier (component plant), is used as a case study to assess the performance of the proposed quality model. The proposed MSQM can help manufacturer adjust their quality plans to reduce quality cost and improve customer satisfaction.

Acknowledgment

The authors would like to appreciate the programming efforts of Ms. Jiang-Wen Lin.

References

[1] Wu C.W., Shu, M.H., Peam, W.L., and Liu, K.H., 2008, Bootstrap Approach for Supplier Selection Based on Production Yield, *International Journal of Production Research*, Vol. 46, No. 18, pp. 5211-5230.

[2] Chan, F.T.S., Kumar, N. Tiwari, M.K., Lau, H.C.W., and Choy, K.L., 2008, Global Supplier Selection: A Fuzzy-AHP Approach, *International Journal of Production Research*, Vol. 46, No. 14, pp. 3825-3857.

[3] Abdel-Malek, L., Kullpattaranirun, T., and Nanthavanij, S., 2005, A Framework for Comparing Outsourcing Strategies in Multi-layered Supply Chains, *International Journal of Production Economics*, Vol. 97, pp. 318-328.

[4] Cooper, R. and Yoshikawa, T., 1994, Inter-organizational Cost Management Systems: The Case of the Tokyo - Yokohama - Kamakura Supplier Chain, *International Journal of Production Economics*, Vol. 37, No. 1, pp. 51-62.

[5] Ghodsypour S.H. and O'Brien C., 2001, The Total Cost of Logistics in Supplier Selection, under Conditions of Multiple Sourcing, Multiple Criteria and Capacity Constraint, *International Journal of Production Economics*, Vol. 73, pp. 15-27.

[6] Tsai W.H. and Hsu, W, 2010, A Novel Hybrid Model based on DEMATEL and ANP for Selecting Cost of Quality Model Development, *Total Quality Management & Business Excellence*, Vol. 21, No. 4, pp. 439-456.

[7] Plewa, M., Kaiser, G., Hartmann, E., 2016, Is Quality still Free? Empirical Evidence on Quality Cost in Modern Manufacturing, Internal Journal of Quality & Reliability Management, Vol. 33, No.9, pp. 1270-1285.

[8] Freeman, J.M., 1995, Estimating Quality Costs, *Journal of the Operational Research Society*, Vol. 46, No. 6, pp. 675-686.

[9] Chang, S.J., Park, Y.H., and Park, E.H., 1996, Quality Costs in Multi-Stage Manufacturing Systems, *Computers & Industrial Engineering*, Vol. 31, No. 1-2, pp. 115-118.

[10] Tagaras, G. and Lee, H.L., 1996, Economic Models for vendor Evaluation with Quality Cost Analysis, *Management Science*, Vol. 42, No. 11, pp. 1531-1543.

[11] Starbird, S.A., 1994, The Effect of Acceptance Sampling and Risk Aversion on the Quality Delivered by Suppliers, *Journal of the Operational Research Society*, Vol. 45, No. 3, pp. 309-320.

[12]. Moskowitz, H. and Plante, R., 1984, Effect of Risk Aversion on Single Sample Attribute Inspection Plans,

Management Science, Vol. 30, No. 10, pp. 1226-1237.

[13] Reyniers, D.J., 1992, Supplier-customer Interaction in Quality Control, *Annals of Operations Research 34*, pp. 307-330.

Integrated maintenance and distribution plan for an Urban Distribution Centre (UDC) using switching strategy

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Nadia Ndhaief[1], Rezg Nidhal[2]
LGIPM-Lorraine University
Metz, France
[1]nadia.ndhaief@univ-lorraine.fr
[2]nidhal.rezg@univ-lorraine.fr

Olivier Bistorin
LGIPM-ICN Business School
Nancy, France
olivier.bistorin@icn-groupe.fr

Abstract—In this paper, we present an integrated maintenance policy and urban distribution plan as a combined approach supporting urban distribution centre (UDC). This proposal is motivated by the emergent need of cities to have only one intermediary platform between retailers and customers. It has to satisfy random daily demands during a finite horizon under a service level constraint. Nevertheless, the delivery rate of UDC can be affected by random failure or preventive maintenance of vehicles. In doing so, we consider distribution policy with a strategy of switching between several UDCs to ensure service levels for their allocated urban area. This leads us to develop an approach combining switching strategy and preventive maintenance as a solution to satisfy random demands for UDC in a collaborative environment.

Keywords—Urban Distribution Center (UDC); switching strategy; integrated maintenance.

I. Introduction

Globalization has a significant impact on companies and organizations. All organization are facing considerable challenges dealing with performance, efficiency and economy. Moreover, organizations have to satisfy demand on time, make profit, satisfy criteria of customers, use innovation, etc. To that end, a collaboration is needed to share activities, knowledge and services.

Urban logistics is the process to improve all the logistics flows in urban areas while considering traffic, environment, congestion and safety [2]. The transport of goods in urban areas has known a growth due to the increasing of residents in cities and appearance of new kind of deliveries (e-commerce, home delivery, etc.). Consequently, this growth generates negative outcome as congestion, accident, more pollution to the urban areas and affect negatively the resident quality of life.

Many initiatives attempted to improve logistics flows in urban areas particularly in Urban Distribution Center (UDC) categorized as city terminals. UDC, in other works called Urban Consolidation Center (UCC), is a freight consolidations platform to solve the last mile problems. Here, deliveries are assembled, sorted and consolidated before forwarded for last mile distribution. In this case, many goals are achieved such as reducing vehicles, flow pooling, reducing pollution and reducing congestion [2, 3, 4]. Consequently, these platforms depend on many components (e.g. vehicles, software system, materials handling, etc.). Nevertheless, one-component failure risks reducing the delivery rate. Maintenance and distribution coordination will tackle this issue by defining a strategy for urban distribution centre to organize distributions and maintenance actions.

We do believe that collaboration between different UDCs can be more efficient to satisfy customers in urban area. In this context, platforms facing problems with their capacity or delivery will communicate and coordinate together for a mutual help and an efficient collaboration. Our approach is inspired by the work of Hajej el al. [5] where authors presented an integrate production/maintenance/transportation plan with strategy of switching between several productions sites in order to satisfy demands with a high service level.

Maintenance policy is discussed in several works. Hajej et al. combined production and maintenance plans in order to satisfy random demand [6]. In this case, the failure rate of the manufacturing system depends on the time and the production rate. Dellagi et al. [7] developed a maintenance strategy integrating subcontracting constraint where the system is described as a machine, which can make a call to one subcontractor to complete demand exceeding the available capacity. Moreover, Dellagi et al [8] has extended their work by developing an integrated maintenance policy for a manufacturing system, which can call several subcontractor machines in order to satisfy a constant demand.

In this paper, we will present a problem of an integrated distribution and maintenance policy for an urban distribution centre. This work consists of defining a collaborative approach to allow switching with different logistics platforms to satisfy customers' demands with a high service level. Our approach consists of finding the best distribution and maintenance plan and switching plan to minimize the total cost.

The paper is structured as follows. Section 2 introduces the problem statement. Section 3 presents the mathematical model for searching the best production and maintenance plans using switching strategy. In section 3, we presented a numerical example illustrating the benefits of our approach. Finally, we conclude and discuss future direction in section 4.

II. Problem statement

The paper's scope is about distribution scheduling and maintenance plan for urban distribution centre within a collaboration framework across delivery sites.

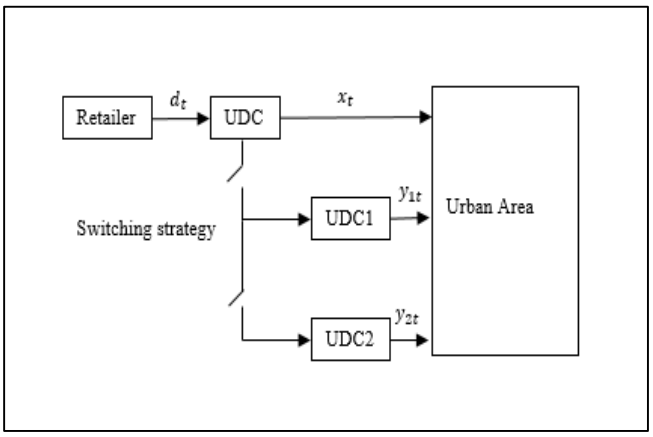


Fig. 1. Problem description

The idea is to define a distribution and maintenance plan for an urban distribution centre (UDC) where it can make a call to another UDC_i (i=1...L) in case of unavailability. So, we identify the amount of deliveries subcontracted to UDC i at each period t (y_{it}). It takes place when the concerned UDC cannot satisfy the total demand. This may happen when the distribution capacity is insufficient or during preventive or corrective maintenance period.

The global demand of the allocated area to the UDC is supposed known in advance at each period t. And the available capacity of UDC_i (Y_{it})is known. A delayed delivery Q_t can be considered if the demand exceeds the sum of the available capacity of the concerned UDC (C_t) and the available capacity of the UDC_i(Y_{it})to switch with.

Our model takes into account that the concerned UDC fails randomly. The paper focuses mainly on the overall unavailability of the system. A periodic preventive maintenance is performed and a minimal repair is carried out at each failure. The degradation is correlated with the delivery rate and time. Figure 1 illustrates the structure of the proposal approach.

III. The Mathematical model

We are given a planning horizon, partitioned equally into H periods. Every period is denoted by t, t= {1, 2... H}. The following parameters are used in the mathematical formulation of the model.

A. Parameters

- L : Number of UDC
- H : Number of periods
- Ct : The cost of transporting one unit of product from UDC to the final destination
- Ctr_i : Cost to subcontract the distribution of one product to platform i.
- Cp : Cost of one product delayed
- Y_{it} : Maximal number of product to be delivered by platform i in each period t
- d_t : Demand at each period
- $C_{\max}$: maximum capacity can be delivered by the main platform
- D_t : The sum between the demand at period t and the delayed products at period t
- Lp : Amount of decreased capacity due to preventive maintenance action
- Lr : Amount of decreased capacity due to corrective maintenance action
- M_p : Preventive maintenance cost
- M_c : Corrective maintenance cost
- Δt : Length of distribution period
- $\lambda_t(k)$: The failure rate at period t ($\forall k \in [0, \Delta t]$)
- λ_{nt} : The nominal failure rate at period t
- mu: monetary unit

B. Decision variables

- x_t : Quantity of deliveries transported from the concerned UDC at each period t
- y_{it} : quantity of products subcontracted to platform i at each period t
- Q_t : quantity of products delayed at each period t
- AF_{it} =1 if UDC i is chosen to ensure deliveries at period t, 0 else
- N : Optimal number of preventive maintenance actions
- T: The optimal period of intervention
- $\zeta_M(x, N)$: The expected number of failures

C. Mathematical model of Transportation with switching strategy

The idea is to minimize the cost of distribution over a finite horizon [0, H]. At each period, the concerned UDC can assume all deliveries or can switch with only one other site. In this case, the choice between UDCi's depends on the available capacity and the cost of subcontracting which depends on the distance between UDC and each of the two UDCi's. The goal of the switching strategy is to satisfy demands on time time. If demand at period t exceed the total capacity and all UDCi's, the unfulfilled demand must be rescheduled for delivery in the next period.". In this case, a penalty will be paid for each product not delivered at time (with $Ct < Ctr_i < Cp$). We assume that the distribution system has known a a known nominal delivery capacity expressed by Ct. This capacity changes according to maintenance operations.

The expression of the total cost of transportation using switching strategy with penalty paid for demand no delivered on time is expressed as follows:

$$Cd = \min \left\{ \begin{array}{l} \sum_t Ct \times x_t + \sum_t \sum_i Ctr_i \times y_{it} \times AF_{it} \\ + \sum_t Cp \times Q_t \end{array} \right\} \quad (1)$$

The following constraint defines that the distribution level cannot exceed the available capacity on delivery at each period

$$0 \le x_t \le C_t \quad \forall t \in H \quad (2)$$

The quantity of subcontracted quantity cannot exceed certain availability rate given by other UDC.

$$0 \le y_{it} \le Y_{it} \quad \forall t \in H \;\; i \in L \quad (3)$$

The value of delayed quantity is given by the following equation.

$$Q_t = \max[0, D_t - x_t - \sum_i y_{it} \times AF_{it}] \quad \forall t \in H \quad (4)$$

This equation gives the sum between the demand and the delayed product in the previous period. This quantity should become the refreshed demand.

$$D_{t+1} = d_{t+1} + Q_t \quad \forall t \in H \quad (5)$$

The following constraint defines that the main UDC can switch with only one other UDC at each period.

$$\sum_i AF_{it} \le 1 \quad \forall t \in H \quad (6)$$

This equation defines binary variables.

$$AF_{it} \in \{0,1\} \quad (7)$$

This equation defines integers variables

$$Q_t, x_t, y_{it} \;\; \text{integers} \quad (8)$$

D. Maintenance Cost

The idea is to develop an optimal maintenance strategy with considering the UDC's degradation according to the distribution plan. In order to know the best time intervals when a preventive maintenance action must be carried out, is important to minimize the total cost of maintenance. Recall the UDC system studied is subject to random failures

The strategy of maintenance adopted here the preventive maintenance policy with minimal repair at failure. Perfect preventive maintenance is performed periodically at times Δt .T and makes the system as good as new. When the site fails between successive preventive maintenance, a minimal repair is performed which restores the system at the same state before failure.

And, at each maintenance action, in any way preventive or corrective actions, it consumes a certain percentage of this distribution rate. So, the preventive maintenance and corrective maintenance consumes respectively Lp=a.Cmax and Lr=b.Cmax.

The analytical expression of the total maintenance cost for the concerned UDC is expressed as follows:

$$C_M(x, N) = M_p \times (N-1) + M_c \times \zeta_M(x, N) \quad (9)$$

where $N \in \{1, 2, 3.....\}$ and $\zeta_M(x, N)$ corresponds to the expected number of failures that occur during the horizon H, considering the distribution rate in each period.

Furthermore, the failure rate of the main UDC increases in each period due to distribution plan. In this part, we are inspired by the work of Hajej et al [6] where the failure in the interval k varies according to distribution rate and it is accumulative in the beginning of each period. Is expressed as following:

$$\lambda_t(k) = \lambda_{t-1}(\Delta t) + \frac{x_t}{C_{\max}} \lambda_n(k) \qquad \forall k \in [0, \Delta t] \quad (10)$$

With $\lambda_0(k=0) = \lambda_0$ and $\Delta\lambda_t(k) = \frac{x_t}{C_{\max}} \lambda_n(k)$

$\lambda_n(k)$ is the nominal failure rate corresponding to the maximal delivery rate.

Using the failure rate expression, the average number of failure over the horizon H.Δt can be presented as following:

$$T^* = \frac{H}{N^*} \text{ intervention period.} \quad (11)$$

$$\zeta_M(x_t, N) = \sum_{j=0}^{N-1} \left[\sum_{m=In(j\times\frac{T}{\Delta t})+1}^{In\left((j+1)\times\frac{T}{\Delta t}\right)} \int_0^{\Delta t} \lambda_m(k)dk + \int_0^{(j+1)\times L - In\left((j+1)\times\frac{T}{\Delta t}\right)\times\Delta t} \lambda_{In\left((j+1)\times\frac{L}{\Delta t}\right)+1}(k)dk \right.$$
$$\left. + \int_{(j+1)\times T}^{\left(In\left((j+1)\times\frac{T}{\Delta t}\right)+1\right)\times\Delta t} \frac{\left(\left(In\left((j+1)\times\frac{T}{\Delta t}\right)+1\right)\right)}{C_{\max}} \times \lambda_n(k)dk \right]$$

As discussed above, the maintenance action will reduce the maximum available capacity. Let N the number of preventive maintenance is equal to H/T* with T*=k × Δt. So, the available capacity on delivery at each period will be calculated as follows:

$$C_t = C_{\max} - Lp - Lr \times \int_0^{\Delta t} \lambda(u + (t-1-(n-1)k)\Delta t)du \quad (12)$$

If the preventive maintenance takes place in period t

$$i.e. \quad t = n \times k \quad \forall n = \{1,..N\}$$

In other periods, the capacity C_t is given by:

$$C_t = C_{\max} - Lr \times \int_0^{\Delta t} \lambda(u + (t-1-(n-1)k)\Delta t)du \qquad (13)$$

$$If \quad (n-1)k+1 \le t \le n \times k - 1 \qquad \forall n = \{1,..N\}$$

IV. Numerical example

To determine the best distribution plan with switching strategy and the optimal number of preventive maintenance actions, the following heuristics has been used.

- Step1: Solve the Mathematical model of Transportation with switching strategy irrespective the maintenance operations so assuming that the UDC delivers with the maximum capacity. To do that we use the mixed integer solver FICO Xpress 8.0.
- Step2: Determine the optimal number of preventive maintenance actions according the distribution plan given by step 1 and then determine the available capacities at each period. We use the software MATHEMATICA.
- Step3: According to the step 2, we integrate these available capacities in our previous model in order to determine the new distribution plan. To do that we use the mixed integer solver FICO Xpress 8.0.

A. Step 1: Transportation with switching strategy plan

We consider an UDC system which has to meet demand at time in the following Table I. The unit transportation cost equals to 10 mu. In this step, we suppose that this UDC deliver with the maximal capacity called C_{max}=13.

Concerning the switching strategy, there are two UDCi's available to complement the service of the UDC if necessary. The available rates for each of the former are depicted in Tables II and III. For the first subcontractor UDC, the unit service cost equals to 15 mu. For the second, the unit service cost equals to 20 mu.

For no delivery product, the penalty cost equals to 50 mu. The number of period H is equals to 8 days.

TABLE I. Mean demand of urban area

t	1	2	3	4	5	6	7	8
d_t	15	17	15	15	15	14	16	14

TABLE II. Available capacity for UDC1

$Y_{1,1}$	$Y_{1,2}$	$Y_{1,3}$	$Y_{1,4}$	$Y_{1,5}$	$Y_{1,6}$	$Y_{1,7}$	$Y_{1,8}$
4	5	1	7	2	2	2	3

TABLE III. Available capacity for UDC2

$Y_{2,1}$	$Y_{2,2}$	$Y_{2,3}$	$Y_{2,4}$	$Y_{2,5}$	$Y_{2,6}$	$Y_{2,7}$	$Y_{2,8}$
0	5	3	8	4	6	9	1

Our mathematical model has binary and integer variables with quadratic term so called quadratic program (QP). These quadratic term appear on the objective functions and constraints creating problems with convexity. In this section, we present first reformulation of our QP into a particular linear programming. Then, we solve this new model using the solver Xpress FICO 8.0.

So, we do a reformulation of following quadratic term of our model by using binary integer linearization techniques. The quadratic terms in our models a product between binary and integer variables. The integer variables are bounded below by 0 and above by big value.

We found in the following equation quadratic term:

- $$Cd = \min \left\{ \sum_t Ct \times x_t + \sum_k \sum_i Ctr_i \times y_{it} \times AF_{it} + \sum_t Cp \times Q_t \right\}$$

  (1)

- $$Q_t = \max[0, D_t - x_t - \sum_i y_{it} \times AF_{it}]$$

  (4)

AF_{it} is binary variable and y_{it} is integer and bounded.

$$0 \le y_{it} \le Y_{it} \quad \forall t \in H \; i \in L \; (3)$$

Then we introduce a new variable

$$B_{it} = y_{it} \times AF_{it} \quad \forall t \in H \; i \in L \; (14)$$

Furthermore, we added others valid inequalities to our model as follow:

$$B_{it} \le Y_{it} \times AF_{it} \quad \forall t \in H \; i \in L \; (15)$$

$$B_{it} \ge 0 \quad \forall t \in H \; i \in L \; (16)$$

$$B_{it} \le y_{it} \quad \forall t \in H \; i \in L \; (17)$$

The new problem is applied to the model. Results are presented in Table IV. Where is used, as already said, the Solver FICO Xpress 8.0. The total cost for the optimal distribution plan is equal to 1320 mu.

TABLE IV. Distribution plan with switching strategy using maximal capacity

Period	1	2	3	4	5	6	7	8

Distribution plan	13	13	13	13	13	13	13	13
Allocation to UDC_1	1	1	0	1	1	1	0	1
Quantity subcontracted to UDC_1	2	4	0	2	2	1	0	1
Allocation to UDC_2	0	0	1	0	0	0	1	0
Quantity subcontracted to UDC_2	0	0	2	0	0	0	3	0
Delayed quantity	0	0	0	0	0	0	0	0

B. Step2: Maintenance strategy

In this numerical example we suppose that our system has a degradation law characterized by a Weibull distribution. The Weibull scale and shape parameters are β=50 and α = 2. The preventive and corrective maintenance cost are respectively M_p= 220 mu, M_c = 1000 mu. In order to solve the maintenance cost, we used the software MATHEMATICA.

Fig 2. Illustrate the total cost of maintenance according to number of preventive maintenance actions. We conclude that the optimal number of preventive maintenance N is equal to 2. Thus, the optimal period of intervention T* is equal to 4. So the total cost of maintenance policy is equal to 1432 mu.

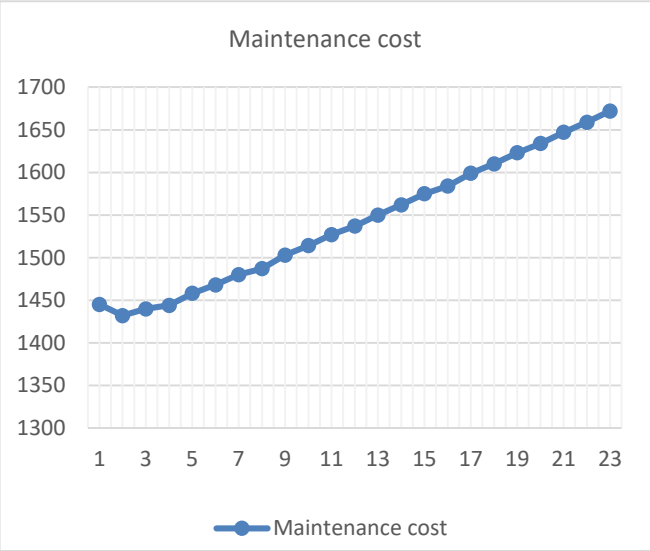


Fig. 2. Average total cost of maintenance according to N

Now, let L_p=1 (a=8%) an L_r=5 (b=39%)defining capacities lost due to maintenance actions. Hence, the Table V. shows the available delivery rate at each period taking into account maintenance actions effects.

TABLE V. The available capacity at each period t

C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
12	12	12	10	12	12	12	10

C. Step 3: The final distribution plan using switching strategy

Now we introduce the values of available capacities computes in step 2 in our programme in the solver FICO Xpress 8.0 and we obtained the following results (Table VI). Thus, the optimal cost of transportation of deliveries taking into account switching strategy and available capacities is equal to 1440 mu.

TABLE VI. Final distribution plan using switching strategy with available capacities

Period	1	2	3	4	5	6	7	8
Distribution plan	12	12	12	10	12	12	12	10
Allocation to UDC_1	1	1	0	1	0	1	0	1
Quantity subcontracted to UDC_1	3	5	0	5	0	2	0	3
Allocation to UDC_2	0	0	1	0	1	0	1	0
Quantity subcontracted to UDC_2	0	0	3	0	3	0	4	0
Delayed quantity	0	0	0	0	0	0	0	1

We can notice according to results that quantity subcontracted had increased and we had delayed quantity in the final period. This is due the decrease in value of available capacities. Consequently, the cost of distribution of delivers increased by 120 mu.

V. Conclusion

In this paper, we present the problem of integrated strategy for an urban distribution center (UDC) subject to a random failure. In order to satisfy the demand at each period, we integrate the switching strategy which the main UDC can switch with other UDC when demand exceed its available capacity. So a collaboration between different UDC and a satisfaction of demand will be achieved.

We have proposed a mathematical model for distribution plan using switching strategy and for maintenance strategy while considering available capacities and allocation constraint. For maintenance strategy, a minimal repair is carried out to restore the system and a periodically perfect preventive maintenance is performed. A numerical example is presented to determine the distribution plan and the optimal

number of preventive maintenance. in doing so, we compute a distribution plan with the maximum capacity on delivery using FICO Xpress 8.0. Then, we determine the optimal number of failure using the optimal distribution plan obtained in step 1 and we calculate the available capacity at each period. Finally, we obtain the distribution plan using this value of capacities.

As future work, we consider a distribution plan for multi urban distribution centers. So all available capacities for switching strategy for all UDCs will be variables. Moreover, we will determine their maintenance.

References

[1] Taniguchi, E. (2012, May). Concept and best practices of city logistics. International Transport Forum.

[2] Panero, M. A., Shin, H. S., & Lopez, D. P. (2011). Urban distribution centers–A Means to reducing freight vehicle miles traveled (No. C-08-23).

[3] Rao, C., Goh, M., Zhao, Y., & Zheng, J. (2015). Locationselection of city logistics centers undersustainability. Transportation Research Part D:Transport and Environment, 36, 29-44.

[4] Crainic, T. G., Ricciardi, N., & Storchi, G. (2009). Modelsfor evaluating and planning city logisticssystems. Transportation science, 43(4), 432-454.

[5] Zied, H., Nidhal, R., & Cormier, G. (2014, April). Integrated transportation and maintenance problem. In Networking, Sensing and Control (ICNSC), 2014 IEEE 11th International Conference on (pp. 474-479). IEEE.

[6] Hajej Z, Dellagi S, Rezg N (2011) Optimal integrated maintenance/production policy for randomly failing systems with variable failurerate. Int J Prod Res 49(19):5695–571218.

[7] Dellagi, S., Rezg, N. and Uie, U.,. Preventive maintenance of manufacturing systems under environmental constraints. International Journal of Production Research, 45 (5), 1233, 2007.

[8] Dellagi, S., Rezg, N., & Gharbi, A. (2010). Optimal maintenance/production policy for a manufacturing system subjected to random failure and calling upon several subcontractors. international journal of management science and engineering management, 5(4), 261-267.

Supply Chain Network Planning

A stochastic model for analyzing the combined effect of demand uncertainty and carbon tax within the context of Green Supply Chain Design

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Amina Chelly
University of Tunis El Manar, National Engineering School of Tunis, UR-OASIS, 1002 Tunis, Tunisie
Atidel B. Hadj-Alouane
University of Tunis El Manar, National Engineering School of Tunis, UR-OASIS, 1002 Tunis, Tunisie
Imen Noiura
Rennes School of Business, 2 rue Robert d'Arbrissel, CS76522, 35065 Rennes CEDEX, France
Yannick Frein
Univ. Grenoble Alpes, CNRS, Grenoble INP*, G-SCOP, F-38000 Grenoble, France

Abstract— Carbon emission legislations are implemented for the purpose of pushing companies to pay more attention to their carbon emissions beyond Supply Chain (SC) activities and encouraging them to invest in cleaner technologies. Within this context, the present work studies the combined effect of carbon legislation and demand uncertainty on one of the strategic decisions related to green supply chain design (GSCD): the selection of production technologies. We consider a stochastic demand and we develop a stochastic newsvendor-based mathematical model for the problem of production planning and technology selection under carbon tax regulation. The majority of works that have studied this problem have developed deterministic models and studied the trade-off between emission reduction and cost increase under carbon legislations. Through a numerical example, we compare the economic and environmental performances of both deterministic and stochastic models. With this example, we show how the carbon legislations tend to push companies to invest in cleaner production technologies when demand uncertainty is considered. The obtained results also illustrate the robustness of the stochastic model to demand uncertainty over the deterministic model.

Keywords: GSCD; carbon tax; carbon emission; economic and environmental performances, technology choice, stochastic demand.

I. Introduction

Global warming has been a serious issue for human beings on the Earth. Based on scientific evidence collected so far, the excessive emissions of Carbone Dioxide (CO_2) on the environment often due to over-developed human activities is one of the main causes behind rapid increase of earth's temperature [22]. Unfortunately, during the past three decades, the emissions of this harmful gas have seriously increased, by about 90%, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total CO_2 emissions increase from 1970 to 2011 [9]. Consequently, the international communities are engaged in developing legally binding international laws for carbon emission reduction [7] [8]. To bring down emissions and drive investments into cleaner production systems, policy-makers either provide incentives or subsidies or impose costs or/and caps on carbon emissions [5] [18]. Common carbon emission regulations, besides the subsidies, include cap, cap-and-trade and tax policies [23]. These different laws of carbon reduction are gradually being adopted by various economies across the world [2]. Meanwhile, a growing chorus of voices that include politicians, economists and business leaders have called for an emission tax policy rather than the cap-and-trade regulation mainly because it is much easier to implement [6]. As a result, 18 countries and one Canadian province have adopted this legislation until 2015 [19]. Following this practice, we also choose to study our problem under the carbon tax legislation.

Faced with such an increasing regulatory pressure, companies are beginning to recognize the importance of greening their SCs by minimizing carbon emissions of their activities [4] [12]. As a result, the concept of low carbon SC has been recently established, and received wide attentions across the business communities and academia within a very short time span. Companies and researchers strive to find new and innovative methods to reduce GHG emissions across all the stages of their GSCs [3] [14]. Investment decisions for environmental considerations still constitute a developing area in the operations research and management science literature [23].

Interestingly, most of the studies within this field were conducted in a deterministic context [23]. However, uncertainty may have a significant impact on the SC performance and thus, it should be considered in the early stages of SC design [10]. In the context of GSCD, both demand and regulations may be subject to uncertainty. Within this scope in mind, this paper intends to assess the impact of demand uncertainty, on certain strategic decisions related to GSCD, particularly under carbon legislation pressure. The objective is to show how the consideration of stochastic demand can modify the strategic decisions in an environment sensitive context. To do so, we consider the problem of technology investment under carbon tax legislation and uncertain demand. We develop a two-stage stochastic model of technology choice, where demand is a random variable.

* Institute of Engineering Univ. Grenoble Alpes

The paper is organized as follows: Section 2 is devoted to a review of existing literature that has dealt with carbon emission reduction through GSCD. We tend to highlight the gap in the literature and position our work and contribution. We dedicate section 3 to the problem statement. Our mathematical stochastic model and the solution method are described in section 4. In section 5, we present numerical experiments that intend to assess the performance of the stochastic model and conduct a comparative analysis against the deterministic modeling approach using some environmental and economic metrics. Finally, we give concluding remarks and new research perspectives.

II. Literature review

We dedicate this section to a review of the most significant studies within the green supply chain management field that deal especially with the GSCD problem under carbon emission considerations. To tackle this problem and help companies take the right strategic decisions within their SC design under carbon emission reduction's context, authors have studied different investment decisions and used various modeling techniques.

In [3], authors used a multi-objective model to assess operational adjustment and technology investment options in terms of cost and carbon emissions reduction. They used the sustainable order quantity model to prove that both options may be effective to lower the impacts of logistics operations under the carbon cap and the carbon tax rules. In [25], a study has addressed the combined impacts of regulation, consumer awareness and competitiveness in the technology investment decision. Authors in [23], analyzed a retailer's joint decisions on inventory replenishment and carbon emission reduction investment under three carbon emission regulation policies mainly carbon cap, tax and cap-and-trade policies. Their findings imply that an investment option may help reduce the retailer's costs under different environmental policies. In [20], authors investigate the technology selection and production decisions problem of a profit-maximizing firm under carbon tax regulation. They found that increasing the tax rate does not necessarily induce the firm to adopt cleaner technology. In [21], an MIP for SC network design problem where the demand for the final product depends on its per unit carbon emission, was proposed. In [24], authors studied the production and emission abatement decisions of a Make-To-Order two-echelon SC under cap-and trade regulation. They investigated the coordination between a manufacturer and a retailer using wholesale price and cost sharing contracts. Authors of [11] studied the environmental investments, capacity expansion, production, and inventory planning options to comply with a cap-and-trade mechanism where the production can be planned using two production technologies: low-carbon and conventional technologies.

It is important to highlight that all these works address the problem of GSCD under carbon emission in a deterministic environment. Considering uncertainty when faced with strategic investment decisions is still a fertile field [23] [14]. Authors of [5], highlighted that the future standing of companies will be seriously affected by the steps they take today in regards to the environment. Authors of [17] were among the first that integrated the demand uncertainty in the problem of technology selection. They studied the problem of investment timing of firms under cap-and-trade and command-and-control regulation based on analytical analysis. They considered investment options in process improvement projects for existing plants. Thus, they made an assumption that the production cost does not change with the investment to reduce emissions. They found that a firm does not always invest earlier as claimed by economists. In [6], authors studied the problem of technologies' selection for new plants. They supposed that the production cost increases with the greenness of the technology. They studied the impact of cap-and-trade and carbon tax regulations on a firm's technology choice and capacity decisions under uncertain demand and carbon price. To solve their model, they used a merit order assumption based on emission price, and derived a closed form solution of the problem. Such assumption may be a strong one for the model's results. Besides, more recently, other authors studied the problem of technology investment under stochastic demand and subject to consumer sensitivity to carbon emission [16]. However, no carbon legislation was taken into account in this model.

From the above analysis, we conclude that the strategic decision of technologies' investment under carbon emission constraints is an important issue within the GSCD problem. Few published works have considered demand uncertainty when studying this problem. We add to the literature by developing and solving a mixed integer stochastic model of technologies selection under a carbon tax and a random demand. Our model allows the company to choose more than one technology, if necessary, in order to balance between profitability and environmental legislation compliance. Particularly, we tend to boost the existing analysis on the trade-off between SC costs' increase and carbon emission reduction, by assessing the regret cost of the strategic decisions made within an environmental and uncertain context. Within these analyses, we contribute in developing the studies of the combined effects of carbon emission legislations and demand uncertainty on the strategic decisions of a SC.

III. Problem statement

We consider the case of a company that faces a stochastic demand and operates under carbon tax pressure. The company has to select production technologies among potential technologies, which differ in terms of carbon emissions as well as costs (fixed and variable) and capacities.

We make the following assumptions to address our problem:

- The company aims to invest in new technologies to produce one type of product;
- The selling price of the product is an exogenous parameter and independent of the deployed technology. It is imposed by the market competition;
- The technologies emit different amount of carbon dioxide and other GHG that may be converted to carbon dioxide emission. The only environmental criterion that we use to distinguish between environmental performances of available technologies is their unit carbon emission (emission amount per one finished product).

- Greener technology is considered more expensive in terms of fixed investment cost.
- The carbon tax amount is assumed to be deterministic and is applied to each ton of carbon emission;

The demand is uncertain, and is modeled as a random variable. This uncertainty affects the company's performances and may induce additional costs. We consider that the company has to support a shortage cost for non-satisfied demand and should sell its surplus quantity at a given salvage price. On the one hand, if the company does not satisfy its clients' demand, then it will pay a shortage cost. In practice, the company may provide the shortage quantity through outsourcing. Besides, we believe that even if the company does not have this option, the shortage affect its brand and induce a market loss that can be quantified through this shortage penalty. On the other hand, if the company produces more than its clients' demand, it will bear a storage cost if the product can be stored without losing its value. Else, if the product will lose its value beyond a certain period, such as the case of fashionable products, foods, drinks, and cosmetics, the company will be forced to sell the product at reduced prices called salvage prices under discount actions. This problem is known as the newsvendor problem as to describe the general approach of a vendor or a manufacturer that has to take the decision on how much to buy or produce before knowing how much exactly to sell.

The problem of the company is to decide which technologies to install and how much to produce on each installed technology in order to maximize its profit under carbon tax legislation and a random demand. To model this problem, we choose to maximize the expected profit of the company instead of its stochastic profit. With this solution, we will obtain a model that can be solved using deterministic modeling techniques. It should be noted, however, that the resulting model will provide different results from a model that considers a deterministic demand that is estimated to its expected value.

We formulate the problem as a two stage stochastic model. In the first stage, the company takes its investment decisions, choosing which technologies to install. Production technology affects the acquisition cost, the per-unit production cost and the emissions rate. Stage 2 decision includes the production quantities on each installed technology. The second stage problem is formulated according to the newsvendor approch.

IV. Mathematical formulation and solution approach

We consider that the company has to choose technologies to install from a set of F available technologies. Each technology is distinguished by its fixed installation cost (K_f); its variable production cost (C_f); its production capacity(Ca_f); and its unit carbon emission (O_f) (emission amount per one unit of finished product). We refer to the unit carbon tax (money unit/ton CO_2) by (Q_O). The product is sold at a fixed exogenous price (S) aligned with the competitor's price. For unsatisfied demand, the company assumes a unit shortage cost (CR). The salvage price of each unsold unit is denoted by (CS). The demand is modeled as a random variable ($\widetilde{D}$) with a mean value $E(\widetilde{D})$ and a cumulative distribution function ($F_{\widetilde{D}}$). Two main decisions are studied, and are represented by the following variables:

- z_f : binary variable equals to one if the technology f is installed and equal to 0 otherwise ;
- x_f : quantity to produce on each installed technology f.

The company aims to maximize its expected profit. We denote the objective function by $\prod$.

$\prod$= Max {Expected value (Total revenue - production costs - emission cost - shortage cost + salvage revenue) - investment cost}.

$$\prod = Max\ \{ E[\min(X, \widetilde{D})].S - [\sum_{f=1}^{F} x_f.C_f - e_O.Q_O - CR.E[max(0, \widetilde{D} - X)] + CS.E[\max(0; X - \widetilde{D})]] - \sum_{f=1}^{F} K_f.z_f\} \quad (1)$$

- ***Problem constraints***

$$X = \sum_{f=1}^{F} x_f \quad (2)$$

$$e_O = \sum_{f=1}^{F} (O_f.x_f) \quad (3)$$

$$x_f \leq Ca_f.z_f \qquad \text{for each } f \text{ in } [1..F] \quad (4)$$

$$x_f\ ; e_m \geq 0 \quad (5)$$

$$z_f \in \{0\ ;\ 1\} \quad (6)$$

Constraints (2) calculates the total quantity to produce through the different installed technologies. Constraint (3) determines the total amount of carbon dioxide emitted through the total production quantity within the different installed technologies. The relationship between the manufactured quantity of the finished product (continuous variable x_f) and the selection of the manufacturing process (binary variable z_f) is given by constraint (4). Indeed, if manufacturing process f is not selected ($z_f = 0$) then $x_f = 0$. Constraint (4) insures that the quantity to be produced on each technology does not exceed its capacity. Finally, constraints (5) and (6) are for variable non-negativity and binary restrictions, respectively.

After presenting the stochastic model, we symplify it through expressing the analytical expressions of the expected values figuring in $\prod$. We deduce the simplification of the model using the following equations as in [1].

- $E[\min(X, \widetilde{D})] = X - E[\max(0, X - \widetilde{D})]$ (7)
- $E[\max(0, X - \widetilde{D})] = \int_0^X F_{\widetilde{D}}(t)dt$ (8)
  where $F_{\widetilde{D}}$ is the cumulative distribution function of the demand variable $\widetilde{D}$
- $E[\max(0, X - \widetilde{D})] - E[\max(0, \widetilde{D} - X)] = X - E(\widetilde{D})$ (9)

$$\prod = Max\ \{[(S + CR).X - (CR + S - CS).\int_0^X F_{\widetilde{D}}(t)dt - CR.E(\widetilde{D})] - \sum_{f=1}^{F} x_f.C_f - e_O.Q_O - \sum_{f=1}^{F} K_f.z_f\} \quad (10)$$

The proposed model is a mixed integer nonlinear (MINL) model with a nonlinear objective function and a set of linear constraints. This model is nonlinear due to the integral of the cumulative distribution function in the objective function. This integral can be further detailed depending on the distribution of the demand [1]. According to [1], the two most suitable

distributions used to model market demands are the uniform and the normal distributions. We choose to model our demand using a uniform random variable defined on the interval [*a, b*]. Using such a stochastic variable and replacing its cumulative distribution function in $\prod$, we obtain a mixed integer quadratic model.

$$\prod = Max\{[(S+CR).X-(CR+S-CS).(\frac{1}{b-a}.(\frac{X^2}{2}-a.X+\frac{a^2}{2}))-CR.(\frac{b-a}{2})]-\sum_{f=1}^{F}x_f.C_f-e_O.Q_O-\sum_{f=1}^{F}K_f.z_f\} \quad (11)$$

Then, we use the common piecewise linearization technique to linearize this model through linearizing the function $Y(X)=\frac{1}{b-a}.(\frac{X^2}{2}-a.X+\frac{a^2}{2})$; we subdivide the interval *[a, b]* into *I* intervals; and we introduce new decision variables x_{fi} instead of x_f, that is the quantity to produce on the technology *f* when the interval *i* is selected. Then we can write on each interval *i* of those *I* intervals, $Y(X)=coeff(i)*x_{fi}+const(i)$ *[1].*

To calculate $coeff(i)$, and $const(i)$, we take two values x_{i1} and x_{i2} on the interval i, then we obtain :

$$coeff(i)=\frac{Y_{1i}-Y_{2i}}{x_{1i}-x_{2i}}; \quad (12)$$

$$const(i)=Y_{1i}-coeff(i)*x_{1i} \quad (13)$$

The new linearized objective function is:

$$\prod = Max\{[(S+CR).\sum_{i=1}^{I}\sum_{f=1}^{F}x_{fi}-(CR+S-CS).\sum_{i=1}^{I}(coeff(i)*\sum_{f=1}^{F}x_{fi}+const(i)*y_i)-CR.E(\widetilde{D})]-[\sum_{i=1}^{I}(\sum_{f=1}^{F}x_{fi}.C_f)-e_O.Q_O]-\sum_{f=1}^{F}K_f.z_f\} \quad (14)$$

We introduce, a binary varibale y_i which is equal to 1 if the interval *i* is selected and 0 otherwise. This binary variable guaranties to select the right interval for each value of the production quantity, and, as a result, the appropriate linear approximation of the expression $E[\max(0, X-\widetilde{D})]$.

The changes to the different existing constraints and additional constraints to the new MILP are summarized in the following equations.

MILP model's constraints

$$X=\sum_{i=1}^{I}\sum_{f=1}^{F}x_{fi} \quad (15)$$

$$e_O=\sum_{i=1}^{I}(\sum_{f=1}^{F}O_f.x_{fi}) \quad (16)$$

$$x_{fi}\leq Ca_f.z_f \quad \text{for each } i \text{ in } [1..I], \text{ for each } f \text{ in } [1..F] \quad (17)$$

Constraint (15) replaces constraint (2) ; it calcultes the total quantity to produce ; constraint (16) replaces constraint (3) and constraints (17) replaces constraints (4).

$$\sum_{i=1}^{I}y_i=1 \quad (18)$$

The constraint (18) guaranties the selection of only one interval *i* based on the value of x_{fi} ;

$$\sum_{f=1}^{F}x_{fi}\leq \text{upper } (i)*y_i \text{ for each } i \text{ dans } [1..I] \quad (19)$$

$$\sum_{f=1}^{F}x_{fi}\geq \text{lower } (i)*y_i \text{ for each } i \text{ dans } [1..I] \quad (20)$$

Constraints (19) and (20) determine boundaries for every interval.

$$x_{fi}, e_m \geq 0 \quad (21)$$

$$z_f ; y_i \in \{0 ; 1\} \quad (22)$$

Constraints (21) and (22) replace constraints (5) and (6).

V. Numerical example and analysis

In this section, we aim to highlight the impact of demand uncertainty on SC decisions (in terms of technology selection) and to measure the impact of this uncertainty on companies' performances (in terms of cost and emissions).

Our numerical data are derived from a textile industry database which is located in a country where we suppose that a carbon tax legislation is implemented with a rate of Q_O= 20 (€/ ton CO_2). We consider a set of 4 potential technologies (*F)* numbered from 1, the most pollutant, to 4 the greenest technology. The unit carbon emission O_f (ton/ one unit of final product) are 0.04, 0.03, 0.02, and 0.005 respectively for each tehnology. The technology investment costs K_f are 22000, 27000, 30000, 33000 respectively for each tehnology. Technology's variable production cost (per unit) $C_f$are 0.2, 0.5, 0.8, 1 respectively. Technology's production capcity Ca_f are 110000, 130000, 150000, 200000 respectively. Selling price *S*= 3€. Salvage price *CS*= 0.2€. Shortage cost *CR= 2.5*€. The Demand $\widetilde{D}$ is supposed to be uniformly distributed on [100 000; 150 000]. Using this numerical example we solve our stochatic model as well as a deterministic model of the problem using the CPLEX 12.6 solver. By denoting *D* as the demand variable, taken to be $D=E(\widehat{D})$, the deterministic model is presented as follows:

$$\prod = Max\{D.S-\sum_{f=1}^{F}x_f.C_f-e_O.Q_O+-\sum_{f=1}^{F}K_f.z_f\} \quad (21)$$

Problem constraints

$$\sum_{f=1}^{F}x_f \leq D \quad (23)$$

$$e_O=\sum_{f=1}^{F}(O_f.x_f) \quad (24)$$

$$x_f \leq Ca_f.z_f \quad \text{for each } f \text{ in } [1..F] \quad (25)$$

$$x_f ; e_m \geq 0 \quad (26)$$

$$z_f \in \{0 ; 1\} \quad (27)$$

As mentioned before, we aim to compare between deterministic and stochastic models performances using some metrics that will be further explained in the following paragraphs. For each metric, we present the relative diffrence *(RD)* between deterministic and stochastic results using the formula (28).

$$RD=\frac{(deterministic\ resulat-stochastic\ resulat)*100}{deterministic\ resulat} \quad (28)$$

We assume that the company may take its strategic decisions once the demand becomes known. This is a non-realistic and idealistic situation. Then, we solve 6 deterministic models that correspond to 6 demand scenarios (*Sc*). From scenario 1 to

scenario 6, the demand is respectively equal to 100000, 110000, 120000, 130000, 140000, and 150000. We refer to those models as "Wait and See models (*W&S*)". The objective is to compare the profit reached through the decisions of those *W&S* models comparing to the profit obtained through holding the decisions of the deterministic (*DM*) and stochastic (*SM*) models respectively under each of those scenarios. Then, to calculate the deterministic regret (*DR)*, we evaluate the expected profit if the company holds the decisions obtained when solving the deterministic model, which corresponds to producing 125000 units using technology 2, under each scenario of the demand realization. We compare it to the profit that the company would reach if it has the opportunity to hold decisions of *W&S* model for every demand realization. The same approach is applied to determine the stochastic regret (*SR*). The obtained stochastic model decisions are to produce 140000 units using technology 4. Then, we evaluate the eventual profit if the company holds those stochastic decisions under each demand scenario. The following tables summarize the above computations and the obtained results.

TABLE 1: Uncertainty cost evaluation

Demand sceanrio	*W & S*	*W&S* Profit (M €)	Profit of *DM*(M €)	*DR* (M €)	Profit of *SM* (M €)	*SR* (M €)
10.10^4	1	168	165.5	2.5	125	43
11.10^4	1	187	185.5	1.5	152	35
12.10^4	1	206	205.5	0.5	179	27
13.10^4	2	220	197.5	22.5	206	14
14.10^4	4	233	161.5	71.5	233	
15.10^4	4	252	125.5	126.5	204	48
Average Profit / Regret		211	173.5	37. 5	183	27.8

TABLE 2: Performance analysis framework

Metrics	Stochastic	Deterministic	Relative difference
Expected Profit (M €)	185.4	210.5	-12,4%
Expected Revenue (M €)	372	375	-0,8%
Pollution cost (M €)	14	100	-86,0%
Average Regret (M €)	27.83	37.5	-25,8%
Maximum regret (M €)	48	126.5	-62,1%
Unit Carbon emission (Kg/unit)	5	40	-87,5%
Produced quantity	140000	125000	12,0%
Emissions quantity (ton)	700	5000	-86,0%

We deduce from the above results that the demand uncertainty has a great impact on strategic decisions of the GSCD problems under environmental considerations. The profit of the stochastic model is 12.4% less than the profit estimated by the more optimistic deterministic model. Actually, under the same carbon tax and problem's parameters, the deterministic model chooses to invest in technology 2, which has moderate environmental performances, but the stochastic model chooses the installation of the greenest technology (technology 4). We note that the investment cost of technology 2 is 22% less than the cost of technology 4. When the demand is uniformly distributed on [100000, 150000], the company finds it more profitable to produce more than the mean value to avoid the shortage costs. However, when the produced quantity increases, the total carbon emission increases too. Therefore, to optimize the carbon tax costs, the choice of technology 4, which results in a product's unit carbon emission that is 87.5% less than its carbon footprint when using technology 2, is justified despite its high investment cost. Using our models and based on these experimentations, we show that, the carbon tax legislation pushes companies to invest in cleaner production technology to develop the required flexibility in response to demand uncertainty under increasingly strict environmental constraints. It is clear that under the stochastic model's decisions, the environmental performances of the company outperform the environmental performances of the deterministic decisions. The stochastic decisions allows for reducing the total carbon emission, the carbon cost on the SC. However, if we only focus on the expected profit, we conclude that the stochastic model has worse economic performances than the deterministic one. This is why we introduced the additional economic metrics to assess the uncertainty costs for both models. We recall that the expected deterministic regret is 37 500€, whereas the expected stochastic regret is only 27 800€. Therefore, even though the expected profit of the deterministic model is 12.4% higher, its expected regret is 25.8% higher. Furthermore, under the worst scenario for the deterministic decisions, the maximum deterministic regret may reach 60% of the expected deterministic profit. However, the maximum stochastic regret is only 26% of the expected stochastic profit. Consequently, the adoption of the deterministic solution may lead to a higher regret under the worst scenario. Within this performance assessment framework, we tried to provide a comprehensive performance analysis tool to prove that stochastic modeling approaches are more suitable for resolving such strategic problems and delivering robust decision support tools to solve GSCD problems under strict environmental legislations.

VI. Conclusion

The GSCD problems have constituted an attractive research field which aim is to provide companies with decision support tools to redesign their supply chains in order to comply with legislations while optimizing their economic performances. However, the majority of the developped models within this field, were deterministe models ignoring uncertain parmeters, such as the demand. To that end, we studied the combined effect of carbon emission tax and stochastic demand on the

technology investment problem. We focused on the comparison between the deterministic and the stochastic models performances to highlight the impact of uncertainty on strategic decisions in such a specific environmental context. We developed a stochastic model to select the technologies to install and determine the production plan based on the newsvendor problem approach. In the solution method, we linearized a simplified version of the two-stage stochastic model using a piecewise linearization technique for the uniform random variable. To analyze the impact of uncertainty on the strategic decisions, we adopted a framework of performance assessment using a set of economic and environmental metrics. We found that the stochastic model is environmentally better and more robust face to uncertainty. Its decisions allowed the company to enhance its environmental performances while being more flexible to comply with environmental legislations. The generic two-stage model that we developed in this paper can be applied to other strategic decisions with in the context of GSCD such as location decisions or supplier selection under carbon tax. Moreover, it can be extended to include other types of carbon dioxide legislations as the Cap or the Cap and Trade. Besides, other random variables, namely the normal demand distribution may be tested to check if the same conclusions are obtained. Furthermore, the only uncertainty source studied in this model was the demand as it influences the production capacities and emissions quantities; however, under the evolving environment of the SC, other sources of uncertainty, such as legislation uncertainty, may have to be taken into consideration when making strategic decisions. Finally, we studied the carbon emission effect on the SC through only legislation; however, the customer awareness and competition are also crucial to this context.

