Proceedings SOR

Rupnik V. and L. Bogataj (Editors): The 1st Symposium on Operational Research, SOR'93. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1993, 310 pp.

Rupnik V. and M. Bogataj (Editors): The 2nd International Symposium on Operational Research in Slovenia, SOR'94. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1994, 275 pp.

Rupnik V. and M. Bogataj (Editors): The 3rd International Symposium on Operational Research in Slovenia, SOR'95. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1995, 175 pp.

Rupnik V., L. Zadnik Stirn and S. Drobne (Editors.): The 4th International Symposium on Operational Research in Slovenia, SOR'97. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1997, 366 pp. ISBN 961-6165-05-4.

Rupnik V., L. Zadnik Stirn and S. Drobne (Editors.): The 5th International Symposium on Operational Research SOR '99, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1999, 300 pp. ISBN 961-6165-08-9.

Lenart L., L. Zadnik Stirn and S. Drobne (Editors.): The 6th International Symposium on Operational Research SOR '01, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2001, 403 pp. ISBN 961-6165-12-7.

Zadnik Stirn L., M. Bastič and S. Drobne (Editors): The 7th International Symposium on Operational Research SOR'03, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2003, 424 pp. ISBN 961-6165-15-1.

Zadnik Stirn L. and S. Drobne (Editors): The 8th International Symposium on Operational Research SOR'05, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2005, 426 pp. ISBN 961-6165-20-8.

Zadnik Stirn L. and S. Drobne (Editors): The 9th International Symposium on Operational Research SOR'07, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2007, 460 pp. ISBN 978-961-6165-25-9.

Zadnik Stirn L., J. Žerovnik, S. Drobne and A. Lisec (Editors): The 10th International Symposium on Operational Research SOR'09, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2009, 604 pp. ISBN 978-961-6165-30-3.

Zadnik Stirn L., J. Žerovnik, J. Povh, S. Drobne and A. Lisec (Editors): The 11th International Symposium on Operational Research SOR'11, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2011, 358 pp. ISBN 978-961-6165-35-8.

Zadnik Stirn L., J. Žerovnik, J. Povh, S. Drobne and A. Lisec (Editors): The 12th International Symposium on Operational Research SOR'13, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2013, 390 pp. ISBN 978-961-6165-40-2.

Proceedings SOR'15

Proceedings of the 13th International Symposium on OPERATIONAL RESEARCH

SOR '15

Bled, Slovenia
September 23-25, 2015

Edited by:
L. Zadnik Stirn • J. Žerovnik • M. Kljajić Borštnar • S. Drobne

SOR '15 Proceedings

The 13th International Symposium on Operational Research in Slovenia

Bled, SLOVENIA, September 23 - 25, 2015

Edited by:

L. Zadnik Stirn, J. Žerovnik, M. Kljajić Borštar and S. Drobne

Slovenian Society INFORMATIKA (SDI)
Section for Operational Research (SOR)

Proceedings of the 13th International Symposium on Operational Research
SOR'15 in Slovenia, Bled, September 23 - 25, 2015.

Organiser : Slovenian Society Informatika – Section for Operational Research, SI-1000 Ljubljana, Litostrojska cesta 54, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

Co-organiser : University of Maribor, Faculty of Organizational Sciences, SI-4000 Kranj, Kidričeva cesta 55a, Slovenia (http://www.fov.um.si/)

First published in Slovenia in 2015 by Slovenian Society Informatika – Section for Operational Research, SI 1000 Ljubljana, Litostrojska cesta 54, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

519.8(082)
519.8:005.745(082)
519.81:519.233.3/.5(082)

INTERNATIONAL Symposium on Operational Research in Slovenia (13 ; 2015 ; Bled)
SOR '15 proceedings / The 13th International Symposium on Operational Research in Slovenia, Bled, Slovenia, September 23-25, 2015 ; [organiser] Slovenian Society Informatika (SDI), Section for Operational Research (SOR) ; edited by L. Zadnik Stirn ... [et al.]. - Ljubljana : Slovenian Society Informatika, Section for Operational Research, 2015