References

[1] Baghaliana, A., Rezapourc, S., Farahani, Z, R. Robust supply chain network design with service level against disruptions and demand uncertainties: A real-life case. *European Journal of Operational Research*, 227(1), 2013, 199–215.

[2] Benjaafar, S., Li, Y., Daskin, M. Carbon footprint and the management of SCs: Insights from simple models, *IEEE Trans. Autom. Sci. Eng*, 10, 2013, 99–116.

[3] Bouchery, Y., Ghaffari, A., Jemai, Z. and Dallery, Y. Adjust or invest: what is the best option to green a SC? *Laboratoire Genie Industriel, Ecole Centrale Paris*, 2011.

[4] Chaabane, A., Ramudhin, A., Paquet, M. Design of sustainable scs under the emission-trading scheme, *International Journal of Production Economics*, 135, 2012, 37-49.

[5] Diabat, A., Simchi-Levi, D. A carbon-capped SC network problem, *IEEE International Conference on IEEM*, Hong Kong, 2009, 523-527.

[6] Drake, D., Kleindorfer, P.R. and Van Wassenhove, L.N. Technology choice and capacity investment under emissions regulation, *INSEAD, Fontainebleau*, France, Working Paper.

[7] Du, S., Hu, L., and Wang, L.. Low-carbon supply policies and SC performance with carbon concerned demand, *Annals of Operations Research*, 2015, DOI 10.1007/S10479-015-1988-0.

[8] Duan, H.B., Zhu, L., and Fan, Y. Optimal carbon taxes in carbon-constrained china: a logistic-induced energy economic hybrid model, *Energy*, 69, 2014, 345-356.

[9] EPA. U.S. Environmental Protection Agency, *Greenhouse Gas Equivalencies Calculator*, Accessed April 2013,. Http://Www.Epa. Gov/Cleanenergy/Energy-Resources/Calculator.Html.

[10] Eskandarpour, M., Dejax, P., Miemczyk, J., Péton, O., Sustainable SC network design: an optimization-oriented review, Omega, 54, 2015, 11-32.

[11] Garcia-Alvarado, M., Paquet., M., Chaabane, A. Joint strategic and tactical planning under the dynamics of a cap-and-trade scheme, *IFAC-Papers Online*, 49(12), 2016, 622–627.

[12] Hammami, R., Nouira, I., Frein, Y. Carbon emissions in a multi-echelon production-inventory model with lead time constraints, *International Journal of Production Economics,* 164, 2015, 292–307.

[13] He, P., Zhang, W., Xu, X., Bian, Y. Production lot sizing and carbon emissions under cap-and-trade and carbon tax regulations, *Journal of Cleaner Production*, 103, 2015, 241-248.

[14] Huang, Y., Wang, K., Zhang, T., Pang, C. Green SC coordination with greenhouse gases emissions management: a game-theoretic approach, *Journal of Cleaner Production,* 2015, doi: 10.1016/j.jclepro, .05.137.

[15] IPCC, Report of the Intergovernmental Panel on Climate Change., Climate Change 2014: mitigation of climate change., Contribution of working group III to the fifth assessment, *Cambridge University Press*, Cambridge, United Kingdom and New York, NY, USA.

[16] Jiang, W., Chen, X. Optimal strategies for manufacturer with strategic customer behavior under carbon emissions-sensitive random demand, *Industrial Management & Data Systems,* 116(4), 2016, 759 – 776.

[17] Jiang, Y., and Klabjan, D. Optimal emissions reduction investment under green house gas emissions regulations. Northwestern university Working Paper, *Northwestern University*, 2012.

[18] Kleindorfer, P.R., Singhal, K., Van Wassenhove, L.N, Sustainable operations management. *Production and Operations Management Poms*, 14(4), 2005, 482–492.

[19] Kossoy, A., Peszko, G., Oppermann, K., Prytz, N., Klein, N., Blok, K., Lam, L., Wong, L., Borkent, B. State and trends of carbon pricing, World Bank, Washington, DC, 2015, Doi: 10.1596/ 978-1-4648-0725-12015.

[20] Krass, D., T. Nedorezov, A. Ovchinnikov. Environmental taxes and the choice of green technology, *Prod. Oper. Manag*. 22(5), 2013, 1035–1055.

[21] Nouira, I., Hammami, R., Frein, Y., Temponi, C. Design of forward supply chains: impact of a carbon emissions-sensitive demand, *Int. J. Production Economics*, 173, 2016, 80–98.

[22] Pachauri, K.M., Reisinger, A. Climate Change 2007 Synthesis Report, *Report of the Intergovernmental Panel on Climate Change*, IPCC. Available On : Https://Www.Ipcc.Ch/Pdf/Assessment-Report/Ar4/Syr/Ar4_Syr_Full_Report.Pdf;

[23] Toptal,A., Özlü, H.,Konur, D. Joint decisions on inventory replenishment and emission reduction investment under different emission regulations, *International Journal of Production Research,* 52(1), 2014, 243-269.

[24] Xu, X., He, P., Xu, H., Zhan, O. SC coordination with green technology under cap-and-trade regulation, *Int. J. Production Economics*, 183(B), 2016, 433–442.

[25] Yalabik, B. and Fairchild, R.J. Customer, regulatory, and competitive pressure as drivers of environmental innovation, *International Journal Of Production Economics*, 131(2), 2011, 519–527.

Lagrangian-Relaxation based Bounding Approaches and Fast Deterministic Heuristics for a Class of Multi-Stage Fixed-Charge Transportation Problems

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

S.A. Ezhil
Doctoral student,
Department of Management Studies, IIT Madras,
Chennai, 600036, INDIA
Email: ezhilsa@gmail.com

Chandrasekharan Rajendran
Professor,
Department of Management Studies, IIT Madras,
Chennai, 600036, INDIA
Email: craj@iitm.ac.in

***Abstract*—Fixed-Charge Transportation Problem (FCTP) is an important problem in supply chains. The FCTP is classified as NP-hard, and is typically solved using the heuristic methods for large-sized problem instances. The FCTP can include multiple intermediate stages, and it is called Multi-Stage FCTP (FCTP:MS) in a supply chain. In this work, we propose two new Lagrangian-Relaxation (LR) approaches for obtaining lower bounds for the FCTP:MS. The salient features of the proposed LR approaches are that the initial iteration of the LR approaches is equivalent to the Linear Programming (LP) relaxation problem, and the subsequent lower bounds get always tightened in each iteration, and eventually converge to dual optimum. The two lower bounds obtained by the proposed two LR approaches are found to be superior when compared to the LP-relaxation, the conventional LR approach, and an existing lower bound approach. Computationally fast deterministic heuristics based on the proposed LR approaches are developed in order to get upper bounds, and they are compared with the existing heuristic method in the literature.**

I. Introduction

Distribution costs contribute to a significant portion of the total supply chain costs. It becomes crucial to make distribution more efficient by bringing more value in a supply chain logistics. Logistic automation systems need methodologies to make faster operational decisions and also provide analytic information for the quality of these decisions. The proposed approaches can provide quick and feasible upper bound solutions and also lower bounds for the Fixed-Charge Transportation Problems (FCTP) in a supply chain.

In the real world, many distribution problems are modeled as FCTP because the transportation costs consist of both fixed and variable components. The FCTP is an extension to the transportation problem when a fixed shipment cost exists in a route, in addition to the variable cost. The FCTP was first formulated in [1]. The FCTP is an NP-hard problem, see [2]; thus, large-sized problem instances are computationally difficult to solve. Reference [3] solved the FCTP using an exact algorithm by ranking the extreme points which is suitable for small-sized FCTP problem instances. Reference [4] proposed a heuristic which are able to solve large-sized FCTP problem instances with minimal computational effort. Reference [5] developed a compact representation for the FCTP and used it in evolutionary algorithms. Reference [6] employed a similar FCTP representation used in [5], and proposed a genetic algorithm which is found to be good for large-sized FCTP problem instances. Reference [7] proposed a methodology for obtaining a lower bound for the FCTP by using quadratic approximation of the objective function.

The Multi-Stage FCTP (called FCTP:MS in this paper) is addressed in this work which extends the Single-Stage FCTP (called FCTP:SS in this paper) to a supply chain. In the FCTP:MS, items are shipped from a set of sources to a set of demands through multiple intermediary transportation stages. The FCTP:MS can be solved as mixed integer programming problems; but they are solvable only for small-sized problem instances due to the enormous computational effort needed. Heuristic methods are often employed for solving the FCTP and their variations in practice. While the Single-Stage FCTP is considered widely in the literature, only limited research is available for the FCTP:MS. One such research attempt on the FCTP:MS is now discussed. Reference [8] that proposed two genetic algorithms for the Two-Stage FCTP (called FCTP:TS in this paper), namely, Two-Stage Genetic Algorithm (TSGA) and Single-Stage Genetic Algorithm (SSGA). The SSGA was found to be better performing, in which the Two-Stage FCTP problem is converted to an equivalent Single-Stage FCTP problem and then solved.

The main motivation for this work is to present new Lagrangian Relaxation (LR) approaches for the FCTP:MS which can provide lower bounds on the optimal solution for evaluating the performance of the heuristic solutions. Feasible solutions can be derived from the solutions obtained in each iteration of the proposed LR approach because the binding constraint, when relaxed, does not affect the feasibility of the FCTP. The corresponding feasible and quick (i.e., computationally fast) upper-bound solutions can be derived from each LR iteration which can further be improved by using a heuristic. The contribution of the paper are two fold: (1) we propose two new LR approaches that ensure that their corresponding lower bounds keep increasing and eventually reach the dual optimum; (2) we are in a position to derive computationally fast upper bounds on the actual total transportation cost from the feasible shipment plans obtained for each of the proposed

LR approaches.

II. Mathematical formulation

See the supplemental material for the framework of Multi-Stage Transportation problem in a supply chain setting considered in this paper.

A. Parameters/Indices

N	number of transportation stages.
i	index for stages in the Multi-Stage transportation problem in a supply chain setting. /* Note: $i = 1, .., N$; when $i = 1$, the stage refers to the supply at the first stage, and when $i = N$, the stage refers to the demand at the last stage.*/
k', k'', k''', k''''	indices for nodes. /* Note: nodes in general refer to all those involved in the transportation */
K_i	number of nodes in stage i.
$C_{(i,k'),(i+1,k'')}$	variable cost per unit for transporting from node k' in stage i to node k'' in stage $i+1$.
$F_{(i,k'),(i+1,k'')}$	fixed cost for transporting from node k' in stage i to node k'' in stage $i+1$.
$S_{k'}$	supply available in node k' in stage 1.
$D_{k''}$	demand in node k'' in stage N.

B. Decision variables

$X_{(i,k'),(i+1,k'')}$	quantity shipped from node k' in stage i to node k'' in stage $i+1$.
$\delta_{(i,k'),(i+1,k'')}$	an indicator (binary) variable that takes value 1 if there is a shipment from node k' in stage i to node k'' in stage $i+1$; 0 otherwise.

C. Problem definition and mathematical model for the FCTP with the consideration of multiple stages in a supply chain

The FCTP:MS can be stated as a distribution problem with N number of stages in which available supplies at the first stage are shipped through the nodes present in the intermediate stages to fulfil the demands at the last stage. Variable cost $C_{(i,k'),(i+1,k'')}$ is incurred per item for the shipment, and fixed cost $F_{(i,k'),(i+1,k'')}$ is incurred if a route through node k' in stage i to node k'' in stage $i+1$ (denoted in general) is used for the shipment. The capacity of nodes in the intermediate stages ($2 \leq i \leq N-1$) is assumed to be unlimited. The Single-Stage FCTP is denoted as FCTP:SS, and the Two-Stage FCTP is denoted as FCTP:TS.

The objective of the FCTP:MS is to find the shipment quantities from the supplies to the demands through intermediary stages of nodes, so that the overall transportation costs (both fixed and variable costs) are minimized:

$$Min\ Z = \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} + \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')}. \quad (1)$$

The constraints are now presented. $M_{(i,k'),(i+1,k'')}$ is a large value. It is to be noted that the absence of binding constraint (5) (linking shipment plan and corresponding binary variables) does not affect the feasibility of a solution as it only enforces that the corresponding fixed cost of a particular route is added to the total cost when there is a shipment.

$$\sum_{k''=1}^{K_2} X_{(1,k'),(2,k'')} \leq S_{k'}, \qquad k' = 1, 2, .., K_1. \quad (2)$$

$$\sum_{k'=1}^{K_{N-1}} X_{(N-1,k'),(N,k'')} = D_{k''}, \qquad k'' = 1, 2, .., K_N. \quad (3)$$

$$\sum_{k'=1}^{K_{i-1}} X_{(i-1,k'),(i,k'')} = \sum_{k'''=1}^{K_{i+1}} X_{(i,k''),(i+1,k''')}, \quad i = 2, .., N-1 \quad \text{and} \quad k'' = 1, 2, .., K_i. \quad (4)$$

$$X_{(i,k'),(i+1,k'')} \leq M_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')}, \quad k' = 1, 2, .., K_i, \quad k'' = 1, 2, .., K_{i+1} \quad \text{and} \quad i = 1, .., N-1. \quad (5)$$

$$X_{(i,k'),(i+1,k'')} \geq 0, \qquad i = 1, .., N-1. \quad (6)$$

The total demand in stage N is assumed not to exceed the total supply available in stage 1 as follows (a constraint on parameters).

$$\sum_{k'=1}^{K_1} S_{k'} \geq \sum_{k''=1}^{K_N} D_{k''}. \quad (7)$$

For the FCTP:SS with $N = 2$, we set the values of $M_{(1,k'),(2,k'')}$ in this paper (in order to get tighter bounds) as follows:

$$M_{(1,k'),(2,k'')} = min\{S_{k'}; D_{k''}\}, \quad k' = 1, 2, .., K_1 \quad \text{and} \quad k'' = 1, 2, .., K_2. \quad (8)$$

For the FCTP:MS, we set the values of $M_{(i,k'),(i+1,k'')}$ in this paper (in order to get tighter bounds) using (9a), (9b) and (9c) as follows:

$$M_{(1,k'),(2,k'')} = S_{k'}, \quad k' = 1, 2, .., K_1 \quad \text{and} \quad k'' = 1, 2, .., K_2; \quad (9a)$$

$$M_{(i,k'),(i+1,k'')} = \sum_{k'''=1}^{K_N} D_{k'''}, \quad k' = 1, 2, .., K_i, \quad k'' = 1, 2, .., K_{i+1} \quad \text{and} \quad i = 2, .., N-2; \quad \text{and} \quad (9b)$$

$$M_{(N-1,k'),(N,k'')} = D_{k''}, \quad k' = 1, 2, .., K_{N-1} \quad \text{and} \quad k'' = 1, 2, .., K_N. \quad (9c)$$

Let us define FCTP-LP:MS as FCTP:MS problem with integrality property of $\delta_{(i,k'),(i+1,k'')}$ being relaxed in the interval [0,1]. Similarly, we define FCTP-LP:SS and FCTP-LP:TS as LP-relaxed problems of FCTP:SS and FCTP:TS respectively.

III. LAGRANGIAN-RELAXATION APPROACHES

A. Conventional Lagrangian-Relaxation approach

We formulate the Lagrangian Dual Problem (LDP) by relaxing the binding constraint (5) and bringing it into the objective function, as given below:

$$\begin{aligned}\text{Min } Z_{CLR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} \lambda_{(i,k'),(i+1,k'')} \times \big(X_{(i,k'),(i+1,k'')} \\ &- M_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')}\big), \\ &\text{with} \quad \lambda_{(i,k'),(i+1,k'')} > 0. \end{aligned} \tag{10}$$

In the conventional LR approach, the LDP is solved using subgradiant procedure (see [9], [10]). We have set $\lambda_{(i,k'),(i+1,k'')}$ in the range (0, 0.5). See the supplemental material in the link http://bit.ly/dropbox-MSFCTP for the details of implementation of the conventional LR approach in our work.

B. Proposed two LR approaches

In order to avoid the term $(X_{(i,k'),(i+1,k'')} - M_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')})$ in the objective function (10) becoming negative, we introduce a variable $\Theta_{(i,k'),(i+1,k'')}$ (that is ≥ 0) using (11), which is active only when the respective constraint (see (5)) is not satisfied :

$$\begin{aligned}\Theta_{(i,k'),(i+1,k'')} \geq (&X_{(i,k'),(i+1,k'')} - M_{(i,k'),(i+1,k'')} \\ &\times \delta_{(i,k'),(i+1,k'')}).\end{aligned} \tag{11}$$

We formulate the Proposed Lagrangian Dual Problem (PLDP) as follows:

$$\begin{aligned}\text{Min } Z_{LR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} \lambda_{(i,k'),(i+1,k'')} \times \Theta_{(i,k'),(i+1,k'')}.\end{aligned} \tag{12}$$

As the binding constraint (5) is relaxed in the PLDP, there is no guarantee that it is satisfied. The following cases can be considered based on this observation.

Case 1: When $X_{(i,k'),(i+1,k'')} > 0$ and the binding constraint (5) is not satisfied (i.e., when $\delta_{(i,k'),(i+1,k'')}$ is set to 0), we have

$\Theta_{(i,k'),(i+1,k'')} = X_{(i,k'),(i+1,k'')}$ (because of the minimization objective function).

Case 2: When $X_{(i,k'),(i+1,k'')} > 0$ and the binding constraint (5) is satisfied, we have

$\delta_{(i,k'),(i+1,k'')} = 1$ and $\Theta_{(i,k'),(i+1,k'')} = 0$.

Case 3: When $X_{(i,k'),(i+1,k'')} = 0$, we have

$\delta_{(i,k'),(i+1,k'')} = 0$ and $\Theta_{(i,k'),(i+1,k'')} = 0$.

Lemma 1: The Proposed Lagrangian Dual Problem (PLDP) is equivalent to the LP-relaxed problem FCTP-LP:MS when all the Lagrangian multipliers $\lambda_{(i,k'),(i+1,k'')}$ are set to $(F_{(i,k'),(i+1,k'')}/M_{(i,k'),(i+1,k'')})$.

Proof: Substituting $\lambda_{(i,k'),(i+1,k'')} = \dfrac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}}$ in the PLDP, we get the following:

$$\begin{aligned}\text{Min } Z_{LR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}} \times \Theta_{(i,k'),(i+1,k'')}.\end{aligned} \tag{13}$$

Based on Case 1 and Case 3, we get the following:

$$\begin{aligned}\text{Min } Z_{LR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} \big(C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')}\big) \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}} \times X_{(i,k'),(i+1,k'')}.\end{aligned} \tag{14}$$

Equation (14) represents the equivalent of the FCTP:MS (minimization) problem in which the integrality property of $\delta_{(i,k'),(i+1,k'')}$ is relaxed and set in the interval [0,1].

Lemma 2: The proposed LR approaches converge to dual optimum as the values of LR multipliers $\lambda_{(i,k'),(i+1,k'')}$ keep increasing during each iteration and eventually reach the value of the fixed cost $F_{(i,k'),(i+1,k'')}$ of the respective route; at this point the PLDP is equivalent to the FCTP:MS problem.

Proof: Substituting $\lambda_{(i,k'),(i+1,k'')} = F_{(i,k'),(i+1,k'')}$ in the PLDP, we get the following:

$$\begin{aligned}\text{Min } Z_{LR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} (C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')}\big) \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \Theta_{(i,k'),(i+1,k'')}.\end{aligned} \tag{15}$$

Based on Case 2 and Case 3, we get the following:

$$\begin{aligned}\text{Min } Z_{LR} = &\sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} (C_{(i,k'),(i+1,k'')} \times X_{(i,k'),(i+1,k'')} \\ &+ \sum_{i=1}^{N-1}\sum_{k'=1}^{K_i}\sum_{k''=1}^{K_{i+1}} F_{(i,k'),(i+1,k'')} \times \delta_{(i,k'),(i+1,k'')}\big).\end{aligned} \tag{16}$$

Now, we observe that based on Lemma 1 and Lemma 2, the limits for the Lagrangian multipliers range are given by $F_{(i,k'),(i+1,k'')}/M_{(i,k'),(i+1,k'')} \leq \lambda_{(i,k'),(i+1,k'')} \leq F_{(i,k'),(i+1,k'')}$.

Salient features of the proposed LR approaches (called LRLB1 and LRLB2) are as follows. We are able to define the lower and upper limits for the Lagrangian multipliers which make LR iterations computationally effective. The values for the initial set of Lagrangian multipliers are given by their respective lower limits. The initial iteration of the PLDP is equivalent to the LP-relaxed problem, called FCTP-LP:MS. So the LR procedure starts with the solution given by LP-relaxation bound. The PLDP is equivalent to the FCTP:MS when the Lagrangian multipliers are assigned with their respective upper limits. The lower bound on Z gets always tightened in each iteration. The reasons are that $\Theta_{(i,k'),(i+1,k'')}$ is always ≥ 0; and $\lambda_{(i,k'),(i+1,k'')}$ is always > 0, and keeps increasing during the iterations.

1) *Proposed LR approach*: *LRLB*1: In each iteration, we increase the value of Lagrangian multipliers from their respective lower limits based on the slope obtained from these defined limits (in the proposed LR approach, called LRLB1). We now present the formal procedure of LR approach, LRLB1.

Step 1: Set the current iteration $iter$ to 0. Set the total number of LR iterations tot_iter equal to $\sum_{i=1}^{N-1}(K_i \times K_{i+1})$.

Step 2: We set the initial set of Lagrangian multipliers with their respective lower limits as follows:

$$\lambda^0_{(i,k'),(i+1,k'')} = \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}}. \quad (17)$$

Step 3: Solve the PLDP with the current set of Lagrangian multipliers $\lambda^{iter}_{(i,k'),(i+1,k'')}$ to get the shipment quantities $X_{(i,k'),(i+1,k'')}$ and the solution value Z_{LR} that is a lower bound on Z. Note that $X_{(i,k'),(i+1,k'')}$ represents a feasible solution to the original problem satisfying all the shipment constraints. Using the shipment quantities $X_{(i,k'),(i+1,k'')}$ in the shipment plan, we can compute the corresponding total transportation cost Z_{UB} by considering the associated fixed costs and variable costs.

Step 4: If $\Theta_{(i,k'),(i+1,k'')} = 0$ for all the respective routes, then the current solution is optimal (with $Z_{LR} = Z_{UB}$), and terminate the procedure; else proceed to Step 5.

Step 5: Update the Lagrangian multipliers for the next iteration as given below:

$$\lambda^{iter+1}_{(i,k'),(i+1,k'')} = \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}} + (F_{(i,k'),(i+1,k'')} - \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}}) \times \frac{iter}{tot_iter}. \quad (18)$$

Step 6: Increment the current iteration $iter$ by one and go to Step 3 to solve the PLDP with the updated set of Lagrangian multipliers. It is to be noted that when $iter$ value becomes equal to tot_iter, the PLDP becomes the FCTP:MS problem.

TABLE I. THE PERFORMANCE OF THE PROPOSED LR APPROACHES COMPARED WITH THE LP RELAXATION AND THE CONVENTIONAL LR APPROACH

Problem instance	LP relaxation	Conventional LR approach	LRLB1	LRLB2
N370 A (50×100) (FCTP:SS)	979,219.53	570,838.74	1,011,879.72	***1,013,540.29***
20×20×20 (FCTP:TS)	386,097.41	79,330.00	***408,919.24***	407,890.98
20×20×20×20 (FCTP:MS)	423,669.84	122,945.00	500,512.88	***505,825.27***

Note:

1. Lower bound for LP-relaxed problem FCTP-LP is obtained by relaxing integrality property of $\delta_{(i,k'),(i+1,k'')}$ in the interval [0,1]. LP-relaxed problem FCTP-LP takes less than 1 second for all problem instances.
2. The conventional LR approach and the proposed LR approaches, LRLB1 and LRLB2, are run up to 3600 seconds for problem instances with respect to FCTP:SS, FCTP:TS and FCTP:MS. A lower bound in bold and italics indicates that it is the best lower bound obtained for the given problem instance.
3. FCTP:SS problem instances are denoted as $(n1 \times n2)$, and taken from [5]. FCTP:TS problem instances are denoted as $n1 \times n2 \times n3$. FCTP:MS problem instances are denoted as $n1 \times n2 \times n3 \times n4$. $n1$, $n2$, $n3$ and $n4$ represent the number of nodes in the first, second, third and fourth stages respectively.

2) *Proposed LR approach*: *LRLB*2: The proposed LR approach, LRLB2, is developed with the premise that once a binding constraint (5) is not satisfied, the corresponding Lagrangian multiplier would be assigned with its respective upper limit, thus preventing the same constraint getting violated in the subsequent iterations of LR. The Lagrangian multipliers are updated for the next iteration when $\Theta_{(i,k'),(i+1,k'')} > 0$ and the corresponding $\lambda_{(i,k'),(i+1,k'')}$ is set as $F_{(i,k'),(i+1,k'')}$. This forms Step 5 in LRLB2. Other than this, the steps of the proposed LR approach, LRLB1, hold for LRLB2 as well.

The numerical examples for FCTP:SS, FCTP:TS and FCTP:MS for the proposed LR approaches are given in link: http://bit.ly/dropbox-MSFCTP (in order to save space in this paper).

IV. COMPUTATIONAL STUDY OF THE PROPOSED LR APPROACHES FOR FCTP:SS, FCTP:TS AND FCTP:MS

In this section, we detail the experiments that we have carried out, and analyze the lower bounds obtained from the proposed LR approaches, LRLB1 and LRLB2. Problem instances for the FCTP:SS, the FCTP:TS and the FCTP:MS are used for evaluation (note: we use the standard problem instances given in [5] in the case of FCTP:SS). The experiments have been performed on a PC with Intel CORE i7 2.9 GHz processor and 8 GB of RAM using ILOG CPLEX solver.

We have performed a pilot study for comparing the conventional LR approach and the proposed LR approaches, LRLB1 and LRLB2. The lower bounds obtained by running these approaches up to 3600 seconds are given in Table I, thus showing that our proposed LR approaches, LRLB1 and LRLB2, are better than the conventional LR approach in terms of obtaining tighter lower bounds on Z. Therefore, the conventional LR approach is not included for further experimentation. Reference [7] used a problem instance (4×3) for FCTP:SS and reported the lower bound value as 334.24. For the same problem instance, we obtain 350 as a lower bound which is optimal solution value as reported in Table II, thereby showing the superiority of our LR approaches. We have subsequently used larger benchmark problem instances

TABLE II. THE PERFORMANCE OF THE PROPOSED LR APPROACHES COMPARED WITH THE QUADRATIC FORMULATION (SEE [7])

Problem instance	FCTP-LP:SS	Quadratic formulation	LRLB1:SS	LRLB2:SS	Optimal
4 × 3	321.67	334.24	***350***	***350***	350

Note:

1. Lower bound for LP-relaxed problem FCTP-LP:SS is obtained by relaxing integrality property of $\delta_{(i,k'),(i+1,k'')}$ in the interval [0,1]. LP-relaxed problem FCTP-LP:SS takes less than 1 second for all problem instances.
2. Proposed LR approaches, LRLB1:SS and LRLB2:SS, are executed for less than 10 seconds for 4 × 3 problem instance.
3. The lower bound for the quadratic formulation is reported as in [7].

which involve 50 nodes in the first stage and 100 nodes in the second stage (taken from [5]). For FCTP:SS, we denote the LR approach LRLB1 as LRLB1:SS, and LR approach LRLB2 as LRLB2:SS. For the FCTP:TS, we find that many large-sized problem instances are not available in the literature. So, for performance evaluation, we have randomly generated large-sized problem instances. For FCTP:TS, we denote the LR approach LRLB1 as LRLB1:TS and LR approach LRLB2 as LRLB2:TS. We have used randomly generated large-sized problem instances for the FCTP:MS as well. For FCTP:MS, we denote the LR approach LRLB1 as LRLB1:MS and LR approach LRLB2 as LRLB2:MS. The results for a sample set of problem instances are presented in Table III. We find from the computational study that the proposed LR approaches, LRLB1 and LRLB2, yield lower bounds on Z that are fairly closer to the optimal solution and tighter than the conventional LR approach and the LP-relaxation approach. Solving the FCTP-LP:SS, FCTP-LP:TS and FCTP-LP:MS take less than a second for each considered problem instances. All the considered problem instances are provided in the link for future researches: http://bit.ly/dropbox-MSFCTP.

V. PROPOSED TWO LR-BASED HEURISTICS

The FCTP:MS becomes equivalent to the Multi-Stage Transportation Problem when the binding constraint (5) is excluded. As this constraint only determines whether a fixed cost is accounted based on the shipment in a route, any feasible solution to a Multi-Stage Transportation Problem is feasible to the FCTP:MS. Instead of excluding the binding constraint (5), the integrality property of the $\delta_{(i,k'),(i+1,k'')}$ variable is relaxed in the interval [0,1] in FCTP-LP:MS. It can be seen that the Multi-Stage Transportation Problem and the FCTP-LP:MS are easier to solve than the FCTP:MS. As evident from Lemma 1 and Lemma 2, initial iteration of the proposed LR approaches is equivalent to the FCTP-LP:MS and gradually transforms to the FCTP:MS during the later iterations. The feasible (upper bound) solution to the FCTP:MS can be derived from the solution that we obtain from the PLDP by reckoning with the fixed cost with respect to each shipment (satisfying the binding constraint) in the objective function. Feasible (upper bound) solutions are obtained with respect to the shipments $X_{(i,k'),(i+1,k'')}$ from each iteration of the proposed LR approaches. Every such solution is given as the input to the improvement heuristic proposed in [4] in the case of FCTP:SS and FCTP:TS. However, we do not implement any improvement heuristic for the FCTP:MS, when number of intermediate stages ≥ 2. When we have $\Theta_{(i,k'),(i+1,k'')} = 0$ for all the routes in an LR iteration, then the current solution of the PLDP is optimal to the FCTP:MS. The heuristics are terminated at this point.

TABLE III. THE PERFORMANCE OF THE PROPOSED LR APPROACHES

Problem instance	FCTP-LP:SS	LRLB1:SS	LRLB2:SS	Optimal
N370 A (50 × 100)	979,219.53	1,011,879.72	***1,013,540.29***	*
N370 B (50 × 100)	988,308.12	1,019,214.95	***1,021,745.13***	*
N370 C (50 × 100)	965,430.94	***1,006,172.67***	1,002,428.89	*
N370 E (50 × 100)	961,651.31	994,945.35	***1,000,269.71***	*
N370 0 (50 × 100)	972,305.74	1,003,513.75	***1,009,225.13***	*
N370 1 (50 × 100)	961,764.01	992,843.75	***993,652.80***	*
N370 2 (50 × 100)	961,184.20	***1,000,531.10***	997,092.30	*
Problem instance	**FCTP-LP:TS**	**LRLB1:TS**	**LRLB2:TS**	**Optimal**
8×10×10	244,139.86	272,447.06	***272,541.75***	272,617
12×15×20	327,252.17	***361,010.69***	359,963.76	*
15×10×20	382,592.42	404,586.04	***404,886.55***	408,617
16×17×18	363,187.33	***383,964.65***	383,395.80	388,342
20×15×15	365,143.01	394,023.79	***394,719.51***	*
Problem instance	**FCTP-LP:MS**	**LRLB1:MS**	**LRLB2:MS**	**Optimal**
12×13×14×15	354,813.25	414,708.77	***426,900.04***	427,989
15×15×5×10	358,950.95	417,701.07	***427,911.29***	427,913
16×20×15×18	380,324.12	454,816.66	***460,582.91***	*
18×17×16×20	399,854.33	473,489.59	***479,205.17***	*
20×10×15×25	471,366.40	533,272.57	***538,192.19***	538,246

Note:

1. Lower bound for LP-relaxed problems are obtained by relaxing integrality property of $\delta_{(i,k'),(i+1,k'')}$ in the interval [0,1] and takes less than 1 second for all problem instances.
2. Proposed LR approaches, LRLB1 and LRLB2, are run upto 3600 seconds for all problem instances. A lower bound in bold and italics indicates that it is the best lower bound obtained for the given problem instance.
3. For FCTP:SS, we denote the LR approach LRLB1 as LRLB1:SS and LR approach LRLB2 as LRLB2:SS. For FCTP:TS, we denote the LR approach LRLB1 as LRLB1:TS and LR approach LRLB2 as LRLB2:TS. For FCTP:MS, we denote the LR approach LRLB1 as LRLB1:MS and LR approach LRLB2 as LRLB2:MS.
4. We attempt to obtain an optimal solution by solving the problem instance upto 3600 seconds, and report as *, if not obtained within this timeframe.

1) *Proposed LR-based generic heuristic*: *LRBH*1: LR-based heuristic, LRBH1, is based on the proposed LR approach, LRLB1. In order to reduce the computational effort required for each iteration, we introduce $scale_factor$ which is multiplied, while getting the new Lagrangian multipliers, where $0 < scale_factor \leq 1$. We redefine the use of $scale_factor$ as follows:

$$\lambda^{iter+1}_{(i,k'),(i+1,k'')} = \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}} + ((F_{(i,k'),(i+1,k'')} - \frac{F_{(i,k'),(i+1,k'')}}{M_{(i,k'),(i+1,k'')}}) \times \frac{iter}{tot_iter} \times scale_factor). \quad (19)$$

In LR-based heuristic, LRBH1, upper-bound solutions are derived in each iteration of the proposed LR approach, LRLB1 subsequently (as explained earlier).

2) *Proposed LR-based generic heuristic*: *LRBH*2: In a similar way, LR-based heuristic, LRBH2, derives the upper-bound solution from each iteration of the proposed LR approach, LRLB2, and improves the solution by using a heuristic (as explained earlier). We denote the proposed heuristic LRBH1 as LRBH1:SS, LRBH1:TS and LRBH1:MS for FCTP:SS, FCTP:TS and FCTP:MS respectively. In the same way, we denote the proposed heuristic LRBH2 as LRBH2:SS, LRBH2:TS and LRBH2:MS for FCTP:SS, FCTP:TS and FCTP:MS respectively.

TABLE IV. THE PERFORMANCE OF THE PROPOSED LR-BASED HEURISTICS FOR THE FCTP:SS

Problem instance (50 × 100)	HGA	LRBH1:SS	LRBH2:SS
N370 A	1,293,618	***1,290,458***	1,299,399
N370 B	1,279,256	***1,266,999***	1,289,289
N370 C	1,269,708	***1,263,710***	1,268,533
N370 E	***1,250,614***	1,251,927	1,263,818
N370 0	1,280,014	***1,270,903***	***1,270,903***
N370 1	1,268,345	***1,265,484***	1,269,239
N370 2	1,261,008	***1,247,617***	1,275,807

Note:

1. Proposed LR-based heuristics, LRBH1:SS and LRBH2:SS, and the genetic algorithm named HGA, proposed in [6], are run up to 100 seconds. An upper bound (in bold and italics) indicates that it is the best upper bound.

TABLE V. THE PERFORMANCE OF THE PROPOSED LR-BASED HEURISTICS FOR THE FCTP:TS

Problem instance	SSGA	LRBH1:TS	LRBH2:TS
8×10×10	272,965	***272,617***	***272,617***
12×15×20	383,926	382,087	***376,828***
15×10×20	421,858	417,376	***416,945***
16×17×18	597,716	406,014	***402,849***
20×15×15	664,816	419,930	***414,914***

Note:

1. Proposed LR-based heuristics, LRBH1:TS and LRBH2:TS, and the genetic algorithm named SSGA, proposed in [8], are run up to 100 seconds. An upper bound (in bold and italics) indicates that it is the best upper bound obtained.

VI. COMPUTATIONAL STUDY OF THE PROPOSED LR-BASED HEURISTICS FOR FCTP:SS, FCTP:TS AND FCTP:MS

In this section, we present the details of the performance of the proposed LR-based heuristics, LRBH1 and LRBH2. Problem instances for the FCTP:SS (taken from [5]), FCTP:TS and FCTP:MS are used for evaluation. As for the heuristic approach LRBH1:SS, the value of the constant $scale_factor$ is set to 0.001 (a small value is chosen with respect to problem instances in [5]) in our study. The results for a sample set of problem instances are presented in Table IV. We compare the upper bounds obtained by the proposed LR-based heuristics, LRBH1:SS and LRBH2:SS, with genetic algorithm (named HGA) proposed in [6]. We find that the proposed LR-based heuristics are superior compared to HGA, when executed up to 100 seconds. So the proposed LR-based heuristics are well suited, while running with minimal CPU time. As for the heuristic approach LRBH1:TS, the value of the constant $scale_factor$ is set to 0.1 in our study. The results for the FCTP:TS for a sample set of problem instances are presented in Table V. We compare the upper bounds obtained by the proposed LR-based heuristics, LRBH1:TS and LRBH2:TS, with the genetic algorithm (named SSGA) proposed in [8]. We find that the proposed LR-based heuristics are better, when compared to SSGA, when run up to 100 seconds. As for the heuristic approach LRBH1:MS, the value of the constant $scale_factor$ is set to 0.1 in our study. The results for the proposed LR-based heuristics, LRBH1:MS and LRBH2:MS, for a sample set of problem instances are presented in Table VI. We find that the proposed LR-based heuristics, LRBH1:MS and LRBH2:MS, are competitive to each other for the FCTP:MS.

TABLE VI. THE PERFORMANCE OF THE PROPOSED LR-BASED HEURISTICS FOR THE FCTP:MS

Problem instance	LRBH1:MS	LRBH2:MS
12×13×14×15	464,306	***438,570***
15×15×5×10	435,171	***432,258***
16×20×15×18	511,601	***501,657***
18×17×16×20	***527,058***	545,985
20×10×15×25	612,179	***568,700***

Note:

1. Proposed LR-based heuristics, LRBH1:MS and LRBH2:MS, are run up to 100 seconds. An upper bound (in bold and italics) indicates that it is the best upper bound.

VII. CONCLUSION

Many transportation companies have started employing logistic automation systems which need techniques for making faster operational decisions and analyse the quality of these decisions. We have considered the Multi-Stage FCTP problem in a supply chain keeping these implications in mind. We have proposed two LR approaches, LRLB1 and LRLB2, for the Multi-Stage FCTP which can serve in getting good lower bounds. The lower bounds obtained by the proposed LR approaches, LRLB1 and LRLB2, are found to be superior when compared to the LP-relaxation, the conventional LR approach, and an existing lower bound approach for the FCTP:SS problem (see [7]). We have also proposed two heuristics, LRBH1 and LRBH2, based on these LR approaches which are better suited and effective, when run for shorter durations for the large-sized problem instances. The proposed LR approaches and heuristics can be used to get solutions that can be given as input to meta-heuristics (that require larger CPU times) for solving the Multi-Stage FCTP. The proposed approaches will find application in solving quickly the real-life large-scale transportation problems encountered in supply chains.

REFERENCES

[1] Hirsch WM and Dantzig GB (1968). The fixed charge problem. *Naval Research Logistics Quarterly*, **15**: 413-24.

[2] Schrenk S, Finke G and Cung VD (2011). Two classical transportation problems revisited: Pure constant fixed charges and the paradox. *Mathematical and Computer Modelling*, **54**: 2306-2315.

[3] Murty KG (1968). Solving the fixed charge problem by ranking the extreme points. *Operations Research*, **16**: 268-279.

[4] Raj KAAD and Rajendran C (2009). Fast heuristic algorithms to solve a single-stage Fixed-Charge Transportation Problem. *International Journal of Operational Research*, **6**: 304-329.

[5] Eckert C and Gottlieb J (2002). Direct representation and variation operators for the fixed charge transportation problem. *International Conference on Parallel Problem Solving from Nature, Springer Berlin Heidelberg*, **2439**: 77-87.

[6] Raj KAAD and Rajendran C (2012a). A hybrid genetic algorithm for solving single-stage fixed-charge transportation problems. *Technology Operation Management*, **2**: 1-15.

[7] Adlakha V and Kowalski K (2014). On lower bounds for the fixed charge problem. *Computers & Operations Research*, **52**:105-112.

[8] Raj KAAD and Rajendran C (2012b). A genetic algorithm for solving the fixed-charge transportation problem: Two-stage problem. *Computers & Operational Research*, **39**: 2016-2032.

[9] Fisher ML (1981). The Lagrangian relaxation method for solving integer programming problems. *Management Science*, **27**: 1-18.

[10] Reeves, C.R, 1993. Modern Heuristic Techniques for Combinatorial Problems. *John Wiley and Sons, New York*.

Modeling and optimization of biomass supply chain with two types of feedstock suppliers

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Duc Huy NGUYEN*, Haoxun CHEN†, Nengmin WANG‡
*Laboratory of Industrial Systems Optimization, Charles Delaunay Institute, University of technology of Troyes, France
Email: duc_huy.nguyen@utt.fr
†Laboratory of Industrial Systems Optimization, Charles Delaunay Institute, University of technology of Troyes, France
Email: haoxun.chen@utt.fr
‡ School of management, Xi'an Jiaotong University, China
Email: wangnm@xjtu.edu.cn

***Abstract*—This paper deals with the stabilization of feedstock supply in a lignocellulosic biomass supply chain. We propose a flexible model which allows for optimizing supplier selection and transportation plan. This is a mixed integer linear programming (MILP) model for simultaneously minimizing the total costs comprising purchase costs, pre-traitement costs, conversion costs, transportation costs and holding costs. This model could handle complex logistics operations in the supply chain to provide optimal decisions that include supplier selection and transportation plan. A numerical example is presented at the end to illustrate the features implemented in the model.**

I. Introduction

The European Union's energy strategy 2020 aims to reduce its greenhouse gas emissions and increase the share of renewable energy. Biomass is considered one of the best alternatives to fossil fuels. As a renewable resource, it will play an important role in the energy transition and appears to be a very promising solution to sustainable production of clean energy in the future.

An effective biomass supply chain management could affect significantly the biofuel price and is one of the major drivers for sustainability of renewable energy. In others words, biomass feedstock availability and transportation costs (due to the low energy density of biomass) are critical to the efficiency of renewable energy production, according to [1]. One possible explanation is that the scale expansion leads to a significant increase of transportation costs of biomass feedstock and could cause economically inefficiency of a renewable energy system.

An effective supply chain management could contribute to the reduction of production costs of biofuels and increased competitiveness related to fossil energy. This can be partially achieved by stabilizing the feedstock supply and designing an optimal transportation plan.

The sustainability of biomass feedstock supply chain can be achieved by using a mix of feedstocks such as corn, wheat and switchgrass, which can be harvested/collected at different time periods of each year. Another possibility is to establish a biomass storage facility in a bio-refinery site to reduce the dependency of feedstock seasonality.

Lignocellulosic biomass such as corn, wheat and switchgrass can be used as feedstock and can be interchangeable because their compositions are very similar in terms of percentage of lignin, cellulose and hemicellulose ([2]). The utilization of three types of feedstock has a very minimal effect on biorefinery production costs and bioethanol yields.

Typically, local/regional suppliers are always prioritized because of short transportation distances and delivery times. In case of supply disruption due to a bad weather condition, international/national market sources might be sought to meet the biofuel production plan. For the stabilization of local/regional feedstock supply, an extra price must be paid to farmers with a contract to incentivize the production of energy crop and collection of crop residues. This extra price corresponds to the opportunity cost of an alternative crop, according to [3].

The research of [4] classified the feedstock, biofuels and the related production processes into three generations. Up to now, only the first generation of biofuels has reached the industrial stage. In this generation, biofuels are produced from food crops through ordinary technologies such as abstracting oils for biodiesel or producing bioethanol through fermentation. However, the main issue with the first generation biofuels is their competition with food security.

The second generation biofuels are in the industrial take-off phase. They use lignocellulosic biomass such as wood, organic waste, food crop waste and specific biomass crops. Unlike the first generation, they can eliminate the main problem of the production of the first generation biofuels.

The third generation biofuels are produced from algae. Their possibility of scaling up has not been yet proved at this time because of high costs and complicated production processes.

In this study, we focus on the management of a lignocellulosic biomass supply chain against the variations of feedstock supply. We want to address issues related to the following questions:

- How can we stabilize the feedstock supply and design an optimal transportation plan? What are potential solutions ?
- How can we develop an optimization model to evaluate the economic impact of the feedstock supply and transportation plan on a biomass supply chain?

This paper is organized as follows: Section 2 describes the main characteristics of biomass supply chains and gives a literature review. In Section 3, the problem and its model are introduced. A numerical study is presented and its results are analyzed in Section 4. Section 5 concludes this paper with some remarks for future research.

II. Literature review

A. Biomass Supply Chain

The research of [5] listed all phases required to supply biomass from its production point to a service station (gas station) such as: ground preparation and plantation, cultivation, harvesting, handling, storage, conversion, transportation and utilization of the fuel at the service station.

Unlike a traditional supply chain, the biomass supply chains differ significantly in terms of their structures and characteristics. Biomass supply chains are usually characterized by the deterioration of biomass and seasonal fluctuation of supply. The degradation of biomass leads to a stock loss, which is estimated to be 1-2% of the material stocked per month under the ambient storage, according to [5]. Moreover, biomass supply chains are characterized by uncertainties related to weather condition, seasonal harvest, biomass's physical and chemical characteristics such as density, long term-contracts with potential suppliers, transportation and local distribution infrastructure and government policies ([6], [7]).

B. Decision-making models

Different authors [8], [9] categorized decision-making in supply chain management into three levels according to the degree of importance and the planing horizon such as strategic, tactical and operational decisions.

Most researchers focused on the design and management of supply chains at the strategic level. The decision-making at this level involves long-term decisions such as selecting production technologies, biomass preprocessing techniques, determining the facilities'capacity and location, the optimal configuration of logistics network, establishing long-term contracts with suppliers. In the research of [10], the authors proposed a mixed integer linear programming (MILP) model which can be used to design the optimal supply chain network to manage logistics operations against fluctuations of biomass availability and hydrogen demand. The research of [11] developed a MILP model called Biomass Value Chain Model (BVCM) which takes account of the economic and environmental impacts associated such as: optimal use of land, biomass resources and technologies with respect to different objectives, scenarios and constraints. [12] proposed a bioefuel supply chain model that minimizes operational costs and the biomass quality-related costs.

Tactical decisions are those such as transportation mode selection or supplier selection. The research developed by [13] proposed an operational scheduling model to find the optimal number of machines and workers to meet resource requirements planning. In the research of [14] proposed a multi-product and multi-period biomass supply chain by taking account of the harvest equipment and transport mode. In addition, [15] studied a mixed integer linear programming (MILP) model for the optimal design of a bioethanol supply chain in northern Italy.

The operational level corresponds to short-term decisions, weekly and daily, which concern harvest schedule, production schedule, inventory control and vehicle routing problems. In the study of [16], a new concept was presented, called "distance potential", which could be applied to design the biomass supply chain and resolve vehicle routing problems. [17] studied a transportation planning problem in a forest biomass supply chain under work-time constraints and predetermined orders.

C. Stabilization of feedstock supply

According to [18], there are several types of contracts which could enable us to stabilize the feedstock supply, are the 1 to 3 year-contracts between biofuel producers and biomass suppliers. The contract duration depends on the negotiation capacity and suppliers' track records .

Based on Stackelberg Game Theory, [19] proposed a contract model, in which the biofuel production company coordinated the farmer's production and the contract quantity. Considering various market uncertainties and different combinations of biomass characteristics (low value, high cost of supply), they concluded that option contract improve the overall benefits of the supply chain and create a win-win situation between biorefineries and farmers.

Similarly, [20] studied a stochastic dynamic programming model to understand the essential role of the "inflexibility premium" in long-term relationship between biorefineries and farmers. The inflexibility premium is defined as an increase in the purchase price of an agricultural product that allows a farmer to sign a long-term contract. Using game theory, the research of [21] focused on the structure of a contract between biorefineries and biomass suppliers in different biomass markets. They concluded that an optimal contracting strategy can reduce risks of biomass supply variations.

To the best of our knowledge, none of these studies pesent a multi-period supply chain model to cope with the stabilization of feedstock supply and the design of transportation plan. For this reason, we present a multi-period mixed integer programming model that can be applied to capture most activities in biomass supply chain, including feedstock supply, storage, transportation, conversion...Our model is quite flexible, allowing to model supplier selection and transportation planning.

III. Problem description and model formulation

In this study, the production infrastructure such as pretreatment facilities and biorefineries have been located. Our work is focused on how to select feedstock suppliers and establish a mid/long-term contract with them to ensure the stabilization of feedstock supply. We assume that biofuel producers seek to establish a long-term relationship with certain local/regional suppliers in order to stabilize the feedstock supply. The unit purchase price and the minimum delivery quantity are fixed under a long-term contract during a planning horizon. In the worst case, biofuel producers could purchase an additional amount of biomass from international/national market (if available) to avoid unmet demand. We assume that purchase price from international/national market includes the unit transportation cost.

We assume that lignocellulosic biomass such as corn, wheat and switchgrass can be used as feedstock and is interchangeable with minimal effect on biorefinery production costs and bioethanol yields.

The unit transportation cost per trip from local/regional suppliers to biorefineries is pre-computed based on their distances and fixed costs associated with trucks.

In the following, we formulate a MILP model that considers the selection of suppliers with a mid/long-term contract, inventory management and transportation planning. We consider a planning horizon of T periods, each period could be a week, a month or a trimester. Let I be a set of biomass suppliers, P be a set of pretraitment facilities and B be a set of biorefineries. An arc (i, j) stands for a path from node i to node j in the supply chain network considered. Each node corresponds to a supplier, a pretreatment facility or a biorefinery.

For convenience, all notations used in the model formulation are summarized in Table I. All continuous decision variables are non-negative. The MILP model is presented as follows:

Objective function:

$$\begin{aligned} min \quad z =& \sum_{t\in T}\sum_{(i,j)\in A} nbrT(i,j,t) * Vcost(i,j) && (1)\\ &+\sum_{t\in T}\sum_{p\in P}\eta^P Q^P(p,t)tcostP(p) && (2)\\ &+\sum_{t\in T}\sum_{b\in B}\Theta(b,t)pcostB(b) && (3)\\ &+\sum_{t\in T}\sum_{b\in B}\eta^B z(b,t)costB(B) && (4)\\ &+\sum_{t\in T}\sum_{p\in P} hcost^P(p)W(p,t) && (5)\\ &+\sum_{t\in T}\sum_{b\in B} hcost^B_{av}(b)V_{av}(b,t) && (6)\\ &+\sum_{t\in T}\sum_{b\in B} hcost^B_{ap}(b)V_{ap} && (7)\\ &+\sum_{t\in T}\sum_{i\in I} Q^I(i,t) * costA^I(i) && (8)\\ &+\sum_{t\in T}\sum_{m\in M}\sum_{b\in B} x(m,b,t) * costA^M(m) && (9) \end{aligned}$$

Constraints on suppliers

$$\begin{aligned} &Q^I(i,t) = \sum_{(i,p)\in A(i,'*')} x(i,p,t) && (10)\\ &Q^I(i,t) \le avail^I(i,t).\varphi(i) && (11)\\ &Q_{min}(i)\varphi(i) \le \sum_{t\in T} Q^I(i,t) && (12) \end{aligned}$$

Constraints on market

$$\sum_{b\in B} x(m,b,t) \le avail^M(m,t) \quad (13)$$

Constraints on preprocessing sites

$$\begin{aligned} &Q^P(p,t) = \sum_{(i,p)\in A('*',p)} x(i,p,t) && (14)\\ &W(p,t) = \lambda^P(p)W(p,t-1) + \eta^P Q^P(p,t)\\ &- \sum_{(p,b)\in A(p,'*')} x(p,b,t) && (15)\\ &\Gamma^{min}_p(p) \le Q^P(p,t) \le \Gamma^{max}_p(p) && (16)\\ &W^{min}(p) \le W(p,t) \le W^{max}(p) && (17) \end{aligned}$$

Constrains on refinery sites

$$\begin{aligned} V_{av}(b,t) =& \lambda^B(b)V_{av}(b,t-1) + \sum_{(p,b)\in A('*',b)} x(p,b,t)\\ &+ \sum_{m\in M} x(m,b,t) - z(b,t) && (18)\\ V_{ap}(b,t) =& V_{ap}(b,t-1) + \eta^B z(b,t)\\ &+ shortage(b,t) - demand(b,t) && (19)\\ \Gamma^{max}_b(b) \le& z(b,t) \le \Gamma^{max}_b(b) && (20)\\ V^{min}_{av}(b) \le& V^B_{av}(b,t) \le V^{max}_{av}(b) && (21)\\ V^{min}_{ap}(b) \le& V^B_{ap}(b,t) \le V^{max}_{ap}(b) && (22) \end{aligned}$$

Constraints on transportation

$$x(i,j,t) \le nbrT(i,j,t) * CapV(i,j) \quad (23)$$

The objective function is to minimize the total cost, composed of nine costs: transportation costs (1st term), pretraitement costs (2nd term), shortage costs (3rd term), biofuel conversion costs (4th terms) holding costs(5-7th terms), purchase costs (8-9th terms).

Equation (10) defines the total amount of feedstock is transported from supplier i to pretreatment facilities in each period. Equation (14) defines the total amount of biomass is transported from pretreatment facility p to refineries in each period. Note that the decision variables $Q^I(i,t)$ and $Q^P(p,t)$ are redundant but they are introduced for improving the readability of the model.

Inequality (11) states that there is no feedstock supply unless there is a mid/long-term contract between supplier i and biorefinery b. This inequality also ensures that the total purchase amount must not exceed the available amount in each period.

Inequality (13) shows that the purchase amount of biomass is also limited by the market's supply availability. Constraints (15), (18) and (19) are the inventory balance constraints for each stock at preprocessing facility and biorefinery, respectively. These constraints ensure that no more biomass or biofuel is delivered or processed at a location than available amount in stock. Constraint (19) ensures that biofuel demand is satisfied as much as possible, and unsatisfied demand is lost. Shortage cost is incurred for any unit lost. Constraints (17), (21) (22) show inventory capacity in each conversion site.

Constraint (23) shows the number of trips required for transporting a given amount of biomass from node i to node j in each period. The capacity of each preprocessing/refinery plant is respected via constraints (16) and (20).

TABLE I. LIST OF PARAMETERS AND VARIABLES

Set	
T	Set of periods
I	Set of suppliers
P	Set of preprocessing facilities
B	Set of biorefineries
A	Set of arcs
Parameters	
$avail^I(i,t)$	Amount of biomass (tons) available at supplier i in period t
$costA^I(i)$	Unit purchase cost of biomass from supplier i (\$/ton)
$Q_{min}(i)$	Minimum amount of biomass (tons) under contract with supplier i during planning horizon T
	International/national market
$avail^M(m,t)$	Amount of biomass (tons) available at market m in period t
$costA^M(m)$	Unit purchase cost of biomass from market m (\$/ton)
	Preprocessing site
$\Gamma_p^{max}(p)$	Maximum amount of biomass (tons) that could be preprocessed at preprocessing facility p
$\Gamma_p^{min}(p)$	Minimum amount of biomass (tons) that could be preprocessed at preprocessing facility p
$W^{max}(p)$	Maximum biomass storage capacity (tons) at preprocessing facility p in each period
$W^{min}(p)$	Minimum biomass storage capacity (tons) at preprocessing facility p in each period
$hcostP(p)$	Biomass inventory holding cost at preprocessing facility p (\$/ton)
$tcostP(p)$	Biomass preprocessing cost of preprocessing facility p (\$/ton)
$\lambda^P(p)$	Ratio of biomass storage loss at preprocessing facility p in each period
$\eta^P(p)$	Preprocessing rate at preprocessing facility p in each period
	Biorefinery site:
$\Gamma_b^{max}(b)$	Maximum amount of biomass (tons) that could be processed by biorefinery b in each period
$\Gamma_b^{min}(b)$	Minimum amount of biomass (tons) that could be processed by biorefinery b in each period
$V_{av}^{max}(b)$	Maximum biomass storage capacity (tons) of biorefinery b in each period
$V_{av}^{mmin}(b)$	Minimum biomass storage capacity(tons)of biorefinery b in each period
$V_{ap}^{max}(b)$	Maximum biofuel storage capacity(litres) of biorefinery b in each period
$V_{ap}^{mmin}(b)$	Minimum biofuel storage capacity (litres) of biorefinery b in each period
$pcost(b)$	Unit penalty cost for unmet biofuel demand at biorefinery b (\$/tons)
$tcostB(b)$	Biofuel production cost at biorefinery b (\$/tons)
$hcostB_{av}(b)$	Unit inventory holding cost of biomass at biorefinery site b in each period (\$ / ton)
$hcost_{ap}^B(b)$	Unit inventory holding cost of biofuel at biorefinery b in each period (\$/ ton)
$\eta^B(b)$	Biofuel yield rate of lignocellulosic biomass for biorefinery b (litre/ton)
$\lambda^B(b)$	Ratio of biomass storage loss at biorefinery b in each period
$demand(b,t)$	Demand of biofuel(litre) at biorefinery b in period t
	Transportation:
$CapV(i,j)$	Transportation capacity from node i to node j (tons/truck)
$Vcost(i,j)$	Transportation cost of each trip of a truck from node i to node j (\$ /trip)
Decision variables	
	Continuous variables
$x(i,j,t)$	Amount of biomass transported from node i to node j in each period t
$W(p,t)$	Biomass inventory level (tons) at preprocessing facility p in period t
$V_{av}^B(b,t)$	Inventory level of biomass at biorefinery b in period t
$V_{ap}^B(b,t)$	Inventory level of biofuel at biorefinery b in period t
$Q^P(p,t)$	Amount of biomass preprocessed (tons) at preprocessing facility p in period t
$z(b,t)$	Amount of biomass required (tons) to produce biofuel at refinery b in period t
$shortage(b,t)$	Shortage of biofuel (litres) at biorefinery b in period t
	Integer variables
$nbrT(i,j,t)$	Number of truck trips from node i to node j in period t
	Binary variables
$\varphi(i)$	Equal to 1 if a mid/long-term contract between refinery b and supplier i is established, 0 otherwise

TABLE II. INPUT PARAMETERS

Input parameters	Value
Bioethanol yield rate (litre/tonne)	313.00
Penalty cost for ummet bioethanol demand ($/litre)	1.06
Bioethanol production cost ($/litre)	0.20
Maximum annual production capacity of biorefinery(MLPY)	380.00
Minimum annual production capacity of biorefinery(MLPY)	190.00
Preprocessing cost at preprocessing facility ($/tonne)	13.39
Inventory cost for biomass in time period t ($/tonne)	0.90
Ratio of biomass storage loss in time period	0.01
Biomass purchase price ($/tonne)	35.00
Maximum annual pretraitement capacity (tonS)	2 000 000.00
Maximum biomass storge capacity (tons) of biorefinery	3 500 000.00
Maximum biomass storge capacity (tons) of pretraitement facility	3 500 000.00

This model allows the user to formulate complex biomass supply chains by adding a new nodes (suppliers, preprocessing/refinery plants) and arcs linking these nodes.

IV. NUMERICAL STUDY

In the numerical study, we consider a biomass supply chain with 1 refinery, 1 pre-traitement site, 10 farms within radius of 50km, over a three year-horizon divided into 12 periods. One period in this study corresponds to 3 months. The biofuel producer seeks to sign a three years contract with several local feedstock suppliers (switchgrass). Note that switchgrass is harvested once a year in autumn.

With regard to suppliers, amount of biomass available in each supplier can be estimated from land area for switchgrass cultivation and its yield. We also use parameter data from [22] and other sources such as literature reviews and research articles to validate our model. These parameters are presented in Table II. We precomputed transportation cost per trip based on distances, variable transportation cost and fixed cost.

The model size increases when the numbers of periods, suppliers, pretraitment facilities and biorefineries increase. Based on historical data (19981-2005) in [22], we generate annual biofuel demand randomly from Normal probability distribution $N(2133, 4138)$ and truncated on the interval $(2006, 2280)$. For testing the model with different instances, we adjust biofuel demand according to the scalability of feedstock suppliers.

Fig 1 represents the distribution of costs in the biomass suppy chain over planning horizon T= 12. The optimal value of total cost in the supply chain is $36 716 326.24 when a biomass supply chain includes 1 refinery, 1 pre-traitement site, 10 farms over 12 time periods.

It is obvious that transportation costs play an important role in the biomass supply chain with approximatively 24% of total system costs. This result shows the importance of transportation operation in biomass supply chains. Holding costs is only 2% because in this study the amount of biomass available for biofuel production is higher than the demand, so the storage of a large amount of biomass at biorefinery seems to be not critical.

One important aspect of the model is the running time which depends on the number of periods, suppliers, pretraitment facilities and biorefineries. We further tested this model

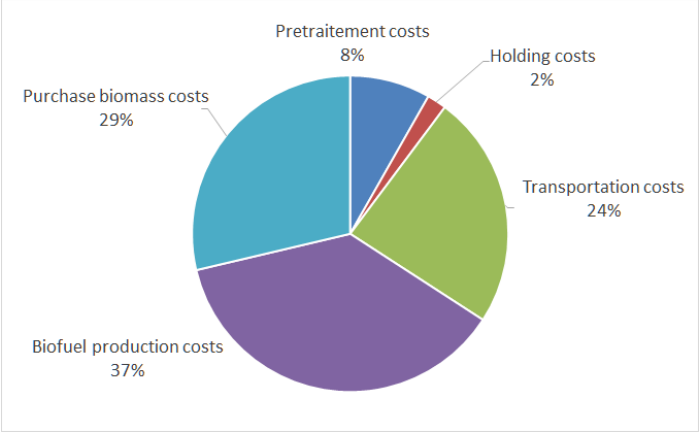


Fig. 1. Distribution of cost in supply chain

for different instances when I= 10, 20, 50, 100; P= 1, 2 and T= 4, 12. All instances were resolved by the MIP solver of GUROBI 6.5 in Python 3 on a HP computer with Intel Core i5-4210M CPU 2.66 GHz and 4.0 GB RAM. The size, error (%) and execution time of these instances are summarized in Table III. In fact, the error is a relative gap between the lower and upper objective bound. The results illustrate that this model can be solved quickly and exactly with a commercial solver GUROBI.

TABLE III. INSTANCES

Instances	I	P	T	Gap (%)	CPU time (s)
1	50	1	4	0.09	0.14
2	100	1	4	0.06	0.30
3	10	1	12	0.08	0.50
4	20	1	12	0.01	0.05
5	50	1	12	0,06	0.12
6	100	1	12	0.08	9.29
7	20	2	12	0,04	10.16
8	50	2	12	0,05	10.50
9	100	2	12	0.05	14.50
10	100	4	12	0.05	24.50

V. CONCLUSION

We developed a mixed integer linear programming (MILP) model for simultaneously minimizing the total system costs.

The model integrates all supply chain logistics such as transportation plan, purchase activities, inventory control, pretraitement and production to provide an optimal solution. The case presented has been solved optimally with different instances in a reasonable computation time.

We hope our study could provide a decision-making support tool for biomass supply chain management. The next step of our work is to extend our model to stochastic environments. This extended model would need to incorporate uncertainty and different sustainability criteria.

REFERENCES

[1] S. D. Ekşioğlu, A. Acharya, L. E. Leightley, and S. Arora, "Analyzing the design and management of biomass-to-biorefinery supply chain," *Computers & Industrial Engineering*, vol. 57, no. 4, pp. 1342–1352, 2009.

[2] Z. Li, Y. Liu, W. Liao, S. Chen, and R. S. Zemetra, "Bioethanol production using genetically modified and mutant wheat and barley straws," *biomass and bioenergy*, vol. 35, no. 1, pp. 542–548, 2011.

[3] A. M. Argo, E. C. Tan, D. Inman, M. H. Langholtz, L. M. Eaton, J. J. Jacobson, C. T. Wright, D. J. Muth, M. M. Wu, Y.-W. Chiu *et al.*, "Investigation of biochemical biorefinery sizing and environmental sustainability impacts for conventional bale system and advanced uniform biomass logistics designs," *Biofuels, Bioproducts and Biorefining*, vol. 7, no. 3, pp. 282–302, 2013.

[4] A. T. E. Pérez, M. Camargo, P. C. N. Rincón, and M. A. Marchant, "Key challenges and requirements for sustainable and industrialized biorefinery supply chain design and management: A bibliographic analysis," *Renewable and Sustainable Energy Reviews*, vol. 69, pp. 350–359, 2017.

[5] A. A. Rentizelas, A. J. Tolis, and I. P. Tatsiopoulos, "Logistics issues of biomass: the storage problem and the multi-biomass supply chain," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 4, pp. 887–894, 2009.

[6] B. H. Ba, C. Prins, and C. Prodhon, "Models for optimization and performance evaluation of biomass supply chains: An operations research perspective," *Renewable Energy*, vol. 87, pp. 977–989, 2016.

[7] S. Gold and S. Seuring, "Supply chain and logistics issues of bio-energy production," *Journal of Cleaner Production*, vol. 19, no. 1, pp. 32–42, 2011.

[8] J. Mula, D. Peidro, M. Díaz-Madroñero, and E. Vicens, "Mathematical programming models for supply chain production and transport planning," *European Journal of Operational Research*, vol. 204, no. 3, pp. 377–390, 2010.

[9] E. Iakovou, A. Karagiannidis, D. Vlachos, A. Toka, and A. Malamakis, "Waste biomass-to-energy supply chain management: a critical synthesis," *Waste management*, vol. 30, no. 10, pp. 1860–1870, 2010.

[10] Y.-b. Woo, S. Cho, J. Kim, and B. S. Kim, "Optimization-based approach for strategic design and operation of a biomass-to-hydrogen supply chain," *International Journal of Hydrogen Energy*, vol. 41, no. 12, pp. 5405–5418, 2016.

[11] S. Samsatli, N. J. Samsatli, and N. Shah, "Bvcm: a comprehensive and flexible toolkit for whole system biomass value chain analysis and optimisation–mathematical formulation," *Applied Energy*, vol. 147, pp. 131–160, 2015.

[12] K. K. Castillo-Villar, H. Minor-Popocatl, and E. Webb, "Quantifying the impact of feedstock quality on the design of bioenergy supply chain networks," *Energies*, vol. 9, no. 3, p. 203, 2016.

[13] M. M. Morales-Chvez, S.-M. Jos A, and A. S. William, "A mixed-integer linear programming model for harvesting, loading and transporting sugarcane. a case study in peru," *DYNA*, vol. 83, no. 195, pp. 173–179, 2016.

[14] B. H. Ba, C. Prins, and C. Prodhon, "Modèle tactique dapprovisionnement en biomasse pour une bioraffinerie de proximité."

[15] O. Akgul, A. Zamboni, F. Bezzo, N. Shah, and L. G. Papageorgiou, "Optimization-based approaches for bioethanol supply chains," *Industrial & Engineering Chemistry Research*, vol. 50, no. 9, pp. 4927–4938, 2010.

[16] J.-M. Li, A.-H. Li, P. S. Varbanov, and Z.-Y. Liu, "Distance potential concept and its applications to the design of regional biomass supply chains and solving vehicle routing problems," *Journal of Cleaner Production*, 2017.

[17] S.-K. Han and G. E. Murphy, "Solving a woody biomass truck scheduling problem for a transport company in western oregon, usa," *Biomass and Bioenergy*, vol. 44, pp. 47–55, 2012.

[18] F.-J. DANIEL and M. TOMSON, "La structuration socio-conomique de lapprovisionnement en biomasse nergie en alsace," 2015.

[19] H. Gong, Y. Zhang, and J. Li, "Coordination mechanism by option contract in the biomass supply chain organized by company and farmer," in *Automation and Logistics (ICAL), 2010 IEEE International Conference on*. IEEE, 2010, pp. 71–75.

[20] R. Jalili, "Long term contracts and farm inflexibility premium in the production of cellulosic ethanol," Ph.D. dissertation, University of British Columbia, 2008.

[21] R. Golecha and J. Gan, "Optimal contracting structure between cellulosic biorefineries and farmers to reduce the impact of biomass supply variation: game theoretic analysis," *Biofuels, Bioproducts and Biorefining*, 2016.

[22] A. Osmani and J. Zhang, "Stochastic optimization of a multi-feedstock lignocellulosic-based bioethanol supply chain under multiple uncertainties," *Energy*, vol. 59, pp. 157–172, 2013.

Joint location of facilities and mobile resources: an application to the sector of public works

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Agri Utama Noor Permadi
IMT Atlantique,
4 rue Alfred Kastler, 44300 Nantes, France
Institute of Technology Bandung
Jl. Ganesha No.10, 40132 Bandung, Indonesia
agri-utama.noor-permadi@imt-atlantique.net

Axel Grimault
IMT Atlantique,
LS2N, UMR CNRS 6004
4 rue Alfred Kastler
44300 Nantes, France
axel.grimault@imt-atlantique.fr

Olivier Péton
IMT Atlantique,
LS2N, UMR CNRS 6004
4 rue Alfred Kastler
44300 Nantes, France
olivier.peton@imt-atlantique.fr

***Abstract*—We model the problem of locating logistics platforms and planning a set of tasks. A subset of tasks requires the use of mobile equipment, i.e. machines that can be moved from one logistics platform to another. Thus, the optimization problem includes two types of decisions: (i) strategic decisions include the location of new platforms and mobile platforms, and (ii) tactical decisions include the flow of materials through the network.**

The company has a large set of clients with irregular demand, with a deterministic but variable demand during a subset of time periods. We consider a time horizon of one year, decomposed into time periods of one month. The objective function to be minimised is the sum of the fixed set-up cost of platforms, production and delivery costs and the cost of moving the mobile resources from one platform to another. The optimization problem is modelled as a mixed integer linear program (MILP) and solved with IBM Ilog Cplex 12.6.3. We present numerical experiments for a case study in the sector of public works.

I. Introduction

This paper introduces a supply chain network design problem in the field of public works. Due to the implementation of government legislation, social responsibility, environmental concern, economic benefits and customer awareness, industries are under a great pressure not only to provide environmentally friendly products but also to recover end-of-use products. The accelerating expansion of road construction and maintenance cause an important amount of recyclable waste, particularly, reclaimed asphalt pavement and recyclable aggregate.

We consider the case of a public works company that provides materials to a large set of construction sites from a small set of logistics platforms. These platforms are used for storage, consolidation of shipments, production, transformation and recycling of materials. Since the location of construction sites can vary a lot from one year to another, the logistics network of the company, i.e. the location of platforms, must be much more flexible than in most industrial activities.

The process of recycling materials is performed by machines which can moved from one logistics platform to another. Machines receives used materials from clients, transforms one part into recycled materials, that are sent back to clients and into non-recyclable materials (waste) that are sent to disposal facilities (landfills). The goal of this study is twofold: (i) select a subset of logistics platforms from a larger list of candidate locations, and (ii) locate an existing set of mobile machines on these platforms. Without loss of generality, we consider a time horizon of one year, decomposed into time periods of one month.

This study belongs to the category of multi-period facility location problems [4], [8] for closed-loop supply chains [1], [2]. It presents some similarities with the Mobile Facility Location Problem (MFLP), in which both clients and facilities are mobile [3]. To the best of our knowledge, the location of mobile resources placed on facilities has been studied in very few papers. Jena et al. [6], [5], [7] study the location and relocation of working camps in the forestry sector in Canada. The camps are composed of trailers that host workers or various facilities. Trailers can be relocated from one camp to another. The authors propose a network model to manage opened and closed trailers at each location. Relocation of trailers is modelled by the use of a central node referred as a hub node: the flow of relocated trailers is first sent to the hub, and then further distributed to another location. In this paper, we consider the location of mobile machines for production or recycling of goods. Therefore, the main contribution of this paper compared with Jena et al. is to include inventory management decisions associated with the location of logistics platforms and machines.

This paper is organised as follows. Section II presents the problem and a mathematical model is described in Section III. Section IV describes the instances that we randomly generated according to demand values, costs and logistic data provided by a public works company in Western France. Section V presents computational results. Conclusions and perspectives are given in Section VI.

II. Problem Settings

A. Facilities and Material Flows

The problem is defined on a graph with four types of facilities: suppliers, logistics platforms, clients and landfills.

Suppliers are the source of all material flows. They can supply clients by means of direct shipments. They can also supply logistics platforms. The capacities of suppliers are assumed to be infinite.

Logistics platforms are used to store materials, collect surpluses and recyclable materials from clients, process materials (e.g. cleaning, sorting, crushing, screening), and recycle goods. There are many reasons for using logistics platforms instead of supplying the clients directly. (i) Some clients are not accessible by large trucks; platforms help split large shipments into smaller ones. (ii) Some clients may require a large quantity of materials in very short time, building stock on a platform is then necessary. (iii) Material can benefit from safety and processing operations at platforms, not at clients. (iv) Clients need to move materials (e.g. earth) to the platforms. Therefore, itineraries between platform and customers generally avoid empty return trips.

In practice, platforms dedicate special areas for sensitive materials (e.g. clinker, asbestos) and the remaining space can be shared among non-sensitive materials. Therefore, we assume that platforms have an overall storage capacity and a maximum capacity level specific for each material. We also consider a flow limitation due to the systematic use of weighbridge by trucks entering and leaving platforms.

Clients are construction sites with both an ingoing demand for construction materials (e.g. aggregate, asphalt, gravel) and an outgoing demand corresponding to the removal of materials from construction sites (e.g. earth, recyclable asphalt or aggregate). The duration of construction projects varies from a few days to a few months. Clients' demands are considered deterministic but irregular: their value can change in each time period; there can even be periods with no demand.

Landfills are large facilities (infinite capacity) that receive all waste, i.e. materials that cannot be recycled. We distinguish two sources of waste. First, recycling materials at platforms results in a proportion of new product and a proportion of waste. Second, clients also generate waste which is sent directly to disposal facilities as soon as a full truckload is reached. This avoids unnecessary and expensive unloading/loading operations at any intermediate facility. This flow cannot be optimised. Thus, we do not consider it in the present paper.

For the sake of simplification, we assume only one single commodity, which can be found in three states (new, used and waste), hereby called products. Figure 1 represents all possible types of arcs in the graph representing the supply chain. New product is sent from suppliers to clients, either directly or via the platforms. Used materiel is moved from clients to platforms. Machines, located at platforms, transform used product into a proportion of recycled product (that has the same characteristic as new product) and a proportion of waste product, which is sent to landfills. Both the new and recycled products can be transported from one platform to another.

Figure 2 shows an example of material flow in a network with 20 clients. It can be seen that clients $c5$, $c6$ and $c15$ are delivered directly by suppliers, 11 customers are delivered by platforms and 5 clients are not delivered at the current time period. Platforms $p5$, $p8$ and $p11$ are active and machines are positioned at platforms $p5$ and $p11$.

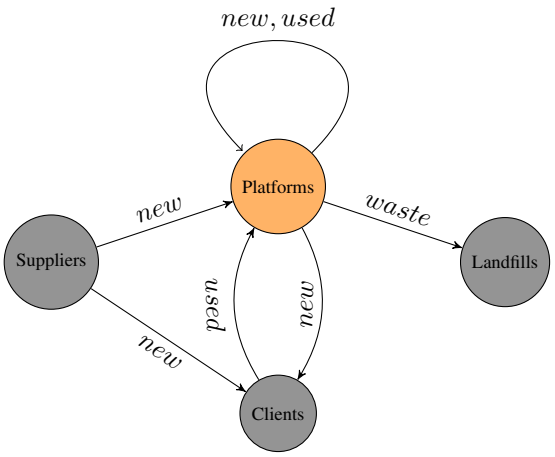
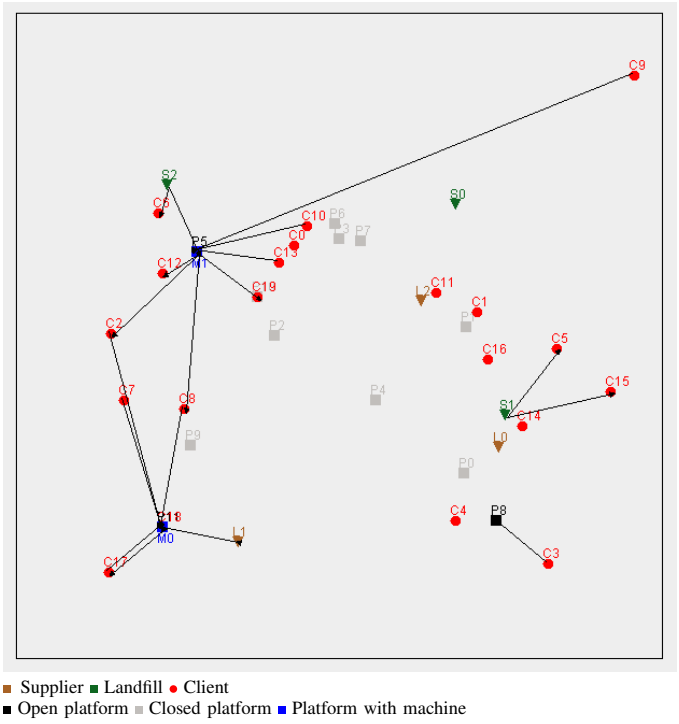


Fig. 1. Interactions between facilities



Supplier Landfill Client
Open platform Closed platform Platform with machine

Fig. 2. Material flow in the logistics network

B. Machines

Civil engineering companies own a set of machines dedicated to specific tasks (e.g. cement plants, asphalt plants, crushers). These machines can be fixed or mobile. The first case can be handled by our mathematical model by setting an arbitrarily large moving cost. In the latter case, machines can be transported from one platform to another. Transportation cost is a lump sum corresponding with the cost of disassembling and reassembling costs, and organizing abnormal transport for a fixed duration (costs related to the distance travelled is negligible). These machines transform used product into new product. Since their capacity is generally high, they can work only if a sufficient inventory of used product is available at the platform. Thus, moving a machine from one platform to another avoids moving much used product. The

transport of a machine only takes a few days and it has a negligible impact on the availability of the machine. However, in order to avoid too frequent moves, we assume that a machine must stay at least τ time periods at a platform once installed.

C. Business Constraints

The main goal of our optimization problem is to locate logistics platforms and machines. We assume a fixed set of suppliers and landfills with infinite capacity. These facilities are open during the whole planning horizon. Given a set of candidate platforms, our goal is to select, in each time period, a subset of active platforms. Contrary to many facility location problems, which assume long term location decisions, logistics platforms can be temporarily used, and thus can be opened or closed more than once over the time horizon (see also [6]). Opening and closing platforms is subject to a quite high fixed cost, so that optimal solutions have the characteristic of avoiding too many moves over the time horizon. Platforms can be closed only if there is no inventory left. At each time period, a maximum number of operating platforms is assumed to be known. Generally, for public works company, there is enough space to store raw materials which are not considered as finished goods with much added-value. This non-significant inventory cost is included in the fixed cost of opening a platform.

All machines are moveable, but must be positioned on active platforms only. The number of mobile machines is known a priori and does not vary during the planning horizon. We assume that the decision of investing in new machines is out of the scope of the present work. However, our model can be used as a simulation tool to test various values for the number of machines.

The real structure of the transportation costs is quite complex, due to several factors: all trips starting at a platform are performed by vehicles owned by the company, which are generally cheaper than outsourced trips. Trips originating form platforms enable shipment consolidation. Finally, trips between two platforms or between a platform and a client are more likely to have a non-empty return trip than trips originating at suppliers or trips terminating at landfills (see Figure 1). Moreover, these trips are always performed by a private fleet of trucks from the company, with cheaper transportation costs on these arcs. In this study, we slightly simplify the real-life costs and consider two categories of costs: all trips between suppliers and clients or between platforms and landfills are penalised by a factor $\beta > 1$.

One remarkable assumption is that we consider only materials with cheap (sometimes even negative) value, which storage does not cause any depreciation. Thus we do not consider any inventory cost.

III. Mathematical Formulation

A. Notations

We consider the set J of logistics platforms, the set S of suppliers and the set L of landfills. The maximum number of logistics platforms that can be operated simultaneously is denoted $\bar{J}$. The set of clients (construction sites), machines and times periods are denoted I, M and $T = 0, 1, \ldots, |T|$, respectively. We also define $T^* = T \backslash \{0\}$. The optimization problem is defined on a graph which set of vertices is I, S, J, L and the set of arcs can be partitioned into 3 sets as illustrated in Figure 1: $\mathcal{A}$ includes all arcs for new product, $\mathcal{A}'$ includes all arcs for used product and $\mathcal{A}''$ includes all arcs for waste.

Clients' demand is twofold. The ingoing demand d_{it} of client $i \in I$ at time period $t \in T$ concerns either new or recycled products. The outgoing demand d'_{it} concerns materials sent from clients to platforms.

When processed by machine $m \in M$, a rate α_m of the used product is recycled into new product and the remaining quantity, is waste product that must be sent to a landfill.

All facilities have limited storage and production capacities. This overall storage capacity (shared between all products) is denoted as $\bar{S}_j$ for all $j \in J$. Storage capacity for the new, used and waste products at facility $j \in I \cup J$ ars denoted $\bar{s}_j$ and $\bar{s}'_j$ respectively. Storage capacity of waste at $j \in J$ is denoted $\bar{s}''_j$. Storage and production capacity at suppliers and landfills is assumed to be infinite.

The flow limitation at platform $j \in J$, i.e. the maximum number of units of all products entering or leaving the platform during one time period, is denoted κ_j. The machine $m \in M$ is able to process a quantity π_m of used product per time period.

We consider the following costs. All platforms $j \in J$ have a fixed opening cost f_j^+ and a fixed closing cost f_j^-. We assume that the set of machines is defined *a priori* and will not change within the time horizon T. Machine $m \in M$ can be moved from a logistics facility to another, with a fixed moving cost f_m independent of the distance between the facilities. Machine $m \in M$ has a production cost c_m per unit of new product processed. Platforms $j \in J$ have a fixed running cost f_j per time period. Finally, each unit of any product transported on arc (u, v) has a cost c_{uv}.

The mathematical model includes five sets of *binary variables*: Let $y_{jt} = 1$ indicate if platform $j \in J$ is open at time period $t \in T$, and 0 otherwise. Variables v_{jt}^+ and v_{jt}^- are set at value 1 if platform $j \in J$ opens (respectively closes) at time period t, and 0 otherwise. Variables $x_{mjt} = 1$ are set at value 1 if machine $m \in M$ is operating at platform $j \in J$ at time period $t \in T$, and 0 otherwise. Finally, the variables z_{mt} are set at value 1 if machine $m \in M$ is moved from one platform to another at time period $t \in T$, and 0 otherwise.

The following continuous decision variables model the product flows. u_{mjt} represents the quantity of used product transformed into new product at platform $j \in J$ by machine m in time period t. q_{uvt}, q'_{ijt} and q''_{ijt} represents the flow of new product, used product and waste, respectively, on arc (u, v) at time period $t \in T$. Finally, s_{jt}, s'_{jt} and s''_{jt} represent the inventory of new product, used product and waste, respectively, at facility $j \in I \cup J$ in the end of time period $t \in T$. Note that, in order to avoid artificial excessive storage in the end of the time period, we impose a target value of these inventories at time period $|T|$. Hence, the initial state of the network (open platforms, location of machines, inventories) at the beginning of the first time period is assumed to be known, as well as inventories at time period $|T|$.

There are initial condition for several variables. Those

variables are y_{jt}, v^+_{jt}, v^-_{jt}, x_{mjt}, s_{jt}, s'_{jt} and s''_{jt}. These initial condition have index $t = 0$.

B. Mixed Integer Linear Programme

The optimization problem can be modelled as the following Mixed Integer Linear Programme (MILP):

$$\min z = F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8, \quad (1)$$

where

$$F1 = \sum_{t\in T}\sum_{j\in J}\sum_{m\in M} c_m\, u_{mjt}$$
$$F2 = \sum_{t\in T}\sum_{j\in J} f_j\, y_{jt}$$
$$F3 = \sum_{t\in T}\sum_{j\in J} f^+_j\, v^+_{jt}$$
$$F4 = \sum_{t\in T}\sum_{j\in J} f^-_j\, v^-_{jt}$$
$$F5 = \sum_{t\in T}\sum_{m\in M} f_m\, z_{mt}$$
$$F6 = \sum_{t\in T}\sum_{(u,v)\in\mathcal{A}} c_{uv} q_{uvt}$$
$$F7 = \sum_{t\in T}\sum_{(u,v)\in\mathcal{A}'} c_{uv} q'_{uvt}$$
$$F8 = \sum_{t\in T}\sum_{(u,v)\in\mathcal{A}''} c_{uv} q''_{uvt}.$$

$$x_{mjt} \le y_{jt} \qquad \forall m\in M, j\in J, t\in T \quad (2)$$
$$\sum_{j\in J} y_{jt} \le \bar{J} \qquad \forall t\in T \quad (3)$$
$$\sum_{j\in J} x_{mjt} = 1 \qquad \forall m\in M, t\in T \quad (4)$$

$$s_{jt} = s_{j,t-1} + \sum_{j'\in S\cup J} q_{j'jt} + \sum_{m\in M}\alpha_m\, u_{mjt} - \sum_{i\in I\cup J} q_{jit} \qquad \forall j\in J, t\in T^* \quad (5)$$

$$s'_{jt} = s'_{j,t-1} + \sum_{i\in I\cup J} q'_{ijt} - \sum_{m\in M} u_{mjt} - \sum_{j'\in J} q'_{jj't} \qquad \forall j\in J, t\in T^* \quad (6)$$

$$s''_{jt} = s''_{j,t-1} + \sum_{m\in M}(1-\alpha_m)u_{mjt} - \sum_{l\in L} q'' w_{jlt} \qquad \forall j\in J, t\in T^* \quad (7)$$

$$s_{it} = s_{i,t-1} + \sum_{j\in S\cup J} q_{jit} - d_{it} \qquad \forall i\in I, t\in T^* \quad (8)$$
$$s'_{it} = s'_{i,t-1} + d'_{it} - \sum_{j\in J} q'_{ijt} \qquad \forall i\in I, t\in T^* \quad (9)$$

$$\sum_{s\in S} q_{sjt} + \sum_{i\in I}(q'_{ijt} + q_{jit}) + \sum_{l\in L} q'' w_{jlt} \le \bar{\kappa}_j\, y_{jt} \qquad \forall j\in J, t\in T \quad (10)$$

$$u_{mjt} \le \pi_m\, x_{mjt} \qquad \forall j\in J, m\in M, t\in T \quad (11)$$
$$s_{jt} \le \bar{s}_j\, y_{jt} \qquad \forall j\in J, t\in T \quad (12)$$
$$s'_{jt} \le \bar{s}'_j\, y_{jt} \qquad \forall j\in J, t\in T \quad (13)$$
$$s''_{jt} \le \bar{s}''_j\, y_{jt} \qquad \forall j\in J, t\in T \quad (14)$$
$$s_{it} \le \bar{s}_i \qquad \forall i\in I, t\in T \quad (15)$$
$$s'_{it} \le \bar{s}'_i \qquad \forall i\in I, t\in T \quad (16)$$
$$s_{jt} + s'_{jt} + s''_{jt} \le \bar{S}_j \qquad \forall j\in J, t\in T \quad (17)$$
$$s_{it} + s'_{it} \le \bar{S}_i \qquad \forall i\in I, t\in T \quad (18)$$
$$y_{jt} - y_{j,t-1} \le v^+_{jt} \qquad \forall t\in T^*, j\in J \quad (19)$$
$$y_{j,t-1} - y_{jt} \le v^-_{jt} \qquad \forall t\in T^*, j\in J \quad (20)$$
$$|x_{mjt} - x_{mj,t-1}| \le z_{mt} \qquad \forall t\in T^*, j\in J, m\in M \quad (21)$$
$$\sum_{t}^{t+\tau} z_{mt} \le 1 \qquad \forall t\in T-\tau, m\in M \quad (22)$$
$$q_{jit} \le \mathcal{M}\, y_{jt} \qquad \forall i\in I, j\in J, t\in T \quad (23)$$
$$q'_{ijt} \le \mathcal{M}'\, y_{jt} \qquad \forall i\in I, j\in J, t\in T \quad (24)$$
$$q''_{jlt} \le \mathcal{M}''\, y_{jt} \qquad \forall j\in J, l\in L, t\in T \quad (25)$$

$$y_{jt}, v^+_{jt}, v^-_{jt} \in \{0,1\} \qquad \forall t\in T, j\in J, \quad (26)$$
$$x_{mjt}, z_{mt} \in \{0,1\} \qquad \forall t\in T, j\in J, m\in M \quad (27)$$
$$u_{mjt} \ge 0 \qquad \forall t\in T, j\in J, m\in M \quad (28)$$
$$s_{jt}, s'_{jt}, s''_{jt}, s_{it}, s'_{it} \ge 0 \qquad \forall i\in I, j\in J, t\in T \quad (29)$$
$$q_{jit}, q'_{ijt}, q''_{jlt} \ge 0 \qquad \forall i\in I, j\in S\cup J, l\in L, t\in T \quad (30)$$

The objective function (1) represents the sum of the overall logistics costs. This includes production costs ($F1$) at platforms, fixed costs ($F2$) of running platforms, fixed costs ($F3$) of opening platforms, fixed costs ($F4$) of closing platforms, fixed costs ($F5$) of moving machines and transportation costs ($F6, F7$ and $F8$) for each product.

Constraints (2) state that all machines can be positioned at open platforms only. Constraints (3) model the maximum number of open platforms in each time period. Constraints (4) force each of the $|M|$ machines to be located at a single platform in every time period.

Inventory balance constraints (5) to (7) are defined for each platform $j \in J$, and all products. The inventory of the considered product at time period $t \in T$ equals the inventory at the preceding time period, plus ingoing flows, minus outgoing flows and local use of the commodity. Inventory balance constraints (8) and (9) are defined for each client $i \in I$ and the new product and used product, respectively.

Capacity constraints (10) limit the sum of ingoing and outgoing flows of platforms to the value κ_j if the platform $j \in J$ is open, and to 0 otherwise. Capacity constraints (11) limit the quantity of used product recycled at platform $j \in J$. It does not exceed the capacity π_m of machine $m \in M$ if this machine is installed at platform $j \in J$.

Capacity constraints (12) to (14) limit the value of inventory of new product, used product and waste, respectively, at any platform $j \in J$. They also impose that close platforms have no inventory. Constraints (15) and (16) indicate the maximum capacity for new product, used product and waste, respectively, at any client $i \in I$. Constraints (17) and (18) represent the overall capacity constraints at platforms and clients, respectively.

Constraints (19) and (20) are used to detect the opening or the closure of a platform. These constraints are used to assign the value of variables v_{jt}^{+} and v_{jt}^{-}, that are used in the objective function only. Similarly, constraints (21) are used to detect all machine movements, i.e. set variables z_{mt}. These constraints can be linearised as $-z_{mt} \leq x_{mjt} - x_{mj,t-1} \leq z_{mt}$. Constraints (22) state that machines cannot be moved more than once every τ time periods.

Constraints (23 to 25) establish relationships between the continuous variables q_{jit} and the decision variables y_{jt}: a flow between a platform and a client can only happen if that platform is open at that time period. The values $\mathcal{M}$, $\mathcal{M}'$ and $\mathcal{M}''$ are arbitrarily large values that can be set at the minimal value between the platform capacities ($\bar{s}_j$, $\bar{s}'_j$ and $\bar{s}''_j$) and the sum of customers demand until the end of the time horizon.

Finally, constraints (26) and (27) define the binary variables and constraints (28) to (30) are non-negativity constraints for all continuous variables.

IV. Generation of Instances

A. Parameters

We generated the value of all parameters according to information collected from a public works company in Western France. All values were randomly generated according to a uniform distribution on a given interval. Since the model is mainly used in a tactical fashion, all costs are assumed constant over the planning horizon. We assume a planning horizon of one year, divided into 12 months period. Thus, $|T| = 12$. The value of parameter τ has been set at 3.

The cost v_{jt}^{+} of opening a new platform $j \in J$ highly depends on its size and the equipment installed on it. It is randomly generated in the interval $[60\,000, 200\,000]$. The cost of closing a platform is randomly generated in the interval $[20\,000, 60\,000]$. The cost of running a platform is set at 18000 per time period. Each platform $j \in J$ has an overall capacity $\bar{S}_j$, randomly generated in the interval $[30\,000, 60\,000]$, and a maximum capacity for each product: $\bar{s}_j = \bar{s}'_j = 50\%\bar{S}_j$ and $\bar{s}''_j = 20\%\bar{S}_j$. The maximum number of open platforms $\bar{J}$ is defined as 60% of $|J|$.

The cost of moving a machine from one place to another is 10 000. The set M is assumed to comprise homogeneous machines. The capacity of a machine is π_m = 50 000 in each time period. The conversion rate is set at $\alpha = 0.7$ and the production cost c_m is set at 3 per unit processed.

We consider two types of clients: 90% are *regular client* and each has demand d_{it} in the interval $[1\,000, 2\,000]$, and 10% are *large clients* and each having demand d_{it} in the interval $[2\,000, 5\,000]$. The outgoing demand of clients is generated between 20% and 60% of their ingoing demand. The maximum inventories s_i and s'_i at clients were set at 2.5 times the value of client's demand.

Finally, the penalty factor β for transportation costs has been set at 1.8

B. Facilities

All facilities locations were generated on a 300×300 grid. A common practice is to settle temporary logistics platforms in the immediate vicinity of large clients. Thus, we first generated candidate platforms at the same location as the largest clients (with a limit of 5 platforms), then we generated all remaining platforms.

Many public works are located near populated areas and logistics platforms are generally positioned in the same areas. Thus, we generated clusters of facility location, i.e. areas in the 300×300 with a larger concentration of facility locations. First, 4 or 5 distinct clusters were randomly chosen (the choice between 4 and 5 is also random). Then, 60% of all platforms and clients were randomly generated in the chosen cluster. The remaining 40% of nodes are scattered randomly throughout the grid. Suppliers and landfills were generated randomly.

V. Computational Results

A. Summary of Computational Experiments

All experiments were run on a Xeon X5650 2.67GHz with 64GB of memory. The MILP was solved with IBM Ilog Cplex 12.6.3, with default settings and a time limitation of 5 hours.

From the mathematical model, it can be observed that the number of binary variables does not depends on the number $|I|$ of clients. This parameter influences the number of continuous variables and constraints only. The number of binary variables is mainly determined by $|J|$. Thus, we first generated instances with a small number of clients (20), 5 suppliers, 5 landfills and a growing number of candidate platforms (from 6 to 12). Then, we generated 8 additional instances with 12 candidate platforms, and from 30 to 100 clients. For each instance, we run two scenarios, with 2 and 3 machines respectively.

Tables I and II present the main characteristics of the mathematical models corresponding to all instance as well as the computational results obtained with Cplex. Columns 1 and 2 detail the number of clients and platforms considered. Columns 3 to 6 report the number of rows, columns, nonzeros and binary variables, respectively, after size reduction by the presolve process of Cplex. Columns 7 and 8 indicate the computing times and the optimality gaps.

Small-sized instances can be solved easily with Cplex, but instances with at least 40 clients and 12 time periods cannot be solved to optimality within 5 hours.

Table III reports the number of machine moves in each instance, for scenarios with 2 and 3 machines respectively. Columns 3 and 4 report the number of moves for machines 1 and 2 respectively, in scenario with 2 machines. Columns 5 to 7 report the number of moves for each machines in scenario with 3 machines. From this table, it seems that the number of machine moves tends to increase when the number of clients increases. This can be explained by capacity restrictions: when new large clients appear, the machines are repositioned in order

TABLE I. RESULTS WITH 2 MACHINES

$\|I\|$	$\|J\|$	rows	columns	nonzeros	binaries	CPU (in s)	gap
20	6	4 274	4 588	20 257	372	39	0%
20	7	5 168	5 712	25 259	430	255	0%
20	8	5 432	6 206	27 700	488	204	0%
20	9	6 111	7 202	32 143	546	362	0%
20	10	6 823	8 263	36 918	604	1 710	0%
20	11	7 277	9 102	40 728	662	830	0%
20	12	7 860	10 125	45 309	720	475	0%
30	12	9 994	12 307	55 684	720	5 900	0%
40	12	13 208	15 627	71 028	720	18 000	4.9%
50	12	15 249	17 664	80 839	720	18 000	4.5%
60	12	18 011	20 494	94 188	720	18 000	2.9%
70	12	20 571	23 105	106 517	720	18 000	2.4%
80	12	22 458	25 005	115 723	720	18 000	6.0%
90	12	25 830	28 484	131 757	720	18 000	8.0%
100	12	28 449	31 138	144 340	720	18 000	8.5%

TABLE II. RESULTS WITH 3 MACHINES

$\|I\|$	$\|J\|$	rows	columns	nonzeros	binaries	CPU (in s)	gap
20	6	4 472	4 632	20 819	456	82	0%
20	7	5 229	5 586	25 181	526	188	0%
20	8	5 720	6 307	28 582	596	58	0%
20	9	6 664	7 539	34 146	666	769	0%
20	10	7 356	8 563	38 801	736	103	0%
20	11	7 676	9 244	41 954	806	279	0%
20	12	8 270	10 272	46 564	876	6 191	0%
30	12	11 054	13 123	59 899	876	8 014	0%
40	12	13 653	15 752	72 371	876	18 000	0.7%
50	12	15 527	17 638	81 547	876	18 000	5.0%
60	12	18 236	20 404	94 538	876	18 000	2.2%
70	12	20 637	22 884	106 161	876	18 000	4.4%
80	12	23 844	26 134	121 366	876	18 000	3.0%
90	12	25 519	27 850	129 707	876	18 000	4.1%
100	12	28 531	30 943	144 090	876	18 000	3.2%

to adapt to demand variation. The number of machines also seems to influence the number of moves. With 3 machines, the geographical area is better covered, so that less moves are required.

TABLE III. MACHINE MOVES

$\|I\|$	$\|J\|$	2 Machines		3 Machines		
		m_1	m_2	m_1	m_2	m_3
20	6	3	3	2	1	2
20	7	2	2	1	1	2
20	8	2	2	1	2	2
20	9	4	3	1	0	1
20	10	2	2	2	1	2
20	11	2	2	1	1	0
20	12	2	1	2	1	2
30	12	4	2	2	1	2
40	12	4	4	1	3	2
50	12	4	3	4	4	3
60	12	3	3	3	3	3
70	12	4	4	4	3	3
80	12	3	4	4	3	3
90	12	4	4	3	2	4
100	12	3	4	4	2	3
Average		*2.97*		*2.16*		

VI. CONCLUSION AND FURTHER RESEARCH

In this paper, we introduced a supply chain network design problem in which two kinds of location decisions have to be taken: select a subset of logistics platforms and locate a set of mobile tools that must be installed on platforms. By nature, this problem is dynamic (the location decisions can vary along time). A first contribution of this work was to propose a mathematical formulation for this optimization problem. This work can be extended is several directions. First of all, the mathematical model could be reinforced by introducing valid inequalities. Intelligent pre-processing is likely to fix many decision variables at value 0. The current model can be extended to more complex networks. In particular, we consider only one type of machine, and three states of the same commodity, whereas real-life public works include several commodities. This assumption is restrictive as soon as several commodities share the same equipment or capacity.

Observing that machines move more often than platforms, another possible extension is to introduce separate time-scales for the two location decisions, as proposed in [9].

A second contribution of this work was to exhibit the practical difficulty of the optimization problem through computational experiments with a state-of-the art MILP solver. Our observations have to be confirmed by further experiments on a larger set of instances.

In further research, we plan to use metaheuristic approaches based on large neighborhoods, by iteratively removing and re-building all variables associated with given time periods, machines or platforms.

REFERENCES

[1] E. Akalı, S. Çetinkaya, and H. Üster. Network design for reverse and closed-loop supply chains: an annotated bibliography of models and solution approaches. *Networks*, 53(3):231–248, 2009.

[2] N. Aras, T. Boyacı, and V. Verter. Designing the reverse logistics network. In M. Ferguson and G. Souza, editors, *Closed Loop Supply Chains: New Developments to Improve the Sustainability of Business Practices*, pages 67–98. Taylor and Francis, CRC Press, 2010.

[3] E. D. Demaine, M. T. Hajiaghayi, H. Mahini, A. S. Sayedir, S. Oveisgharan, and M. Zadimoghaddam. Minimizing movement. *ACM Transactions on Algorithms (TALG)*, 5(3):30, 2009.

[4] R. Z. Farahani, M. Abedian, and S. Sharahi. *Dynamic Facility Location Problem*, pages 347–372. Physica-Verlag HD, Heidelberg, 2009.

[5] S.D. Jena, J.-F. Cordeau, and B. Gendron. Dynamic facility location with generalized modular capacities. *Transportation Science*, 49(3):484–499, 2015.

[6] S.D. Jena, J.-F. Cordeau, and B. Gendron. Modeling and solving a logging camp location problem. *Annals of Operations Research*, 232(1), 2015.

[7] S.D. Jena, J.-F. Cordeau, and B. Gendron. Solving a dynamic facility location problem with partial closing and reopening. *Computers & Operations Research*, 67:143 – 154, 2016.

[8] S. Nickel and F. S. da Gama. *Multi-Period Facility Location*, pages 289–310. Springer International Publishing, Cham, 2015.

[9] M. I. G. Salema, A. P. B. Póvoa, and A. Q. Novais. A strategic and tactical model for closed-loop supply chains. *OR Spectrum*, 31(3):573–599, 2009.