ISBN 978-961-6165-45-7
1. Slovensko društvo Informatika. Sekcija za operacijske raziskave
281141504

Proceedings of the 13th International Symposium on Operational Research in Slovenia (SOR'15) is cited in: ISI (Index to Scientific & Technical Proceedings on CD-ROM and ISI/ISTP&B online database), Current Mathematical Publications, Mathematical Review, MathSci, Zentralblatt für Mathematic / Mathematics Abstracts, MATH on STN International, CompactMath, INSPEC, Journal of Economic Literature

Technical editor : Samo Drobne
Designed by : Samo Drobne
Printed by : Statistical Office of the Republik of Slovenia, Ljubljana, Slovenia

The 13th International Symposium on Operational Research in Slovenia - SOR '15
Bled, SLOVENIA, September 23 - 25, 2015

Program Committee:

L. Zadnik Stirn, University of Ljubljana, Biotechnical Faculty, Ljubljana, Slovenia, Chair
M. Kljajić Borštnar, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia, Co-Chair
J. Žerovnik, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia, Co-Chair
Z. Babić, University of Split, Faculty of Economics, Department for Quantitative Methods, Split, Croatia
M. Bastič, University of Maribor, Faculty of Business and Economics, Maribor, Slovenia
M. Bogataj, MEDIFAS-CERRISK, Šempeter pri Gorici, Slovenia
M. Bohanec, Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
K. Cechlárová, P. J. Šafárik University, Institute of Mathematics, Košice, Slovakia
T. Csendes, University of Szeged, Department of Applied Informatics, Szeged, Hungary
S. Drobne, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia
K. Dumičić, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
M. Ferjan, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
W. Gutjahr, University of Vienna, Department of Statistics and Decision Support Systems, Vienna, Austria
J. Jablonsky, University of Economics, Faculty of Informatics and Statistics, Prague, Czech Republic
D. Jukić, University of Osijek, Department of Mathematics, Osijek, Croatia
S. Klavžar, University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
D. Kofjač, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
J. Kušar, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia
A. Lisec, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia
Z. Lukač, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
L. Neralić, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
M. Pejić Bach, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
M. Petitjean, University Paris 7, INSERM, Paris, France
U. Pferschy, University of Graz, Department of Statistics and Operations Research, Graz, Austria
J. Povh, Faculty of Information Studies, Novo mesto, Slovenia
V. Rajkovič, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
M. S. Rauner, University of Vienna, Depart. of Innovation and Technology Management, Vienna, Austria
R. Sotirov, Department of Econometrics and Operations Research, Tilburg University, The Netherlands
K. Šorić, Zagreb School of Economics and Management, Zagreb, Croatia
D. Škulj, University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia
P. Šparl, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
T. Trzaskalik, University of Economics, Department of Operational Research Katowice, Poland
G. W. Weber, Middle East Technical University, Institute of Applied Mathematics, Ankara, Turkey
M. Zekić Sušac, University of Osijek, Faculty of Economics, Croatia

Organizing Committee:

S. Drobne, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia, Chair
P. Gorjanc, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia, Co-Chair
M. Bogataj, The Mediterranean Institute for Advanced Studies, Šempeter pri Gorici, Slovenia
M. Kljajić Borštnar, University of Maribor, Faculty of Organizational Science, Kranj, Slovenia
J. Povh, Faculty of Information Studies, Novo mesto, Slovenia
L. Zadnik Stirn, University of Ljubljana, Biotechnical Faculty, Ljubljana, Slovenia
J. Žerovnik, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

Chairs:

V. Bahovec, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
D. Battini, University of Padua, Department of Management and Engineering, Padua, Italy
M. Bogataj, MEDIFAS-CERRISK, Šempeter pri Gorici, Slovenia
M. Bohanec, Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
D. Bokal, University of Maribor; Faculty of Natural Sciences and Mathematics, Maribor, Slovenia
K. Cechlárová, P. J. Šafárik University, Institute of Mathematics, Košice, Slovakia
S. Drobne, University of Ljubljana, Faculty of Civil Engineering and Geodesy, Ljubljana, Slovenia
K. Dumičić, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
L. Ferbar, University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia
B. Filipič, Jožef Stefan Institute, Department of Intelligent Systems, Ljubljana, Slovenia
H. Gaspars-Wieloch, Poznan University of Economics, Dept. of Operations Research, Poznan, Poland
R. W. Grubbström, Linköping Institute of Technology, Linköping, Sweden
M. Kljajić Borštnar, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
M. Kobal, Arctur d.o.o., Nova Gorica, Slovenia
N. Kurnoga, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
J. Kušar, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia
S. Pivac, University of Split, Faculty of Economics, Split, Croatia
J. Povh, Faculty of Information Studies, Novo mesto, Slovenia
V. Rajković, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
J. Šilc, Jožef Stefan Institute, Computer Systems Department, Ljubljana, Slovenia
T. Trzaskalik, University of Economics in Katowice, Faculty of Informatics and Communication, Katowice, Poland
L. Zadnik Stirn, University of Ljubljana, Biotechical Faculty, Ljubljana, Slovenia
M. Zekić-Susac, University of Josip Juraj Strossmayer in Osijek, Faculty of Economics, Osijek, Croatia
J. Žerovnik, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

Preface

This volume, Proceedings of The 13th International Symposium on Operations Research, called SOR'15, contains papers presented at SOR'15 (http://sor15.fov.uni-mb.si/) that was organized by Slovenian Society INFORMATIKA (SDI), Section for Operations Research (SOR) and University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia, held in Bled, Slovenia, from September 23 to September 25, 2015. The volume contains blindly reviewed papers or abstracts of talks presented at the symposium.

The opening address at SOR'15 was given by Prof. Dr. L. Zadnik Stirn, President of the Slovenian Section of Operations Research, Prof. Dr. Maja Makovec Brenčič, Minister of Education, Science and Sport, Prof. Dr. József Györkös, Director of Slovenian Research Agency, Mr. Niko Schlamberger, the President of Slovenian Society Informatika, Prof. Dr. Marko Ferjan, the Dean of Faculty of Organizational Sciences, Dr. Sarah Fores, The Association of European Operational Research Societies (EURO), and presidents/representatives of a number of Operations Research Societies from abroad.

SOR'15 is the scientific event in the area of operations research, another one in the traditional series of the biannual international OR conferences, organized in Slovenia by SDI-SOR. It is a continuity of twelve previous symposia. The main objective of SOR'15 is to advance knowledge, interest and education in OR in Slovenia, in Europe and worldwide in order to build the intellectual and social capital that are essential in maintaining the identity of OR, especially at a time when interdisciplinary collaboration is proclaimed as significantly important in resolving problems facing the current challenging times. Further, by joining IFORS and EURO, the SDI-SOR agreed to work together with diverse disciplines, i.e. to balance the depth of theoretical knowledge in OR and the understanding of theory, methods and problems in other areas within and beyond OR. We believe that SOR'15 creates the advantage of these objectives, contributes to the quality and reputation of OR by presenting and exchanging new developments, opinions, experiences in the OR theory and practice.