Vehicle Routing

A Particle Swarm Optimization for The Multi-Depots Pick-up and Delivery Problems with Time Windows and Multi-Vehicles

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Essia Ben Alaïa[*], Imen Harbaoui Dridi[*]
[*]LACCS : Université de Tunis El Manar, Ecole Nationale des Ingénieurs de Tunis, LR11ES20, Laboratoire d'Analyse Conception et Commande des Systèmes, Tunisie.
essia.alaia@enit.rnu.tn , imen.harbaoui@issatkr.rnu.tn,

Hanen Bouchriha[*], Pierre Borne[**]
[**]CRIStAL : Centre de Recherche en Informatique, Signal et Automatique de Lille. Ecole Centrale de Lille Villeneuve d'Ascq, France
hanen.bouchriha@enit.rnu.tn, pierre.borne@centralelille.fr

Abstract— In this paper, we present our approach developed for resolution of multi-vehicles, multi-depots pick-up and delivery problems with time windows (m-MDPDPTW) based on Particle Swarm Optimization (PSO). The m-MDPDPTW is an optimization vehicles routing problem which must meet requests for transport between suppliers and customers for the purpose of satisfying precedence, capacity and time constraints.
The goal is to give a satisfying solution to the m-MDPDPTW minimizing the total distance travelled. For this, we have developed an approach that uses an algorithm PSO.
The performance of the new approach is tested on data sets instances of Li & Lim's PDPTW benchmark problems. The results obtained were compared with other works and this indicated that our proposed approach gave better results.

Keywords— vehicles routing; pickup and delivery; optimization; PSO; m-MDPDPTW.

I. Introduction

The Vehicle Routing Problem (VRP) is considered one of the most studied in combinatorial optimization, and has been widely used over the last few decades. The growing interest of researchers for this type of problem is mainly due to its many real applications in logistics, for the transport of persons, goods or products.

The variant of VRP, which will be treated, is noted multi-vehicles and Multi-Depot Pickup and Delivery Problem with Time Windows: m-MDPDPTW.

In the m-MDPDPTW, we define a pair of origin (suppliers) and destination (customers) for the pickup and delivery of the goods. Each customer must be served exactly once, by a single vehicle, and should not be visited before his supplier.

Each vehicle is assigned to a single depot, which is the beginning and the end of its route, and must carry without exceeding its capacity.

For the resolution of such a complex problem, combining several variants (m-MDPDPTW), it is necessary to use meta-heuristic methods. These methods allow finding feasible solutions, in a reasonable calculation time. Our approach developed, based on PSO, must plan the itineraries of a fleet of vehicles in order to serve a set of nodes, minimizing the total distance travelled, while respecting the different constraints of the problem.

II. Literature Review

A. Pick-up and Delivery Problem with Time Windows (PDPTW)

Several studies introduce the characteristics that distinguish the PDP from the standard VRP and present a literature review of this type of problem and the methods proposed for its resolution [1]; [2]; [3] and [4]. This problem is often treated with time windows [5].

The authors in [6] have developed a genetic algorithm for the resolution of PDPTW with multiple vehicles. The considered objective is the minimization of the total travel distance and the total transport cost. This work has been extended, proposing a new GA based on the Pareto dominance method to provide a set of solutions that minimizes: total travel cost, total tardiness time and the vehicles number [7]; [8].

Another approach has been proposed by these same authors for the resolution of the multi-criteria m-PDPTW based on the aggregation method and minimizing the compromise between the total travel cost, the total tardiness time and the vehicles number [9]. This algorithm has been treated in the dynamic case in [10].

The authors in [11] propose a hybrid algorithm for the m-PDPTW. The simulated annealing algorithm is used to reduce the number of routes and the large neighborhood search algorithm (LNS) is added to reduce the Total travel cost.

Nagata and Kobayashi in 2011 present a memetic algorithm for the resolution of PDPTW with an efficient crossover operator [12].

A tabu method, a genetic algorithm based on chromosome permutation, a "split" procedure and a local search are proposed by [13] to solve a particular problem of

PDPTW. This Problem considers several requests: the delivery of medicines and medical care pharmacy at home patients and the pick-up of biological samples and unused drugs at the patients.

In [14], the authors present a solution approach based on particle swarm optimization (PSO) in which a local search is performed by variable neighborhood descent algorithm (VND). This approach was developed to solve the vehicle routing problem with simultaneous pickup and delivery (VRPSPD).

A hybrid particle swarm optimization (HPSO) is proposed by [15] to solve the multi-objective PDPTW. This algorithm adds particles neighbor information to diversify the particle swarm and use the variable neighborhood search (VNS) to enhance the convergence speed.

B. Multi-Depots Problems

Ombuki and Hanshar introduce a new genetic algorithm for optimization of the MDVRP with capacity constraints and restrictions on the traveled distance. These authors use indirect coding and adaptive mutation operator inter-depots for the affectation of customers. [16]

For the resolution of the MDVRP [17] use a stochastic approach based on the GA and the fuzzy logic to adapt the crossover and mutation rates. They consider the total traveled cost and the time spent in the objective function.

In [18] the authors present two meta-heuristics for the resolution of the MDVRP which are based on a local research and a hybrid genetic algorithm. These approaches allow to have an optimal choice of the vehicle allocation, the depots position and the determination of the first customer to visit in the route.

[19] propose an approach which is based on the combination of Genetic Algorithm (GA) with the clustering algorithm for the optimization of the m-MDPDP. The main contribution in this work is to find new depot locations in order to obtain feasible solution (routes) for the m-MDPDP.

The PSO is also used to solve the vehicle routing problem with multi-depot. [20] propose an improved PSO for the multi-depot vehicle routing problem with time windows. Another PSO algorithm is proposed for solving the practical case of multi-depot vehicle routing problem with simultaneous pickup and delivery and time window [21].

We are interested in using the PSO to solve our problem. In fact, it is based on a solutions population and the search for optimums is done by updating generations involved. However, the PSO has no evolutionary operator such as the crossover and the mutation, and in its research strategy, it gives more importance to cooperation than to competition between individuals. Contrary to most meta-heuristics, the PSO is a method that was conceived initially to solve optimization problems with continuous variables.

In the PSO algorithm, the speed is the basic mechanism that directs research in promising areas of solution space. This speed allows us to update the particles positions. In most applications, the particles positions represent the solutions of the studied problem, but in our case, the solution to m-MDPDPTW is decoded from the new position of the particle.

This paper is organized as follows. The problem formulation and the mathematical model for the m-MDPDPTW is described in Section 3. Section 4, proposes our approach developed, based on PSO, for solving a m-MDPDPTW to minimize the total travel distance. Section 5 validates the proposed approach by numerical example. Finally, the concluding remarks and further research are included in Section 6.

III. Problem Description

A. Problem Variants

Our problem is characterized by the following parameters :

$\boldsymbol{L}$: Set of depots, $\{1,...,dep\}$;

$\boldsymbol{H}$: Set of nodes (pick-up and delivery), $\{1,\ldots,n\}$;

$\boldsymbol{H^+}$: Sets of pick-up nodes, $\{1,\ldots,n/2\}$;

$\boldsymbol{H^-}$: Sets of delivery nodes, $\{1,\ldots,n/2\}$;

$\boldsymbol{H_c}$: Set of couples: delivery and pickup, $\{1,\ldots,n/2\}$;

$\boldsymbol{Ci}$: The couple (c_i, f_i) the pick-up node (f_i) with its corresponding delivery (c_i), $\forall i \in \{1,\ldots,n/2\}$;

$\boldsymbol{V_m}$: Set of available vehicles from depot m (i.e. the maximum number of vehicles from depot m that can be used for pick-up and delivery), $\{V_1,\ldots,V_{dep}\}$

$\boldsymbol{d_{ij}}$: Euclidean distance between node i and j;

$\boldsymbol{K}$: The total number of vehicles available for all the depots;

$\boldsymbol{k}$: Vehicle index in each depot m $\{1,...,V_m\}$;

$\boldsymbol{q_i}$: Goods quantity request of the node i, (if $qi < 0$ it is a delivery node else if $qi > 0$ it is a pick-up node);

$\boldsymbol{t^k_{ij}}$: Time taken by the vehicle k to travel from node i to node j;

$\boldsymbol{Q}$: The maximum capacity of a vehicle;

y_i^k : The load of vehicle k after leaving the node i;

$\boldsymbol{ET_i}$: The earliest time that node i can be visited by a vehicle;

$\boldsymbol{LT_i}$: The latest permitted time to leave node i;

$\boldsymbol{S_i}$: Service time at node i;

$\boldsymbol{A_i}$: Arrival time of the assigned vehicle at the node i;

$\boldsymbol{D_i}$: Departure time of the vehicle from the node i;

$\boldsymbol{W_i}$: Waiting time of the vehicle at node i;

The mathematical programming formulation of m-MDPDPTW requires a decision variable, which models the sequence in which vehicle visits nodes. It's defined as follows:

$$x_{ij}^{mk} \begin{cases} =1 & \text{if vehicle k originates from depot m travel along arc (i, j)} \\ =0 & \text{Sinon} \end{cases}$$

B. Mathematical Model

The function to minimize is given as follows:

$$Minimize \quad f = \sum_{m \in L} \sum_{k \in V_m} \sum_{i \in (H \cup m)} \sum_{j \in (H \cup m)} d_{ij}\, x_{ij}^{mk} \quad (1)$$

Subject to :

- Each service node is visited only once by exactly one vehicle under the following constraints :

$$\sum_{m \in L} \sum_{i \in H \cup L} \sum_{k \in V_m} x_{ij}^{mk} = 1 \qquad (\forall j \in H \cup L) \qquad (2)$$

$$\sum_{m \in L} \sum_{j \in H \cup L} \sum_{k \in V_m} x_{ij}^{mk} = 1 \qquad (\forall i \in H \cup L) \qquad (3)$$

- Each vehicle departs and returns at the same depot such as satisfies the following constraints :

$$\sum_{j \in H} x_{ij}^{mk} = \sum_{j \in H} x_{ji}^{mk} \qquad (\forall i = m \in L \text{ et } k \in V_m) \quad (4)$$

- Capacity constraints: Vehicle capacity constraints (5) and (6) impose that all the vehicles which leave and return to depot are unloaded. :

$$x_{ij}^{mk} = 1 \Rightarrow y_i^k = 0 \quad (\forall i \in L, j \in H \text{ et } k \in V_m) \quad (5)$$

$$x_{ji}^{mk} = 1 \Rightarrow y_i^k = 0 \quad (\forall i \in L, j \in H \text{ et } k \in V_m) \quad (6)$$

- For each vehicle of each depot, the load of vehicle *k* leaving node ***j*** is:

$$x_{ij}^{mk} = 1 \Rightarrow y_j^k = y_i^k + q_j (\forall i, j \in H \text{ et } k \in V_m) \quad (7)$$

- The equation (8) assures that at any time the load, on the vehicle *k*, must not exceed the vehicle capacity :

$$0 \le y_i^k \le Q \quad (\forall i \in H \text{ et } k \in V_m) \qquad (8)$$

- The arrival time A_j at a destination node ***j*** :

$$x_{ij}^{mk} = 1 \quad \Rightarrow \quad A_j = D_i + t_{ij}^k \quad (\forall k \in V_m) \qquad (9)$$

- The departure time D_i from the node ***i*** is:

$$D_i = A_i + s_i \quad (\forall i \in H) \qquad (10)$$

- The departure and service times at every depot are defined as :

$$D_i = s_i = 0 \quad (\forall i \in L) \qquad (11)$$

- The service of each customer must start within the applicable time window. If the vehicle arrives before this earliest time, it must wait.

$$A_i \ge ET_i \quad (\forall i \in H) \qquad (12)$$

$$D_i \le LT_i \quad (\forall i \in H) \qquad (13)$$

- The pick-up node (f_i) of every couple ***i*** must be visited before the corresponding delivery node (c_i) as:

$$D_{f_i} < D_{c_i} \quad (\forall i \in H_c, f_i \in H^+ \text{ et } c_i \in H^-) \quad (14)$$

IV. Particle Swarm Optimization for the m-MDPDPTW

A. General principle of the algorithm PSO proposed

The principle of the PSO is to start from an initial swarm and to apply a research strategy based on the cooperation of its *Ne* particles. The search for optimums is done by generating several generations. In every generation, a potential solution to the problem is generated, and then evaluated to record the best solutions found. The solution to our m-MDPDPTW represents the best solution found on all generations.

In the PSO algorithm, speed is the basic mechanism that directs research in the promising regions of the solutions space. This speed allows us to update the particles positions. In most applications, the particles positions represent the solutions of the problem studied, but in our case the solution to m-MDPDPTW is decoded from the new particle position.

The travel strategy of a particle *i* is illustrated in Fig.1.

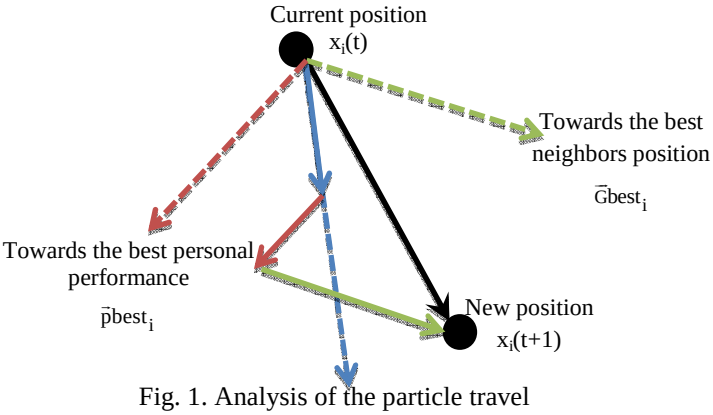


Fig. 1. Analysis of the particle travel

B. Creation of the initial swarm $P_{node/depot}$

The first step is to assign the nodes to the depots. The nodes are assigned in pairs (customer / supplier) by applying a grouping phase which ensures that each couple belongs to a single initial depot.[22]

In order to improve the solutions quality of our initial swarm $P_{node/depot}$, we have elaborated for its construction a heuristic that minimizes the total distance traveled, taking into account the different characteristics of the problem. The principle is to choose, randomly a starting node, in each depot, from those that are assigned to it. Then, to calculate the distance between the latter and all the nodes belonging to the same depot, so that it can be followed by the one closest to it. This procedure is repeated until the solutions number are equal to the particles number.

Fig.2 shows a particles example of the swarm $P_{node/depot}$ with 20 nodes and 3 depots of index 21, 22, 23.

Depot1	**21**	13	4	8	17	10	5	12	20	**21**
Depot2	**22**	7	18	3	6	19	2	14	16	**22**
Depot 3	**23**	1	15	11	9	**23**				

Fig. 2. Particles example of swarm $P_{node/depot}$

C. Decoding of the initial solutions: routing phase for the creation of the swarm $P_{node/vehicle/depot}$

In order to decode the passage of vehicles from particles of the swarm $P_{node/depot}$, we shall need to create a new type of swarm vehicle/depot ($P_{vehicle/depot}$) of size N_e.

This one will indicate to each particle of the swarm $P_{node/depot}$ the number of nodes visited by each vehicle available in each depot.

Once the swarm $P_{node/vehicle/depot}$ is generated, the correction heuristics of precedence and capacity are applied in order to have feasible solutions. [22]

However, a new heuristic correction is added to these two last ones in order to verify the belonging of each couple to the same route.

The swarm $P_{node/vehicle/depot}$ contains particles to the m-MDPDPTW, representing in each depot the visits order of node for each vehicle.

Fig.3 and Fig.4 show an example of the swarm $P_{vehicle/depot}$ and $P_{node/vehicle/depot}$.

	k_1	k_2	k_3
Depot1	6	2	0
Depot2	4	4	
Depot3	4	0	

Fig. 3. Example of swarm $P_{vehicle/depot}$

Depot 1	**21**	8	13	4	17	10	5	**21**	12	20	**21**
Depot 2	**22**	19	7	2	18	**22**	3	6	14	16	**22**
Depot 3	**23**	1	15	11	9	**23**					

Fig.4 . Example of swarm $P_{node/vehicle/depot}$

For this example we have a vehicles total number equal to 7 and 10 couples customer / supplier which are defined as follows:

{C_1 (13,8), C_2 (10,4), C_3 (20,12), C_4 (5,17), C_5 (7,19), C_6 (18,2), C_7(14,3), C_8 (16,6), C_9 (9,1), C_{10} (11,15)};

D. Initializing and updating speed and position

We consider that the particles movement is controlled by the limitation of their traveled maximum distance during an iteration. Thus, in order to escape the deviation problem of the particles from the search space, we use a technique of interval confinement. The speed of each particle is initialized by values between *0* and *n* (*n* is the nodes number to be visited).

The update of the speed and the best position is obtained by the particle ($\vec{P}best_i$: personal best) and by the swarm ($\vec{G}best$: global best), using equations (15), (16) and (17). The new position of the particle is calculated from equation (18).

$$\vec{P}best_i(t+1) = \begin{cases} \vec{x}_i(t+1), & \text{if } f(\vec{x}_i(t+1)) \text{ is better than } f(\vec{P}best_i(t)) \\ \vec{P}best_i(t), & \text{else} \end{cases} \quad (15)$$

$$\vec{G}best(t+1) = \arg\min_{\vec{P}best_i} f(\vec{P}best_i(t+1)),\ 1 \le i \le N \quad (16)$$

$$v_{i,j}(t+1) = \begin{cases} w\, v_{i,j}(t) + \\ c_1 r_{1_{i,j}}(t)(pbest_{i,j}(t) - x_{ij}(t)) + \\ c_2 r_{2_{i,j}}(t)(gbest_j(t) - x_{i,j}(t)) \end{cases} \quad (17)$$

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1) \quad (18)$$

With:

$x_{i,j}(t)$: The position of the particle *i* in dimension *j* at time *t*;

$v_{i,j}(t)$: The speed of the particle *i* in dimension *j* at time *t*;

$pbest_{i,j}(t)$: The best position obtained by the particle *i* in dimension *j* at times *t*;

$gbest_j(t)$: The best position known by the particle *i* at time *t*;

c1, *c2*: Acceleration coefficients;

r1, *r2*: Random numbers drawn uniformly in [0, 1], at each iteration *t* and for each dimension *j*;

$w\vec{v}_i$: The inertia component of the movement;

$c_1 r_1(pbest_i - \vec{x}_i)$: The cognitive component of the particle movement (moving to its best known position);

$c_2 r_2(gbest_i - \vec{x}_i)$: The social component of the particle movement (closer to the best position of its neighbors).

E. Decoding the new particle position

The elements of the new particle position after its updating do not reveal directly the nodes index in the route. A decoding phase of the new visit order is therefore necessary in order to find the solutions adapted to m-MDPDPTW.

We consider that the speed is defined as a probability vector, where the value of each element corresponds to the probability of its permutation in the route.

The decoding phase of the new particles gives us the new permutation of the nodes order in the route.

In the same route, nodes will be visited from the nearest to the furthest. If there are several nodes that are in the same position, then they will be visited according to the speed of their movement. The node with the highest speed will be visited first.

We therefore reorder in ascending order of the values of $\vec{x}_i(t+1)$. These values are subsequently replaced by the corresponding node index to build the new visit order of the nodes.

The structure of the adaptation algorithm of the PSO for the m-MDPDPTW optimization is illustrated in Fig.5.

n: nodes number to visit ; f: function to minimize ; *T*: maximum number of iterations; *dep*: number of depot;

1. Initialization: t=0; x_{max}=n; x_{min}=0;
2. Initializatio*n :* N_e : Swarm size ; c_1; c_2; r_1; r_2;
3. Generating N_e initial particles of $P_{node/depot}$
4. Application of the routing phase for the creation of initial solutions $P_{node/vehicle/depot}$;
5. Application of the corrections heuristics of precedence, capacity and belonging of each couple to the same route ;
6. Evaluate the N_e initial particles of $P_{node/vehicle/depot}$ (1);
7. Initialize the speed of each particle by random values between [0; n];
8. Initialize the position of each particle $\vec{x}(0) = P_{node/depot}$
9. Initialize $\vec{P}best_i(0) = P_{node/vehicle/depot}[i]$;
10. Initialize the best solution found by the swarm $\vec{G}best(0) = \arg\min_{\vec{P}best_i} f(\vec{P}best_i(0));$
11. While (t<T) do ;
12. For i=*1* to N_e do ;
13. Speed Update (17);
14. Update of the new position (18);
15. Verify the new particle does not leave the search space:

    $\text{if } (\vec{x}_i(t+1) > x_{max}) \text{ then } \vec{x}_i(t+1) = x_{max};$

    $\text{else if } (\vec{x}_i(t+1) < x_{min}) \text{ then } \vec{x}_i(t+1) = x_{min};$
16. Decoding the new particles position;
17. Generating new solutions: Routing phase;
18. Apply corrections heuristics;
19. Evaluate the N_e particles of $P_{node/vehicle/depot}$ (1);
20. Save the best result found by the particle;

For i=*1* to N_e do ;

$\text{if } f(P_{node/vehicle/depot}(i)) \le f(pbest_i(t)) \text{ then}$

$Pbest_i(t+1) = f(x(t+1))$

$\text{else } Pbest_i(t+1) = Pbest_i(t)$

End For;

21. Update the best solution found by the sawarm $\vec{G}best(t+1) = \arg\min_{\vec{P}best_i} f(\vec{P}best_i(t+1));$
22. End For;
23. End While;

Save the best solution found on all generation;

Fig.5 . PSO algorithm for the m-MDPDPTW optimization

To find the optimal solution, we browse the swarm $P_{node/vehicle/depot}$ particle by particle, to determine that which minimizes our objective function representing the total distance travelled (1).

V. Computational Results

For the parameters values of the PSO algorithm and after several tests, we choose the following values: c_1 = 0: 2, r_1 = 0: 3, c_2 = 0: 2, r_2 = 0: 5, w varies between [0.8 ; 0.5]. The new position of the particle is calculated from (18).

The simulation results are obtained using benchmark problem instances generated by Li and Lim. [23]

Table I gives the location of the different depots considered in this simulation for each type of Li and Lim problems and are available in the time window [0; 600].

The method of insertion new depots has been proposed in our work [19]. This method solves our m-MDPDPTW with an optimal choice of the depots number.

TABLE I. Location of the different depots

Instances	Coordinates of the 1st depot	Coordinates of the 2nd depot
LC	(40 , 50)	(34 , 32)
LR	(35 , 35)	(60 , 40)
LRC	(40 , 50)	(65 , 30)

To compare our results we adapted to our problem the discreete technique of the PSO de3veloped by [24].

Table II shows the results of our simulation using the parameters of the problem LC101, LC102, LC201, LC202, LR101, LR102, LR201, LR202, LRC101, LRC102, LRC201, LRC202.

TABLE II. Comparison of the results of the two PSO approaches using the Li and Lim instances to two depots

Instances	Best Distance Our PSO Approach	Best Distance Discreet PSO	Evolution rate (%)
LC101	1839,3962	1940,205	5.196
LC102	1606,801	1982,140	18.936
LC201	947,914	958,613	1.116
LC202	1472,292	1563,263	5.819
LR101	2137,17	2054,441	--
LR102	2128,639	1999,699	--
LR201	2060,405	2077,726	0.834
LR202	2497,547	2568,205	2.751
LRC101	2513,376	2593,103	3.075
LRC102	2567,438	2530,135	--
LRC201	2937,202	2972,344	1.182
LRC202	2829,999	2859,729	1.04

We observe that the values of the total distance traveled given by our approach based on the PSO are better than those given by the discreete PSO.

This last only gave a low improvement of the fitness for instances LR101, LR102 and LRC 102.

By calculating the percentage of variation between the two approaches, also known as the evolution rate, we observe that our approach represents a reduction of distance which can go up to 18%. These results justify the efficiency of our approach developed.

VI. Conclusions

In this paper, we have presented our approach to solve the m-MDPDPTW, based on Particle Swarm Optimization for the minimization of the total distance traveled.

We proposed a brief literature review on the PDPTW and the Multi-Depots Problems. The mathematical formulation of our problem was subsequently presented. Then, we have detailed the use of PSO algorithm to determine the solution which minimizes our objective function. Simulation was presented in a last part by using benchmark's data.

For our future work, we propose to study the multi-objective optimization by comparing the genetic algorithms and the Particle Swarm Optimization.

References

[1] Savelsbergh, M. W. P. and SOL, M. (1995). The general pickup and delivery problem. Transportation Science, 29(1):17–29.

[2] Berbeglia, G., Cordeau, J. F., Gribkovskaia, I. and Laporte, G. (2007). Static pickup and delivery problems : a classification scheme and survey. Top, 15(1):1–31.

[3] Parragh, S. N.; Doerner, K.. H. R. (2008). A survey on pickup and delivery problems. Transportation between customers and depot. Journal fur Betriebswirtschaft, 58(1):21–51.

[4] Parragh, S. N., Doerner, K. and Hartl, R. F. (2010). Variable neighborhood search for the dial-a-ride problem. Computers and Operations Research, 37(6):1129–1138.

[5] Kumar, S. N. and Panneerselvam, R. (2012). A survey on the vehicle routing problem and its variants. Intelligent Information Management, 4(3):66–74.

[6] Harbaoui, D. I., Kammarti, R., Borne, P. and Ksouri, M. (2008). Un algorithme genetique pour le probleme de ramassage et de livraison avec fenetres de temps a plusieurs vehicules. In CIFA, Roumanie.

[7] Harbaoui, D. I., Kammarti, R., Borne, P. and Ksouri, M. (2009a). Academic approche multicritere pour le probleme de ramassage et de livraison avec fenetres de temps a plusieurs vehicules. In LT IEEE, Tunisie.

[8] Harbaoui, D. I., Kammarti, R., Borne, P. and Ksouri, M. (2009b). Genetic algorithm for mulicriteria optimization of a multi-pickup and delivery problem with time windows. In IFAC, Russia, pages 1521–1526.

[9] Harbaoui, D. I., Kammarti, R., Ksouri, M. and Borne, P. (2011). Multi-objective optimization for the mpdptw : Aggregation methode with use of genetic algorithm and lower bounds. International Journal of Computers, Communications and Control (IJCCC), 6(2):246–257.

[10] I. Harbaoui Dridi, E. Ben Alaia and P. Borne. Heuristic Approach for The Optimization of The Dynamic Multi-Vehicle Pickup and Delivery Problem with Time Windows. IEEE International Conference on Industrial Engineering and Systems Management IESM 2015, 21-23 October, Seville, Spain, 2015.

[11] Bent, R. and Hentenryck, P. V. (2006). A two-stage hybrid algorithm for pickup and delivery vehicle routing problems with time windows. Computers and Biomedical Research Operations Research.

[12] Nagata, Y. and Kobayashi, S. (2011). A memetic algorithm for the pickup and delivery problem with time windows using selective route exchange crossover, volume 6238 de LNCS, pages 536–545. Springer Berlin Heidelberg.

[13] Liu, R., Xie, X., Augusto, V. et Rodriguez, C. (2013). Heuristic algorithms fora vehicle routing problem with simultaneous delivery and pickup andtime windows in home health care. European Journal of Operational Research, 230(3):475–486.

[14] Fatma Pinar Goksal, Fulya Aaltiparmak and Ismail Karaoglan. A Hybrid Particle Swarm Optimization for Vehicle Routing Problem with Simultaneous Pickup and Delivery. Computers & Industrial Engineering 2012.

[15] Shuilong Zou, Jin L and Xueqian Li. A Hybrid Particle Swarm Optimization Algorithm for Multi-Objective Pickup and Delivery Problem with Time Windows. Journal of Computers, vol. 8, no. 10, october 2013.

[16] Ombuki, B. B. and Hansharin, F. (2009). Using genetic algorithms for multi depot vehicle routing, volume 161 de Studies in Computational Intelligence, pages 77–99. Springer Berlin Heidelberg.

[17] Lau, H. C.W., Chan, T. M., Tsui,W. T. and Pang.,W. K. (2010). Application of genetic algorithms to solve the multidepot vehicle routing problem. IEEE Transactions on Automation Science and Engineering, 7(2):383–392.

[18] Vidal, T., Crainic, T. G., Gendreau, M. and Prins, C. (2014). Implicit depot assignments and rotations in vehicle routing heuristics. European Journal of Operational Research, 237(1):15–28.

[19] E. Ben Alaia, I. Harbaoui Dridi, H. Bouchriha and P. Borne. Insertion of new depot locations for the optimization of multi-vehicles Multi-Depots Pickup and Delivery Problems using Genetic Algorithm. IEEE International Conference on Industrial Engineering and Systems Management IESM 2015, Seville, Spain, 2015.

[20] Lei Wen and Fanhua Meng. An Improved PSO for the Multi-Depot Vehicle Routing Problem with Time Windows. Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008. PACIIA '08.

[21] P. Sombuntham and V. Kachitvichayanukul. A Particle Swarm Optimization Algorithm for Multi-depot Vehicle Routing problem with Pickup and Delivery Requests. The International MultiConference of Engineers and Computer Scientists, IMECS 2010.

[22] E. Ben Alaia, I. Harbaoui Dridi, H. Bouchriha and P. Borne. Genetic Algorithm for Multi-Criteria Optimization of Multi-Depots Pickup and Delivery Problem with Time Windows and multi vehicles. Acta Polytechnica Hungarica, vol. 12, no. 8, pp. 155-174, 2015.

[23] Li H. and Lim A. A metaheuristic for the pickup and delivery problem with time windows. In IEEE International Conference on Tools with Artificial Intelligence, volume 13, pages 160– 167, 2001.

[24] Osadciw, L. et Veeramachaneni, K. (2009). Particle Swarm Optimization, chapitre Particle Swarms for Continuous, Binary, and Discrete Search Spaces, pages 451–460. In-Tech.

A Multi objective optimization approach for the m-MDPDPTW

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Essia Ben Alaïa [**], Hajer Ben Mahmoud [*]
[*]LISI : Université de Carthage, Institut National des Sciences Appliquées et de Technologie, Laboratoire d'Informatique pour les Systèmes Industriels, Tunisia.

essia.benalaia@enit.rnu.tn, hajerbenmahmoud@hotmail.com

Imen Harbaoui Dridi[**], Taieb Ben Romthane[*]
[**]LACCS : Université de Tunis El Manar, Ecole Nationale des Ingénieurs de Tunis, Laboratoire d'Analyse Conception et Commande des Systèmes, Tunisia.

imen.harbaoui@issatkr.rnu.tn, tbr.tim@panet.tn

Abstract— In this paper, we consider multi-vehicles, Multi-Depot, Pickup and Delivery Problem with Time Windows (m-MDPDPTW). The principle of the m-MDPDPTW is to serve a set of geographically dispersed customers (pickup and delivery nodes) with known demands by minimizing cost vehicle routes originating and terminating at a depot.

Our objective is to find a set of best vehicle routes in order to satisfy different transportation requests minimizing the sum of travel distance, the vehicle fleet and the total tardiness time. All requests must be performed without violating any of the instance specific constraints: the precedence, capacity constraints and the customer time window stipulated at each location.

In this context, we present an improving multi objective optimization approach based on Genetic Algorithm (GA), Pareto dominance concept and a decision support tool based on Analytic Hierarchy Process (AHP) to solve m-MDPDPTW.A benchmark dataset is used to validate the performance of proposed algorithm. Comparing with prior works, promising results indicate that the proposed approach may hold high potentials for generating powerful tool for solving m-MDPDPTW.

Keywords— multi-depots; pick-up and delivery problem with time window; genetic algorithm; Pareto dominance concept; multi-objective optimization, Analytic Hierarchy Process, decision making process.

I. Introduction

Recently, the studies on the different variants of the well-known Vehicle Routing Problem (VRP) increase rapidly in order to develop many real applications in logistics and retail distribution. In this paper, we study a multi-vehicles, Multi-Depot, Pickup and Delivery Problem with Time Windows, which we note m-MDPDPTW. Even for a small problem size, this m-MDPDPTW is clearly NP Hard (Non-deterministic Polynomial-time Hard) since it is a combination of several variants of VRP, which falls into the category of NP Hard problems [1]. Our objective is to find a set of best vehicle routes in order to satisfy different transportation requests minimizing multi objective: the sum of travel distance, the number of vehicles and the total tardiness time.

We consider a homogeneous fleet of vehicles, with limited capacity, based on several depots. Each load must be transported by one vehicle between a set of origin (pick-up nodes) and destination points (delivery nodes). A time window is associated with each pick-up and delivery node defining an interval where in the nodes have to be supplied. Furthermore, we assume that the initial depot is unloaded (does not contain goods) and each vehicle route must starts and ends at this depot. Our best solutions (a set of routes) are feasible if the different constraints of the problem are respected which are: (1) the capacity constraints impose that at any duration in the route, the load on the vehicle must not exceed its maximum capacity and all the vehicles which leave and return to depot are unloaded; (2) the precedence constraints must respect that for each request, the origin (the pickup node) must precede the destination (the delivery node); (3) in the time constraints, we use a soft time window: every vehicle arriving before the lower limit of the time window causes additional waiting time on the route. A later service, after the upper bound of the time window of node, does not affect the feasibility of the solution, but causes additional tardiness time on the route which is penalized by adding a value to the objective function.

Generally, the exact approaches to solve this type of problems are inefficient [2]. However, to find good solutions in a reasonable computational time, we focused on meta-heuristic techniques especially the Genetic Algorithm (GA) approach, which is considered as a reasonable choice in finding solutions for our large scale instances. In the past several years, a large variety of multi objective algorithms including efficient heuristics and meta-heuristics have already been designed to solve a different combination of some variants of VRP [3]. Among the modern optimization techniques used mostly to solve these very complicated problems, we find the meta-heuristics approach such as GA which is based upon the evolution process. By generating various generations of feasible solutions, the GA tries to move toward the best solution point of the problem.

In literature, several recent multi-objective algorithms were inspired either by Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) [4, 5], or by Strength Pareto Evolutionary Algorithm 2 (SPEA2) [6]. Harbaoui Dridi et al. [7] proposes to solve the pickup and delivery problem with time windows using a GA with aggregated methods, which

bring the multi-objective problem into a single objective problem. A solution to the VRP problem using heuristics methods is proposed by Nagy et al. [8] to solve simultaneous VRP with pickup and delivery for single and multiple depots. A new multi-objective adaptive GA with multiple Pareto fronts called "aGAME" using multi-Pareto-ranking was proposed by Rudolph et al. in [9]. Pisacane finds in [10] the set of efficient solutions for approximating the Pareto Fronts to solve the multi objective Multi vehicle Pickup and Delivery problem with time Windows. Zinflou et al. present in [11] a new algorithm called GISMOO, which hybridizes GA and artificial immune systems. Ombuki-Berman and Hanshar use GA in [12] to solve a multi depot vehicle routing problem. Wang et al. [13] design a new GA for MDVRPTW with multi-type vehicles limits. In [14], the authors use the combination of GA search and fuzzy logic techniques to adjust the crossover and mutation rate for solving VRP with multiple depots, customers and products. For detailed information about other multi objective optimization using evolutionary algorithms readers are encouraged to refer to [15, 16].

In this paper, an MOO approach based on GA with Pareto approach and AHP method is proposed to solve m-MDPDPTW. This approach gives a best vehicle routes in order to satisfy the preference of DM and several transportation requests minimizing different objective and respecting various constraints. The framework of this approach is described in Section 2. The proposed MOO approach is presented in Section 3. In order to illustrate the application of the suggested approach, a case study and experimental results are exposed in Section 4. The last section is devoted to the discussions and to the conclusion.

II. Motivations and Problem Framework

The Multi-Objectives Optimization Problem (MOO) aims to optimize simultaneously several contradictory criteria. Therefore, the selection of one solution over another requires a good knowledge of the Decision-Maker's (DM) preferences and the use of a decision-making tool. An approach to MOO solving consists in searching a set of the satisfactory solutions for which a balance between the objectives to be optimized, is achieved. The latter is used to evaluate the best route by: (1) Affect couples (customer, supplier) in the depots and the routes in the vehicles, (2) Identifying and evaluating attainable routes, (3) Determining the set of best routes for the m-MDPDPTW by minimizing three criteria: the total drive distance, the number of vehicles used and the total tardiness time. Once the set of best routes is determined, the information on the solutions that are required to be affined is transferred to the decision support tool. At this stage, our approach is applied to help the DM in selecting the best solution, which is considered a Multi-Criteria Decision Making (MCDM) problem. The approach framework is represented by Fig. 1.

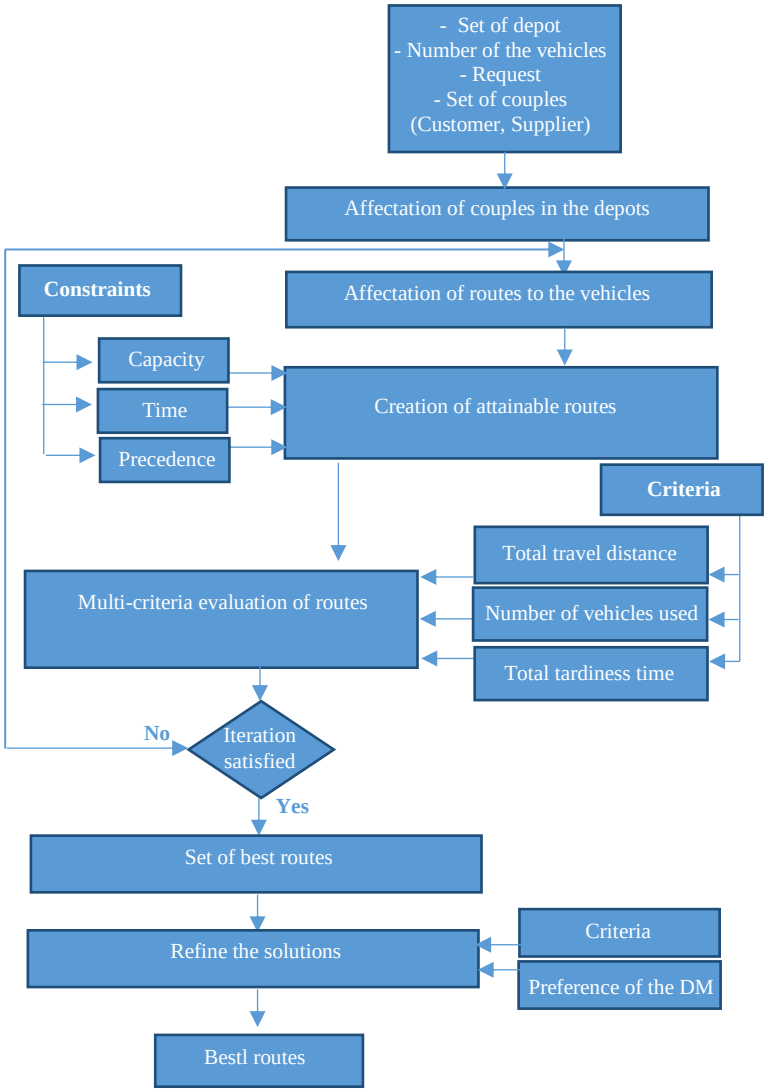


Fig. 1. The framework of the proposed MOO approach

III. Multi-Objective Optimization (MOO) approach for the m-MDPDPTW

A. Mathematical Model

Our problem takes into account the above variables and parameters:

L : Set of depots $\{1,...,dep\}$;

H : Set of nodes (pick-up and delivery) $\{1,2,...,n\}$;

H^+ : Sets of pick-up nodes;

H^- : Sets of delivery nodes;

H_c : Set of couples: delivery and pickup;

C_i : The couple (c_i, f_i): the pick-up node (f_i) with its corresponding delivery node (c_i), $\forall i \in \{1,...,(n/2)\}$;

V_m : Set of available vehicles from depot m $\{V_1,...,V_{dep}\}$;

d_{ij} : Euclidean distance between node i and j;

K : The total number of vehicles available for all the depots;

q_i : Goods quantity request of the node i, (if $q_i < 0$ it is a delivery node else if $q_i < 0$ it is a pick-up node);

t_{ij}^k : Time taken by the vehicle k to travel from node i to j;

Q : The maximum capacity of a vehicle;
y_i^k : The load of vehicle k after leaving the node i;
ET_i : The earliest time to serve node i;
LT_i : The latest permitted time to leave node i;
S_i : Service time at node i;
A_i : Arrival time of the assigned vehicle at the node i;
D_i : Departure time of the vehicle from the node i;
W_i : Waiting time of the vehicle at node i;
T_i : Tardiness time of the vehicle at node i;

A decision variable is defined as follows:

$$x_{ij}^{mk}\begin{cases}=1 & \text{if vehicle k originates from depot m travel along arc (i, j)}\\ =0 & \text{Sinon}\end{cases}$$

The m-MDPDPTW considered in this study aims to minimize total travel distance, total tardiness time and number of vehicles used. The objective function is formulated as:

$$f=\begin{cases}\sum_{m\in L}\sum_{k\in V_m}\sum_{i\in(H\cup m)}\sum_{j\in(H\cup m)} d_{ij}x_{ij}^{mk},\\ \sum_{i\in(H\cup L)}\max(0,D_i-LT_i),\\ K\end{cases} \quad (1)$$

Subject to:

$$\sum_{m\in L}\sum_{i\in H\cup L}\sum_{k\in V_m} x_{ij}^{mk}=1 \qquad (\forall j\in H\cup L) \quad (2)$$

$$\sum_{m\in L}\sum_{j\in H\cup L}\sum_{k\in V_m} x_{ij}^{mk}=1 \qquad (\forall i\in H\cup L) \quad (3)$$

$$\sum_{j\in H} x_{ij}^{mk}=\sum_{j\in H} x_{ji}^{mk} \quad (\forall i=m\in L \text{ and } k\in V_m) \quad (4)$$

$$x_{ij}^{mk}=1\Rightarrow y_i^k=0 \ (\forall i\in L,\ j\in H,\ k\in V_m \text{ and } m\in L) \quad (5)$$

$$x_{ji}^{mk}=1\Rightarrow y_i^k=0 \quad (\forall i,m\in L,\ j\in H \text{ and } k\in V_m) \quad (6)$$

$$x_{ij}^{mk}=1\Rightarrow y_j^k=y_i^k+q_j \ (\forall i,j\in H,m\in L \text{ and } k\in V_m) \quad (7)$$

$$0<y_i^k\le Q \qquad (\forall i\in H \text{ and } k\in V_m) \quad (8)$$

$$x_{ij}^{mk}=1\Rightarrow A_j=D_i+t_{ij}^k \ (\forall i,j\in H,m\in L \text{ and } k\in V_m) \quad (9)$$

$$D_i=A_i+S_i \qquad (\forall i\in H) \quad (10)$$

$$D_i=S_i=0 \qquad (\forall i\in L) \quad (11)$$

$$ET_i>A_i \ \Rightarrow \ W_i=ET_i-A_i \qquad (\forall i\in H) \quad (12)$$

$$T_i=\max(0,D_i-LT_i) \qquad (\forall i\in H) \quad (13)$$

$$D_{f_i}<D_{c_i} \quad (\forall i\in H_c,\ f_i\in H^+ \text{ and } c_i\in H^-) \quad (14)$$

In this formulation, constraints (2) and (3) ensure that each request is served once by the same vehicle, while constraint (4) guarantees that each vehicle starts and ends its route at the same depot. Constraints (5) and (6) impose that all the vehicles which leave and return to depot are unloaded. For each vehicle of each depot, the load of vehicle k leaving node i to j is defined in (7), while capacity constraint (8) guarantees that at any time the load, on the vehicle k, must not exceed the vehicle capacity. Each node i has time interval [ETi, LTi] in which service at location i must take place. These time windows define in constraints (9) to (11) the arrival time, the departure time, service times at every depot, respectively. If the vehicle arrives at customer before the beginning of the applicable service time, a waiting time is calculated according the equation (12). And if the departure time from a node is later than its latest time of service we calculate a tardiness time by equation (13). Finally, the precedence constraints (14) ensure that the pick-up node *(f_i)* of every couple i must be visited before the corresponding delivery node *(c_i)*.

B. GA with Pareto approach and AHP for the m-MDPTW

The proposed MOO model is developed within nine phases and is described by the flowchart shown in Fig. 2.

Phase 1: The depot-grouping phase

The process rules of depot grouping phase is explained in details in the following reference [17]. At the end of this strategy, each couple (pick-up node and its associated delivery node) are assigned to the nearest depot.

Phase 2: Generation of an initial solution

A group of initials solutions is randomly generated, constructing the first population named $P_{couple/depot}$ containing N individuals [18]. The chromosomes of the solution are encoded using path representation. Our initial solution is a sequence of genes encoded in integers. Each gene identifies a node and the order of the genes gives for each vehicle the order in which these couples will be visited. Each coded individual contains both customers and suppliers. We chose to indicate the start and the end of each path by the depot indices. The index " 0 " is not used throughout the work.

Phase 3: Use of GA operators

The genetic operators are used to create new population $P_{couple/depot}$ containing 2N individuals. Its first part represents an exact copy of the N individuals of $P_{couple/depot}$ created in phase 2, while the remaining 50 percent are created by applying GA operators on the first part. In our case, we select two parent chromosomes from population of step 2 by using tournament selection. For recombination, it is difficult to determine the most effective crossover method in advance. Therefore, we designed a crossover operator suitable for the permutation coding, considering a multi depots individuals. It is a combination between the one point crossover and the uniform crossover. The principle is to generate randomly a binary mask. Each bit of this mask corresponds to one depot in the chromosome. The bit 1 indicates that the depot will be copied: the genes of the first parent will be copied into the second child and those of the second parent will be copied into the first child. The bit 0 means that the depot will be crossed: we randomly choose a crossing position for both parents. The part before this position will be copied into both child. The next part will be completed in the first child by the genes that were not inherited from the first parent but rearranged according to their order of appearance in the second parent. We apply the same procedure to complete the second child using the order of appearance in the first parent. This procedure is repeated until the crossover rate chosen has

been reached equal to 0.8. For diversification, we apply swap mutation with a rate equal to 0.2.

Phase 4: The routing phase: population $P_{node/vehicle/depot}$

For each depot, the number of vehicles used and the order of delivery and pick-up within each route are specified by the population named $P_{node/vehicle/depot}$. To construct this population type we should follow three steps: knowing the number of vehicles available in each depot, we start by creating 2N individuals of a new population that indicates the number of nodes visited by each vehicle, named $P_{vehicle/depot}$. After, we affect visited couples to vehicles by coding each individual of the population $P_{couple/depot}$ created in phase 3 by an individual of the population $P_{vehicle/depot}$. This new population, named $P_{couple/vehicle/depot}$, verifies that each couple belongs exactly to same route. To find, for each depot the order in which all pick-up and delivery nodes are visited, we develop a heuristic algorithm. Its principle is to choose randomly, for each route, a starting node from the assigned couples. Then, we follow it by the closest node. Our heuristic minimizes the total distance traveled for each individual of $P_{node/vehicle/depot}$.

Phase 5: Heuristics for creation of feasible solutions

Each individual of population $P_{node/vehicle/depot}$ must respect different constraints. The precedence constraint ensures that each delivery point, on the same route, is not visited before its supplier. The capacity correction constraint ensures that the total load of the vehicle must be smaller than or equal to the maximum capacity of the vehicle. The heuristics algorithms for precedence and capacity corrections procedures, to transform each individual into feasible solution, are explained in details in [19].

Phase 6: The Pareto optimality concept for the multi-criteria evaluation

Our m-MDPDPTW problem has more than one objective to minimize which are: f_1= the total travel distance; f_2= the total tardiness time and f_3= the number of vehicles. To differentiate individuals that are clearly superior to others, we must find a compromise between these 3 criteria. We seek to determine the Pareto front that contains a subset of the Pareto best solution. These solutions are named non-dominated solutions and represent the best ones of our problem. According the Pareto concept, in our case an individual $I_1 \in P_{Pareto}$ dominates another individual $I_2 \in P_{Pareto}$ if:

$$f_1(I_1) \le f_1(I_2) \text{ and } f_2(I_2) \le f_3(I_1) \text{ and } f_3(I_1) \le f_3(I_2) \quad (15)$$

Two individuals (I_1, I_2) are non-dominated by one over the other if they check the following system:

$$\begin{cases} f_1(I_1) > f_1(I_2),\ f_2(I_1) \le f_2(I_2),\ f_3(I_1) \le f_3(I_2) \\ f_1(I_1) \le f_1(I_2),\ f_2(I_1) > f_2(I_2),\ f_3(I_1) \le f_3(I_2) \\ f_1(I_1) \le f_1(I_2),\ f_2(I_1) \le f_2(I_2),\ f_3(I_1) > f_3(I_2) \\ f_1(I_1) > f_1(I_2),\ f_2(I_1) > f_2(I_2),\ f_3(I_1) \le f_3(I_2) \\ f_1(I_1) > f_1(I_2),\ f_2(I_1) \le f_2(I_2),\ f_3(I_1) > f_3(I_2) \\ f_1(I_1) \le f_1(I_2),\ f_2(I_1) > f_2(I_2),\ f_3(I_1) > f_3(I_2) \end{cases} \quad (16)$$

We calculate for every individual of $P_{node/vehicle/depot}$ fitness values (f_1, f_2 and f_3), representing the Pareto population noted P_{Pareto}. After the Pareto dominance strategy is used to attribute a rank to each individual of $P_{node/vehicle/depot}$: The individuals which are non-dominated by all the population will have the first rank. These ranks stratify the population into preference categories.

Phase 7: Update a set of best routes

The individuals with Pareto rank = 1 are the best in the current population. At each iteration, these solutions are recorded and an update is made to keep the non-dominated solutions compared to the previous generation.

Phase 8: Update initial population by the best N solutions

We use the elitism strategy that is for each generation, the best N non dominated solutions of the old population are copied in the new initial population. Lower ranks are preferable, and the individuals within rank 1 are the best in the current population [20].