SOR'15 was highlighted by a distinguished set of six keynote speakers. The first part of the Proceedings SOR'15 comprises invited abstracts, presented by six outstanding scientists: Professor Dr. Aharon Ben-Tal, William Davidson Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Haifa, Israel, Assoc. Professor Dr. Sergio Cabello, University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia, Dr. Stefano Cozzini, Centro DEMOCRITOS, Istituto Officina dei Materiali CNR-IOM, Trieste, Italy, Assist. Professor Dr. Nebojša Gvozdenović, University of Novi Sad, Faculty of Economics, Subotica, Serbia, Professor Dr. Gerhard Wilhelm Weber, Middle East Technical University, Institute of Applied Mathematics, Ankara, Turkey, and Professor Dr. Ou Tang, Linköping University, Department of Management and Engineering, Linköping, Sweden. Proceedings includes 93 papers or abstracts written by 191 authors. Most of the authors of the contributed papers came from Slovenia (76), then from Croatia (42), Slovak Republic (14), Czech Republic (9), Turkey (9), Italy (8), Greece (5), Israel (4), Poland (4), Spain (3), Sweden (3), Bulgaria (2), France (2), Hong Kong (2), Iran (2), Russian Federation (2), Hungary (1), Ireland (1), Serbia (1) and The Netherlands (1). The papers published in the Proceedings are divided into Plenary Lectures (6 abstracts), four special sessions: Qualitative Multiple Criteria Decision Making (6 papers), Inventory Research (7), Metaheuristic Optimization (7), and Big Data (4), and nine sessions: Mathematical Programming and Optimization (7), Graphs and their Applications (5), Multiple Criteria Decision Making (5), Econometric Models and Statistics (10), Production

(7), Finance and Investments (7), Location and Transport (7), Environment and Human Resources (9), OR Perspectives (6).

The Proceedings of the previous twelve International Symposia on Operations Research organized by the Slovenian Section of Operations Research are indexed in the following secondary and tertiary publications: Current Mathematical Publications, Mathematical Review, Zentralblatt fuer Mathematik/Mathematics Abstracts, MATH on STN International and CompactMath, INSPEC. The Proceedings SOR'15 are expected to be covered by the same bibliographic databases.

The success of the scientific events at SOR'15 and the present proceedings should be seen as a result of joint effort. On behalf of the organizers we would like to express our sincere thanks to all who have supported us in preparing the event. We would not have succeeded in attracting so many distinguished speakers from all over the world without the engagement and the advice of active members of the Slovenian Section of Operations Research. Many thanks to them. Further, we would like to express our deepest gratitude to prominent keynote speakers, to the members of the Program and Organizing Committees, to the referees who raised the quality of the SOR'15 by their useful suggestions, section's chairs, and to all the numerous people - far too many to be listed here individually - who helped in carrying out The 13th International Symposium on Operations Research SOR'15 and in putting together these Proceedings. Last but not least, we appreciate the authors' efforts in preparing and presenting the papers, which made The 13th Symposium on Operations Research SOR'15 successful.

We would like to express a special gratitude to The Association of European Operational Research Societies (EURO) for a financial support.

Bled, September 23, 2015

Lidija Zadnik Stirn
Janez Žerovnik
Mirjana Kljajić Borštnar
Samo Drobne
(Editors)

Contents

Session 8: Environment and Human Resources *471*

Session 9: OR Perspectives *527*

APPENDIX *561*

Author index

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Plenary Lectures

ROBUST SOLUTIONS OF UNCERTAIN OPTIMIZATION PROBLEMS UNDER AMBIGUOUS STOCHASTIC DATA

Aharon Ben-Tal
William Davidson Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology
Technion city, 32000 Haifa, Israel
abental@ie.technion.ac.il

We show how robust optimization (RO) can provide tractable, safe approximation to probabilistic constraints (chance constraints) even under partial information (ambiguity) on the random parameters.

In particular, we address the case where the only available information is on means and dispersion measures. Unlike previous attempts where the dispersion measure is the *variance*, here we derive *tight* approximations when the dispersion measure is the MAD (mean absolute deviation). The theory is applied to problems in portfolio selection, inventory management and antenna array design.