Phase 9: Use AHP method to select one solution according to the choice of the decision-maker

When the number of generations is reached, the set of the best solution (created in phase 7) is found. Following the generation of this solution set, we have developed a decision support tool in order to refine the multi criteria resolution of m-MDPDPTW and to select the single solution satisfying the preferences of Decision Maker (DM). Introduced by Thomas Saaty [21], AHP is one of the most widely used multiple criteria decision-making tools. The use of AHP in this paper is motivated by the ability of this method to organize the problem of selecting the best route into a hierarchical structure of objective, criteria, and alternatives. Moreover, using a set of pairwise comparisons, the AHP enables obtaining the important weights of the criteria, and the relative performance scores of the alternatives related to each individual criterion. Furthermore, it provides a mechanism for checking and improving the evaluations consistency, which set AHP apart from other multi-criteria methods [22]. The application of this method for selecting the best route is done within four steps: (1) Define the problem and determine the overall goal: select the best route satisfying the preferences of DM; (2) Structure the problem in a hierarchy of different levels constituting the goal, the criteria C_j ($j=1,..,p$), the sub-criteria and the alternatives S_i ($i=1,\ldots,n$) are presented in Table I; (3) Weigh the criteria, sub criteria and alternatives as a function of their importance for the corresponding element of the higher level by developing pairwise comparison matrices for each element using Saaty 9-point scale; (4) Calculate the Weighted Arithmetic Mean (WAM) for each decision alternative. Choose the one with the highest score. The obtained decision matrix is represented in Table I, where w_j is the weight of criterion j with $w_j \ge 0$, the sum of the weights is equal to *1* and X_i is defined as follows:

$$X_i = \sum_{j=1}^{p} (x_{ij} * w_j) \quad (16)$$

TABLE I. The decision matrix

Criteria weights	c_1 (w_1)	c_2 (w_2)	... ...	c_p (w_p)	
Alternatives					*WAM*
S1	x_{11}	x_{12}	...	x_{1p}	X_1
S2	x_{21}	x_{22}	...	x_{2p}	*X2*
.			.		.
.			.		.
.			.		.
Sn	x_{n1}	x_{n2}	...	x_{np}	X_n

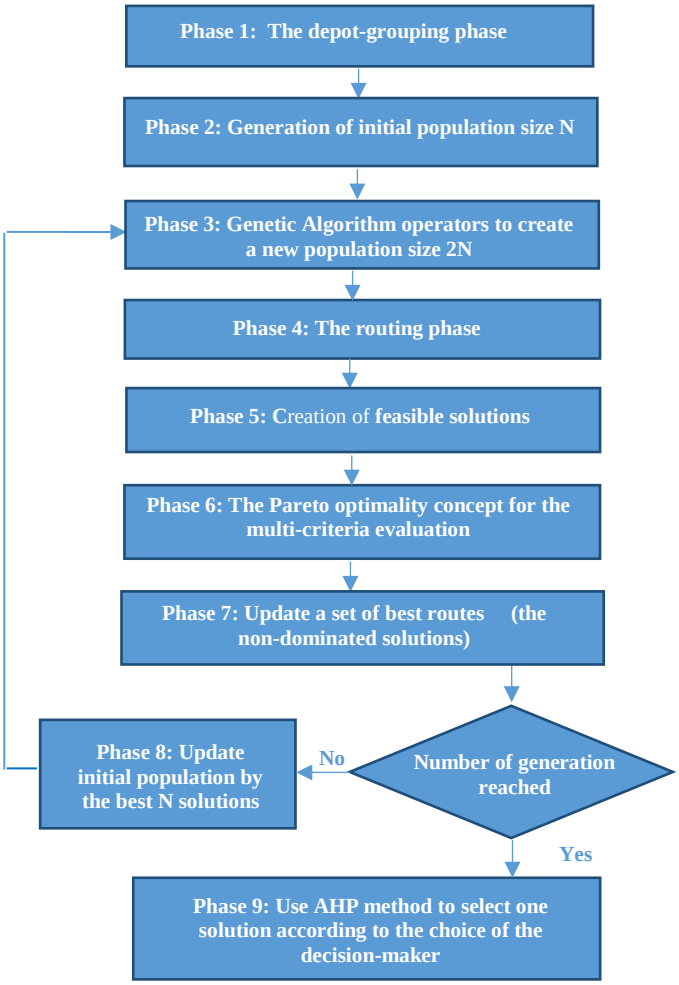


Fig. 2. The flowchart of the proposed MOO approach

IV. Case study

The efficiency and feasibility of the proposed MOO approach would be demonstrated by the following numerical example in this section. Our algorithm described in section 3.2 is coded in C language using Microsoft Visual Studio 2010 and executed on personal computer Intel Core i7, CPU 2.50 GHz and 6.00 GB memory (RAM) under the operating system Windows 8 Professional. We use a Li & Lim's [23] benchmark instances as testing data that consider just one depot for the PDPTW problem. To adapt our m-MDPDPTW problem to this benchmark test instances, we add a multi depots locations [24]. The experimental problem considered the LC101 instances which have a partially clustered pickup delivery node. In this study, we have 102 customers with 25 vehicles having as 20 total capacity based at 3 depots. The parameters details of added depots in this paper can be seen in Table II.

TABLE II. Details of added locations depots

Indices of Depots	*Coordinate X*	*Coordinate Y*
103	40	50
104	0	40
105	88	35

The result of MOO approach using AG and Pareto strategy, minimizing three objectives is presented in Table III.

TABLE III. The Pareto best solution of the LC101 instances

Solutions	f_1	f_2	f_3
S1	1924.663	71390.07	11
S2	1713.649	72114.192	9
S3	1837.556	107388.461	6
S4	1903.189	77750.510	9
S5	1895.835	79811.345	8
S6	1889.771	70861.223	11

In Table III f_1, f_2 and f_3 denote the total travel distance, the total tardiness time and the number of vehicles respectively. From this table, it shows that by using MOO approach there are 6 better solutions in LC101 instances. The best result is found after 50 runs for a population size equal to 200.

We note that the best solution for satisfying the objectivef1 is found by S2. For minimize the second objective, the best solution is found by S6. On the other hand, S3 is the best solution found to minimize the number of vehicles. Following this diversification of solutions, we cannot make a decision if we want to consider three objectives at the same time. For this reason, we will have to apply the preferences of the decision-makers by the AHP method. According to the choice of the decision-maker, the priorities of the criteria are as follows: The distance traveled is a *little more important* than the delay, the delay is *much more important* than the number of vehicles used, the number of vehicles used is *less important* than the distance traveled.

The score for each action will then be obtained for each criterion. We can then assign an overall score to each and it remains to store these notes to obtain a final ranking of the solutions. The results obtained for the route evaluation using the AHP method are shown in Table IV.

TABLE IV. Global priority vector

Solutions	Distance traveled (0.586)	Delay (0.332)	Number of vehicles (0.145)	Global priority vector
S1	0.188*0.586	1.490*0.332	0.444*0.145	0.669
S2	3.237*0.586	1.075*0.332	0.789*0.145	**2.368**
S3	1.241*0.586	0.125*0.332	2.169*0.145	1.083
S4	0.339*0.586	0.577*0.332	0.789*0.145	0.504
S5	0.431*0.586	0.370*0.332	1.363*0.145	0.573
S6	0.529*0.586	2.358*0.332	0.444*0.145	1.157

Note that the highest note *2.368*, found by the AHP method, is for the second solution *S2* which corresponds to the preferences of the decision maker.

V. Discussions and conclusion

The paper presents a MOO approach based on GA with Pareto approach and AHP method for the m-MDPTW. The GA and Pareto approach give a set of best vehicle routes in order to satisfy different transportation requests minimizing different objectives and respecting the precedence, capacity and time constraints. Therefore, the AHP method provides the criteria weights, the partial scores and the overall scores for the select the one best route satisfying the preference of DM. The application of the proposed MOO approach in the modified Li & Lim's [23] benchmark instances particularly LC101 serves the best route that provides the best compromise between the different criteria (the sum of travel distance, the vehicle number and the total tardiness time).In future work, we will be interested in developing a Particle Swarm Optimization (PSO) which gives one best solution and compare it with MOO approach.

References

[1] J. Lenstra and A. RinnooyKan: Complexity of the Vehicle Routing and Scheduling Problems, Networks international journal, Vol. 11, pp. 221-228, Springer 1981.

[2] Martin Desrochers, Jacques Desrosiers and Marius Solomon. A New Optimization Algorithm for the Vehicle Routing Problem with Time Windows Operations Research, Vol. 40, No. 2. (Mar. - Apr., 1992), pp. 342-354.

[3] Yahia Zare Mehrjerdi. Multiple-Criteria Decision-Making Combined with VRP: A Categorized Bibliographic Study, International Journal of Supply and Operations Management, IJSOM, August2015, Volume 2, Issue 2, pp. 798-820.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A Fast and Elitist Multi objective Genetic Algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation, Vol. 6, No. 2, pp. 182-197, 2002.

[5] N. Velasco, P. Dejax, C. Guéret and C. Prins. A Non-dominated Sorting Genetic Algorithm for a Bi-Objective Pickup and Delivery Problem, Engineering Optimization, 44, pp. 305-325, 2012.

[6] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm: Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problem EUROGEN, Athens, Greece, pp. 95-100, 2002.

[7] I. Harbaoui Dridi, R. Kammarti, P. Borne and M. Ksouri. Multi-Objective Optimization for the m-PDPTW: Aggregation Method With Use of Genetic Algorithm and Lower Bounds, International Journal of Computers, Communications & Control (IJCCC), E-ISSN 1841-9844, Vol. VI, No. 2, pp. 246-257, 2011.

[8] G. Nagy, and S. Salhi. Heuristic Algorithms for Single and Multiple Depot Vehicle Routing Problems with Pickups and Deliveries, European Journal of Operational Research, Vol. 162, pp. 126-141, Elsevier 2005.

[9] G. Rudolph, O. Schütze, C. Grimme and H. Trautmann. A Multi Objective Evolutionary Algorithm Guided by Averaged Hausdorff Distance to Aspiration Sets In EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation, pp. 261-273, Springer 2014.

[10] O. Pisacane, F.Guerriero and F.Pezzella. The multi objective Multi vehicle Pickup and Delivery problem with time Windows, Social and Behavioral Sciences Journal, Volume 111, 5 February 2014, Pages 203-212.

[11] A. Zinflou, C. Gagné, and M. Gravel. GISMOO: A New Hybrid Genetic/Immune Strategy for Multiple-Objective Optimization, Computers & Operations Research, Vol. 39, No. 9, pp. 1951-1968, 2012.

[12] B. Ombuki-Berman, and T. H. Franklin. Using Genetic Algorithms for Multi-Depot Vehicle Routing, Bio-inspired Algorithms for the Vehicle Routing Problem, Vol. 161, pp. 77-99, Springer 2009.

[13] X. Wang, C. Xu, and H. Shang. Multi-Depot Vehicle Routing Problem with Time Windows and Multi-Type Vehicle Number Limits and its Genetic Algorithm, 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-5, Dalian, China, 2008.

[14] H. C. W. Lau, T. M. Chan, W. T. Tsui, and W. K. Pang. Application of Genetic Algorithms to Solve the Multi Depot Vehicle Routing Problem, IEEE transactions on automation science and engineering, Vol. 7(2), 2010.

[15] C. A. Coello, Gary B. Lamont and David A. Van Veldhuizen. Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd Edition, Springer 2007.

[16] K. Deb. Multi objective Optimization Using Evolutionary Algorithms. Chichester, U.K.: Wiley, 2001.

[17] Essia Ben Alaia, Imen Harbaoui Dridi, Hanen Bouchriha and Pierre Borne. Optimization of the Multi-Depot and Multi-Vehicle Pickup and Delivery Problem with Time Windows using Genetic Algorithm. IEEE International Conference on Control, Decision and Information Technologies CODIT 2013, 6-8 May, Hammamet. Tunisie, 2013.

[18] Essia Ben Alaia, Imen Harbaoui Dridi, Hanen Bouchriha and Pierre Borne. Genetic Algorithm with Pareto Front selection for Multi-Criteria Optimization of Multi-Depots and multi-vehicle Pickup and Delivery Problems with Time Windows. IEEE International Conference on Sciences and Techniques of Automatic control and computer engineering, STA 2014, 21-23 December, Hammamet, Tunisie, 2014.

[19] Essia Ben Alaia, Imen Harbaoui Dridi, Hanen Bouchriha and Pierre Borne. Genetic Algorithm for Multi-Criteria Optimization of Multi-Depots Pickup and Delivery Problem with Time Windows and multi vehicles. Acta Polytechnica Hungarica, vol. 12, no 8, pp. 155-174, 2015.

[20] DEB, K. Multi-objective evolutionary optimization: Past, present and future, chapitre Evolutionary Design and Manufacture, pages 225–236. Springer, 2000.

[21] T. L. Saaty, \How to make a decision: the analytic hierarchy process," European journal of operational research, vol. 48, no. 1, pp. {9, 26}, 1990.

[22] M. Alexander, Decision-making using the analytic hierarchy process (ahp) and sas/imlR," Available from (Last checked 28th July 2014), 2012.

[23] LI, H. and LIM, A. (2001). A metaheuristic for the pickup and delivery problem with time windows. In IEEE International Conference on Tools with Artificial Intelligence, pages 160–167.

[24] Essia Ben Alaia, Imen Harbaoui Dridi, Hanen Bouchriha and Pierre Borne. Insertion of new depot locations for the optimization of multi-vehicles Multi-Depots Pickup and Delivery Problems using Genetic Algorithm. IEEE International Conference on Industrial Engineering and Systems Management IESM 2015, 21-23 October, Seville, Spain, 2015.

Freight Buses in Three-Tiered Distribution Systems for City Logistics: Modeling and Evaluation

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Zheng Chang
University of Technology of Troyes
ICD-LOSI, UMR CNRS 6281
Troyes, France
zheng.chang@utt.fr

Haoxun Chen
University of Technology of Troyes
ICD-LOSI, UMR CNRS 6281
Troyes, France

Farouk Yalaoui
University of Technology of Troyes
ICD-LOSI, UMR CNRS 6281
Troyes, France

Abstract—In this paper, we put forward the concept of freight bus, which is a new public transport means that can replace city freighters operated by different private logistics companies in a city. In order to evaluate the effectiveness of freight buses for city logistics, we build a mathematical programming model for the routing planning of freight buses. Based on the model, we compare a distribution system with freight buses and the corresponding system without freight bus. Preliminary numerical results on randomly generated instances show that the system with freight buses can significantly reduce transportation costs compared with the system without freight buses.

Keywords—City Logistics; Three-Tiered Distribution System; Freight Bus; Vehicle Routing

I. Introduction

With the rapid development of e-commerce and urbanization, traditional one tier and two-tiered systems, cannot satisfy the market demand in metropolitan areas. This gives rise to three-tiered distribution systems for city logistics in many metropolitan areas. The first tier is between city distribution centers (CDCs) and satellite platforms. The second tier is between satellite platforms and depots. The third tier is between depots and customers. These three tiers have different types of vehicles employed: urban vehicles, city freighters and electro-tricycles. Such three-tiered distribution systems evolved from two-tiered distribution systems. As a result of the appearance of three-tiered distribution systems, many city freighters were born and circulate in the center of a city, which makes the city traffic more and more congested. The traffic control for city freighters, on the one hand, is difficult to implement, and on the other hand cannot improve the efficiency of city logistics significantly. What's more, the market opportunity of city logistics has been creating more and more private logistics companies to operate their city freighters in big cities. These problems are affecting the efficiency of city logistics. The problems of traffic congestion and air pollution in metropolitan areas are becoming more and more severe.

The research of city logistics aims at optimally planning, controlling and managing logistics flows in urban areas in an integrated and coordinated manner. The discussion of the challenges and perspectives in city logistics can be found in Crainic et al. (2009), Benjelloun and Crainic (2008), Benjelloun et al. (2008), Dablanc (2007), Russo and Comi (2004), and Taniguchi et al. (2001). Traffic management and control in a metropolitan transportation network is typically formulated as a bi-level optimization problem (Patriksson and Rockafeller, 2002). The impact of city logistics on the environment is also considered, and models have been proposed to measure and minimize CO2 emissions in urban areas (Taniguchi and Van Der Heijden, 2000, Taniguchi and Thompson, 2002).

Most of strategies or systems proposed for city logistics and traffic management, have a main objective to achieve an optimal trade-off between ensuring optimal productivity, reliable customer service and reducing environmental impacts, air pollution emissions, energy consumption and traffic congestion. Earlier strategies only involve a single-tier system, in which inbound freight is first consolidated at one or several city distribution centers (CDCs), located at the city limits, and is then distributed to the customers inside the city (Benjelloun et al., 2009). Later strategies, involving two-tiered systems (Crainic et al., 2004, 2009), typically appropriate for large cities, are proposes based on consolidated distribution concept.

In 1973 Shize (1973) first put forward the joint distribution which has been proved to be an effective way for city logistics. Joint distribution promotes enterprises with similar functions to use common facilities and equipment such as warehouses, logistics platforms, and vehicles, through which small orders of goods for delivery can be consolidated into large-volume distribution to achieve the economics of scale in transportation and other logistics services. Gill and Allerheiligen (1996) pointed out the necessity of cooperation through joint distribution among members in distribution channels and illustrated the effectiveness of joint distribution. Zhao and Li (2009) proposed a model for cost sharing among small companies implementing joint distribution.

Motivated by joint distribution, in this paper, we put forward the concept of freight bus, which is a new public transport means that can replace city freighters belonging to different private logistics companies operated in a city. Freight bus has some advantages compared with city freighter. Firstly, freight bus can realize joint distribution of different third-party logistics companies, and can thus save city logistics costs and reduce the air pollution; Secondly, because of having a regular schedule, freight bus can improve the timeliness and accuracy of logistics services; Thirdly, replacing private city freighters by freight buses can facilitate the traffic control in a city and

reduce the traffic congestion. Finally, freight bus can improve the utilization rate of special lanes reserved for buses.

In order to evaluate the effectiveness of freight bus, we build a mathematical programming model for the routing planning of freight buses. In order to verify the effectiveness of freight bus, based on the model, we compare a distribution system with freight buses and the corresponding system without freight bus. Preliminary numerical results on randomly generated instances show that the system with freight buses can significantly reduce transportation costs compared with the system without freight buses.

II. Three-tiered Distribution System

The three-tiered distribution system evolved from the two-tiered system (Crainic et al., 2004, 2009). The first tier is between city distribution centers (CDCs) and satellite platforms. The second tier is between satellite platforms and depots. The third tier is between depots and customers. These three tiers have different types of vehicles employed: urban vehicles, city freighters and electro-tricycles. The following two figures show their differences.

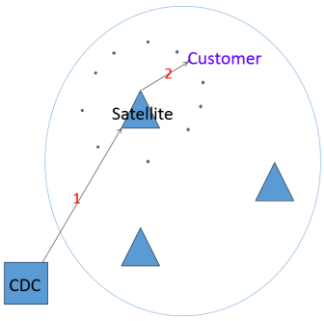


Fig. 1. Two-tiered system

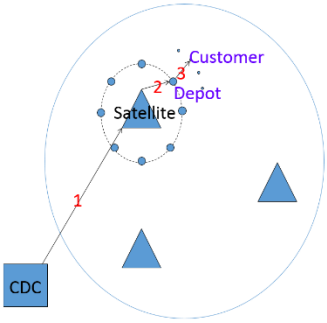


Fig.2. Three-tiered system

City distribution centers (CDCs) are usually situated on the outskirts of a city. All the inbound freights are consolidated at CDCs at first. The goods coming from the CDCs are then transferred to satellites platforms (satellites for short) where goods are consolidated into full or almost full vehicles which are adapted for utilization in dense city zones. Finally, the goods coming from the satellites are transferred to depots where goods are consolidated into vehicles which are used to directly serve the customers. Each depot in the system may be a smart cabinet for temporary storage the goods.

In the two-tiered distribution system, there are usually a lot of satellites in the center of a city. At first, the urban vehicles deliver the goods from a CDC to multiple satellites. Then the electro-tricycles deliver the goods from satellites to customers. The follow table gives the vehicles of different tiers:

TABLE Ⅰ. Vehicles used in different tiers

	Two-tiered system	Three-tiered system
First tier	urban vehicle	urban vehicle
Second tier	electro-tricycle	city freighter
Third tier	×	electro-tricycle

We can see by adopting the three-tiered distribution system, the running distance of tricycles can be greatly reduced, and it will greatly reduce the labor costs. However, accompanied with the three-tiered distribution system, we can see that a lot of city freighters were born and circulate in the city center. The market opportunity of city logistics has been creating more and more city freighters in the city center. As a result, the traffic congestion in the city center becomes more and more severe. The traffic control for city freighters, on the one hand, is difficult to implement, and on the other hand cannot improve the efficiency of city logistics significantly. In order to solve these problems, we put forward the concept of freight bus.

III. Freight Bus

In the three-tiered distribution system, due to market competition, the second tier has used different city freighters from different private logistics companies. In this paper, we propose a new solution to replace the city freighters from different private logistics companies, that is, the local government of a city builds freight buses in the center of the city.

Here, we define freight bus as follows: Freight bus is a public transport vehicle in a city which provides services for city logistics. Like a bus, each freight bus has a fixed schedule and route and can use the special lanes reserved for buses.

The new three-tiered distribution system with freight buses can be described as follows. The first tier is between CDCs and satellites, urban vehicles are used to transport goods between them. The second tier is between satellites and depots, freight buses replace all city freighters from different private logistics companies. The third tier is between depots and customers. The electro-tricycles deliver the goods between them. The figure below illustrates the freight bus in the three-tied distribution system.

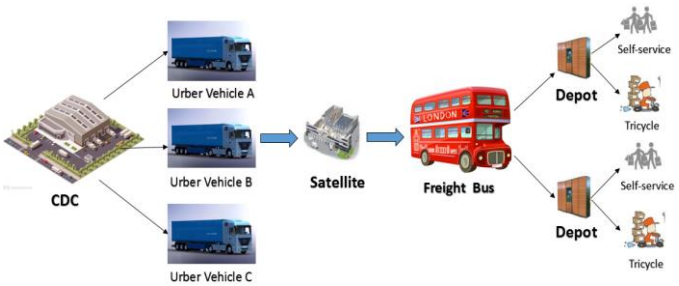


Fig.3. Freight bus in the three-tiered distribution system

Freight bus has some advantages compared with city freighter. Firstly, freight bus can realize joint distribution of different third-party logistics companies, and can thus save city logistics costs and reduce the air pollution; Secondly, because

of having a regular schedule, freight bus can improve the timeliness and accuracy of logistics services; Thirdly, replacing private city freighters by freight buses can facilitate traffic control in a city and reduce the traffic congestion. In addition, freight bus can improve the utilization rate of special lanes reserved for buses.

Freight bus should have a fixed schedule and route. Each depot in the system is a freight bus station, it may be a smart cabinet for temporary storage of the goods. The follow is an example of freight bus lines.

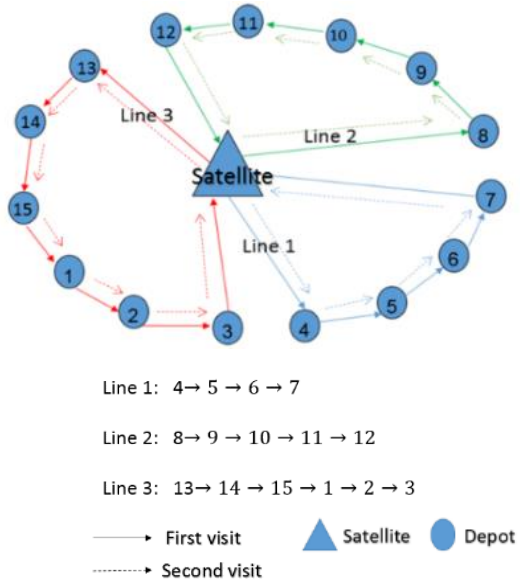


Fig.4. An example of freight bus lines

IV. Mathematic Model for Freight Bus Routing

A. Problem description

In last section, we put forward the concept of freight bus. As the first step to prove the concept, we need to verify whether the distribution system with freight bus can make cost-savings. In order to do so, we need a mathematical model for freight buses. In the model, the freight buses deliver goods from a satellite to a set of depots. For simplicity, we don't consider the interactions of the freight buses with the first tier urban vehicles and the third tier electro-tricycles. We consider a time horizon of M periods (M > 1) and assume that each freight bus visits its served depots once in each period and the demand of delivery of goods to each depot in each period is known. In addition, we assume each freight bus has a fixed route in the time horizon, and every depot must be served by one freight bus in each period. The total delivery demand of each depot must be met at the end of the time horizon, but it is possible that the demand of a depot in a period is totally or partially met in a later period because of the capacity limitation of a freight bus.

The main parameters of the model are defined as follows:

- s The satellite where a freight bus leaves from and returns to.
- V Set of freight buses.
- G Set of depots.
- U The capacity of each freight bus.
- C_{ij} The operating cost for a freight bus to be charged when it travels from the node i to j ($i, j \in \{s\} \cup G$).
- M The number of time period we consider in the mathematic model.
- $d_i(k)$ The demand of depot $i \in G$ during the k-th period. $k \in \{1,2 \dots M\}$

We assume that the satellite s serves P (P is an integer) depots in the system. The distance between the node i and j is D_{ij}. Freight buses load the goods at the satellite s, and deliver them to depots. The operating cost of freight bus from node i to j is calculated by $C_{ij} = C * D_{ij}$ (C is the unit distance operating cost of freight bus). There are N (N is an integer) freight buses operated for the satellite s, and the capacity of every freight bus is U.

In each period, each freight bus leaves from the satellite s, visits its served depots and returns to the satellite. What's more, every period, all the goods must be unloaded from the freight bus before it returns to the satellite s. It is assumed that the demand $d_i(k)$ of each depot i in each period $k \in \{1,2 \dots M\}$ is known. In each period, the delivery of freight bus v at each depot i can be the demand of that period plus the demand (the partial demand) of an earlier period but cannot be the demand of a later period.

We need to plan the vehicle route for each freight bus v, and the delivery quantity of every freight bus at each depot in each period. Our objective is to minimize the operating costs of all the freight buses in the M periods.

B. Mathematic model

In this subsection, we propose a mathematic model for freight buses. With this mathematic model, we can calculate the whole operating costs of the system with freight bus in the planning horizon. At the same time, we can also get the optimal routes of freight buses.

The model of freight bus is a special periodic vehicle routing problem. In conventional periodic vehicle routing problem, different depots often require different visit times of vehicles, and each vehicle tends to have different vehicle routes in different visits (Ann Melissa Campbell, 2014). However, in our periodic vehicle routing model of freight bus, different depots in the system have the same visit times of freight bus. For all visits, each freight bus will have a fixed vehicle route. The detailed mathematic model of freight bus is as follow:

Decision Variables

- x_{ij}^v A binary variable which is equal to 1 if the freight bus $v \in V$ goes from i to j ($i, j \in \{s\} \cup G$); 0 otherwise.
- y_i^v A binary variable which is equal to 1 if and only if the depot $i \in G$ is served by the freighter bus $v \in V$; 0 otherwise.
- $d_i^v(k)$ The unloading quantity of the freighter bus $v \in V$ at the depot i on the k-th visit $k \in \{1,2 \dots m\}$; 0 otherwise.

- $Q_i^v(k)$ The load of the freighter bus v ϵV when it just arrives at the point i ϵ{s} ∪ G during the k-th visit kϵ $\{1,2 \dots M\}$.

Objective Function

The objective is to minimize the operating costs of the freight buses in M periods:

$$\text{Minimize } Z = M * \sum_{v\epsilon V} \sum_{i\in\{s\}\cup G} \sum_{j\in\{s\}\cup G} c_{ij} x_{ij}^v$$

Constraints

- $\sum_{j\in G} x_{sj}^v = \sum_{j\in G} x_{js}^v \quad \forall v \in V$ (1)
- $\sum_{i\in\{s\}\cup G} x_{ij}^v = \sum_{i\in\{s\}\cup G} x_{ji}^v \quad \forall j \in G, \forall v \in V$ (2)
- $\sum_{i\in\{s\}\cup G} x_{ij}^v = y_j^v \quad \forall j \in G, \forall v \in V$ (3)
- $\sum_{v\in V} y_j^v = 1 \quad \forall j \in G$ (4)
- $Q_j^v(k) >= Q_i^v(k) + d_j^v(k) - U(1-x_{ij}^v)$

  $\forall i \in G, \forall j \in \{s\} \cup G, \forall v \in V, \forall k \in \{1,2 \dots M\}$ (5)
- $Q_s^v(k) = 0 \quad \forall v \in V, \ \forall k \in \{1,2 \dots M\}$ (6)
- $0 <= Q_i^v(k) <= U \quad \forall i \in G, \forall v \in V, \forall k \in \{1,2 \dots M\}$ (7)
- $\sum_{k=1}^{M} d_i^v(k) = d_i \quad \forall i \in G, \forall v \in V$ (8)
- $\sum_{k=1}^{n} d_i^v(k) <= \sum_{k=1}^{n} d_i(k)$

  $\forall i \in G, \forall v \in V, \forall n \in \{1,2 \dots M\}$ (9)
- $x_{ij}^v \in \{0,1\} \quad y_i^v \in \{0,1\}$

  $\forall i \in \{s\} \cup G, \ \forall j \in \{s\} \cup G, \forall v \in V$ (10)

Constraints (1) indicate that each freight bus leaves from and returns to the satellite. Constraints (2) ensure that each freight bus arriving at a depot has to leave it. Constraints (3) and (4) guarantee that all the depots must be served and each depot is served by at most one freight bus. Constraints (5) (6) and (7) formulate vehicle capacity constraints. Constraints (8) guarantee that all demands at each depot must be met after M times of visit of freight bus. Constraints (9) indicate that each period, the delivery of freight bus v at each depot i can be the demand of that or earlier period but cannot be the demand of later period. Finally, constraints (10) define the variables.

V. Comparison with a System without Freight Bus

A. The system without freight bus

In order to evaluate the effectiveness of the proposed freight bus system, we compare a system with freight bus with the corresponding system without freight bus.

In the system without freight bus, we assume that there are city freighters from two companies A and B, which separately deliver their customers' demands from satellite s to depots (The demands are given in Table 2). Each city freighter from company A or B also visits its served depots once during each period kϵ{1,2⋯M}. And the demand of each company at each depot i in each period must be met by its own city freighters. We assume that all the city freighters from company A and company B have the same capacity and unit distance operating cost. In each period, the city freighters also begin and end their travel at the satellite s. What's more, in each period, each city freighter must unload all the goods on the vehicle before it returns to the satellite s. The total delivery demand of each depot for each company must also be met at the end of the time horizon, and it is also possible that the demand of a depot in a period is totally or partially met in a later period because of the capacity limitation of a city freighter.

But in this system, each city freighter is not required to have a fixed vehicle route in its M visits, and each company must serve its own demands from satellite s to depots. The objective of each company is to minimize the operating costs of its own city freighters. For each company in each period, the mathematic model of the system without freight bus is similar to the model of capacitated vehicle routing problem.

B. Experimental results

In order to verify the mathematical model of freight bus routing, and further through the comparison to prove that the proposed freight bus system is effective, we need to generate instances which are representative for the freight bus system and the system without freight bus.

The model of the system with freight bus and the model of the system without freight bus considered in this paper are both related to the Capacity Vehicle Routing Problem. So when generating the instances, we use some data from the benchmark instances of Capacity Vehicle Routing Problem provided by Augerat et al at http://neo.lcc.uma.es/vrp/. However, since the two models involve multiple periods, we generate the demand of each depot in each period randomly based on the benchmark data.

Twenty sets of instances are generated (Table 3). For all instances, the following data are taken from the benchmark instances: the number of nodes, the coordinates of these nodes, the number of city freighters. The other data are generated randomly or based on the benchmark instances: the demand of each depot in each period, the capacity of each city freighter, the capacity of each freight bus, the number of freight buses, and the number of periods M. For the instances, the number of depots is set to 31 or 38, the number periods is set to 5 or 10, the capacity of each freight bus is set to 100 or 120, and the capacity of city freighters is set to 80 or 100. The capacity of each freight bus is set to one no less than the capacity of each city freighter, because freight bus is used to realize the joint distribution of multiple city freighters. The parameter setting of each instance is a combination of the possible values of these parameters. Since it is assumed that all freight buses and city freighters have the same unit distance operating cost, we simply set C to 1 for all instances.

As an example, for one instance in TABLE Ⅲ, the number of depots is 31, the number of periods is 5, the capacity of each freight bus is 100, the capacity of each city freighter is 100, the number of freight buses is 10, the number of city freighters from company A is 5, the number of city freighters from company B is 5. The figure below presents the coordinates of satellite and depot nodes.

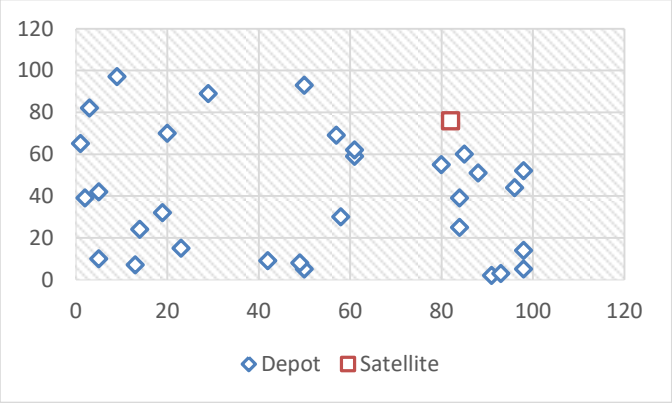


Fig.5. Coordinates of the satellite and depot nodes

Because in the two models, each vehicle will operate M (M is an integer) times in M periods, once each period. So we need to generate the demand of each depot in each period according to the benchmark instances. At first, we generate the demand of company A and the demand of company B at each depot in each period separately, both of them are randomly generated from [1, 20]. Then, we can generate the demand of each depot in each period for the system with freight bus by summing up the two corresponding demands of the two companies. The follow table gives an example of the demand for each depot in one period.

TABLE Ⅱ. An example of the demand for each depot in one period

Depot	Demand of Company A	Demand of Company B	Demand for Freight Bus	Depot	Demand of Company A	Demand of Company B	Demand for Freight Bus
1	5	13	18	17	8	20	28
2	7	14	21	18	11	15	26
3	11	8	19	19	9	16	25
4	3	2	5	20	16	19	35
5	18	10	28	21	4	20	24
6	19	4	23	22	13	14	27
7	14	6	20	23	5	15	20
8	16	3	19	24	17	20	37
9	9	15	24	25	20	16	36
10	16	4	20	26	8	3	11
11	7	7	14	27	6	7	13
12	5	9	14	28	10	9	19
13	1	12	13	29	13	10	23
14	16	13	29	30	12	8	20
15	14	14	28	31	10	13	23
16	13	5	18				

We then use CPLEX to solve the two models for each instance with a preset running time of 1 hour, and compare the results of the two distribution systems. The follow table gives the experimental results.

TABLE Ⅲ. Experimental results of the two distribution systems

No.	Depot	Visit times M	Capacity of freight bus	Capacity of city freighter	Cost with freight bus	Cost without freight bus	Cost saving
1	31	5	100	100	1793	2906	38.3%
2	31	5	100	100	1778	2902	38.7%
3	31	5	100	100	1791	2908	38.4%
4	31	5	120	80	1726	3393	49.1%
5	31	5	120	80	1726	3366	48.7%
6	31	10	100	100	3824	6493	41.1%
7	31	10	100	100	3847	6494	40.8%
8	31	10	100	100	3822	6484	41.1%
9	31	10	120	80	3776	6548	42.3%
10	31	10	120	80	3759	6549	42.6%
11	38	5	100	100	1946	3230	39.8%
12	38	5	100	100	1944	3250	40.2%
13	38	5	100	100	1934	3201	39.6%
14	38	5	120	80	1923	3385	43.2%
15	38	5	120	80	1921	3383	43.2%
16	38	10	100	100	4223	6759	37.5%
17	38	10	100	100	4219	6739	37.4%
18	38	10	100	100	4221	6742	37.4%
19	38	10	120	80	4195	6947	39.6%
20	38	10	120	80	4173	6940	39.9%

From the experimental results, we can see that if we use the proposed freight bus system, the cost saving is from 37.4% to 49.1% with the average cost saving 40.9%. The preliminary numerical results show that the system with freight bus can significantly reduce transportation costs compared with the system without freight bus. Analyzing the impact of different parameter values on cost savings, we see that larger capacity of freight buses can help to increase cost savings, and there is no strong relationship between the number of depots or the number of periods M and cost savings.

VI. Conclusions

In this paper, we put forward the concept of freight bus, which is a new public transport means that can replace city freighters belonging to different private logistics companies operated in a city. Freight bus has some advantages compared with city freighter. Firstly, freight bus can realize joint distribution of different third-party logistics companies, and can thus save city logistics costs and reduce the air pollution; Secondly, because of having a regular schedule, freight bus can improve the timeliness and accuracy of logistics services; Thirdly, replacing private city freighters by freight buses can facilitate the traffic control in a city and reduce the traffic congestion. Finally, freight bus can improve the utilization rate of special lanes reserved for buses.

In order to evaluate the effectiveness of freight bus, we build a mathematical programming model for the routing

planning of freight buses. Based on the model, we compare a distribution system with freight buses and the corresponding system without freight bus. Preliminary numerical results on randomly generated instances show that the system with freight buses can significantly reduce transportation costs compared with the system without freight bus.

However, in this paper, we have not quantitatively analyzed the impact of using freight buses on the improvement of the timeliness of city logistics services. In the future, we need to analyze the impact, propose an effective optimization algorithm to solve the freight bus routing problem proposed in this paper for large instances, and consider more practical issues to operate freight buses in a city.

Acknowledgment

This work is supported by the French National Research Agency (ANR) under the project ANR-14-CE22-0017 entitled "Collaborative Transportation in Urban Distribution".

References

[1] Rodrigo R.A., El-Houssaine A., City Logistics and Traffic Management: Modelling the Inner and Outer Urban Transport Flows in a Two-Tiered System, Transportation Research Procedia 2015, 6, 297 – 312.

[2] Benjelloun, A., T. G. Crainic. Trends, challenges, and perspectives in city logistics. Transportation and Land Use Interaction, Proc.TRANSLU'08. Editura Politecnica Press, Bucharest, Romania, 2008, 269–284.

[3] Benjelloun, A., Bigras, Y., Crainic, T.G., Towards a Taxonomy of City Logistics Systems, Publication CIRRELT, Interuniversity Research Center on Enterprise Networks, Logistics and Transportation, Université de Montréal.

[4] Bogers, E.A.I., and van Zuylen, H.J., The importance of reliability in route choices in freight transport for various actors on various levels. In Proceedings European Transport Conference, 2004, 1-13, Strasbourg Association for European transport.

[5] Crainic,T.G, Ricciardi,N and Storchi,G: Models for Evaluating and Planning City Logistics Systems, Transportation Science 43(4), pp. 432–454, 2009, INFORMS.

[6] Dablanc, L. Goods Transport in Large European Cities: Difficult to Organize, Difficult to Modernize, Transportation Research Part A: Policy and Practice, 2007, 41(3), 280-285.

[7] Kazimi, C., Brownstone, D., Ghosh, A., Golob, T.F., and Van Amelsfoort, D., eds, Willingness-to-Pay to Reduce Commute Time and its Variance: Evidence from the San Diego I-15 Congestion Pricing Project, 2000, 79th Transportation Research Board, Washington D.C., USA.

[8] Russo, F., A. Comi. A state of the art on urban freight distribution at European scale Presentation, The European Conference on Mobility Management (ECOMM 2004). Lyon, France. http://www.epomm.org/.

[9] Sankaran, J.K., Gore, A., Coldwell, I., The impact of road traffic congestion on supply chains: insights from Auckland, New Zealand. International Journal of Logistics 2005, 8(2), 159-180.

[10] Shize Xiuxiong. Introduction of joint distribution. Circulation and economic study, 1973, 8 (3):87-94.

[11] Gill L. E., R. P. Allerheiligen. Co-operation in channels of distribution: physical distribution leads the way International Journal of Physical Distribution & Logistics Management, 1996:49-63.

[12] Zhao Yan-ping, Yan Li. Study on the Choice of Logistics Delivery Model in Small and Medium-sized Manufacturing Enterprises. Operations Research and Management Science, 2009, 18 (5): 163-167.

[13] Ann M. C. Jill H. W. Forty Years of Periodic Vehicle Routing. Operations Research and Management Science, 2014, 63 (1): 2-15.

A Linear Program for the Resource-Constrained Project Scheduling Problem with Routing

(presented at the 7th IESM Conference, October 11 – 13, 2017, Saarbrücken, Germany)

Philippe Lacomme, Alain Quilliot,
Nikolay Tchernev, Marina Vinot
Université Clermont Auvergne,
(LIMOS, UMR CNRS 6158),
63177 Aubière, France
placomme@isima.fr, alain.quilliot@isima.fr,
tchernev@isima.fr, marina.vinot@isima.fr

Aziz Moukrim
Sorbonne Universités, Université de Technologie de
Compiègne, (Heudiasyc UMR CNRS 7253),
60203 Compiègne France
aziz.moukrim@utc.fr

Abstract— This paper concerns the mathematical formulation with the optimal resolution of the Resource-Constrained Project Scheduling Problem with Routing (RCPSPR). Our objective is to propose optimal solutions for small scale instances. Since no instances dealing with the problem are available, computational experiments are carried out on new generated instances. The numerical results provide a set of optimal solutions that can be used to evaluate the performance of afterwards developed metaheuristic approaches.

Keywords—RCPSPR, Linear programming, Scheduling, Routing

I. Introduction to the RCPSPR

A. RCPSPR definition

The RCPSPR (Resource-Constrained Project Scheduling Problem with Routing) is an extension of the RCPSP (Resource-Constrained Project Scheduling Problem) where resources are transported from one pickup activity to a delivery activity using a vehicle. The problem consists in solving both the scheduling problem with activities and the vehicle routing problem minimizing the schedule length. The schedule length C_{max} ($i.e.$ the project makespan) is defined by the end of the last transport operation from one activity to the dummy activity $n+1$). The resource project scheduling part consists in scheduling a set of activities, $V = \{0, \ldots, n+1\}$, where n is the number of non-dummy activities with 0 and $n+1$ two dummy activities. This paper deals with the case of one resource, $q = 1$, with an availability of B units. An activity j is characterized by a duration p_j, a resource requirement b_j and a set of precedence constraints E. The routing part consists in scheduling trips, with a set of vehicles $T = \{1, \ldots, N\}$ sorted in descending order of capacity $c_u, u \in \{1, \ldots, N\}$. Since in this problem the transportation times are not dependent of the vehicle used or of the vehicle load, a transportation time from activity i to j is denoted t_{ij}. A routing solution is defined as a set of trips. Each trip is an ordered sequence of loaded transport operations defined by a pickup activity, a delivery activity, a vehicle, a departure time and an arrival time of the vehicle.

The RCPSPR integrates both scheduling and routing problems, these problems received a lot of attention in the past few years [5] and highlight a key issue in supply chain management. Currently, the integration of routing problems with RCPSP has not been studied, contrary to the flow-shop [6] and the job-shop problems [4].

B. RCPSPR constraints

To obtain solutions of the RCPSPR, three sets of constraints have to be taken into account, the set of constraints from the scheduling problem (RCPSP), the set of constraints from the routing problem and the set of constraints integrating both scheduling ant routing.

To obtain solutions of the scheduling problem the minimal set of constraints to take into account is as follow:

- an activity cannot start before the completion of its predecessors;
- an activity cannot be interrupted;
- at each time period and for each resource, the total activity resource requirement does not exceed the resource availability.

The minimal set of constraints to obtain a routing solution of the RCPSPR is as follow:

- the load of a vehicle cannot exceed the vehicle capacity;
- a vehicle start and end its trip at the depot;
- a vehicle realized only direct transport.

The minimal set of constraints integrating both scheduling ant routing of the RCPSPR is as follow:

- an activity can only start when all the resources required have been delivered by one or several vehicles;
- a vehicle can pickup resources from an activity, if the activity is completed;

II. Linear formulation for the RCPSPR

The flow-based formulation for the standard RCPSP is reminded first below, and then extended in the next section to consider the Resource-Constrained Project Scheduling Problem with Routing (RCPSPR).

A. Linear formulation for the RCPSP

Inspired by the formulation of Balas et al. [3] on the job-shop problem, and on the formulation of Alvarez-Valdes and

Tamarit [1] with the concept of forbidden sets, Artigues et al. [2] proposed a flow-based model for the RCPSP that uses flow variables to manage the resources. The idea is as follows. All resources are available and stored at the dummy activity 0. Each activity uses resources at the beginning of its execution, and once completed, it transfers these resources to the activities that follow. The last activities to be executed finally forwards the resources to the dummy activity $n+1$.

In the flow model, there is three variables:

φ_{ijk} quantity of resource k transferred from activity i to activity j

S_i starting time of activity i

$z_{ij} = \begin{cases} 1 & \text{if the activity i is scheduled before activity j} \\ 0 & \text{otherwise} \end{cases}$

In the linear formulations introduced below, the objective (1) concerns the makespan minimization (*i.e.* minimization of the starting time of the dummy activity $n+1$). Without loss of generality, in this paper dealing with the case of one resource, the indices k are not used.

$$\min S_{n+1} \quad (1)$$

$$z_{ij} = 1 \qquad \forall (i,j) \in E \quad (2)$$

$$S_j \geq S_i + p_i + M(z_{ij} - 1) \qquad \forall (i,j) \in V^2 \quad (3)$$

$$\varphi_{ij} \leq min(b_i, b_j) z_{ij} \qquad \forall (i,j) \in V^2 \quad (4)$$

$$\sum_{i \in V} \varphi_{ij} = b_j \qquad \forall j \in V \quad (5)$$

$$\sum_{j \in V} \varphi_{ij} = b_i \qquad \forall i \in V \quad (6)$$

$$z_{ij} \in \{0,1\}, z_{ii} = 0 \qquad \forall (i,j) \in V^2 \quad (7)$$

$$S_i \geq 0 \qquad \forall i \in V \quad (8)$$

$$\varphi_{ij} \geq 0, \varphi_{ii} = 0 \qquad \forall (i,j) \in V^2 \quad (9)$$

M is some large-enough constant to let S_i and S_j unrealeted when $z_{ij} = 0$. Constraint (2) set the variable z_{ij} to 1, if the activity j is a successor of activity i according to the precedence constraints of the problem. Constraint (3) ensures that two activities i and j are not performed at the same time if $z_{ij} = 1$, in this case, activities j cannot start before the activity i completion. Constraint (4) limits the flow transferred between activities i and j to $\min(b_i, b_j)$ if i precedes j, and to 0 otherwise. Constraints (5) and (6) are the usual inequalities of the flow conservation (respectively the incoming flow on node j and the outgoing flow on node i).

B. RCPSPR formulation

In this section, the resource-constrained project scheduling problem with routing is formulated as a linear programming model taking into account the linear formulation of the RCPSP presented in the previous section. Both part are required to fully define the problem with scheduling and routing coordination.

The global formulation uses five types of continuous decision variables $(\varphi_{ij}, S_i, y_{ijut}, A_{ijut}, B_{ijut})$ and three types of binary variables $(z_{ij}, x_{ijut}, a^u_{ijtpqv})$.

y_{ijut} quantity of resource transported from activity i to activity j by the vehicle u in the t^{th} transport between activity i and j.

A_{ijut} arrival time on activity j of the vehicle u in the t^{th} transport between activity i and j.

B_{ijut} departure time on activity i of the vehicle u in the t^{th} transport between activity i and j.

$x_{ijut} = \begin{cases} 1 & \text{if resources are transported by the vehicle } u \text{ in the } t^{th} \text{ transport between activity } i \text{ and } j. \\ 0 & \text{otherwise} \end{cases}$

$a^u_{ijtpqv} = \begin{cases} 1 & \text{if } t^{th} \text{ transport between activity } i \text{ and } j \text{ is scheduled before the } v^{th} \text{ transport between activity } p \text{ and } q, \text{ both on the same vehicle } u \\ 0 & \text{otherwise} \end{cases}$

The Fig. 1 below permits to represent two transport operations assigned to the same vehicle (u) between activities i and j. In order to differentiate these two transport operations, the subscript on the transport number is needed. The first transport operation $(x_{ijut} = 1)$ and the second $(x_{ijut+1} = 1)$ are scheduled in such a way that $a^u_{ijtijt+1} = 1$.

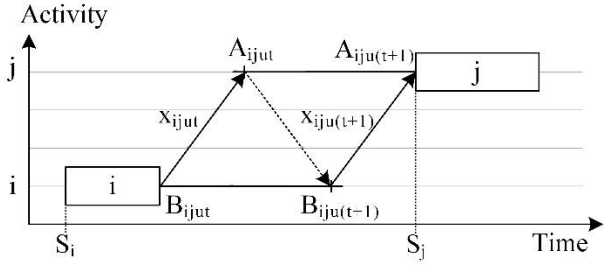


Fig. 1. Illustration of two sequential transport operations.

For convenience, subscripts i, j, p, q are used for activities, u is used for vehicles and t, v are used for the transport number. The transport of all resources from activity i to activity j can required in the worst case (with a vehicle of capacity 1), $t = min(b_i, b_j)$ loaded transport. According to this assumption, for each couple (i,j) of activities, a set $X_{ij} = \{1, \dots, min(b_i, b_j)\}$ can be created to number the transport.

$$\sum_{u\in T}\sum_{t\in X_{ij}} y_{ijut} = \varphi_{ij} \qquad \forall\,(i,j)\in V^2 \tag{10}$$

$$y_{ijut} \le min(b_i, b_j, c_u)\,x_{ijut} \qquad \forall\,(i,j)\in V^2,\ \forall u\in T,\ \forall\, t\in X_{ij} \tag{11}$$

$$x_{ijut} \le y_{ijut} \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{12}$$

$$A_{ijut} \ge B_{ijut} + t_{ij} + (x_{ijut}-1)H \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{13}$$

$$S_j \ge A_{ijut} + (x_{ijut}-1)H \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{14}$$

$$B_{ijut} \ge S_i + p_i + (x_{ijut}-1)H \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{15}$$

$$a^u_{ijtpqv} + a^u_{pqvijt} \le x_{ijut} \qquad \forall\,(i,j,p,q)\in V^4,\ \forall\, u\in T, \forall\,(t,v)\in X^2_{ij} \tag{16}$$

$$a^u_{ijtpqv} + a^u_{pqvijt} \le x_{pquv} \qquad \forall\,(i,j,p,q)\in V^4,\ \forall\, u\in T, \forall\,(t,v)\in X^2_{ij} \tag{17}$$

$$a^u_{ijtpqv} + a^u_{pqvijt} \ge 1 - (2 - x_{ijut} - x_{pquv})H \qquad \forall\,(i,j,p,q)\in V^4,\ \forall\, u\in T, \forall\,(t,v)\in X^2_{ij} \tag{18}$$

$$B_{pquv} \ge A_{ijut} + t_{jp} + (a^u_{ijtpqv}-1)H \qquad \forall\,(i,j,p,q)\in V^4,\ \forall\, u\in T, \forall\,(t,v)\in {X_{ij}}^2 \tag{19}$$

$$y_{ijut+1} \le y_{ijut} \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in \{1,\dots, min(b_i,b_j)-1\} \tag{20}$$

$$y_{ijut} \ge 0 \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{21}$$

$$x_{ijut} \in \{0,1\} \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{22}$$

$$A_{ijut} \ge 0 \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{23}$$

$$B_{ijut} \ge 0 \qquad \forall\,(i,j)\in V^2,\ \forall\, u\in T, \forall\, t\in X_{ij} \tag{24}$$

$$a^u_{ijtpqv} \in \{0,1\} \qquad \forall\,(i,j,p,q)\in V^4,\ \forall\, u\in T, \forall\,(t,v)\in X^2_{ij} \tag{25}$$

H is some large-enough constant. Constraint (10) ensures that all the resources transferred from activity i to activity j, due to φ_{ij}, are transported. The resources transportation can be handled by one or several vehicles $u \in T$ whatever the transport number $t \in X_{ij}$.

Constraint (11) ensures that if the transport operation (i,j,u,t) exists ($x_{ijut}=1$), the quantity of resources transported cannot exceed the possible quantity of resources transferred $min(b_i,b_j)$ and the vehicle capacity c_u.

Constraint (12) link directly the variable x_{ijut} and the variable y_{ijut}. Constraint (13) ensures that the arrival time of a transport operation (i,j,u,t) is greater or equal than the departure time of the same transport operation plus the transportation time if $x_{ijut}=1$. Otherwise, A_{ijut} and B_{ijut} are unrelated.

Constraint (14) ensures that an activity can start only when all the deliveries of the transport operation (i,j,u,t) have been achieved. An example is detailed in Fig. 2 with two transport operations to activity j to deliver resources at times A_{ijut} and A_{kjwv}, before the starting time of the activity S_j. In the other way, constraint (15) ensures that a pickup of a transport operation can occurs if and only if the pickup activity is completed. Figure 1 gives two examples of the constraint (15), with a pickup of one transport operation at time $B_{ijut} \ge S_i + p_i$ and another pickup of one transport operation at time $B_{kjwv} \ge S_k + p_k$. In the following figures, the transport operations are represented by an arrow between two activities.

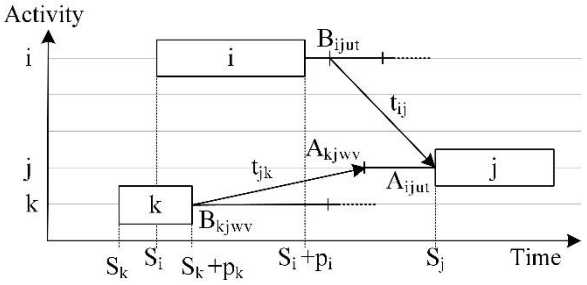


Fig. 2. Illustration of constraints (14) and (15).

Constraints (16), (17) and (18) are disjunctions between transport operations. These constraints ensure that two transport operations cannot be performed at the same time by the same vehicle. If the transport operations (i,j,u,t) and (p,q,u,v) have to be scheduled ($x_{ijut}=x_{pquv}=1$), only two cases are possible. In the first case, (i,j,u,t) precedes (p,q,u,v) and $a^u_{ijtpqv}=1$ or, in the second case, (p,q,u,v) precedes (i,j,u,t) and $a^u_{pqvijt}=1$. Both cases are illustrated in Fig. 3 with the first case on the left and the second case on the right.

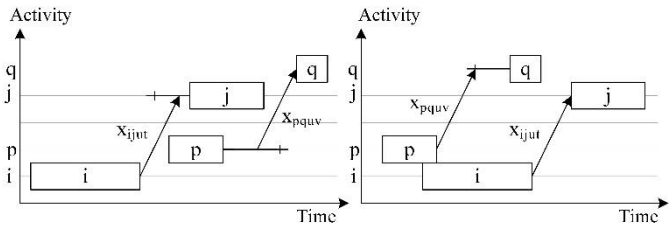


Fig. 3. Illustration of constraints (16), (17) and (18).

Constraint (19), illustrated on Fig. 3, states that, if $a^{u}_{ijtpqv} = 1$, then the departure time of the transport operation (p, q, u, v) is necessarily scheduled after the arrival time of the transport operation (i, j, u, t) plus the time of the unloaded transport operation between them t_{jp}.

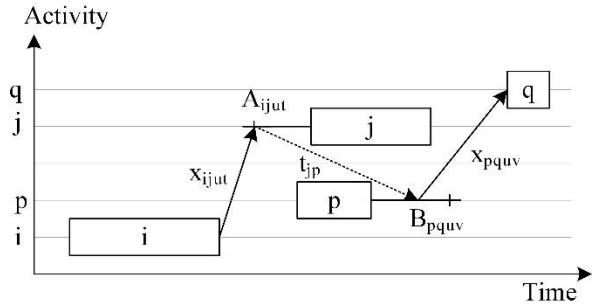


Fig. 3. Illustration of constraints (19).

Constraint (20) permits to delete symmetries created with the subscripts on the transport number. The transport numbers are ordered by decreasing quantity of resources transported in the transport operation.

III. Numerical experiments

A. New set of instances

A new set of small scale instances composed of 18 instances with 6 activities plus the start/end activity is introduced and encompasses two configurations for the location of activities. The resources requirement, the resource availability, the capacity of the vehicles and the ratio are given in the Table 1. The minimal number of operation to schedule $LB(n_{tot})$is given by the column 3 in Table 1. These values represents the n activities to schedule, plus the minimal number of transport operations equal to $n+1$ (with a sequential scheduling).

Table 1
Small scale instances characteristics

Instance	n	LB(n_{tot})	Location	Resource requirements	Resource availability	Vehicle capacity	Ratio
LMQV_U1	6	13	uniform	{4;10;3;3;8;4}	12	{12;12}	1.6
LMQV_U2	6	13	uniform	{4;10;3;3;8;4}	12	{12;12}	1.1
LMQV_U3	6	13	uniform	{4;10;3;3;8;4}	12	{12;12}	1.3
LMQV_U4	6	13	uniform	{2;8;7;8;8;5}	14	{6;5}	0.5
LMQV_U5	6	13	uniform	{2;8;7;8;8;5}	14	{6;5}	0.4
LMQV_U6	6	13	uniform	{2;8;7;8;8;5}	14	{6;5}	0.4
LMQV_U7	6	13	uniform	{5;3;8;2;3;1}	10	{4;2}	0.3
LMQV_U8	6	13	uniform	{5;3;8;2;3;1}	10	{4;2}	0.3
LMQV_U9	6	13	uniform	{5;3;8;2;3;1}	10	{4;2}	0.3
LMQV_C1	6	13	two clusters	{7;7;5;7;5;4}	7	{7;7}	3.0
LMQV_C2	6	13	two clusters	{7;7;5;7;5;4}	7	{7;7}	2.1
LMQV_C3	6	13	two clusters	{7;7;5;7;5;4}	7	{7;7}	3.0
LMQV_C4	6	13	two clusters	{7;1;2;6;2;6}	11	{6;5}	1.1
LMQV_C5	6	13	two clusters	{7;1;2;6;2;6}	11	{6;5}	1.0
LMQV_C6	6	13	two clusters	{7;1;2;6;2;6}	11	{6;5}	1.0
LMQV_C7	6	13	two clusters	{4;10;4;6;3;4}	12	{4;2}	0.8
LMQV_C8	6	13	two clusters	{4;10;4;6;3;4}	12	{4;2}	0.8
LMQV_C9	6	13	two clusters	{4;10;4;6;3;4}	12	{4;2}	0.8

The ratio is defined as the ratio between the average duration of an activity and the average duration of a transport operation per vehicle. A ratio greater than 1 means that the scheduling processing time represents a larger amount of time than the routing and a ratio lower than 1 experienced the reverse.

B. Numerical study of the formulation

All experiments are achieved on a single thread program, using IBMILOG CPLEX 12.6 and Windows 7 as operating system on a Dell Optiflex 9020 with an Intel Core i7-4770 CPU 3.4 GHz and 16 Gb of RAM. With the linear formulation and using CPLEX, the RCPSPR linked to the 18 instances introduce in the previous section can be optimally solved. The solution obtained are presented Table 2 with the following notations:

- r: ratio of the instance
- nc: number of variables in the formulation (columns)
- nr: number of constraints in the formulation (rows)
- Avg: average value
- TT: total CPU time in hours to get the optimal solution

Table 2
RCPSPR optimal resolution

Instance	r	nc	nr	Optimal Solution	TT (h.)
LMQV_U1	1.6	54186	140974	74	5.22
LMQV_U2	1.1	54186	140974	144	4.97
LMQV_U3	1.3	54186	140974	188	6.66
LMQV_U4	0.5	53999	140584	79	>48.00
LMQV_U5	0.4	53999	140584	192	19.77
LMQV_U6	0.4	53999	140584	228	>48.00
LMQV_U7	0.3	13421	34831	97	4.05
LMQV_U8	0.3	13421	34831	197	5.24
LMQV_U9	0.3	13421	34831	218	15.83
LMQV_C1	3.0	84523	220446	70	1.14
LMQV_C2	2.1	84523	220446	118	1.49
LMQV_C3	3.0	84523	220446	143	1.59
LMQV_C4	1.1	21072	54979	33	0.42
LMQV_C5	1.0	21072	54979	68	0.34
LMQV_C6	1.0	21072	54979	72	0.36
LMQV_C7	0.8	27848	72442	44	15.16
LMQV_C8	0.8	27848	72442	90	>48.00
LMQV_C9	0.8	27848	72442	94	>48.00
Avg.	1.1	42508	110709		>24.00

The linear formulation of the RCPSPR, encompasses a formulation with a large number of constraints and variables: for the instance LMQV_U1 for example, the linear formulation is composed of 54 186 variables and 140 974 constraints including binary variables to model the disjunction of activities and transport operations. In average to solve an instance to optimality requires more than one day (24 h) of computational time (Table 2). The optimal solution of the instance LMQV_C9 is detailed in Fig. 4, with a makespan $C_{max} = S_{n+1} = 94$.

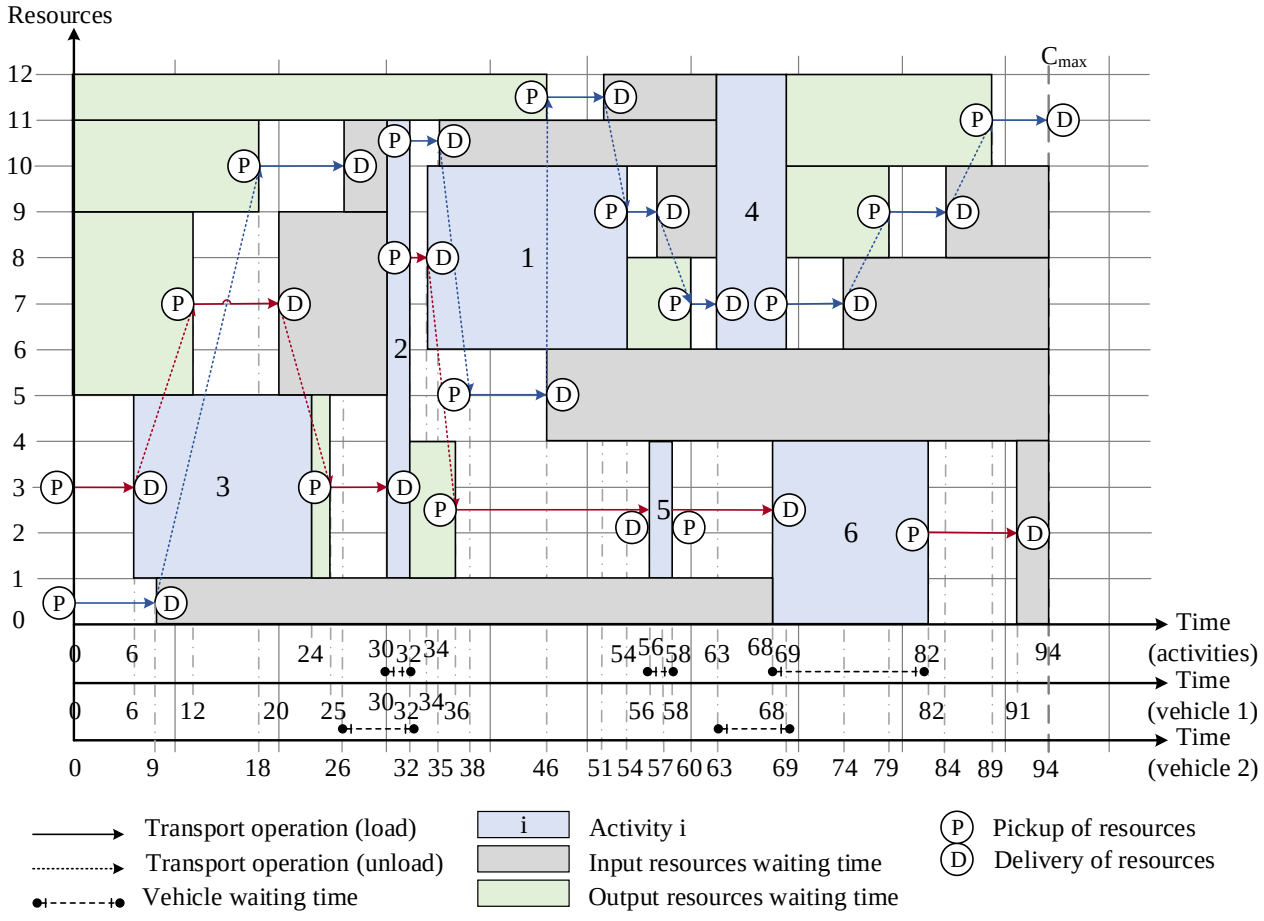


Fig. 4. Optimal solution of instance LMQV_C9.

The Fig. 4 displays the solution as a 2-dimensional Gantt chart where the *x* axis represents the time and the *y* axis represents the resource occupation. In Fig. 4, all the starting times and completion times of the activities are given. The resources waiting times are also represented with either on the activity input or on the output. The trips of the vehicles are displayed with all the transport operations scheduled between activities. All the departure and arrival times linked to pickup and delivery operations on activities are provided in Fig. 4.

For example, in Fig. 4, one can notice that the activity 2 receives its 10 resources from three transport operations. Since all available resources are stored at the dummy activity 0, six resources are coming from dummy activity 0 and four resources come from activity 3. All these resources are then, used by activity 2 between time 30 and 32, when activity 2 is completed the resources come back to an available status. The resources are then transported from activity 2 to four activities (activities 1, 5, 4 and final dummy activity 7).

The solution can also be displayed as a road network, taking into account the activity locations and the vehicles trips, as stressed in Fig. 5. The ten transport operations of the vehicle 1 are displayed and numbered in Fig. 5 (in red), as well as the sixteen transport operations of the vehicle 2 (in bleu).

In the optimal solution of instance LMQV_C9, the linear program have to deal with 6 activities, and 26 transport operations, which represents 32 operations to schedule. Let us note that no transportation cost is defined and the objective consists in makespan minimization only.

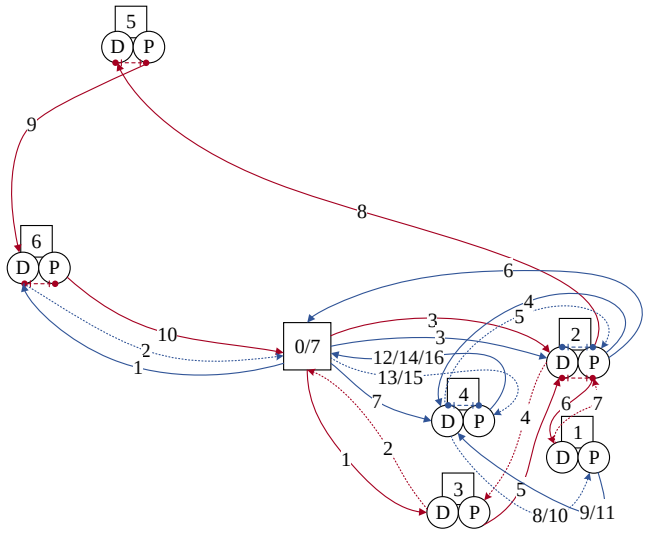


Fig. 5. Trips of the optimal solution of instance LMQV_C9

IV. Concluding remarks

This paper introduce the first linear program formulation for the resource-constrained project scheduling problem with routing (RCPSPR), with multiple vehicles and including the following constraints:

- limited vehicle capacity
- limited number of available resources
- precedence between activities
- direct shipping

The proposed linear model permits to compute optimal solutions to the scheduling/routing problems for small instances (6 activities, but at least 13 operations to schedule including 7 operations to defined, and assigned to a vehicle). Because the linear formulation is intractable for medium scale instances, iterative decomposition methods (including column generation) could be investigated.The objective function is to minimize the schedule length (makespan). In addition to the makespan, another objective could be taken into account in future works, including for example the total riding time. A bi-objective problem minimizing both the schedule length and the total riding time, is an interesting avenue to highlight the benefit of an integrated approach in solving scheduling and routing problems.

V. Acknowledgment

This work was carried out and funded in the framework of the ATHENA project (Reference ANR-13-BS02-0006) and the Labex MS2T. It was supported by the French Government, through the program Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

References

[1] R. Alvarez-Valdès and J.M. Tamarit, “The project scheduling polyhedron: dimension, facets and lifting theorems”, European Journal of Operational Research, vol. 67(2):204–220, 1993.

[2] C. Artigues, P. Michelon and S. Reusser, “Insertion for static and dynamic resource-constrained project scheduling”, European Journal of Operational Research, vol.149, pp. 249-67, 2003.

[3] E. Balas, “Project scheduling with resource constraints”, in E.M.L. Beale, (Ed.), Applications of Mathematical Programming Techniques, pp. 187-200, American Elsevier, 1970.

[4] C. Meinecke and B. Scholz-Reiter, “A heuristic for the integrated productionand distribution scheduling problem”, International Journal of Mechanical, Aerospace, Industrial, Mechatronics and Manufacturing Engineering, vol. 8, pp. 278–285, 2014.

[5] S. Moons, K. Ramaekers, A. Caris and Y. Arda, “Integrating production scheduling and vehicle routing decisions at the operational decision level: A review and discussion”, Computers & Industrial Engineering, vol. 104, pp. 224-245, February 2017.

[6] B. Scholz-Reiter, T. Makuschewitz, A. G. N. Novaes, E. M. Frazzon and O. F. Jr. Lima, “An approach for the sustainable integration of production and transportation scheduling”, International Journal of Logistics Systems and Management, vol. 10, pp. 158–179, 2011.

Mathematical Model for Electric Vehicles Cleaning Problem

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Juan Jiang†, Feifeng Zheng‡, Chengbin Chu*§
*School of Economics & Management, Tongji University, China
Email:mingliu@tongji.edu.cn
†Glorious Sun School of Business & Management, Donghua University, China
Email:kokojiang007@163.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email:ffzheng@dhu.edu.cn
§Centrale Supélec, Laboratoire Génie Industriel, Université Paris-Saclay, France
Email:chengbin.chu@centralesupelec.fr

***Abstract*—This paper studies the electric vehicles cleaning problem, which is receiving increasing attentions in Electric Vehicle Rental Project. Electric vehicles are featured with mobility, environmental and societal benefits, however, the cleaning process may influence the customer experience. Generally, this problem can be considered as a Vehicle Routing Problem (VRP) with Multiple Time Windows and Multiple Locations, in which the latter feature has not been widely studied. The main contributions include: (1) the electric vehicle cleaning model is first studied and (2) the Multiple Time Windows and Multiple locations are simultaneously considered. We first construct an integer programming model for this problem with the objective to minimize the total cleaning costs and travel distance. CPLEX is called to solve this problem with small size instances. As this problem is an extension of VRP, which has been noted NP-hard (Dror et al. 1994), we then plan to apply ant colony algorithm hybridized with 2-opt heuristics for large size instances.**

I. Introduction

Electric vehicles are environmentally friendly transports in the 21st century. With the rapid development of the world economy, energy shortages and environment crisis have become a major problem in human development and survival. The fuel is disposable and could not be renewable after consuming. With the increase of global energy consumption, the earth's mineral energy has been depleting. The environmental problem is also increasingly prominent in the world around the city. Air pollutants occupy about 40% to 70% from the internal combustion engine vehicle emissions. Air pollution caused by automobile emissions is also very serious problem in China. A survey conducted by a panel of experts from the World Bank points out that the pollution situation in major cities is currently most serious all over the world, and 16 of the world's 20 most polluting cities are in China (Gentle 2016). Shown by the large city survey done in North China and East China, 70% of air pollution comes from the car's exhaust emissions (Guo et al. 2014). Fortunately, electric vehicles relying on electric energy do not release harmful gasses in the process of driving. For these reasons, electric vehicles have become an option to solve energy and environmental problems.

Nowadays, electric vehicles rental companies are constantly emerging. In Shanghai, there is a professional company to carry out leasing and sharing for electric vehicles. It is mainly engaged in electric vehicles for group users (B2B, B2B2C) and private users (B2C) of rental service in a short or long time. Electric Vehicle Rental Project (*EVRP*) is one of a business of the Shanghai company, which focuses on carrying out the electric vehicle time-sharing leasing project. *EVRP* is different from the traditional form of vehicle rental, which achieves a new vehicle time-sharing rental service model with the Internet of Things technology. Customers can book electric vehicles at any time. Comparing with the traditional vehicle rental, electric vehicle rental has some characteristics such as meeting the requirement of short-distance travel, high efficiency about vehicle time-sharing, multiple distribution points, and reducing the ownership of motor vehicle. At present, the international community has obtained some successful cases and experiences.

A process of *EVRP* rental for a customer is summarized as follows. (1) Register: a customer registers on the mobile phone application or the official website to become a member before renting an electric vehicle. (2) Rent: a member schedules to book the time to take an electric vehicle on the mobile client or the website. (3) Return: a member must dock the electric vehicle and return it to the designated parking lot with charging devices.

EVRP has facilitated trips for everyone. But in the process of using a leased electric vehicle, it will bring about the hygienic pollution of the electric vehicle. Cleaning electric vehicles at the right time to keep electric vehicles clean is an important issue that *EVRP* needs to consider. If the electric vehicle is dirty, it will cause customers to give up leasing. On the one hand, the company will lose business and reduce revenue. On the other hand, the satisfaction of customer will reduce. High standards of electric vehicle cleaning will reduce such incidents and improve the market competitiveness of enterprises.

We call it cleaning department who takes the responsibility for cleaning the electric vehicle in *EVRP*. Based on the electric vehicle rental process, it is clear that the cleaning department can not clean the electric vehicle when a member is using the electric vehicle. The cleaning department can only clean the

electric vehicle when becomes idle. In addition, an electric vehicle may dock in the different parking lots when it is in different non-rented period.

According to the analysis of *EVRP* cleaning process, we assume that the non-rented time windows of electric vehicles can be given ahead. Hence the cleaning problem can be noted as a Vehicle Routing Problem (*VRP*). The details of the cleaning problem are described as follows. For a certain period of time, the goal of the cleaning trolley is to clean all electric vehicles. Each electric vehicle can be only cleaned once and the cleaning of electric vehicles must be in the non-rented time windows (corresponding to parking lots information). The cleaning trolley can reach in advance but needs to wait. The optimization goal of the problem is to choose an appropriate path to minimize the total costs of the cleaning process.

To the best of our knowledge, we are the first to establish a model of Vehicle Routing Problem with Multiple Time Windows Corresponding to Multiple Locations (*VRPMTW-ML*). The contributions of this paper include: (1) we present a detailed understanding of the process of *EVRP* and electric vehicles cleaning; (2) based on the difference between the cleaning trolley routing problem and the traditional *VRP*, electric vehicles in different non-rented time windows corresponding to different locations and multiple locations constraints are considered; (3) a *VRPMTW-ML* mathematical model is established.

The remainder of this paper is organized as follows. Section II gives a brief literature review. In section III, we describe the concerned problem and propose a mathematical formulation. Section IV states an example to reveal the process of *VRPMTW-ML*. Section V presents a hybrid ant colony algorithm with 2-opt. Finally, Section VI concludes the work.

II. Literature Review

Since Danzig and Ramser proposed *VRP* in 1959, *VRP* in the operations research, computer, logistics, and transportation or other fields has been widely studied and applied. The Vehicle Routing Problem with Time Windows (*VRPTW*) is based on the *VRP* with the time windows in which the customers request to access. *VRP* is NP-hard (Dror et al. (1994)), both *VRPTW* and *VRPMTW* (Vehicle Routing Problem with Multiple Time Windows) are also NP-hard problems.

Many researchers have considered the *VRP* in different scenarios such as single-vehicle, time dependent, multiple-criterion, multiple trips, multiple visits, synchronization and precedence constraints, split deliveries, multiple depot and congestion charge. (e.g., Azi et al. (2007); Balseiro et al. (2011); Belhaiza et al. (2016); Cattaruzza et al. (2016); Daniela et al. (2013); Haddadene et al. (2014); Ho et al. (2004); Polacek et al. (2008); Wen et al. (2015); etc.).

At present, there are relatively few studies about *VRPMTW*. Peng et al. (2012) proposed the time-dependent vehicle routing problem with models of multiple time windows that satisfies First-In First-Out property with constraints of multiple time windows, treated departure times as decision variables and used a three-phase hybrid heuristic to solve it. Daniela et al. (2013) presented a mathematical formulation of *VRPMTW* with account periodic constraints, proposed and implemented an algorithm based on ant colony system to solve it.

Belhaiza et al. (2014) presented a new hybrid variable neighborhood-tabu search heuristic for the *VRPMTW*. They also proposed the minimum value of the backward time slack algorithm for a multiple time windows environment. Then, Belhaiza et al. (2014) presented a hybrid tabu search variable neighborhood (*HVNTS*) heuristic that chose solutions of Pareto non-dominated for satisfying a set of Nash equilibrium conditions for a multiple-agent game theory model. It applied to the *VRPMTW* and tested on three objectives: minimizing the total travel cost (expressed in time units), maximizing the minimal customers' utility, and maximizing the minimal drivers' utility. What's more, Belhaiza (2016) presented a new framework which used a hybrid variable neighborhood tabu search heuristic and chose Pareto non-dominated solutions from the search space of solutions that satisfy a set of Nash equilibrium conditions. They built a multiple-agent game theory model for the optimization of real-life multiple-objective *VRPMTW*.

Summing up, from some existing literature, we can describe *VRPMTW* as follows. A vehicle from a site to serve customers returns to the site when the service is completed (see the left of Figure 1). Since a customer may have multiple non-overlapping time windows, the vehicle must choose one and only one time window to serve the customer. Meanwhile, each customer can only be served by a vehicle and only once. The goal of the optimization is to construct the appropriate path so that the total costs to complete all the requirements is the smallest (or maximize total profit) under the condition of certain satisfying constraints.

However, the current model for *VRPMTW* based on the customer location is always same in the literature (see the left of Figure 1, the location of customer 2 is always same). In practice, the customer may be at different locations in different time windows (see the right of Figure 1, customer 2 has three different locations). Therefore, Vehicle Routing Problem with Multiple Time Windows corresponding to Multiple Locations (*VRPMTW-ML*) has a theoretical and practical significance.

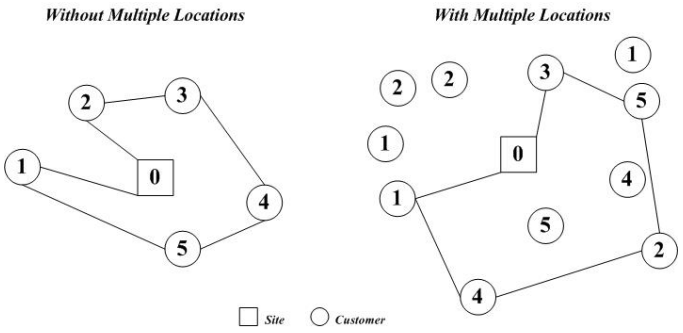


Fig. 1. *VRPMTW* without/with Multiple Locations

III. Problem Statement and Mathematical Model

A. Problem Statement

The problem tackled in this paper can be described as follows. The *EVRP* has a number of electric vehicles that can provide rental services and a number of parking lots for electric vehicles. These electric vehicles can be leased once or more times a day. Then the non-rented time windows can be

defined as the time when not leased and cleanable. And the non-rented time windows is the charging time, it means that an electric vehicle cannot be used like a combustion powered vehicle in a car sharing system. There are long idle times required, which are used for the charging of batteries. The parking lot for each electric vehicle in different non-rented time windows may be same or not. Based on this situation, we can collect information about non-rented time windows and parking lots for each electric vehicle. What's more, the distance between different parking lots and the travel time of cleaning trolley are given in advance. There is only one cleaning trolley for cleaning work (In this paper, we first consider about one cleaning trolley, situation of more cleaning trolley will be considered in future research.), and the goal is to clean all electric vehicles in a certain period of time. For each electric vehicle, any of the non-rented time windows can be chosen to clean it. For example, the charging time of every electric vehicle each time not less than half an hour, so that the electric vehicle can charge to 80% of the electricity. The non-rented time windows of an electric vehicle is: [8: 30, 10: 30], [13: 30, 15: 30], [17: 30, 18: 30], [22: 00, 22: 30], and the four non-rented time windows corresponding to the parking lots code are 1, 2, 2, 3. The cleaning department must select one in these four non-rented time windows such as [13: 30, 15: 30] to go to the parking lot 2 to clean it (see Figure 2). However, in reality, there may be difficult to clean all electric vehicles, and uncleaned electric vehicles will cause the failure of the rental service next time. In order to ensure that cleaning does not result in loss of rental services, the cleaning department must hire a cleaning company to clean the uncleaned electric vehicles which may result in additional cleaning costs.

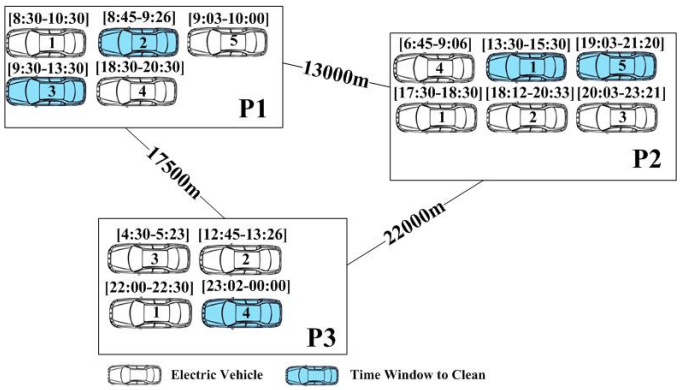


Fig. 2. Problem Statement Describe

It can be seen from the above, the costs of cleaning work consist of two parts. Fixed costs, cleaning trolley costs (such as depreciation, maintenance, etc.) and staff costs (such as wages, benefits, etc.). Variable costs, such as cleaning trolley travel costs between different parking lots, basic cleaning costs for cleaned electric vehicles and additional cleaning costs for uncleaned electric vehicles. Fixed costs are the same every time, and can not be reduced or increased. But for variable costs, it can be optimized. Therefore, we aim to minimize the travel costs and cleaning costs (both for cleaned and uncleaned electric vehicles). Based on the above problem, we establish a mixed integer programming model for *VRPMTW-ML*.

B. Model hypothesis

In order to simplify the problem and establish the model, the following hypothesis are put forward. (1) Cleaning trolley can only clean an electric vehicle at a time. (2) Every electric vehicle is up to cleanliness standard after cleaned. (3) The moving time of the cleaning trolley in the parking lot can be ignored. (4) The cleaning trolley instantaneously arrives when cleaning the first electric vehicle. (5) The cleaning process can not be interrupted. (6) The launch time of cleaning trolley can be ignored.

C. Model formulation

1) Basic Notations: Below we propose a mixed integer programming model of *VRPMTW-ML*. We first define parameters and variables, then present the objective function and relevant constraints.

Indices

i, j : indices of electric vehicles;

n, m : indices of parking lots;

a, b : indices of non-rented time windows;

h : index of the cleaning order of the electric vehicles.

Parameters

I : set of electric vehicles, and $I = \{1, 2, \ldots, |I|\}$;

N : set of parking lots, and $N = \{1, 2, \ldots, |N|\}$;

K : set of the cleaning order of the electric vehicles;

TC_i : the number of time windows for electric vehicle i, where $i \in I$;

e_i^a : the earliest time of electric vehicle i in time window a can be cleaned, where $i \in I, a \in TC_i$;

l_i^a : the latest time of electric vehicle i in time window a can be cleaned, where $i \in I, a \in TC_i$;

loc_i^a : the parking lot of electric vehicle i in time window a, where $i \in I, a \in TC_i$;

t_c : the time of cleaning an electric vehicle;

c_o : the unit distance costs of cleaning trolley travel;

c_c : the costs of cleaning an electric vehicle;

d_{nm} : the distance between the parking lots n and m, where $n, m \in N$;

t_{nm} : the time required for the cleaning trolley to travel between the parking lots n and m, where $n, m \in N$;

P : the additional total costs of extra clean an electric vehicle;

w_i : the weight of additional total costs to clean electric vehicle i, where $i \in I$;

M : a sufficiently large positive integer.

Decision Variables

at_i : the arrival time of cleaning trolley when cleaning electric vehicle i, where $i \in I$;

wt_i : the waiting time of cleaning trolley before cleaning electric vehicle i, where $i \in I$;

H_i : the cleaning order of the electric vehicle i, where $i \in I$;

R_i : =1 if electric vehicle i has been cleaned; =0 otherwise, where $i \in I$;

X_{ih} : =1 if electric vehicle i is cleaned in the hth position; =0 otherwise, where $i \in I, h \in K$;

Y_{ij} : =1 if electric vehicle j is immediately cleaned after electric vehicle i; =0 otherwise, where $i, j \in I$, and $i \neq j$;

Z_i^a : =1 if electric vehicle i is cleaned in the time window a; =0 otherwise, where $i \in I, a \in TC_i$;

λ_{ij} : =1 if electric vehicle i is cleaned earlier than electric vehicle j; =0 otherwise, where $i, j \in I$, and $i \neq j$;

$U_{i,j}^{a,b}$: =1 if electric vehicle i is cleaned in the time window a, and electric vehicle j is cleaned in the time window b; =0 otherwise, where $i, j \in I, a \in TC_i, b \in TC_j$, and $i \neq j$;

$V_{i,j}^{a,b}$: =1 if electric vehicle i is cleaned in the time window a, and electric vehicle j is cleaned in the time window b, and electric vehicle j is following electric vehicle i; =0 otherwise, where $i, j \in I, a \in TC_i, b \in TC_j$, and $i \neq j$.

2) Mathematical Model: With the above consideration, the *VRPMTW-ML* is formulated in a mixed integer programming model as follows.

$$\min\ Z = \sum_{i\in I}\sum_{j\in I}\sum_{a\in TC_i}\sum_{b\in TC_j} V_{i,j}^{a,b}\cdot d_{loc_i^a loc_j^b}\cdot c_o + \sum_{i\in I} R_i\cdot c_c + \sum_{i\in I}(1-R_i)\cdot w_i\cdot P, \quad i\neq j$$

$$at_i \le \sum_{a\in TC_i} Z_i^a \cdot l_i^a, \quad i \in I \tag{1}$$

$$\sum_{a\in TC_i} Z_i^a \cdot e_i^a \le at_i + wt_i, \quad i \in I \tag{2}$$

$$at_i + wt_i \le \sum_{a\in TC_i} Z_i^a \cdot l_i^a, \quad i \in I \tag{3}$$

$$\sum_{a\in TC_i} Z_i^a \le 1, \quad i \in I \tag{4}$$

$$wt_i \ge \sum_{a\in TC_i} Z_i^a \cdot e_i^a - at_i \quad i \in I \tag{5}$$

$$R_i = \sum_{a\in TC_i} Z_i^a, \quad i \in I \tag{6}$$

$$\sum_{i\in I} X_{ih} \le 1, \quad h \in K \tag{7}$$

$$\sum_{h\in K} X_{ih} = R_i, \quad i \in I \tag{8}$$

$$\sum_{i\in I} X_{ih} \ge \sum_{i\in I} X_{i(h+1)}, \quad h \in K \setminus \{|I|\} \tag{9}$$

$$\sum_{h\in K} X_{ih} \ge Y_{ij}, \quad i,j \in I, i \neq j \tag{10}$$

$$Y_{ij} \ge X_{ih} + X_{j(h+1)} - 1, \quad i,j \in I, i \neq j, h \in K\setminus\{|I|\} \tag{11}$$

$$H_i = \sum_{h\in K} h \cdot X_{ih}, \quad i \in I \tag{12}$$

$$H_j \ge H_i - M(1-\lambda_{ij}) - M(2-R_i-R_j), \quad i,j \in I, i \neq j \tag{13}$$

$$H_j \le H_i + M\cdot\lambda_{ij} + M(2-R_i-R_j), \quad i,j \in I, i \neq j \tag{14}$$

$$H_j - H_i \ge 1 - M(2-\lambda_{ij}-Y_{ij}) - M(2-R_i-R_j), \quad i,j \in I, i \neq j \tag{15}$$

$$H_j - H_i \le 1 + M(2-\lambda_{ij}-Y_{ij}) + M(2-R_i-R_j), \quad i,j \in I, i \neq j \tag{16}$$

$$U_{i,j}^{a,b} \le Z_i^a, \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{17}$$

$$U_{i,j}^{a,b} \le Z_j^b, \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{18}$$

$$U_{i,j}^{a,b} \ge 1 - M(2 - Z_i^a - Z_j^b), \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{19}$$

$$V_{i,j}^{a,b} \le Y_{ij}, \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{20}$$

$$V_{i,j}^{a,b} \le U_{i,j}^{a,b}, \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{21}$$

$$V_{i,j}^{a,b} \ge 1 - M(2 - Y_{ij} - U_{i,j}^{a,b}), \quad i,j \in I, i \neq j, a \in TC_i, b \in TC_j \tag{22}$$

$$at_i + wt_i + t_c + \sum_{a\in TC_i}\sum_{b\in TC_j} t_{loc_i^a loc_j^b}\cdot V_{i,j}^{a,b} - M(1-Y_{ij}) - M(2-R_i-R_j) \le at_j, \quad i,j \in I, i \neq j \tag{23}$$

$$at_i, wt_i, H_i \in Z, i \in I \tag{24}$$

$$R_i, X_{ih}, Y_{ij}, Z_i^a, \lambda_{ij}, U_{i,j}^{a,b}, V_{i,j}^{a,b} \in \{0,1\}, \quad i,j \in I, i \neq j, h \in K, a \in TC_i, b \in TC_j \tag{25}$$

The objective function contains two parts. The former part is the travel costs of cleaning trolley when cleaning electric

vehicles. The latter part is cleaning costs, both basic cleaning costs of electric vehicles which cleaning trolley could clean and additional costs of electric vehicles which cleaning trolley could not clean. Constraints (1)-(5) and (23) represent the constraints of multiple time windows. Constraint (1) indicates that the cleaning trolley arrival time does not allow more than the latest start time of the electric vehicle selected non-rented time window. Constraints (2)-(3) indicate that the start cleaning time of the electric vehicle must be between the earliest start cleaning time and the latest start cleaning time of the selected non-rented time window. Constraint (4) indicates that there is one or no time window can be selected as clean time. Constraint (5) indicates if cleaning trolley reach in advance that needs to wait. Constraint (23) indicates the time relationship of cleaned electric vehicles that between the predecessor and subsequently. Constraint (6) indicates that if a non-rented time window is selected, the electric vehicle i is cleaned. Constraint (7) indicates that the hth task position of cleaning trolley may clean one electric vehicle or not. Constraint (8) indicates that if electric vehicle i is cleaned, then it must be one task of cleaning trolley. Constraint (9) indicates that if the $(h+1)$th task position of cleaning trolley clean one electric vehicle, then the hth task position of cleaning trolley clean one electric vehicle necessarily. Constraint (10) indicates that if the cleaning trolley clean the electric vehicles i and j continuously, the electric vehicle i must be cleaned. Constraints (11)-(16) indicate the constraints of cleaning trolley to clean the electric vehicles i and j continuously. Constraints (17)-(19) are the definition of decision variables $U_{i,j}^{a,b}$. Constraints (20)-(22) are the definition of decision variables $V_{i,j}^{a,b}$. Constraints (24)-(25) are the variable ranges.

IV. Computational experiments

In this section, we will give a set of data to test the *VRPMTW-ML* mathematical model built in section III by CPLEX 12.6. The practical significance is that the cleaning trolley needs to clean all electric vehicles that each electric vehicle has multiple time windows and multiple locations in a certain period time. The computational experiments are conducted with the software MATLAB R2014b on a PC with Intel Core i5, 3GHz processors and 4 GB RAM.

A. values of paprameters

To reveal the process of the *VRPMTW-ML* mathematical model on the schedule of tasks for cleaning electric vehicle, an instance with 10 electric vehicles to be cleaned is given. The values of parameters which are randomly generated are listed in Table I and Table II. Table I shows the distance between the two parking lots, the time for the cleaning trolley to travel between the two parking lots, and there are four parking lots. The weight coefficient of each electric vehicle is shown in Table II. If the weight coefficient is greater, the more extra cleaning costs occur when uncleaned. In addition, multiple time windows corresponding to locations of 10 electric vehicles have been shown in Table II. We define the total costs of extra clean an electric vehicle P = 100 (unit: RMB), the time of clean an electric vehicle t_c = 10 (unit: minute), the costs of unit distance for cleaning trolley travel c_o = 0.001 (unit: RMB), the costs of cleaning an electric vehicle c_c = 2 (unit: RMB). By the way, the charging time of every electric vehicle each time not less than half an hour, so that the electric vehicle can charge to 80% of the electricity. So the length of each non-rented time window is more than 30 minutes.

TABLE I. Values of Parameters for Parking Lots

d_{nm}; t_{nm}	1	2	3	4
1	0; 0	22000; 63	19700; 65	26000; 87
2	22000; 63	0; 0	17300; 51	13000; 59
3	19700; 65	17300; 51	0; 0	17000; 76
4	26000; 87	13000; 59	17000; 76	0; 0

TABLE II. Values of Parameters for Electric Vehicles

i	w_i	$[e_i^a, l_i^a; loc_i^a]$
1	0.83	[0, 31; 1], [108, 148; 2]
2	0.74	[3, 44; 4], [130, 166; 3]
3	0.84	[67, 99; 3], [135, 170; 3]
4	0.73	[8, 40; 4], [72, 105; 4], [132, 165; 4]
5	0.8	[56, 101; 1], [181, 212; 2], [250, 282; 2]
6	0.9	[5, 120; 3], [197, 230; 1], [306, 340; 3], [415, 445; 1], [470, 500; 1]
7	1	[2, 74; 1], [147, 177; 2]
8	0.45	[121, 155; 4], [246, 285; 3]
9	0.77	[7, 69; 3], [143, 175; 2], [255, 285; 1], [385,426; 4]
10	0.78	[33, 87; 4], [215, 295; 1], [307, 430; 2]

B. results of instances

According to the result of the instance by CPLEX in Figure 3, we could know that the order of cleaning for the cleaning trolley is the electric vehicle 2, 10, 4, 1, 9, 7, 5, 8, 6, and electric vehicles 3 is uncleaned. The time window and location of the electric vehicles are: [3, 44; 4], [33, 87; 4], [72, 105; 4], [108, 148; 2], [143, 175; 2], [147, 177; 2], [181, 212; 2], [246, 285; 3], [306, 340; 3]. The result shows total costs is 132.3. The travel costs of cleaning trolley between parking lots follows the routing is 30.3 ((13000 + 17300)* 0.001). Basic cleaning costs is 18 (9*2) and additional cleaning costs is 84 (0.84*100) for cleaning electric vehicles 3.

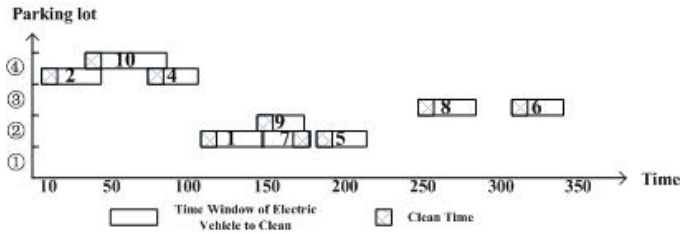


Fig. 3. Result of the instance

The results show that in order to minimize the travel costs of the cleaning trolley, the cleaning trolley clean as many electric vehicles as possible in a parking lot. For example, four electric vehicles (1, 9, 7, 5) are continuously cleaned in the parking lot 2. However, as time windows and the distance, electric vehicle 3 can not be cleaned.

For one thing, the solution consumes more than 5000 seconds when there are 10 electric vehicles and with up to 5 time windows each electric vehicle. Thus, more efficient method is necessary to solve the *VRPMTW-ML*.

V. CONCLUSION

In this paper, we study a Vehicle Routing Problem with Multiple Time Windows and considering the corresponding Multiple Locations. We note electric vehicles cleaning process as a *VRP*, and aim to minimize the travel expense of cleaning trolley and the number of electric vehicles that have not been cleaned. A mixed integer programming model of Vehicle Routing Problem with Multiple Time Windows corresponding to Multiple Locations has been established. Small-scale instances are carried out for the *VRPMTW-ML*.

Future research directions include (1) situation of more cleaning trolleys rather than one cleaning trolley; (2) multiple time windows and multiple locations with uncertainties; and (3) more efficient solution algorithms, such as hybrid ant colony algorithm with 2-opt will be developed for the large-scales.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (71432007, 71531011, 71571134, 71571135, and 71428002), and the Cai Yuanpei Program between the French Ministries of Foreign and European Affairs and the Higher Education and Research and the Chinese Ministry of Education (27927VE). The work was also supported by the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program.

REFERENCES

[1] N. Azi, M. Gendreau, J. Y. Potvin, *An exact algorithm for a single-vehicle routing problem with time windows and multiple routes*. European Journal of Operational Research, 178(3), 755-766, 2007.

[2] S. R. Balseiro, I. Loiseau, J. Ramonet, *An ant colony algorithm hybridized with insertion heuristics for the time dependent vehicle routing problem with time windows*. Computers & Operations Research, 38(6), 954-966, 2011.

[3] S. Belhaiza, *A game theoretic approach for the real-life multiple-criterion vehicle routing problem with multiple time windows*. IEEE Systems Journal, 1-12, 2016.

[4] S. Belhaiza, R. M'Hallah, *A Pareto non-dominated solution approach for the vehicle routing problem with multiple time windows*. IEEE Congress on Evolutionary Computation, 3515-3524, 2014.

[5] S. Belhaiza, P. Hansen, G. Laporte, *A hybrid variable neighborhood tabu search heuristic for the vehicle routing problem with multiple time windows*. Computers & Operations Research, 52, 269-281, 2014.

[6] D. Cattaruzza, N. Absi, D. Feillet, *The multi-trip vehicle routing problem with time windows and release dates*. Transportation Science, 50(2), 177-185, 2016.

[7] F. Daniela, M. Elena, P. Paola, *Ant colony system for a vrp with multiple time windows and multiple visits*. Journal of Interdisciplinary Mathematics, 10(2), 263-284, 2013.

[8] B. G. Dantzig, J. H. Ramser, *The truck dispatching problem*. Management Science, 6(1), 80-91, 1959.

[9] M. Dror, G. Laporte, P. Trudeau, *Vehicle routing with split deliveries*. Discrete Applied Mathematics, 50(3), 239-254, 1994.

[10] P. Gentle, *China's environmental policy and urban development*. Asia Pacific Business Review(2), 2016.

[11] Y. Guo, Z. Wang, H. Kang, C. Gu, G. Wang, X. Zhang, *Impact of automobile vehicles exhaust emissions on metropolitan air quality: analysis study on the air pollution change before and after the spring festival in urumqi city, china*. Huanjing Kexue Xuebao, 34(5), 1109-1117, 2014.

[12] S. R. A. Haddadene, N. Labadie, C. Prodhon, *GRASP for the Vehicle Routing Problem with Time Windows, synchronization and precedence constraints*. International Conference on Wireless and Mobile Computing, NETWORKING and Communications, IEEE, 72-76, 2014.

[13] S. C. Ho, D. Haugland, *A tabu search heuristic for the vehicle routing problem with time windows and split deliveries*. Computers & Operations Research, 31(12), 1947-1964, 2004.

[14] B. Peng, Y. Zhou, *A three-phase hybrid heuristic for time dependent vehicle routing problem with multiple time windows*. Journal of Convergence Information Technology, 7(4), 59-69, 2012.

[15] M. Polacek, S. Benkner, K. F. Doerner, R. F. Hartl, *A cooperative and adaptive variable neighborhood search for the multi depot vehicle routing problem with time windows*. Business Research, 10(2), 613-627, 2008.

[16] L. Wen, R. Eglese, *Minimum cost vrp with time-dependent speed data and congestion charge*. Computers & Operations Research, 56, 41-50, 2015.

Integration of Timetabling, Multi-type Vehicle Scheduling and User Routing in Public Transit Network considering Fuel Consumption

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Ming Liu*, Ying Li†, Feifeng Zheng‡, Feng Chu*§
*School of Economics & Management, Tongji University, China
Email: mingliu@tongji.edu.cn
†Glorious Sun School of Business & Management, Donghua University, China
Email: liying320424@126.com
‡Glorious Sun School of Business & Management, Donghua University, China
Email: ffzheng@dhu.edu.cn
*Laboratory IBISC, University of Evry Val d'Essonne, Evry 91020, France
§Management Engineering Research Center, Xihua University, Chengdu 610039, China
Email: feng.chu@univ-evry.fr

***Abstract*—This paper focuses on the integration of timetabling, multi-type vehicle scheduling and user routing in transit network considering fuel consumption. For the integrated problem, we consider the users' preferences for departure and arrival times and the capacity limits of vehicles. Additionally, the multi-type vehicle and fuel consumption factors are also the key contributions, which are firstly added into the integrated problem. The objective is to minimize the inconvenience of the users (i.e. in-vehicle times, line-transfer penalties and deviation between desired departure and arrival times), penalties of users not served and the cost of line runs and fuel consumption for operators. Finally, we establish a mixed-integer programming model for the integrated problem.**

I. Introduction

In recent decades, the transit network has become huger and huger due to the development of road network and the increasing of vehicles. The complex transit network forms a bigger challenge for the road traffic planning and the vehicle scheduling. Especially for the public transit network, more convenient service for users is an quite important problem, meanwhile, the factors of energy saving and environmental protect are also an unneglected problem.

The transit network timetabling and scheduling problem (TNTSP) includes two subproblems: the transit network timetabling problem (TNTP), vehicle scheduling problem (VSP). The transit network timetabling problem and vehicle scheduling problem in transit network have attached more attention. In this paper, we also consider the user routing problem. For the user routing, we assume that users are interested in minimizing their scheduled inconveniences related to in-vehicle times, line-transfer penalties and deviation between desired departure and arrival time. For the timetabling problem, it needs to consider the potential itinerary and departure and arrival times, then a convenient timetable can be obtained. On the other hand, the convenient timetable can also reduce the waste of public transport resources. For the vehicle scheduling problem, most existing studies assumed that all the vehicles are identical. Laporte et al. (2017) studied the multi-objective integration of timetables, vehicle schedules and user routings, however, they also assumed the identical vehicle factor. In fact, many public transports, such as train, subway or bus, have various configurations or types. Therefore, the speed and fuel consumption are also different, then the various speeds of vehicles will also impact on the users' optimal option for itineraries and the departure times.

In this paper, we focus on the integration of timetabling, multi-type vehicle scheduling and user routing in transit network considering fuel consumption. Our main contributions are as follows. 1) For the integrated problem, we not only pursue transfer coordination but also consider users' preferences for departure and arrival times. 2) Each user need to be treated individually with hard time windows constraints for trip duration, departure and arrival times. 3) We consider multi-type vehicle, vehicle capacity limits and vehicle fuel consumption. 4) We establish a mixed-integer programming model for the integrated problem.

The rest of this paper is organized as follows. In Section 2, a brief literature review of previous work is given. Section 3 describes the problem under consideration, and Section 4 formulates the mathematical model. Finally, some conclusions are given in Section 5.

II. Literature Review

Recent research in transit network has mainly focused on the transit network timetabling problem, vehicle scheduling problem and the integration of the two problems.

The timetabling problem has caught a lot of attention in the field of train and bus scheduling. For the train timetabling problem, Huang (2006) developed ant colony optimization (ACO), considering the trains scheduling in the transit period

between peak period and off-peak period, the conflict resolving and the balance of in and out trains for each depot. Kang et al. (2016) studied the optimization of first train timetables for an urban railway network, which focuses on designing convenient and smooth timetables for morning passengers, then proposed a mixed integer programming (MIP) model for minimizing the number of missed trains. Hassannayebi and Zegordi (2015) focused on the rail rapid transit timetabling problem, aiming at minimizing the total and maximum waiting time of the passengers through optimization of the train timetables for urban rail transit systems. To solve the problem, they developed mixed-integer linear and non-linear programming models, then proposed adaptive and variable neighbourhood search algorithms to tackle large instances. Canca et al. (2016) studied the design and optimization of non-periodic train timetables when passengers demand follows a dynamic behavior along certain planning horizon, then present four different formulations for this problem. Ibarra-Rojas et al. (2015) proposed the multiperiod synchronization bus timetabling (MSBT) problem, which specifies the departure times of the trips of all lines where each line has its own planning periods along the day, with the objective of optimizing synchronization events: maximize passenger transfers and minimize bus bunching along the network.

Compared with the transit network timetabling problem, the research of vehicle scheduling problem has formed rather mature methods and contents. Foster et al. (1976) described an integer programming formulation of the vehicle scheduling problem and illustrated how such a formulation can be extended to incorporate restrictions on work load, coverage and service that occur in real world vehicle scheduling problems. Huisman et al. (2004) focused on dynamic scheduling and considered three measurements throughout the paper, namely the number of vehicles used, the percentage of trips starting late, and a virtual measure for delay costs. Hadjar et al. (2006) considered the multiple depot vehicle scheduling problem (MDVSP) and proposed a branch-and-bound algorithm for solving it that combines column generation, variable fixing, and cutting planes. Haghani and Banihashemi (2012) presented new models for multiple depot vehicle scheduling problem (MDVS) and multiple depot vehicle scheduling problem with route time constraints, considering the route time constraints.

Based on the two aforementioned problems, some literature gradually turn to the integration problem. Ibarra-Rojas et al. (2014) focused the trade-off between the level of service and operating costs in transit system, then presented two integer linear programming models for the timetabling and vehicle scheduling problems, and combined them in a bi-objective integrated model. Ceder (2011) addressed the timetable development and vehicle-scheduling with different vehicles types, and proposed a method to construct timetables with the combination of both even-headway and even-load concepts, then the vehicle-scheduling problem are based on given sets of trips and vehicle types arranged in decreasing order of vehicle cost. Liu and Shen (2007) established a bi-level programming to study the regional bus operating problem. In the upper model, the regional bus vehicle scheduling, which is designed to minimize the total number of the required vehicles and the total time of the deadheading trips, is formulated as a class of the scheduling problem complying strictly with the vehicle chain-running time, the maximum capacity, and the required retaining vehicles in each depot. In the lower model, the objective is to minimize the total transfer time of passengers in every connection stop, and the synchronization coefficient describing the cases of lines-crossing in one connection stop and the satisfaction criteria. Guihaire and Hao (2010) focused on combining important features of these two steps and proposed a simultaneous solution approach to redefine timetables with the objective of bringing improvements to both quality of service and vehicle costs incurred, then introduced an optimization procedure based on Iterated Local Search to solve the problem. Fedorko and Weiszer (2012) presented a selection of crossover operator and determination of maximum number of generations in algorithm for integrated timetable and vehicle scheduling optimization in public transport. Michaelis and Schobel (2009) suggested to reorder the classic sequence of the planning steps: first designing the vehicle routes, then splitting them to lines and finally calculating a (periodic) timetable. Laporte et al. (2017) studied the transit network timetabling and scheduling problem, aiming at determining an optimal timetable for each line of a transit network by establishing departure and arrival times at each station and allocating a vehicle to each timetable.

Some of the above studies only focused on a subproblem, such as timetabling problem or vehicle scheduling problem. In addition, most of the studies about the integration problem ignored the user factor, even assumed that all the vehicles in transit network are identical. In this paper, we develop the integration of timetabling, multi-type vehicle scheduling and user routings, based on Laporte et al. (2017). Specifically, an user can choose different types of vehicles to complete his itinerary. Meanwhile, each user has the desired departure and arrival times.

III. Problem Statement and Mathematical Model

A. Problem Statement

We define a public transportation network (PTN) as a graph $G = (S, A)$ with a set of nodes S representing stations and s set of arcs A. Each arc denotes a direct connection between two stations of S. A directed graph $G_l = (S_l, A_l)$ stands for a public transportation line (PTL), and l belongs to the set of lines $\mathcal{L}$, $S_l \subseteq S$ is the set of stations and $A_l \subseteq A$ contains the direct connections between two stations using line l, the set of direct path lines $\overline{\mathcal{L}} \subseteq \mathcal{L}$ and the set of direct cycle lines $\mathring{\mathcal{L}} \subseteq \mathcal{L}$. Additionally, the set of lines $\mathcal{L}$ can also be divide into two categories: going forward lines set $\overrightarrow{\mathcal{L}}$ and going backward lines set $\overleftarrow{\mathcal{L}}$. Figure 1 shows a PTN with 4 nodes $S = \{1, 2, 3, 4\}$ and 8 arcs $A = \{(1, 2), (2, 1), (2, 4), (2, 3), (3, 2), (3, 4), (4, 2), (4, 3)\}$. We define the set of lines $\mathcal{L} = \{1, 2, 3, 4\}$, and $A_1 = \{(1, 2), (2, 3)\}$, $A_2 = \{(2, 3), (3, 4), (4, 2)\}$, $A_3 = \{(3, 2), (2, 1)\}$, $A_4 = \{(2, 4), (4, 3), (3, 2)\}$. Meanwhile, $\overline{\mathcal{L}} = \{1, 3\}$, $\mathring{\mathcal{L}} = \{2, 4\}$,

$\overrightarrow{\mathcal{L}} = \{1, 2\}$, $\overleftarrow{\mathcal{L}} = \{3, 4\}$. Additionally, users can transfer lines on the nodes (stations) included by no less than two lines, for example, a user departing station 1 along line 1 can transfer to line 4 at their common station 2.

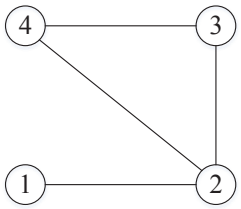


Fig. 1. A PTN with four nodes and eight arcs

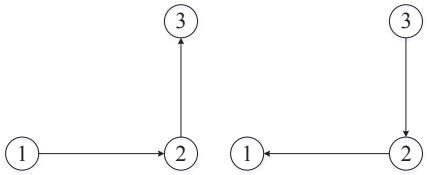


Fig. 2. The going forward line 1 and its opposite direct line 3

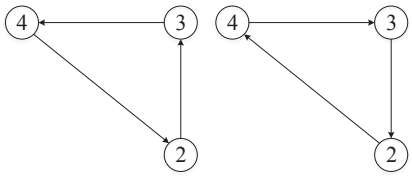


Fig. 3. The going forward line 2 and its opposite direct line 4

For timetabling problem, we first define a line run as the journey of a vehicle making stops for boardings and alightings at every station along the line. Each line is provided two types of vehicle: express-vehicle and general-vehicle, users can choose any one alone this line. The travel time and capacity of two types of vehicles are different, and their cost of implementing a line run in a same line are also different. In addition, when a vehicle allocated to $l \in \overrightarrow{\mathcal{L}}$ can start a line run at any time slot. Once the line run is completed, the associated vehicle becomes part of the fleet size of the line $l + |\overrightarrow{\mathcal{L}}|$ ($l - |\overrightarrow{\mathcal{L}}|$ if $l \in \overleftarrow{\mathcal{L}}$). Circular lines, are similar to path lines except that they have only one terminal station. Therefore, any itinerary that involves traversing the terminal station will require a transfer at that station. Each user has fixed upper and lower bounds associated to the departure and arrival times. Additionally other inconveniences related to in-vehicle times, line-change penalties and deviation between desired departure and arrival times will be taken into account.

Each itinerary offers different travel options according to each combination of the potential timetables from the different lines that can be used for completing a trip. Let Π be the set of itineraries and $\Pi_i \subseteq \Pi$ the subset the itinerary that can be available for user i. Note that $\mathcal{L}_\pi$ is the set of lines related with itinerary $\pi \in \Pi$. Once his path is defined, a user can consider different options of departure times, depending on the combinations of timetables that can be implemented on each line of the path. By $R^e_{i\pi}$ ($R^g_{i\pi}$) we denote the set of options that can be used to serve request i under the itinerary π by a express-vehicle (general-vehicle). We define a strategy as a combination of an itinerary and a potential timetable that a user can choose to travel in a PTN. For the Figure 1, if a user want to travel from station 2 to station 4, he can choose three itineraries: 1) itinerary 1, starting from the station 2 to station 4 along the line 4; 2) itinerary 2, starting from station 2 to station 4 along the line 2; 3) itinerary 3, starting from the station 2 to station 3 along line 1 and transferring to the line 2 at the station 3, then starting the second trip from station 3 to station 4 along line 2. For the itinerary 1 of the user, we will introduce the available options about the departure times. If the traveling time of the express-vehicle and general-vehicle between adjacent stations are 1 and 2 respectively. In addition, t_i^- denotes the earliest time at which user i can start the trip and $t^+_{i+|I|}$ denotes the latest time at which i can reach his destination. If the two parameter are 3 and 8 respectively. Therefore, if user i chooses itinerary 3 by express-vehicle, then the options for departure time and transferring are as follows: 1) starting his trip at time slot 3 and transferring at any of the time slots $\{4, 5, 6, 7\}$; 2) starting his trip at time slot 4 and transferring at any of the time slots $\{5, 6, 7\}$; 3) starting his trip at time slot 5 and transferring at any of the time slots $\{6, 7\}$; 4) starting his trip at time slot 6 and transferring at the time slots 7; so he has 10 options under itinerary 3 by express-vehicle. On the other hand, if he chooses itinerary 3 by general-vehicle, then his options are as follows: 1) starting his trip at time slot 3 and transferring at any of the time slots $\{5, 6\}$; 2) starting his trip at time slot 4 and transferring at the time slots 6; so there are only three options by general-vehicle to carry out his trip in itinerary 3.

The remaining of this paper is based on the following preliminary assumptions.

(1) All express-vehicle (general-vehicle) have the same speed and stopping time at every (non-terminal) station along a line.
(2) Once a line run is completed, the associated vehicle becomes part of the fleet size of the line $l + |\overrightarrow{\mathcal{L}}|$ ($l - |\overrightarrow{\mathcal{L}}|$ if $l \in \overleftarrow{\mathcal{L}}$).
(3) Each user can choose express-vehicle or general-vehicle to carry out his trip, and the express-vehicle has a higher speed than the general-vehicle.
(4) All the users only consider the strategies involving at most one transfer.

B. Model formulation

In this section, we adopt some notions in Laporte et al. (2017). Additionally, our models are also an extension of the models in this paper.

Indices

- s: indices of nodes (stations);
- a: indices of arcs;
- l: indices of lines;
- r: indices of options of departure times for users;
- i: indices of user requests;
- π: indices of itineraries;
- t: indices of time slots;

Input parameters

- G: Graph corresponding to the public bus transit network;
- S: the set of nodes, i.e., public bus stations, $s \in S$;
- A: the set of arcs, $a \in A$;
- $\mathcal{L}$: the set of lines, $l \in \mathcal{L}$;
- S_l: the subset of stations used by line l, $S_l \subseteq S$;
- A_l: the subset of arcs that used by line l, $A_l \subseteq A$;
- $\overline{\mathcal{L}}$: the set of directed path lines, $\overline{\mathcal{L}} \subseteq \mathcal{L}$;
- $\mathring{\mathcal{L}}$: the set of directed cycle lines, $\mathring{\mathcal{L}} \subseteq \mathcal{L}$;
- $\overrightarrow{\mathcal{L}}$: the set of path lines going forward, $\overrightarrow{\mathcal{L}} \subseteq \mathcal{L}$;
- $\overleftarrow{\mathcal{L}}$: the set of path lines going backwards, $\overleftarrow{\mathcal{L}} \subseteq \mathcal{L}$;
- $\mathcal{N}$: the set of stations of all lines, $\mathcal{N} = \{(l,i) : l \in \mathcal{L}, i \in S_l\}$;
- $\mathcal{A}$: the set of arcs of all lines, $\mathcal{A} = \{(l,i,j) : l \in \mathcal{L}, (i,j) \in A_l\}$;
- $\ddot{\mathcal{N}}$: the set of transfer stations;
- $\mathcal{A}^{(tra)}$: the set of transfer edges, $\mathcal{A}^{(tra)} = \{(l,i,n) : l \in \mathcal{L}, i \in S_l, n \in \ddot{\mathcal{N}}\}$;
- T: the set of time slot, $T = \{1, \ldots, |T|\}$, $|T|$: the planning horizon, $t \in T$;
- τ_l^e: the fixed travel time required by a express-vehicle to complete a line run in line l;
- τ_l^g: the fixed travel time required by a general-vehicle to complete a line run in line l;
- Q^e: the capacity of a express-vehicle;
- Q^g: the capacity of a general-vehicle;
- c_l^e: the cost to locate a line run by a express-vehicle in line l, including fuel cost;
- c_l^g: the cost to locate a line run by a general-vehicle in line l, including fuel cost;
- λ^e: the cost of a express-vehicle;
- λ^g: the cost of a general-vehicle;
- ρ: the total available budget to locate line runs by the express-vehicle and general-vehicle;
- $\mathcal{K}^e$: the fleet size of express-vehicle;
- $\mathcal{K}^g$: the fleet size of general-vehicle;
- I: the set of users' transportation requests;
- t_i: the preferred departure times for user request i;
- $t_{i+|I|}$: the preferred arrival times for user request i;
- t_i^+: the earliest times that can serve the user request i;
- $t_{i+|I|}^+$: the latest times that can serve the user request i;
- Π: the set of possible itineraries in the public bus transit network, $\pi \in \Pi$;
- Π_i: the subset of itineraries that user i can choose from, $\Pi_i \subseteq \Pi$;
- $\mathcal{L}_\pi$: the set of lines related the itinerary $\pi \in \Pi$;
- $|\mathcal{L}_\pi|$: the number of lines contained by the itinerary $\pi \in \Pi$;
- $\mathcal{L}_{i\pi}$: the set of lines chosen by user i when itinerary $\pi \in \Pi$ is selected;
- $\mathcal{R}_{i\pi}^e$: the set of options available for user i by express-vehicle when itinerary $\pi \in \Pi$ is selected, $r \in \mathcal{R}_{i\pi}^e$;
- $\mathcal{R}_{i\pi}^g$: the set of options available for user i by general-vehicle when itinerary $\pi \in \Pi$ is selected, $r \in \mathcal{R}_{i\pi}^g$;
- $m_{\pi a}$: a binary parameter equals to 1 if the itinerary $\pi \in \Pi$ contains arc $a \in A$, 0 otherwise;
- $\varphi_{i\pi r}^e$: the cost that user i chooses express-vehicle under the itinerary π and option r;
- $\varphi_{i\pi r}^g$: the cost that user i chooses general-vehicle under the itinerary π and option r;
- $t_{i\pi rl}^e$: the departure time slot of a express-vehicle in line l when user i chooses the itinerary π and the option r;
- $t_{i\pi rl}^g$: the departure time slot of a general-vehicle in line l when user i chooses the itinerary π and the option r;

Decision Variables

- ρ_l^e: the number of line runs to be deployed in line l by express-vehicle;
- ρ_l^g: the number of line runs to be deployed in line l by general-vehicle;
- $\mathcal{K}_l^e$: the number of express-vehicle initially allocated to line l;
- $\mathcal{K}_l^g$: the number of general-vehicle initially allocated to line l;
- x_{lt}^e: a binary variable equals to 1 if a express-vehicle starts a line run in line l at time slot t, 0 otherwise;
- x_{lt}^g: a binary variable equals to 1 if a general-vehicle starts a line run in line l at time slot t, 0 otherwise;
- $y_{i\pi r}^e$: a binary variable equals to 1 if user i chooses express-vehicle under the itinerary π and option r, 0 otherwise;
- $y_{i\pi r}^g$: a binary variable equals to 1 if user i chooses general-vehicle under the itinerary π and option r, 0 otherwise;

Mathematical Model: The model for the problem can be expressed as follows. the objective function:

$$\begin{aligned}\min\{w_1 \sum_{i\in I}\sum_{\pi\in\Pi_i}(\sum_{r\in\mathcal{R}_{i\pi}^e}\varphi_{i\pi r}^e y_{i\pi r}^e + \sum_{r\in\mathcal{R}_{i\pi}^g}\varphi_{i\pi r}^g y_{i\pi r}^g) \\ + w_2\sum_{i\in I}[1 - \sum_{\pi\in\Pi_i}(\sum_{r\in\mathcal{R}_{i\pi}^e} y_{i\pi r}^e + \sum_{r\in\mathcal{R}_{i\pi}^g} y_{i\pi r}^g)] \\ + w_3\sum_{l\in\mathcal{L}}\sum_{t\in T}(c_l^e x_{lt}^e + c_l^g x_{lt}^g) + w_4\sum_{l\in\mathcal{L}}(\lambda^e\mathcal{K}_l^e + \lambda^g\mathcal{K}_l^g)\}\end{aligned} \quad (1)$$

The objective function is to minimize the total user inconvenience. The first term denotes the inconvenience related to in-vehicle times (express-vehicle or general-vehicle), the penalties of changing lines and deviation between desired departure and arrival time, which are all about the users who have been served in transit network. The second term computes

the penalty cost of users not assigned to any service in transit network. The third term computes the cost of line runs and the last term computes the cost of fleet size. The first and second terms are user-oriented, while the the third and last terms are operator-oriented.

$$\sum_{\pi\in\Pi_i}(\sum_{r\in\mathcal{R}^e_{i\pi}} y^e_{i\pi r}+\sum_{r\in\mathcal{R}^g_{i\pi}} y^g_{i\pi r})\le 1,\quad i\in I \tag{2}$$

Constraint (2) ensures each user i can choose no more than one strategy, specifically, a user can only choose more than one combination of an itinerary and a potential timetable of express-vehicle or general-vehicle vehicle.

$$\sum_{t\in T} x^e_{lt}\le \rho^e_l,\quad l\in\mathcal{L} \tag{3}$$

$$\sum_{t\in T} x^g_{lt}\le \rho^g_l,\quad l\in\mathcal{L} \tag{4}$$

Constraint (3) guarantees no more than ρ^e_l line runs by express-vehicle can be located on each line l. Similarly, Constraint (4) guarantees no more than ρ^g_l line runs by general-vehicle can be located on each line l.

$$|\mathcal{L}_\pi| y^e_{i\pi r}\le \sum_{l\in\mathcal{L}_\pi} x^e_{lt^e_{i\pi rl}},\quad i\in I,\pi\in\Pi_i,r\in\mathcal{R}^e_{i\pi} \tag{5}$$

$$|\mathcal{L}_\pi| y^g_{i\pi r}\le \sum_{l\in\mathcal{L}_\pi} x^g_{lt^g_{i\pi rl}},\quad i\in I,\pi\in\Pi_i,r\in\mathcal{R}^g_{i\pi} \tag{6}$$

Constraint (5)-(6) ensure no request will be allocated to a strategy that cannot be carried out with the available line runs. i.e., if the right term equals to 0, then the left term must be 0;

$$\sum_{l\in\mathcal{L}}(c^e_l\rho^e_l+c^g_l\rho^g_l)\le\rho \tag{7}$$

Constraint (7) guarantees the total cost of incurred by all the line runs by express-vehicle and general-vehicle cannot exceed the total available budget.

$$\sum_{i\in I}\sum_{\pi\in\Pi_i:l\in\mathcal{L}_\pi}\sum_{r\in\mathcal{R}^e_{i\pi}:t^e_{i\pi rl}=t} m_{\pi a}y^e_{i\pi r}\le Q^e x^e_{lt}, \tag{8}$$
$$l\in\mathcal{L},a\in A_l,t\in T$$

$$\sum_{i\in I}\sum_{\pi\in\Pi_i:l\in\mathcal{L}_\pi}\sum_{r\in\mathcal{R}^g_{i\pi}:t^g_{i\pi rl}=t} m_{\pi a}y^g_{i\pi r}\le Q^g x^g_{lt}, \tag{9}$$
$$l\in\mathcal{L},a\in A_l,t\in T$$

Constraint (8)-(9) consider the capacity limit of the express-vehicle and general-vehicle, ensuring the total number of users use an same arc by the same type vehicle that cannot exceed the vehicle capacity of the express-vehicle or general-vehicle.

$$\sum_{l\in\mathcal{L}_\pi}(x^e_{lt^e_{i\pi r_1 l}}+x^g_{lt^g_{i\pi r_2 l}})\le(|\mathcal{L}_\pi|-1) \tag{10}$$
$$+\sum_{\pi'\in\Pi_i}\sum_{r'\in\mathcal{R}^e_{i\pi}} y^e_{i\pi' r'},$$
$$i\in I,\pi\in\Pi_i,r_1\in\mathcal{R}^e_{i\pi},r_2\in\mathcal{R}^g_{i\pi}$$

$$\sum_{l\in\mathcal{L}_\pi}(x^e_{lt^e_{i\pi r_1 l}}+x^g_{lt^g_{i\pi r_2 l}})\le(|\mathcal{L}_\pi|-1) \tag{11}$$
$$+\sum_{\pi'\in\Pi_i}\sum_{r'\in\mathcal{R}^g_{i\pi}} y^g_{i\pi' r'},$$
$$i\in I,\pi\in\Pi_i,r_1\in\mathcal{R}^e_{i\pi},r_2\in\mathcal{R}^g_{i\pi}$$

Constraint (10)-(11) ensure that if there exists a given itinerary and a vehicle type for a given user, then the user must be served.

$$\sum_{l\in\mathcal{L}_\pi}(x^e_{lt^e_{i\pi r_1 l}}+x^g_{lt^g_{i\pi r_2 l}})+\sum_{\pi'\in\Pi_i}\sum_{r'\in\mathcal{R}^e_{i\pi}:\varphi^e_{i\pi' r'}>\varphi^e_{i\pi r_1}} y^e_{i\pi' r'} \tag{12}$$
$$+\sum_{\pi'\in\Pi_i}\sum_{r'\in\mathcal{R}^g_{i\pi}:\varphi^g_{i\pi' r'}>\varphi^g_{i\pi r_2}} y^g_{i\pi' r'}\le|\mathcal{L}_\pi|,$$
$$i\in I,\pi\in\Pi_i,r_1\in\mathcal{R}^e_{i\pi},r_2\in\mathcal{R}^g_{i\pi}$$

Constraint (12) guarantees that if there exists a given itinerary and a vehicle type for a given user, then the user will not choose another combination of itinerary and vehicle type with higher cost.

$$\sum_{t'=1}^{\tau^e_l} x^e_{lt'}\le\mathcal{K}^e_l,\quad l\in\mathcal{L} \tag{13}$$

$$\sum_{t'=1}^{\tau^g_l} x^g_{lt'}\le\mathcal{K}^g_l,\quad l\in\mathcal{L} \tag{14}$$

Constraint (13) ensures that the total number of line runs by express-vehicle cannot exceed the fleet size of express-vehicle in line l during τ^e_l, which is time required to complete a line run along line l. Specifically, during the time interval $[1,\tau^e_l]$, there are no more than $\mathcal{K}^e_l$ express-vehicle. Similarly, no more than $\mathcal{K}^g_l$ general-vehicle can be located on a same arc simultaneously.

$$\sum_{t'=1}^{t} x^e_{lt'}-\sum_{t'=1}^{t-\tau^e_l} x^e_{l+|\overrightarrow{\mathcal{L}}|,t'}\le\mathcal{K}^e_l, \tag{15}$$
$$l\in\overrightarrow{\mathcal{L}},t\in[\tau^e_l,|T|-\tau^e_l]$$

$$\sum_{t'=1}^{t} x^g_{lt'}-\sum_{t'=1}^{t-\tau^g_l} x^g_{l+|\overrightarrow{\mathcal{L}}|,t'}\le\mathcal{K}^g_l, \tag{16}$$
$$l\in\overrightarrow{\mathcal{L}},t\in[\tau^g_l,|T|-\tau^g_l]$$

Constraint (15) (Constrain (16)) describes the difference between the number of express-vehicle (general-vehicle) that have departed from line l and the number of express-vehicle (general-vehicle) that have arrival at l cannot exceed the fleet size of express-vehicle (general-vehicle) in line l.

$$\sum_{t'=1}^{t} x^e_{lt'}-\sum_{t'=1}^{t-\tau^e_l} x^e_{l-|\overleftarrow{\mathcal{L}}|,t'}\le\mathcal{K}^e_l, \tag{17}$$
$$l\in\overleftarrow{\mathcal{L}},t\in[\tau^e_l,|T|-\tau^e_l]$$

$$\sum_{t'=1}^{t} x^g_{lt'}-\sum_{t'=1}^{t-\tau^g_l} x^g_{l-|\overleftarrow{\mathcal{L}}|,t'}\le\mathcal{K}^g_l, \tag{18}$$
$$l\in\overleftarrow{\mathcal{L}},t\in[\tau^g_l,|T|-\tau^g_l]$$

Constraint (17) (Constrain (18)) ensures the difference between the number of express-vehicle (general-vehicle) that have arrival at line l and the number of express-vehicle (general-vehicle) have departed from line l that cannot exceed the fleet size of express-vehicle (general-vehicle) in line l.

$$\sum_{l\in\mathcal{L}} \mathcal{K}_l^e \leq \mathcal{K}^e \tag{19}$$

$$\sum_{l\in\mathcal{L}} \mathcal{K}_l^g \leq \mathcal{K}^g \tag{20}$$

Constraint (19) (Constrain (20)) states the total number of express-vehicle general-vehiclein all lines, i.e. the total fleet size of express-vehicle general-vehicle, cannot exceed the number of available express-vehicle general-vehicle.

$$\rho_l^e, \rho_l^g \in Z^+, \quad l \in \mathcal{L} \tag{21}$$

$$\mathcal{K}_l^e, \mathcal{K}_l^g \in Z^+, \quad l \in \mathcal{L} \tag{22}$$

$$x_{lt}^e, x_{lt}^g \in \{0,1\}, \quad l \in \mathcal{L}, t \in T \tag{23}$$

$$y_{i\pi r}^e \in \{0,1\}, \quad i \in I, \pi \in \Pi_i, r \in \mathcal{R}_{i\pi}^e \tag{24}$$

$$y_{i\pi r}^g \in \{0,1\}, \quad i \in I, \pi \in \Pi_i, r \in \mathcal{R}_{i\pi}^g \tag{25}$$

IV. Conclusion

The transit network planning can't ignore users' preference and environmental factor. Therefore, this paper focuses on the integration of timetabling, multi-type vehicle scheduling and user routing in transit network considering fuel consumption. For the integrated problem, we not only pursues transfer coordination but also users' preferences for departure and arrival times. Each user need to be treated individually with hard time windows constraints for trip duration, departure and arrival times. Due to the capacity limits of vehicles, there may be some users not assigned to any service in transit network. Additionally, the multi-type vehicle and fuel consumption factors are also the key contributions. The objective is to minimize the inconvenience of the users (i.e. in-vehicle times, line-transfer penalties and deviation between desired departure and arrival times), penalties of users not served and the cost of line runs and fuel consumption for operators.

For the integrated problem, we also establish a mixed-integer programming model for the integrated problem and solve it via CPLEX for small-size instances. From some numerical experiments, we get some conclusion: 1) If the express-vehicle and general-vehicle are identical, the problem in this paper will degenerate into the problem proposed in Laporte et al. (2017). 2) The express-vehicle with a higher speed have a higher efficiency, then the fuel consumption of the former is less from the long term. 3) When the difference between the ticket price of the two types of vehicle is not big enough, users will be more partial to the express-vehicle.

Acknowledgment

This work was supported by the National Nature Science Foundation of China (71432007, 71531011, 71571134, 71571135, and 71428002), and the Cai Yuanpei Program between the French Ministries of Foreign and European Affairs and the Higher Education and Research and the Chinese Ministry of Education (27927VE). The work was also supported by the Fundamental Research Funds for the Central Universities, and Shanghai Pujiang program.

References

[1] D. Canca, E. Barrena, G. Laporte, L. Coelho, *Rail rapid transit timetabling problem under dynamic passenger demand: Exact formulations*. Or2016 International Conference of the German Operations Research Society, 2016.

[2] A. Ceder, *Optimal multi-vehicle type transit timetabling and vehicle scheduling*. Procedia - Social and Behavioral Sciences, 20(6), 19-30, 2011.

[3] G. Fedorko, M. Weiszer, *Configuring the parameters of multi-objective evolutionary algorithm for integrated timetabling and vehicle scheduling in public transport*. Carpathian Logistics Congress, 1-6, 2012.

[4] B. A. Foster, D. M. Ryan, *An integer programming approach to the vehicle scheduling problem*. Journal of the Operational Research Society, 27(2), 367-384, 1976.

[5] V. Guihaire, J. K. Hao, *Transit network re-timetabling and vehicle scheduling*. Communications in Computer & Information Science, 14, 135-144, 2008.

[6] A. Hadjar, O. Marcotte, F. Soumis, *A branch-and-cut algorithm for the multiple depot vehicle scheduling problem*. Operations Research, 54(1), 130-149, 2006.

[7] A. Haghani, M. Banihashemi, *Heuristic approaches for solving large-scale bus transit vehicle scheduling problem with route time constraints*. Transportation Research Part A Policy & Practice, 36(4), 309-333, 2002.

[8] E. Hassannayebi, S. H. Zegordi, *Variable and adaptive neighbourhood search algorithms for rail rapid transit timetabling problem*. Computers & Operations Research, 2015.

[9] J. Huang, *Using ant colony optimization to solve train timetabling problem of mass rapid transit*. Advances in Intelligent Systems Research, 2006.

[10] Huisman D, Freling R, Wagelmans APM , *A robust solution approach to the dynamic vehicle scheduling problem*. Transportation Science, 38(4), 447-458, 2004.

[11] O. J. Ibarra-Rojas, R. Giesen, Y. A. Rios-Solis, *An integrated approach for timetabling and vehicle scheduling problems to analyze the trade-off between level of service and operating costs of transit networks*. Transportation Research Part B Methodological, 70(7), 35-46, 2014.

[12] O. J. Ibarra-Rojas, F. Lopez-Irarragorri, Y. A. Rios-Solis, *Multiperiod synchronization bus timetabling*, 2015.

[13] L. Kang, X. Zhu, H. Sun, J. Puchinger, M. Ruthmair, B. Hu, *Modeling the first train timetabling problem with minimal missed trains and synchronization time differences in subway networks*. Transportation Research Part B, 93, 17-36, 2016.

[14] G. Laporte, F. A. Ortega, M. A. Pozo, J Puerto, *Multi-objective integration of timetables, vehicle schedules and user routings in a transit network*. Transportation Research Part B, 98, 94-112, 2017.

[15] Z. G. Liu, J. S. Shen, *Regional bus operation bi-level programming model integrating timetabling and vehicle scheduling*.Systems Engineering-Theory & Practice, 27(11), 135-141, 2007.

[16] M. Michaelis, A. Schobel, *Integrating line planning, timetabling, and vehicle scheduling: a customer-oriented heuristic*. Public Transport, 1(3), 211, 2009.

Rich Vehicle Routing Problem with loading constraints and resources synchronization.

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Mayssoun Messaoudi*‡, Issmail Elhallaoui‡*, Mohamed Anouar Jamali§, Louis-Martin Rousseau†*, Adil Tahir*‡,
*Mathematics and Industrial Engineering Department, Ecole Polytechnique de Montréal.
Montréal, Québec, Canada H3C 3A7
Email: mayssoun.messaoudi@gmail.com
†Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT).
Montréal, Québec, Canada H3C 3A7
‡Group for Research in Decision Analysis (GERAD)
Montréal, Québec, Canada H3C 3A7
§SNTL Group, 10050 Rabat, Morocco

***Abstract*—In this paper we consider a real-life combinatorial problem arising in logistics and transportation applications. Our overall work is aimed at developping a decision support tool for a Third Party Logistics provider. We are interested in optimizing both the routing of vehicles and the loading of goods. The separate problems are known to be NP-hard and it's very difficult to combine them in practice. We propose to solve a new variant of rich vehicle routing problem which incorporates real endogenous and exogenous constraints. The mathematical model is solved using column generation approach, where the subproblem is a shortest path problem with resource constraints. The numerical results based on a real test case are discussed.**

I. Introduction

The supply chain is an integration of interdependent links that move goods and services from the supplier to the end customer while responding to the cost-quality-time triad. The economic reality of the markets and the increasing complexity of value chains have imposed the growth of new practices allowing the efficient management of operations and flows. In this perspective the role of logistics providers has grown. Indeed, it acts as a service integrator within an extended value chain, with a multitude of stakeholders with different processes, constraints and objectives. SNTL Group, a Moroccan Third Party Logistics (3PL) Provider, plays a leading role in the African ecosystem by offering a range of value added services in supply chain management. It manages the global logistics of a wide range of customers operating in different sectorsIt, throug a logistics platform of 28 ha, and a fleet of 300 vehicles.

However, in a context of pull flow, transport optimization has become a major concern in commerce and industry. Indeed, customers demand frequent deliveries, in small quantities, while imposing specific constraints such as time windows, adapted vehicles, or even the type of packaging. Moreover, in an emerging country such as Morocco, logistics costs account for 20% of GDP, of which 30% is represented by the cost of road transport, and therefore the integration of adapted and resilient optimization solutions proves to be difficult. That said, the number of contributions in the literature shows that the transport networks optimization has been a concern for the scientific community for six decades.

Our objective is to maximize the profit of the logistics provider, case of our study, through an integrated resolution of a rich vehicle routing problem (VRP) with loading constraints and a synchronization of resources. Our contribution essentially lies in the structure of attributes that faithfully reflect the complex reality of operations, the integrated resolution approach, the consideration of business rules and human solutions in the proposed model, and ultimately providing useful solutions in industrial environment and real life applications.

The popularity of vehicle routing problems is mainly due to their wide application in logistics, and in passenger and freight transport. First introduced by [1] and generalized five years later by [2], the VRP is one of the major problems of operational research and is the subject of an intensive literature. The classical VRP is an NP-hard problem in combinatorial optimization problem [3]. A classical VRP is used to design minimal cost routes for a fleet of identical vehicles such that each customer is visited exactly once, the capacity is respected and each vehicle starts and ends its route at the depot. VRP problems have become increasingly complex by integrating difficult constraints from the practical environment [4]. Exact optimization methods are difficult to solve those problems in acceptable CPU times, hence, heuristics and meta-heuristics are most often used. In order to incorporate real-life complexities such as time windows, heterogonous fleet, time-dependent travel time, etc., several variants of VRP have emerged, which are listed and widely detailed in [5].

The VRP with loading constraints reflects the reality in industrial loaders and logistics providers, which is why several contributions have addressed this class of problems since $2000s$, by looking for optimization methods able of solving realistic problems, and integrating several attributes. The VRP with two-dimensional loading constraints (2L-VRP) deals with the problem integrating two issues: Construction of

vehicles routes, and feasible allocation of transported goods within the vehicle loading area. The objective is to build feasible routes to visit a sequence of customers, so that the capacity of the vehicles is respected, customer requests are met and the loading plan is feasible. In view of the complexity of the problem, the loading problem is often considered as a subproblem, solved using heuristics. Another approach uses exact methods of the type *branch-and-bound* and the lower bounds found in the literature [6], [7]. Viewing the extreme difficulty of the integrated problem, exact methods like *branch-and-cut*, or *branch-and-price* are very restricted in the literature, while heuristic approaches are more encountered [8].

Over the past decades, column generation (CG) has gained considerable popularity in solving practical decision-making problems. A wide variety of vehicles routing has been successfully solved by this method whose applications are reported in [9]. The CG is an exact method of solving large linear problems which are often decomposed into a *master problem* (MP) and a *subproblem* (SP) called also *pricing*, it's also known as *branch-and- price* when embedded in a *branch-and-bound*.

We find that in most of the applications, the MP of the column generation method is formulated as set partitioning where the variables refer to vehicles routes, generated by one or more several SP corresponding to a shortest path problem with resource constraints (SPPRC) which is usually solved using a dynamic programming algorithm. The SPPRC is defined on a directed graph [10], it consists in finding a least-cost path from the source to the sink, while respecting the constraints on resource consumption. According to [11], the GC owes its effectiveness to the SPPRC that allows the modeling of the difficult constraints. Several algorithms assure an effective resolution of the problem. The dynamic programming algorithm introduced by [12] solves the SPPRC in three steps : i) Path extension step, ii) filtering step, ii) dominance step.

Although the vehicle routing problems are widely studied, the solutions proposed don't tackle industrial problems as a whole, but consider in most cases simplified and truncated problems.

This article is composed of 5 sections. First, we present a description of the problem arising in a real-life logistics privider (section 2). Next, we present our mathematical model (section 3). The branch-and-price solution method is discussed (section 4) and the numerical results are presented (section4). Finally, concluding remarks are discussed in the last section.

II. Problem description

A logistics provider needs to satisfy the demand of its customers located through a national scale, which are served by a heterogeneous fleet of type $k \in K$ from the regional logistics platform $\{0\}$. Profit is made if the daily orders are satisfied while respecting the capacity of the vehicles, the time windows and the loading plan. The latter is feasible if it allows feasible positioning of goods inside the vehicle, using the appropriate resources and taking into account the sequence of visits and also the specific characterizations of items related to shape, fragility, stability, orientation, etc.

A range of endogenous and exogenous constraints is encountered in practice, that penalize not only the provider's operational efficiency but also its profit.

The first classical constraint is the time windows, which means that the customer must be visited only during a time slot he chooses. In addition, household appliance customers require that their orders be prepared at specific time windows, resulting in off-peak hours between preparation orders and a resources immobilization.
According to practice and business rules, the fleet is heterogeneous in terms of type, capacity and costs. In addition, some of the vehicles belong to external transporters.
The trucks are loaded manually, guided by the human reflex and experience feedback. Loading the transported goods should take into account their different dimensions, shapes and characteristics. Moreover, this complex operation can begin only if the orders to be delivered are prepared, so the departure time of vehicles varies according to the destination and the availability of resources. Consequently, significant costs of waiting and resources immobilization are generated, which penalizes efficiency and profit. Hence the need for synchronization of resources becomes primordial.
Finally, we notice that in our case, customer demand is received during the operational day. The commands are prepared in real time, and then the vehicles are assigned manually based on a usual mass flow and common business rules.

After analyzing the fundamental problems the logistics provider faces on a daily basis, we propose in the following a new operational research model capable of improving the operational performance and the financial profitability of this actor.

III. Model formulation

Our objective is summarized as follows: to maximize the profit of the vehicles routing using integrated resolution approach of two difficult problems: VRP with time windows and heterogeneous fleet, and a two-dimensional loading problem with resources synchronization.

Customers are divided into two subsets: N_O indicates potential clients with $[a_i, b_i]$ time windows, and N_C indicates optional clients with no time windows, indeed, if their insertion in the route is not profitable their supply can be postponed without exceeding time period of $[T, T+2]$.

A k -type vehicle is characterized by a capacity q_k defined by $(length\ L, width\ l, height\ H, weight\ p)$. A demand of m items is characterized by a weight $p = \sum_{j \in m} p_j$ and a volume $v = \sum_{j \in m} l_j \times L_J \times H_j$.

A. Notations

We introduce the notations used in the mathematical modeling of the problem.

- c_r: Cost of route $r \in R$;
- d_i: Demand of client $i \in N$;
- E: Fixed size loading crew;
- K: Set of vehicle type k;

- N: Set of customers;
- $N_C \subset N$: Subset of optional customers;
- $N_O \subset N$: Subset of potential customers;
- Ω: Set of feasible routes;
- q_k: Capacity of k-type vehicles ;
- s_i: Service time of customer i;
- T: Operational day;
- t: Time period of T;

B. Master Problem

The master problem (P) is a set partitioning model with side constraints. Considering that a demand of volume v requires t loading time units and E_t crews, this model allows to manage the resources and to plan the starting times of loading operations and departure for each vehicle. We will also be able to evaluate the waiting times at the sub-problem level.

1) Variables:

$$\theta_r = \begin{cases} 1 & \text{if } r \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

$$\zeta_i = \begin{cases} 1 & \text{if } i \in N_C \text{ is not visited} \\ 0 & \text{otherwise} \end{cases}$$

$$e_t = \begin{cases} \text{Missing resources at } t \text{ if } e_t \geq 0 \\ \text{Excess resources at } t \text{ if } e_t \leq 0 \end{cases}$$

$$f_t = \begin{cases} 0 & \text{if } e_t \leq 0 \\ \neq 0 & \text{otherwise} \end{cases}$$

2) Parameters:

p_r: Profit incurred by r

c_i: Cost of not visiting $i \in N_C$

η_t: Cost of recruiting additional resources at t

a_{ir}: Is equal to 1 if i is visited by r, and 0 otherwise.

E_r^t: Crew assigned to r at t

d_t: Available resources at t

3) Mathematical model: We propose the following integer mathematical model:

$$(P) \quad \max_{\theta,\zeta,e} \sum_{r\in\Omega} p_r\theta_r - \sum_{i\in N_C} c_i\zeta_i - \sum_{t\in\mathbb{T}} \eta_t f_t \tag{3.1}$$

s.c.

$$\sum_{r\in\Omega} a_{ir}\theta_r = 1 \quad , \quad \forall i \in N_O \tag{3.2}$$

$$\sum_{r\in\Omega} a_{ir}\theta_r + \zeta_i = 1 \quad , \quad \forall i \in N_C \tag{3.3}$$

$$\sum_{r\in\Omega} E_r^t\theta_r - e_t = d_t \quad , \quad \forall t \in \mathbb{T} \tag{3.4}$$

$$f_t \geq e_t \quad , \quad \forall t \in \mathbb{T} \tag{3.5}$$

$$f_t \geq 0 \quad , \quad \forall t \in \mathbb{T} \tag{3.6}$$

$$\theta_r = \{0,1\} \quad , \quad \forall r \in \Omega \tag{3.7}$$

$$\zeta_i = \{0,1\} \quad , \quad \forall i \in N_C \tag{3.8}$$

$$e_t \in \mathbb{R} \quad , \quad \forall t \in \mathbb{T} \tag{3.9}$$

The objective function 3.1 maximizes the profit incurred by the feasible route r which is penalized by the cost c_i of the non-visit of an optional customer, as well as by the cost η_t of the recruitment of additional resources. The constraints 3.2 impose that potential customers are visited exactly once, while constraints 3.3 express the choice to visit or not an optional customer. Binding constraints 3.4 manage the use of resources for each period t. The constraints 3.5 and 3.6 allow to penalize the objective only in case of lack of the resources in period t. The choice of binary variables is imposed by 3.7 and 3.8. (LP) is the linear relaxation of (P), obtained by substituting 3.7 and 3.8 respectively by $0 \leq \theta_r \leq 1 \quad \forall r \in \Omega$ and $0 \leq \zeta_i \leq 1 \quad \forall i \in N_C$.

C. SPPRC

The subproblem (SP) arises as Shortest Path Problem with Resource Constraints (SPPRC). It searches for an element of Ω such that:

$$\bar{p}_r = p_r - \sum_{i\in N_O} a_{ir}\pi_i - \sum_{i\in N_C} a_{ir}\sigma_i - \sum_{t\in T} E_r^t\gamma_t \quad > 0$$

The dual variables associated with the 3.2, 3.3 and 3.4 constraints are respectively $(\pi, \sigma, \gamma) \in (R^n \times R^m \times R)$. If all columns have negative reduced cost, the algorithm stops and an optimal solution is obtained for (LP).

The SPPRC is modeled as a pickup problem. Indeed, in our network, we consider that an empty truck leaves the source, checks customers that it can visit until reaching a volume v. The vehicle then leaves the depot with this LIFO loading plan.

1) Network structure: The graph is constructed according to the actual structure of the physical network. By means of a pre-processing (oriented axial delivery, removal of long arcs, appropriate traffic rules, etc.), the obtained graph is acyclic. We proceeded by static clustering of customers in the form of regions. For future work, we will opt for dynamic segmentation using, for example, K-means Clustering algorithm.

There is a SP per k-type of vehicle, modeled on a directed graph $G^k = (V^k, A^k)$, where V and A are non-empty sets of nodes and arcs, respectively. V contains $|N|$ + 2 nodes, one for each customer $i \in N$, (s,t) corresponds to the depot. A contains departure arcs $(s,j), \quad \forall j \in N$, arrival arcs $(i,t), \quad \forall i \in N$ and travel arcs $(i,j), \quad \forall (i,j) \in N$ so that cutomer j can be visited immediately after customer i by a feasible route.
For each vehicle of type k, let R^n be the set of resources with $|R|$ = n. To each node $i \in V^k$, we associate n resource windows $[{a_i}^r, {b_i}^r], \quad r \in R$ limiting the consumption of resources used to reach i.

2) Ressource constraints: As a first step, we consider the following resource constraints:

- Time: This resource allows to control the visiting hours of each customer in order to respect the imposed time windows, including the loading time at the depot, the customer service time and the travel time. Note that this resource is also used to manage temporal accessibility constraints for certain types of vehicles in the city center;
- Working time: this resource calculates the legal time worked by the driver. Volume: This resource handles part of the loading plan. The volume of items must not exceed the carrying capacity of the vehicle.
- Length: This resource calculates the Total length of the tour, and ensures that the driver does not exceed the authorized distance.

3) Extension function: A function f_k^r allows the vehicle k to go from i to j. This extension function is defined by the specific attributes, and it decides to extend or not a partial path to new nodes. The label is extended to i as long as: i) the capacity of the vehicle k does not exceed the volume v_i of the i customer's order; (ii) the daily working time does not exceed 8 h; iii) the vehicle must deliver the customer during the time window if it exists.

The complexity of the algorithm depends on the number of resources, hence the importance of proper use of aggregation and dominance rules to speed up resolution and avoid memory problems.

IV. Solution method

Our problem is suitable to be solved using column generation (CG). For each iteration, the first step consists in solving the restricted master problem (RMP), which makes it possible to obtain a feasible solution for P, and a dual solution which is also optimal for LP if all columns have a positive reduced cost. The second step consists in solving the SP, which will generate new columns enriching the RMP. The algorithm stops if all columns have a negative reduced cost. At the root node of the enumeration tree, if the optimal solution is fractional, separation and evaluation strategy *branch-and-bound* is applied. At each node, we solve the LP by generating columns.

At this stage, we use two well-known column generation based heuristics [13] :

- A diving heuristic (DH) is a depth-first heuristic search in the branch-and-price tree. At each node, a branch is chosen heuristically. The process stops when the solution of the MP is integer.
- A restricted master heuristic (RMH) consists of solving by column generation the MP at the root node. Integrality constraints are then added on the last restricted master problem RMP variables. This integer RMP is then solved by a commercial MIP solver.

V. Numerical results

Our computational study involves a base model without resource constraints, that can easily be expanded or changed to include several attributes as crew capacity, routing decisions, and more loading constraints. We tested our model on instances covering 6 operational days. Being the potential customer of the company, we considered the case of Samsung. The instances contain until 186 clients located throughout the country, with different time windows. We have heterogeneous fleet with 7 types of vehicles. The fleet size is unlimited since the company outsources transport.

We made our test runs on a Linux machine (Intel Xeon E3-1226 processor clocked at 3.3GHz). The algorithms were implemented using IBM CPLEX commercial solver (version 12.4). The SPPRC was solved by dynamic programming using the Boost library, version 1.54. The first run (DH) used a diving heuristic. The second (RMH) used restricted master heuristic. In the table above, we present our numerical results as follows:for each instance defined by (name, number of customers or orders) we present time, number of iterations, cost, and the number of vehicles used.

TABLE I. Real instances computational results

Instances		DH				RMH				Manual Solution		
Name	Orders	Iter	Time(s)	Cost(MAD)	N.Veh	Iter	Time(s)	Cost(MAD)	N.Veh	Time(s)	Cost(MAD)	N.Veh
J1	129	25	1.0	-12052.4	23	10	1.9	-10265.9	22	1800	-18230.05	11
J2	151	23	4.1	11321.9	26	10	18.5	11684.5	25	2100	-5960.34	27
J3	161	14	0.8	10084.2	34	7	2.9	11552.4	33	2400	-12790	31
J4	163	26	2.2	-21179.9	25	7	10.5	-18909.4	24	2100	-97066	12
J5	179	42	22.3	-16672.3	28	13	42.9	-15259	28	1800	-19840.66	10
J6	186	31	9.6	8415.47	28	11	30.5	8926.71	27	2400	-16485	30

In the table, the positive and negative cost values represent respectively profit and loss. We notice that both heuristics found a better solution in less resolution time which does not exceed 45 sec. This is very useful within a logistics and transportation company operating in a competitive and dynamic environment. Indeed, this helps planners to adapt to different practical uncertainties (such as the arrival of a new order, breakdown or unavailability of trucks during peak periods, etc.). We also conclude that the number of subproblems solved are very small, which is justified by the generation of a large number of columns. The costs of both DH and RMH solutions are always better than the manual planning. Indeed, in some instances losses have been minimized up to 5 times. We point out that the strategic plan of the company requires it to practice very uncompetitive selling prices. Indeed, Samsung is a potential customer of the electric household sector that allows the company to position itself. New price policy will be adopted in 2018. Nevertheless, our solution achieved a spectacular economy of costs.

The number of vehicles in the manual solution is sometimes lower, mainly due to the use of courier-express services in manual solutions, and whose purchase cost is extremely high. In fine, the solutions found by DH and RMH are very comparable, which DH is slightly faster than RMH. Finally, we notice that the average loading rate reaches 78% while one route allows to visit 6 clients on average.

VI. Conclusion

This numerical study allowed us to test the effectiveness of our model, especially the SPPRC which allowed managing several exogenous constraints. Also, the DH and RMH methods are very promising, which encourages us to perform further test cases with larger instances and incorporate more difficult constraints, namely loading constraints and resource synchronization Through the broad literature, we have shown that rich VRPs combined with loading problems are extremely difficult and challenging, and are often solved by means of approximate models. Our proposal is part of the new generation of rich VRPs arising in practice, and integrating increasingly strong and realistic constraints to serve specific economic sectors.

Acknowledgment

This work has been supported by Tamayuz Supply Chain (Excellence and Innovation Center of SNTL). The authors would like to express their deep appreciation for the help and the real data provided.

References

[1] G. B. Dantzig and J. H. Ramser, "The truck dispatching problem," *Management science*, vol. 6, no. 1, pp. 80–91, 1959.

[2] G. U. Clarke and J. W. Wright, "Scheduling of vehicles from a central depot to a number of delivery points," *Operations research*, vol. 12, no. 4, pp. 568–581, 1964.

[3] J. K. Lenstra and A. H. G. Kan, "Complexity of vehicle routing and scheduling problems," *Networks*, vol. 11, no. 2, pp. 221–227, 1981.

[4] L. C. Coelho, J. Renaud, and G. Laporte, "Road-based goods transportation: a survey of real-world applications from 2000 to 2015," Technical report, Technical Report FSA-2015-007, Québec, Canada, Tech. Rep., 2015.

[5] P. Toth and D. Vigo, *Vehicle routing: problems, methods, and applications*. Siam, 2014, vol. 18.

[6] S. Martello and D. Vigo, "Exact solution of the two-dimensional finite bin packing problem," *Management science*, vol. 44, no. 3, pp. 388–399, 1998.

[7] S. Martello, D. Pisinger, and D. Vigo, "The three-dimensional bin packing problem," *Operations Research*, vol. 48, no. 2, pp. 256–267, 2000.

[8] D. Pisinger and M. Sigurd, "Using decomposition techniques and constraint programming for solving the two-dimensional bin-packing problem," pp. 36–51, 2007.

[9] M. E. Lübbecke and J. Desrosiers, "Selected topics in column generation," *Operations Research*, vol. 53, no. 6, pp. 1007–1023, 2005.

[10] S. Irnich and G. Desaulniers, "Shortest path problems with resource constraints," in *Column generation*. Springer, 2005, pp. 33–65.

[11] J. Desrosiers and M. E. Lübbecke, "A primer in column generation," in *Column generation*. Springer, 2005, pp. 1–32.

[12] M. Desrochers and F. Soumis, "A generalized permanent labelling algorithm for the shortest path problem with time windows," *INFOR: Information Systems and Operational Research*, vol. 26, no. 3, pp. 191–212, 1988.

[13] C. Joncour, S. Michel, R. Sadykov, D. Sverdlov, and F. Vanderbeck, "Column generation based primal heuristics," *Electronic Notes in Discrete Mathematics*, vol. 36, pp. 695–702, 2010.

A DCA Approach for the Stochastic Shortest Path Problem in Vehicle Routing

(presented at the 7^{th} IESM Conference, October 11–13, 2017, Saarbrücken, Germany)

Vinh Thanh Ho*, Hoai An Le Thi† and Ahmed Zidna‡
* Theoretical and Applied Computer Science Laboratory, University of Lorraine,
Ile du Saulcy, 57045 Metz Cedex 01, France
Email: vinh-thanh.ho@univ-lorraine.fr
† Theoretical and Applied Computer Science Laboratory, University of Lorraine,
Ile du Saulcy, 57045 Metz Cedex 01, France
Email: hoai-an.le-thi@univ-lorraine.fr
‡ Theoretical and Applied Computer Science Laboratory, University of Lorraine,
Ile du Saulcy, 57045 Metz Cedex 01, France
Email: ahmed.zidna@univ-lorraine.fr

***Abstract*—The paper concerns the Stochastic Shortest Path (SSP) problem for a single independent vehicle on a road network using the probability tail model-based criterion. Specifically, this SSP problem aims to search an optimal path that maximizes the probability of reaching destination before a particular deadline. In fact, the maximization problem is reformulated as a cardinality minimization problem (cardinality of a vector is the number of nonzero elements in that vector) with zero-one variables. Recently, some algorithms have been proposed by approximating the cardinality term due to its discontinuity, however, without treating zero-one variables efficiently. In this paper, we develop a DC (Difference of Convex functions) programming and DCA (DC Algorithm) based approach for solving this cardinality problem. We investigate first a DC approximation approach for the cardinality term and then an exact penalty technique for the zero-one variables. The resulting optimization reformulation can be expressed as a DC program for which DCA is applied. Numerical experiments in a road network with different given deadlines show the efficiency of the proposed algorithm in terms of both the quality of solutions and the rapidity when compared with the existing algorithms.**

I. Introduction

The Stochastic Shortest Path (SSP) problem in vehicle routing on a given road network is the problem of finding the optimal path in which the length of road links is random [1]–[3]. As one fundamental problem in network studies, the SSP problem has attracted attention from researchers in many areas of, e.g., transportation engineering, computer science and operations research. In this paper, we consider the SSP problem for a single independent vehicle where the length of road links is represented by their travel time. In fact, the randomness of travel time is due to many uncertain traffic conditions, e.g., the quality of vehicle, traffic jam, weather and so on. In such a stochastic environment, there are many different criteria for an optimal path, e.g., least expected travel [4], mean-risk model [5], path optimality index [2], probability tail (PT) model [1], [6], [7], which are reviewed in detail in [8]. Due to perspicuity for drivers and solving daily issues, we are interested in SPP problems using the PT model-based criterion in which the optimal path is the one that maximizes the probability of reaching destination before a given deadline, which are known as the arriving-on-time problems. An optimization formulation of PT model-based SSP problems on a road network can be expressed as the following.

A. An optimization formulation of PT model-based SSP problems

A road network can be modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ where $\mathcal{V} = \{1, \ldots, n\}$ represents the set of nodes and $\mathcal{A} \subseteq \{(v_1, v_2) : v_1, v_2 \in \mathcal{V}, v_1 \neq v_2\}$ represents the set of arcs with the size of m, (v_1, v_2) represents an arc from node v_1 to node v_2. As mentioned above, the goal of the PT model-based SSP problem is to find the path $\mathbf{x}$ that maximizes the probability of arriving at the destination d from the origin o not later than a given deadline T (more precisely, T is the remaining time to deadline) based on the distributed travel time samples $\mathbf{w}$. The corresponding optimization formulation can be written as follows [9].

$$\max_{\mathbf{x}} \ \mathbb{P}(\mathbf{w}^\top \mathbf{x} \leq T) \tag{1}$$

s.t. $\forall v_1 \in \mathcal{V}$:

$$\sum_{\substack{v_2 \in \mathcal{V}, \\ (v_1, v_2) \in \mathcal{A}}} x(v_1, v_2) - \sum_{\substack{v_2 \in \mathcal{V}, \\ (v_2, v_1) \in \mathcal{A}}} x(v_2, v_1) = \begin{cases} 1 & \text{if } v_1 = o, \\ -1 & \text{if } v_1 = d, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbb{P}(X)$ denotes the probability of a random variable X, $\mathbf{x} \in \{0, 1\}^m$ such that its each component $x(v_1, v_2)$ refers to an arc $(v_1, v_2) \in \mathcal{A}$ and $x(v_1, v_2) = 1$ if and only if (v_1, v_2) is on the concerned path. This formulation is equivalent to the following problem:

$$\begin{aligned} &\min_{\mathbf{x}} \ \mathbb{P}(\mathbf{w}^\top \mathbf{x} > T) \\ &\text{s.t. } \mathbf{M}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \in \{0, 1\}^m, \end{aligned} \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{n \times m}$ is a node-arc incidence matrix (see, e.g., [10] for more details), $\mathbf{b} \in \mathbb{R}^n$ whose elements are zero except the two elements corresponding to node o and node d, which are 1 and -1, respectively.

B. Related works

Most related studies for solving the problem (2) are based on the common assumptions about specific distributions of travel times, correlations between travel times on different road links, deadlines (see, e.g., [1], [6], [7], [11], [12] for more details). However, on real road networks, these assumptions are hard to be satisfied. Without working on these assumptions, Cao et al. [13] proposed a data-driven approach based on the travel time samples of all road links and solved the problem (2) by reformulating it as a cardinality minimization problem with the zero-one variables as follows.

$$\min \{\texttt{Card}(C(\mathbf{x})) : \mathbf{x} \in \{0,1\}^m,\ \mathbf{M}\mathbf{x} = \mathbf{b}\}, \qquad (3)$$

where $\texttt{Card}(\mathbf{z})$ is the number of nonzero elements in $\mathbf{z}$, vector $C(\mathbf{x}) = \left([\mathbf{w}_1^\top \mathbf{x} - T]^+, \ldots, [\mathbf{w}_S^\top \mathbf{x} - T]^+\right)$, $\{\mathbf{w}_i\}_{i=1,\ldots,S}$ denotes the set of travel time samples on all arcs with the size of S and $[\cdot]^+ = \max\{0, \cdot\}$. There exists a double difficulty in solving this cardinality minimization problem: how to treat, first, the cardinality term and, second, the zero-one variables. In [13], the authors approximated the cardinality term by ℓ_p-norm ($0 < p \leq 1$) functions, the logarithmic approximation function and their combination, and thus, proposed the so-called ℓ_1-norm algorithm and reweighted-ℓ_1 algorithms. The ℓ_1-norm algorithm only solves one Mixed Integer Linear Program (MILP), while the reweighted-ℓ_1 algorithms require solving successive MILPs. There is one limitation of this work: expensive computations in solving MILPs. In [14], the authors avoided solving MILPs by relaxing them to linear programs as well as using partial Lagrange multiplier techniques and a property of incidence matrix $\mathbf{M}$ to guarantee that their obtained solutions are zero-one. Concerning the choice of good approximation of the cardinality term, it is known that there exist approximation functions in [15] which have been proved to be more efficient than those in [13], [14].

In the optimization and machine learning community, the cardinality minimization problem (3) with the continuous variables instead of zero-one variables is known as an ℓ_0-norm problem (see, e.g, [15]) where ℓ_0-norm is defined as $\texttt{Card}$. A variety of works relating to ℓ_0-norm problems have been recently reviewed in the seminal work of Le Thi et al. [15]. According to the way to treat the ℓ_0-norm, they are divided into three main categories: convex approximation, nonconvex approximation and nonconvex exact reformulation. Le Thi et al. [15] developed a unified DC (Difference of Convex functions) approximation and DCA (DC Algorithm) based approach for the ℓ_0-norm problem and indicated that corresponding DCA schemes cover all standard algorithms in nonconvex approximation approaches as special versions. Although there are many studies for solving the ℓ_0-norm problems, these problems are actually hard to solve due to ℓ_0-norm. In this paper, we follow the continuous DC approximation function in [15] to tackle the discontinuous $\texttt{Card}$ term.

C. Our contributions

We develop a DCA approach for solving the problem (3) via two following main steps with respect to the $\texttt{Card}$ term and the zero-one variables. First, we use a DC approximation of the $\texttt{Card}$ term, which results in approximating the problem (3) to a DC minimization problem with zero-one variables. Second, we exploit an exact penalty technique in DC programming for treating these zero-one variables, which has been widely studied in, e.g., [16]–[18]. Consequently, the resulting combinatorial optimization problem is equivalently reformulated as a continuous problem of the standard form of a DC program. Thus, a DCA-based algorithm is designed for solving the considered SSP problem. In addition, as an illustrative experiment for the arriving on time problem, we demonstrate the efficiency of our approach in terms of both the quality of obtained paths and the rapidity on a road network in comparison with the ℓ_1-norm and the reweighted-ℓ_1 algorithm.

The rest of the paper is organized as follows. An introduction of DC programming and DCA is given in Sect. II. In Sect. III, we first show how to reformulate the cardinality minimization problem as a DC program, and then design a DCA-based algorithm for the resulting optimization reformulation. Sect. IV reports the numerical results on a simulated road network which is followed by some conclusions in Sect. V.

II. An introduction to DC programming and DCA

DC programming and DCA, introduced by Pham Dinh Tao in a preliminary form in 1985 and extensively developed by Le Thi Hoai An and Pham Dinh Tao since 1994, constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization. They address DC programs of the form

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in \mathbb{R}^p\} \quad (P_{dc})$$

where $g, h \in \Gamma_0(\mathbb{R}^p)$, the set of all lower semicontinuous proper convex functions on $\mathbb{R}^p$. Such a function f is called DC function, and $g - h$, DC decomposition of f while g and h are DC components of f. A closed convex constraint set C can be incorporated in the objective function of (P_{dc}) by adding the characteristic function of C ($\chi_C(x) = 0$ if $x \in C$, $+\infty$ otherwise) to the first DC component g of the DC objective function f. A DC program (P_{dc}) is called polyhedral DC program when either g or h is a polyhedral convex function (i.e. it is the sum of the characteristic function of a nonempty polyhedral convex set and the pointwise supremum of a finite collection of affine functions). Note that a polyhedral convex function is almost always differentiable.

Recall that, for $\varphi \in \Gamma_0(\mathbb{R}^p)$ and $x \in \text{dom } \varphi := \{u \in \mathbb{R}^p : \varphi(u) < +\infty\}$, the subdifferential of φ at x, denoted $\partial\varphi(x)$, is defined as

$$\partial\varphi(x) := \{y \in \mathbb{R}^p : \varphi(u) \geq \varphi(x) + \langle u - x, y\rangle, \forall u \in \mathbb{R}^p\}$$

which is a closed convex set in $\mathbb{R}^p$. Denote dom $\partial\varphi = \{x \in \mathbb{R}^p : \partial\varphi(x) \neq \emptyset\}$. It generalizes the concept of derivative in the sense that φ is differentiable at x if and only if $\partial\varphi(x)$ is reduced to a singleton set, i.e. $\partial\varphi(x) = \{\nabla\varphi(x)\}$. Each $y \in \partial\varphi(x)$ is called a subgradient of φ at x.

The necessary local optimality condition for (P_{dc}) is

$$\partial h(x^*) \subset \partial g(x^*). \qquad (4)$$

This condition (4) is also sufficient for many important classes of DC programs, for example, for polyhedral DC programs, or when function f is locally convex at x^* [19], [20].

A point x^* that satisfies the generalized Kuhn-Tucker condition

$$\partial g(x^*) \cap \partial h(x^*) \neq \emptyset$$

is called a critical point of $g - h$. It follows that if h is polyhedral convex, then a critical point of $g - h$ is almost always a local solution to (P_{dc}).

The DCA has the quite simple interpretation: each iteration k of DCA approximates the concave part $-h$ by its affine majorization (taking $y^k \in \partial h(x^k)$) and minimizes the resulting convex function.

Generic DCA scheme

Initialization: Let $x^0 \in \mathbb{R}^p$ be a best guess, $k = 0$.

repeat

1. Compute $y^k \in \partial h(x^k)$.
2. Compute $x^{k+1} \in \text{argmin}\{g(x) - h(x^k) - \langle x - x^k, y^k\rangle : x \in \mathbb{R}^p\}$ (P_k).
3. $k = k + 1$.

until convergence of $\{x^k\}$.

Convergence properties of the DCA and its theoretical basis are described in [19]–[21]. However, it is worthwhile to mention the following properties:

• DCA is a descent method *without linesearch* (the sequence $\{g(x^k) - h(x^k)\}$ is decreasing) but with global convergence (i.e. it converges from an arbitrary starting point).

• If $g(x^{k+1}) - h(x^{k+1}) = g(x^k) - h(x^k)$, then x^k is a critical point of $g - h$. In this case, DCA terminates at kth iteration.

• If the optimal value α of problem (P_{dc}) is finite and the infinite sequence $\{x^k\}$ is bounded, then every limit point x^* of this sequence is a critical point of $g - h$.

• DCA has a linear convergence for general DC programs. Especially, for polyhedral DC programs, the sequence $\{x^k\}$ contains finitely many elements and the algorithm convergences to a solution in a finite number of iterations.

In the past years, DCA has been successfully applied in several works of various fields among them machine learning, financial optimization, supply chain management, etc (see, e.g., [15], [18]–[29] and references in [30]).

For a complete study of DC programming and DCA the reader is referred to [19]–[21], [26] and the list of references in [30].

III. Solution methods by DC programming and DCA

A. A DC reformulation of the cardinality problem (3)

By introducing the slack variable $\mathbf{y}$, the following proposition (whose proof is evident) will justify our considered optimization formulation (5) whose feasible set is bounded.

Proposition 1: If $\mathbf{x}^*$ is an optimal solution to the problem (3), then $(\mathbf{x}^*, \mathbf{y}^*)$ is an optimal solution to the following problem

$$\min \left\{ \begin{array}{c} \mathbf{Mx} = \mathbf{b},\ \mathbf{x} \in \{0,1\}^m, \mathbf{y} \geq 0, \\ \texttt{Card}(\mathbf{y}) : y_i \geq \mathbf{w}_i^\top \mathbf{x} - T,\ \forall i = 1, \ldots, S, \\ y_i \leq [||\mathbf{w}_i||_1 - T]^+,\ \forall i = 1, \ldots, S \end{array} \right\}, \quad (5)$$

where $\mathbf{y} = (y_1, \ldots, y_S) \in \mathbb{R}^S$, $y_i^* = [\mathbf{w}_i^\top \mathbf{x}^* - T]^+,\ \forall i = 1, \ldots, S$. Conversely, if $(\mathbf{x}^*, \mathbf{y}^*)$ is an optimal solution to (5), then $\mathbf{x}^*$ is an optimal solution to (3).

Note that the upper bounds for the variable $\mathbf{y}$ are added in (5) due to the fact that for all $\mathbf{x} \in \{0,1\}^m$, $0 \leq [\mathbf{w}_i^\top \mathbf{x} - T]^+ \leq [||\mathbf{w}_i||_1 - T]^+,\ \forall i = 1, \ldots, S$. This boundedness of the feasible set is useful for the exact penalty techniques as well as the nonconvex exact reformulation approach for the problem (3) as discussed in Sect. I-B.

In this paper, we use a continuous DC approximation function of the `Card` term, specifically the exponential concave approximation [31], which is proven to be efficient in pratice (see, e.g., [15], [31]–[33]), as follows.

$$\texttt{Card}(\mathbf{y}) \approx \sum_{i=1}^{S} r_\theta(y_i),$$

where the concave function $r_\theta(t) = 1 - e^{-\theta|t|}$ for $t \in \mathbb{R}$ and the positive approximation parameter θ. Thus, we derive the following approximate problem

$$\alpha := \min \left\{ \sum_{i=1}^{S} r_\theta(y_i) : \begin{array}{r} \mathbf{Mx} = \mathbf{b},\ \mathbf{x} \in \{0,1\}^m, \mathbf{y} \geq 0, \\ y_i \geq \mathbf{w}_i^\top \mathbf{x} - T,\ \forall i = 1, \ldots, S, \\ y_i \leq [||\mathbf{w}_i||_1 - T]^+, \forall i = 1, \ldots, S \end{array} \right\}. \quad (6)$$

In fact, solving the problem (6) is still difficult due to zero-one variables. Using the exact penalty techniques to treat these variables leads to an exact continuous reformulation of (6) which is of the standard form of a DC program for which DCA is applied. Let $p : [0,1]^m \to \mathbb{R}$ be the penalty function defined by $p(\mathbf{x}) := \sum_{i=1}^{m} \min\{x_i, 1 - x_i\}$. The problem (6) can be rewritten as follows:

$$\alpha = \min \left\{ f(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{S} r_\theta(y_i) : (\mathbf{x}, \mathbf{y}) \in \mathcal{K},\ p(\mathbf{x}) \leq 0 \right\},$$

where

$$\mathcal{K} := \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+S} : \begin{array}{r} \mathbf{Mx} = \mathbf{b},\ \mathbf{x} \in [0,1]^m, \mathbf{y} \geq 0, \\ y_i \geq \mathbf{w}_i^\top \mathbf{x} - T, \forall i = 1, \ldots, S, \\ y_i \leq [||\mathbf{w}_i||_1 - T]^+, \forall i = 1, \ldots, S \end{array} \right\}$$

is the bounded polyhedral convex set defined by $n + S$ linear constraints and upper/lower bound constraints. It leads to the corresponding penalized problem (τ being the positive penalty parameter)

$$\alpha(\tau) := \min \{ f(\mathbf{x}, \mathbf{y}) + \tau p(\mathbf{x}) : (\mathbf{x}, \mathbf{y}) \in \mathcal{K} \}. \quad (7)$$

It follows from Theorem 7 in [16] that there exists $\tau_0 > 0$ such that for all $\tau > \tau_0$, the two problems (6) and (7) are equivalent, in the sense that they have the same optimal value and the same solution set.

Considering the problem (7) with a sufficiently large number τ, we use the following DC decomposition of r_θ [15]:

$$r_\theta(t) = g(t) - h(t),\ \forall t \in \mathbb{R}, \quad (8)$$

where $g(t) = \theta|t|$ and $h(t) = \theta|t| - 1 + e^{-\theta|t|}$. Thus, a DC reformulation of the problem (7) can be expressed as follows.

$$\min \left\{ F(\mathbf{x}, \mathbf{y}) = G(\mathbf{x}, \mathbf{y}) - H(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+S} \right\}, \quad (9)$$

where $G(\mathbf{x},\mathbf{y}) = \chi_{\mathcal{K}}(\mathbf{x},\mathbf{y}) + \sum_{i=1}^{S} g(y_i)$ and $H(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{S} h(y_i) - \tau p(\mathbf{x})$. Obviously, (9) is a polyhedral DC program where the first DC component $G(\mathbf{x},\mathbf{y})$ is polyhedral convex.

B. DCA for solving the polyhedral DC program (9)

According to the generic DCA scheme, at each iteration k, after computing one subgradient $(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k) \in \partial H(\mathbf{x}^k, \mathbf{y}^k)$, the calculation of $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$ is reduced to solve the following convex program

$$\min\left\{G(\mathbf{x},\mathbf{y}) - \left\langle(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k), (\mathbf{x},\mathbf{y})\right\rangle : (\mathbf{x},\mathbf{y}) \in \mathbb{R}^{m+S}\right\},$$

which is equivalent to a linear program

$$\min\left\{\left\langle(-\overline{\mathbf{x}}^k, \theta\mathbf{e} - \overline{\mathbf{y}}^k), (\mathbf{x},\mathbf{y})\right\rangle : (\mathbf{x},\mathbf{y}) \in \mathcal{K}\right\}, \tag{10}$$

where $\mathbf{e}$ is the vector of ones in the appropriate vector space.

Let $q(\mathbf{y}) := \sum_{i=1}^{S} h(y_i)$. We present how to compute a subgradient $(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k) \in \partial H(\mathbf{x}^k, \mathbf{y}^k)$. Since $\mathbf{y}^k \geq 0$, we can choose $\overline{\mathbf{x}}^k \in \tau\partial(-p)(\mathbf{x}^k)$ and $\overline{\mathbf{y}}^k \in \partial q(\mathbf{y}^k)$ as follows. For $i = 1, \ldots, m$,

$$\overline{x}_i^k = \begin{cases} -\tau & \text{if } x_i^k < 1/2, \\ 0 & \text{if } x_i^k = 1/2, \\ \tau & \text{if } x_i^k > 1/2, \end{cases} \tag{11}$$

and for $j = 1, \ldots, S$,

$$\overline{y}_j^k = \begin{cases} \theta(1 - e^{-\theta y_j^k}) & \text{if } y_j^k > 0, \\ 0 & \text{if } y_j^k = 0. \end{cases} \tag{12}$$

DCA applied to (9) can be given by Algorithm 1 (DCA).

Algorithm 1: DCA for solving the problem (9) **(DCA)**
Initialization: Let $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^{m+S}$, $k = 0$.
repeat
1. Compute $(\overline{\mathbf{x}}^k, \overline{\mathbf{y}}^k) \in \partial H(\mathbf{x}^k, \mathbf{y}^k)$ using (11)-(12).
2. Compute $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$, an optimal solution to (10).
3. $k = k + 1$.

until stopping criterion

According to the convergence properties of the generic DCA scheme for polyhedral DC programs in Sec. II, we deduce the following interesting convergence properties of Algorithm 1.

Theorem 1: i) Algorithm 1 generates the sequence $\{(\mathbf{x}^k, \mathbf{y}^k)\}$ containing finitely many elements such that the sequence $\{F(\mathbf{x}^k, \mathbf{y}^k)\}$ is decreasing.

ii) The sequence $\{(\mathbf{x}^k, \mathbf{y}^k)\}$ converges to a critical point $(\mathbf{x}^*, \mathbf{y}^*)$ of (9) after a finite number of iterations.

iii) For a sufficiently large number τ, if there exists an iteration $\overline{k}$ such that $\mathbf{x}^{\overline{k}} \in \{0,1\}^m$, then $\mathbf{x}^l \in \{0,1\}^m$ for all $l > \overline{k}$.

Proof: The properties *i)* and *ii)* are direct consequences of the convergence properties of DCA for a polyhedral DC program. The property *iii)* is proved similarly as Theorem 1 in [23] and Theorem 2 in [17]. ■

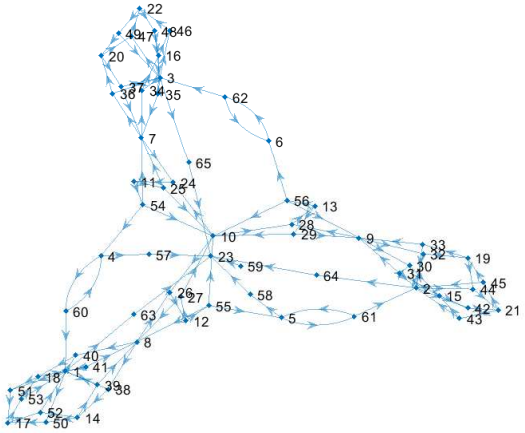


Fig. 1. A road network with 123 road links and 65 intersections [13]

IV. Numerical experiments

In this experiment, we implement the arriving-on-time problem (i.e. maximizing the probability of arriving at the destination d from the origin o not later than the deadline T) for a single vehicle on a road network with 123 road links and 65 intersections of roads as in [13], corresponding to a directed graph with 123 arcs and 65 nodes respectively as Fig. 1 and the length of arcs is represented by the travel time of road links. We study the performance of our DCA algorithm and compare it with two existing algorithms in [13]: ℓ_1-norm, reweighted ℓ_1-norm minimization algorithms (we call ℓ_1, re-ℓ_1 respectively for short). The ℓ_1 algorithm replaces the `Card` term by the ℓ_1-norm function and solves one MILP. The re-ℓ_1 algorithm uses the combination of the ℓ_p-norm ($0 < p < 1$) and logarithmic approximation functions and requires solving successive MILPs. In fact, the approximation in the re-ℓ_1 algorithm is a special version of our DC approximation approach (see [15] for more details). It has been shown in [13] that when comparing between re-ℓ_1 and ℓ_1, in most cases, re-ℓ_1 (resp. ℓ_1) is better than the other in terms of the quality of solutions (resp. the rapidity).

Data set: Numerical experiments are conducted on the data sets (including the training set of size S_1 and test set of size S_2) of travel time samples on each arc generated by some independent random distribution functions (e.g., normal, gamma and lognormal distributions). The (o, d) pair is chosen randomly from the given set. The deadline T is defined by $T = T_1 + \alpha(T_2 - T_1)$ where α is a deadline coefficient, T_2 is the minimum longest travel time for all paths of (o,d) and T_1 is the shortest travel time with respect to the same path (see [13] for more details).

Set up experiments: The parameters S_1, S_2, α are set as follows: $S_1 \in \{100, 500, 1000\}$, $S_2 = 5000$, $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$. The given set of ($o$,$d$) pairs is {(3,7), (7,3), (1,8), (8,1), (10,2), (2,10), (1,10), (10,1), (3,10), (10,3)}. Three random distribution functions are used with different parameters chosen randomly, specifically, normal distribution with mean in {5, 10, 15, 20} and standard deviation in {10, 30, 50, 70}, gamma distribution with shape and scale in {1, 2,

3, 4, 5}, lognormal distribution with mean in {5, 10, 15, 20} and variance in {10, 30, 50, 70}. Our experiment is run 100 times and performed in MATLAB R2016a on a Laptop Intel(R) Core(TM) i7-2720QM CPU @ 2.20GHz of 4GB RAM. The software CPLEX 12.6 is used for solving linear programs in our algorithm, and solving MILPs in the existing algorithms. DCA and re-ℓ_1 use the same zero starting point, their approximation parameter θ is set to 0.01, and their stopping criteria are either $||\mathbf{x}^k-\mathbf{x}^{k-1}||+||\mathbf{y}^k-\mathbf{y}^{k-1}|| \leq \varepsilon(1+||\mathbf{x}^{k-1}||+||\mathbf{y}^{k-1}||)$ or $|F(\mathbf{x}^k,\mathbf{y}^k)-F(\mathbf{x}^{k-1},\mathbf{y}^{k-1})| \leq \varepsilon(1+F(\mathbf{x}^{k-1},\mathbf{y}^{k-1}))$. For re-$\ell_1$, p =0.5. The default tolerance is $\varepsilon = 10^{-4}$. The penalty parameter τ will be updated at each iteration.

Comparison criteria of algorithms: We are interested in the following aspects: accuracy (in %) and CPU time (in seconds). The accuracy of each algorithm is computed on the test set by the percentage of the probability of arriving on time of the path obtained by the algorithm relative to the maximum probability of all possible paths by the enumeration method.

Descriptions of tables of results: The average accuracy and average CPU time obtained by three algorithms DCA, ℓ_1 and re-ℓ_1 on 100 runs with the different deadline coefficients α and random distributions are reported in Tables I, II. The values in 2nd-4th columns of Table I represent the algorithms' accuracy in case the size of training set $S_1 = 100$, 500, 1000, respectively. These descriptions are similar to Table II. Bold values in Tables I, II are the best results in each column.

Comments on numerical results (Tables I, II): we observe from experiments that with suitable penalty parameters, our DCA algorithm always furnishes the zero-one solutions, which is an advantage of DCA when solving the continuous problems. Moreover, in terms of the accuracy, our DCA algorithm is almost always better than the existing algorithms with the different deadline coefficients and different random distributions. The quality of our solution is very high with the accuracy of above 95% and as for normal/gamma distribution, even above 99%, that is to say, our algorithm usually finds the optimal path as the enumeration method. Moreover, in most cases of our DCA, the larger the training set is, the more the accuracy is improved, however for normal/gamma distribution, the accuracy in case $S_1 = 500$ is sufficiently high. Concerning CPU time, DCA is faster than the existing algorithms ℓ_1, re-ℓ_1, respectively, up to 16.2, 18.6 times, which can be explained by the fact that our DCA algorithm only requires solving successive linear programs, while the existing algorithms solve MILPs. In addition, with the good obtained accuracy, DCA runs much faster than the enumeration method up to 211 times.

TABLE I. ACCURACY (%) OF DCA AND ℓ_1-NORM ALGORITHM (ℓ_1), REWEIGHTED-ℓ_1 ALGRORITHM (RE-ℓ_1) ON 100 RUNS WITH THE DIFFERENT DEADLINE COEFFICIENTS α AND RANDOM DISTRIBUTIONS

NORMAL DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
α	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
0.2	**99.16**	99.14	99.14	**99.60**	**99.60**	**99.60**	**99.60**	99.59	99.59
0.4	**99.54**	99.53	99.53	**99.90**	99.89	99.89	**99.90**	99.89	99.89
0.6	**99.73**	99.72	99.72	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**
0.8	**99.81**	99.80	99.80	**100**	**100**	**100**	**100**	**100**	**100**
1.0	**99.87**	99.87	99.87	**100**	**100**	**100**	**100**	**100**	**100**
1.2	**99.91**	99.90	99.90	**100**	**100**	**100**	**100**	**100**	**100**

GAMMA DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
α	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
0.2	**99.77**	99.71	99.71	**99.87**	99.61	99.61	**99.87**	99.83	99.83
0.4	**99.82**	99.78	99.78	**99.75**	99.71	99.71	**99.92**	99.71	99.71
0.6	**99.90**	99.85	99.85	**99.84**	99.80	99.80	**99.84**	99.80	99.80
0.8	**99.94**	99.90	99.90	**99.90**	99.86	99.86	**99.90**	99.86	99.86
1.0	99.86	**99.92**	**99.92**	**99.96**	99.92	99.92	**99.96**	99.92	99.92
1.2	**100**	99.94	99.94	**99.98**	99.94	99.94	**99.98**	99.94	99.94

LOGNORMAL DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
α	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
0.2	**95.98**	95.89	95.89	**96.48**	95.66	95.66	**97.84**	96.99	96.99
0.4	**96.43**	96.21	96.22	**97.01**	96.32	96.32	**98.18**	97.41	97.41
0.6	**96.88**	96.65	96.67	**97.50**	96.80	96.80	**98.51**	97.72	97.72
0.8	**97.32**	97.06	97.07	**97.74**	97.15	97.15	**98.05**	97.99	97.99
1.0	96.82	97.30	**97.31**	97.37	**97.48**	**97.48**	98.19	**98.24**	**98.24**
1.2	**97.28**	96.47	96.48	**97.63**	97.28	97.28	**98.38**	97.91	97.91

* S_1 is the size of training set. Bold values are the best results in each column.

TABLE II. CPU TIME (SECONDS) OF ENUMERATION METHOD (ENUM), DCA, ℓ_1 AND RE-ℓ_1 ON 100 RUNS WITH DIFFERENT RANDOM DISTRIBUTIONS IN CASE $\alpha = 1.2$

NORMAL DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
enum	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
1.81	**0.01**	0.06	0.07	**0.03**	0.35	0.45	**0.10**	1.03	1.20

GAMMA DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
enum	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
2.11	**0.01**	0.10	0.12	**0.04**	0.49	0.43	**0.05**	0.81	0.93

LOGNORMAL DISTRIBUTION

	$S_1 = 100$			$S_1 = 500$			$S_1 = 1000$		
enum	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1	DCA	ℓ_1	re-ℓ_1
2.16	**0.04**	0.22	0.37	**0.14**	0.57	1.21	**0.72**	2.14	5.21

V. CONCLUSIONS

We have intensively studied a nonconvex, nonsmooth programming approach based on DC programming and DCA for solving the SSP problem in vehicle routing for a single independent vehicle, which can be reformulated as a cardinality optimization problem with zero-one variables. Specifically, we have used a DC approximation approach for the cardinality term. Exploiting an exact penalty technique in DC programming, we have dealt with the zero-one variables and obtained an exact reformulation of the corresponding approximate problem. Applying DCA programming and DCA for the resulting reformulation problem, we have developed a fast, efficient DCA scheme for the considered cardinality problem. The numerical results on simulated road networks show that our proposed approach is effective in comparison with the existing algorithms in terms of both the quality of obtained solutions and the rapidity. In the future, we plan to explore the property of incident matrix for our approach, show the efficiency of DCA on larger/real road networks, develop our approach for solving a variety of SSP problems, e.g., SSP problems with Delay Excess Penalty and study these problems with the direction based on nonconvex exact reformulation approach for the cardinality term using the exact penalty techniques.

REFERENCES

[1] H. Frank, "Shortest Paths in Probabilistic Graphs," *Oper. Res.*, vol. 17, no. 4, pp. 583–599, Aug. 1969.

[2] C. E. Sigal, A. A. B. Pritsker, and J. J. Solberg, "The Stochastic Shortest Route Problem," *Oper. Res.*, vol. 28, no. 5, pp. 1122–1129, 1980.

[3] J. J. Martin, "Distribution of The Time Through A Directed, Acyclic Network," *Oper. Res.*, vol. 13, no. 1, pp. 46–66, Feb. 1965.

[4] E. D. Miller-Hooks and H. S. Mahmassani, "Least Expected Time Paths in Stochastic, Time-Varying Transportation Networks," *Transportation Science*, vol. 34, no. 2, pp. 198–215, May 2000.

[5] E. Nikolova, *Approximation Algorithms for Reliable Stochastic Combinatorial Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 338–351.

[6] E. Nikolova, J. A. Kelner, M. Brand, and M. Mitzenmacher, *Stochastic Shortest Paths via Quasi-convex Maximization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 552–563.

[7] Y. Y. Fan, R. E. Kalaba, and J. E. Moore, "Arriving on Time," *Journal of Optimization Theory and Applications*, vol. 127, no. 3, pp. 497–513, 2005.

[8] Z. Cao, "Maximizing the Probability of Arriving on Time : A Stochastic Shortest Path Problem," Ph.D. dissertation, Interdisciplinary Graduate School (IGS), 2017.

[9] J. Cheng, S. Kosuch, and A. Lisser, "Stochastic Shortest Path Problem with Uncertain Delays," in *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems*, 2012, pp. 256–264.

[10] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[11] S. Lim, H. Balakrishnan, D. Gifford, S. Madden, and D. Rus, *Stochastic Motion Planning and Applications to Traffic*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 483–500.

[12] Y. M. Nie and X. Wu, "Shortest Path Problem Considering On-Time Arrival Probability," *Transportation Research Part B: Methodological*, vol. 43, no. 6, pp. 597–613, 2009.

[13] Z. Cao, H. Guo, J. Zhang, D. Niyato, and U. Fastenrath, "Finding The Shortest Path in Stochastic Vehicle Routing: A Cardinality Minimization Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1688–1702, June 2016.

[14] ——, "Improving The Efficiency of Stochastic Vehicle Routing: A Partial Lagrange Multiplier Method," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 3993–4005, June 2016.

[15] H. A. Le Thi, T. Pham Dinh, H. M. Le, and X. T. Vo, "DC Approximation Approaches for Sparse Optimization," *European Journal of Operational Research*, vol. 244, no. 1, pp. 26–46, 2015.

[16] H. A. Le Thi, T. Pham Dinh, and H. V. Ngai, "Exact Penalty and Error Bounds in DC Programming," *Journal of Global Optimization*, vol. 52, no. 3, pp. 509–535, 2012.

[17] H. A. Le Thi, T. Pham Dinh, and M. Le Dung, "Exact Penalty in DC Programming," *Vietnam Journal of Mathematics*, vol. 27, no. 2, pp. 169–178, 1999.

[18] T. Pham Dinh and H. A. Le Thi, *Recent Advances in DC Programming and DCA*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, vol. 8342, pp. 1–37.

[19] H. A. Le Thi and T. Pham Dinh, "The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems," *Annals of Operation Research*, vol. 133, no. 1-4, pp. 23–46, 2005.

[20] T. Pham Dinh and H. A. Le Thi, "Convex Analysis Approach to DC Programming: Theory, Algorithms and Applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[21] ——, "DC Optimization Algorithms for Solving The Trust Region Subproblem," *SIAM Journal of Optimization*, vol. 8, no. 2, pp. 476–505, 1998.

[22] E. Esser, Y. Lou, and J. Xin, "A Method for Finding Structured Sparse Solutions to Non-negative Least Squares Problems with Applications," *SIAM J. Imaging Sciences*, vol. 6, no. 4, pp. 2010–2046, 2013.

[23] H. A. Le Thi, H. M. Le, and T. Pham Dinh, "Feature Selection in Machine Learning: An Exact Penalty Approach Using A Difference of Convex Function Algorithm," *Machine Learning*, vol. 101, no. 1–3, pp. 163–186, 2015.

[24] H. A. Le Thi and M. C. Nguyen, "Self-Organizing Maps by Difference of Convex Functions Optimization," *Data Mining and Knowledge Discovery*, vol. 28, no. 5-6, pp. 1336–1365, 2014.

[25] H. A. Le Thi, M. C. Nguyen, and T. Pham Dinh, "A DC Programming Approach for Finding Communities in Networks," *Neural Computation*, vol. 26, no. 12, pp. 2827–2854, 2014.

[26] H. A. Le Thi and T. Pham Dinh, "Solving A Class of Linearly Constrained Indefinite Quadratic Problems by D.C. Algorithms," *Journal of Global Optimization*, vol. 11, no. 3, pp. 253–285, 1997.

[27] H. A. Le Thi, X. T. Vo, and T. Pham Dinh, "Feature Selection for Linear SVMs under Uncertain Data: Robust Optimization based on Difference of Convex Functions Algorithms," *Neural Networks*, vol. 59, pp. 36–50, 2014.

[28] Y. Liu, X. Shen, and H. Doss, "Multicategory ψ-Learning and Support Vector Machines: Computational Tools," *Journal of Computational and Graphical Statistics,*, vol. 14, no. 1, pp. 219–236, 2005.

[29] T. Schüle, C. Schnörr, S. Weber, and J. Hornegger, "Discrete Tomography by Convex-Concave Regularization and D.C. Programming," *Discrete Appl. Math.*, vol. 151, pp. 229–243, 2005.

[30] H. A. Le Thi, "DC Programming and DCA," 2005. [Online]. Available: http://www.lita.univ-lorraine.fr/$\sim$lethi

[31] P. S. Bradley and O. L. Mangasarian, "Feature Selection via Concave Minimization and Support Vector Machines," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 82–90.

[32] H. A. Le Thi, H. M. Le, V. V. Nguyen, and T. Pham Dinh, "A DC Programming Approach for Feature Selection in Support Vector Machines Learning," *Advances in Data Analysis and Classification*, vol. 2, no. 3, pp. 259–278, 2008.

[33] H. A. Le Thi and D. N. Phan, "DC Programming and DCA for Sparse Optimal Scoring Problem," *Neurocomputing*, vol. 186, pp. 170 –181, 2016.

Author Index

Abdeljaouad, Mohamed Amine 382, 388
Aghezzaf, El-Houssaine 75, 317, 340
Ait El Cadi, Abdessamad 304, 394
Aït-Kadi, Daoud 335, 394
Al-Fandi, Lawrence 169
Alabadleh, Ahmed 51
Albana, Abduh Sayid 474
Amodeo, Lionel 174
Aqlan, Faisal 169
Araújo, Madalena 499, 505
Artiba, Abdelhakim 304, 394
Attia, El Awady 329
Aubourg, Gautier 81

Bahroun, Mohamed 63
Bahroun, Zied 382, 388
Baptiste, Pierre 210
Barbosa-Póvoa, Ana 412
Bauer, Dennis 87
Bauernhansl, Thomas 87
Beauregard, Yvan 57
Behrendt, Fabian 129
Belil, Sabah 480
Ben Alaïa, Essia 623, 629
Ben Ammar, Oussama 462
Ben Mahmoud, Hajer 629
Ben Romthane, Taieb 629
Ben Salem, Ali 26
Benbitour, Mohammed Hichame 69
Benkalai, Imène 210
Benmansour, Rachid 216
Bergmann, Ulf 359
Berrado, Abdelaziz 560
Berrah, Lamia 93
Bertolini, Massimo 437
Bettayeb, Belgacem 462
Billaut, Jean-Charles 273
Bistorin, Olivier 591
Borne, Pierre 623
Botto, Miguel Ayala 574
Boubaker, Selmen 192
Boubaker, Zied 311
Bouchriha, Hanen 399
Boudella, Mohamed El Amine 8
Bouguerra, Soumaya 311

Braun, Oliver 216
Bril El Haouzi, Hind 291
Buchriha, Hanen 623
Bürger, Adrian 156

Caillaud, Emmanuel 198
Calzavara, Martina 32
Carmona-Benítez, Rafael 574
Carrasco, Fernando 220
Castagna, Pierre 364
Cerezo-Narvaez, Alberto 487
Cerqueus, Audrey 1
Chabchoub, Habib 548
Chabouh, Safa 399
Chang, Zheng 635
Charalambis, Alessandro 99, 105
Chatzoglou, Prodromos 111
Chelbi, Anis 311
Chelly, Amina 598
Chen, Chia-Wen 586
Chen, Haoxun 610, 635
Chen, Kuen-Suan 523
Chen, Yan-Chun 468
Chen, Zhuxi 38
Chirinos, Orlando 493
Chu, Chengbin 238, 250, 418, 424, 430, 647
Chu, Feng 244, 653
Clivillé, Vincent 93
Cung, Van-Dat 279

Dafaoui, El Mouloudi 548, 560
Dallery, Yves 8, 69, 192
Dantan, Jean-Yves 352
Dao, Thien-My 57
De Guio, Roland 449
De Vuyst, Stijn 75
Della Croce, Federico 268
Dellagi, Sofiene 346
Delorme, Xavier 2
Deneux, Dominique 135
Denisova, Liudmila 117
Dequeant, Kean 323
Diallo, Claver 317, 340
Dinis, Duarte 412
Dolgui, Alexandre 20, 462
Dorval, Valérie 394
Dubois, Sébastien 449

Eburdy, Romain 141
El Mhamedi, Abderrahman 548, 560

El Mokrini, Asmae 560
El-Assal, Ahmed 329
El-Banna, Mahmoud 51
Elhallaoui, Issmail 659
Elidrissi, Abdelhak 529
Elleuch, Hatem 548
Espinouse, Marie-Laure 279, 323
Etienne, Alain 511
Ezhil, S.A. 604

Felsberger, Andreas 87
Fenies, Pierre 370
Fernandes, Gabriela 499, 505
Fernandez-Viagas, Victor 220
Foroozesh, Nazanin 535
Fotia, Sophia 150
Foulloy, Laurent 93
Framinan, Jose M. 220
Frein, Yannick 474, 598

Galasso, François 81
Gaudichau, Olivier 198
Gharbi, Ali 304
Giard, Vincent 443
Glock, Christoph 32
Godichaud, Matthieu 174
Gomes, Daniela 505
Gondran, Matthieu 226, 232
Grabot, Bernard 81
Grichi, Yosra 57
Grimault, Axel 616
Grosse, Eric 32
Gupta, Jatinder 38
Gurevsky, Evgeny 20
Gürster, Daniel 87

Habchi, Georges 493
Hadj-Alouane, Atidel 554 , 598
Haidar, Mohammad 329
Hajej, Zied 346
Hajmoosaei, Mojtaba 123
Hammami, Ramzi 474
Hammami, Sondes 399
Hanafi, Said 216
Harbaoui, Imen 623, 629
Harbi, Slim 63, 554
He, Junkai 238
Heinicke, Matthias 359
Hennequin, Sophie 529
Heßler, Anja 44

Hipólito, Tomás 574
Ho, Vinh Thanh 664
Hssini, Imane 180
Häberer, Sebastian 129

Jamali, Mohamed Anouar 659
Jamshidi, Afshin 335
Jangnilwong, Saran 204
Jemaî, Ziad 192
Jiang, Juan 647
Joblot, Laurent 135

Kafa, Nadine 560
Kemmoé T., Sylverin 370, 480
Kerbache, Laoucine 105
Khatab, Abdelhakim 317, 340
Kouiss, Khalid 364
Kourentzes, Nikolaos 75
Kumar, Shankar Saravana 14

Lacomme, Philippe 226, 232, 641
Lai, Anh Ngoc 566
Lameche, Khaled 364
Lamothe, Jacques 81
Lamouri, Samir 135, 141
Lamy, Damien 370
Lau, Lina 129
Le Thi, Hoai An 541, 664
Leisten, Rainer 14, 285
Lemaire, Pierre 323
Li, Fan 511
Li, Ying 653
Liu, Ming 238, 244, 250, 418, 424, 430, 647, 653
Liu, Wenchang 580
Liu, Xin 244
Lyu, Jrjung 586
López, Eduardo M. 376

Ma, Andrew 256
Machado, Ricardo 499
Magalhães, Pedro 499
Magali, Pralus 493
Magnani, Florian 198
Maier, Florian 87
Majdouline, Ilias 346
Man, Xiaoyi 418
Marcon, Eric 399
Mechmech, Rihab 63, 554
Meshcheryakov, Vitalii 117
Meskens, Nadine 180

Messaoudene, Zahir 493
Messaoudi, Mayssoun 659
Michailidou, Vasiliki 111
Mifdal, Lahcen 346
Mischkot, Michael 105
Moeuf, Alexandre 141
Monteiro, Thibaud 405
Mosquera Varela, Anderson Fabian 405
Moukrim, Aziz 641
Mourtzis, Dimitris 150
Mousavi, S. Meysam 535
Mura, Enzo 226
Müller, Stephan 156

Nabais, João 574
Najid, Najib 364
Nassehi, Aydin 256
Ndhaief, Nadia 591
Negenborn, Rudy R. 574
Nguyen, Duc Huy 610
Nguyen Canh, Nam 541
Nimis, Jens 156
Noor Permadi, Agri Utama 616
Nouira, Imen 598
Noyel, Mélanie 291

Oberegger, Bernhard 87
Otero-Mateo, Manuel 487

Pacheco Osorio, Adriana Maria 262
Pastor-Fernandez, Andres 487
Paviot, Thomas 135
Pedersen, David Bue 105
Pellerin, Robert 141, 517
Percebois, Christian 123
Perez-Gonzalez, Paz 220
Perrier, Nathalie 517
Pham Thi, Hoai 541
Phan, Dinh Anh 566
Pinaton, Jacques 26
Pinto, Eduardo B. 499
Pirogov, Aleksandr 20
Ponsignon, Thomas 87
Péton, Olivier 616

Quilliot, Alain 641

Rajendran, Chandrasekharan 14, 285, 604
Ramudhin, Amar 335
Rasovska, Ivana 449

Rebaine, Djamal 210
Reiner, Gerald 87
Respen, Jean 186
Rezaei-Malek, Mohammad 352
Rezg, Nidhal 311, 591
Riane, Fouad 180
Ringard, Kivin 273
Romagnoli, Giovanni 437
Rossi, André 20
Rousseau, Louis-Martin 659

Saadani, Nour El Houda 382, 388
Sagaert, Yves R. 75
Sahin, Evren 8, 69, 192
Samuel, Karine 162
Santirojanakul, Sirikorn 204
Sauer, Nathalie 262
Sauvey, Christophe 262
Schwindt, Christoph 44
Sgarbossa, Fabio 32
Shang, Lei 268
Siadat, Ali 198, 352, 511
Song, Qinrui 424
Sui, Yang 250
Sun, Zhihui 430

Ta, Thanh Thuy Tien 273
Tahir, Adil 659
Tamayo, Simon 141
Tan, Kim-Hua 523
Tavakkoli-Moghaddam, Reza 352, 535
Tchernev, Nikolay 226, 232, 370, 480, 641
Thomas, André 291
Thomas, Philippe 291
Tonetti, Marco 99
Tosello, Guido 99, 105
Tran, Hanh Nhi 123
Trigos, Federico 376
Trilling, Lorraine 405
Troncet, Emmanuel 26

Uckelmann, Dieter 437

Venkatadri, Uday 317, 340
Vernadat, François 511
Vialletelle, Philippe 323
Vincent, T'Kindt 268
Vinot, Marina 641
Vlachou, Ekaterini 150
Vo, Thi Le Hoa 566

Wajdi, Trabelsi 262
Wang, Ching-Hsin 523
Wang, Nengmin 610
Wanitwattanakosol, Jirapat 204
Waschneck, Bernd 87
Wei, Tao 297
Widad, Naji 279
Wiener, Patrick 156

Yalaoui, Farouk 635
Yang, Danqin 580
Yazan, Devrim Murat 455
Yazdanpanah, Vahid 455
Yugma, Claude 26

Zaabar, Imen 57
Zacharias, Miriam 285
Zammori, Francesco 437
Zhang, Dan 297
Zheng, Feifeng 238, 244, 250, 418, 424, 430, 647, 653
Zidna, Ahmed 664
Zimmermann, Emmanuel 291
Zouari, Belhassen 382, 388
Zufferey, Nicolas 186