TWO APPLICATIONS OF GEOMETRIC OPTIMIZATION

Sergio Cabello
Faculty of Mathematics and Physics, University of Ljubljana
Jadranska 19, SI-1000 Ljubljana, Slovenia
sergio.cabello@fmf.uni-lj.si

In this paper we will encounter two unrelated problems in the area of Geometric Optimization and some of the algorithmic ideas leading to their efficient solution. The work reported has been done recently with Drago Bokal, David Eppstein, Panos Giannopoulos and Lazar Milinkovic.

Motivated by sensor networks, we will discuss the following minimum separation problem: given a set of disks in the plane and two points s and t, not covered by any of the disks, compute the minimum number of disks one needs to retain so that any path connecting s to t intersects some of the retained disks.

Motivated by trajectory analysis, we will discuss the following problem: given a sequence of points in the plane describing a trajectory, find all maximal subtrajectories with a prescribed hereditary property, like for example, having diameter one.

HPC AS A KEY-ENABLING TECHNOLOGY

Stefano Cozzini
Centro DEMOCRITOS, Istituto Officina dei Materiali CNR-IOM
c/o Scuola Internazionale Superiore di Studi Avanzati (SISSA)
Via Bonomea 265, 34136 Trieste, Italy
stefano.cozzini@sissa.it

High Performance Computing (HPC) has become an enabling-technology for science and industry with implication as well on society in general. The new generation of highly performance computers, together with the innovative approach of Cloud Computing makes now easily available computing resources on demand tailored in any specific need. HPC is becoming more and more a pervasive tool allowing industry and academia to develop new products, services and research which can enforce positions on the competitive worldwide arena. 'Today, to Out-Compute is to Out-Compete' best describes the role of HPC. HPC is also recognized as crucial in addressing the Grand Societal Challenges.

In this paper, after briefly presenting the key concepts of HPC and Cloud Computing I will illustrate opportunities enabled by such technologies. I will also review some of the drawback/difficulties that are being currently addressed by the HPC community.

OR TECHNIQUES IN REAL WORLD VEHICLE ROUTING

Nebojša Gvozdenović
University of Novi Sad, Faculty of Economics
Segedinski put 9-11, 24000 Subotica, Serbia
nebojsa.gvozdenovic@gmail.com

OR solution techniques for vehicle routing problems led to many real world applications during the last 25 years. Recently, vehicle routing software industry offered several SaaS models. In this presentation, we first give a short survey on vehicle routing theory and practice. As a contribution, we present innovative OR techniques that resolve some challenges of the SaaS vehicle routing business. In particular, we present the fastest time and distance matrix generation technique that relies on global map APIs and opens new perspectives for real time routing. Finally, we show our innovative techniques for synchronized planning of production, inventory and transport.

A NEW CONTRIBUTION TO OPERATIONAL RESEARCH: OPTIMAL CONTROL OF STOCHASTIC HYBRID SYSTEMS WITH JUMPS - WITH APPLICATIONS IN FINANCE, ECONOMICS, ENGINEERING AND BIOSCIENCES

Gerhard Wilhelm Weber, Emel Savku, Diogo Pinheiro and Nuno Azevedo
Middle East Technical University, Üniversiteler Mahallesi
Institute of Applied Mathematics
Dumlupınar Bulvarı 1, 06800 Çankaya Ankara, Turkey
gweber@metu.edu.tr

In this paper, we contribute to modern OR by hybrid, e.g., mixed continuous-discrete dynamics of stochastic differential equations with jumps and to its optimal control. These hybrid systems allow the representation of random regime switches and are of growing importance in economics, finance, science and engineering. We introduce two new approaches to this area of stochastic optimal control: one is based on the finding of closed-form solutions, the other one on a discrete-time numerical approximation scheme. The presentation ends with a conclusion and an outlook to future studies.

REMANUFACTURING AND CORE ACQUISITION

Shuoguo Wei and **Ou Tang**
Linköping University, Department of Management and Engineering
Division of Production Economics, 581 83 Linköping, Sweden
ou.tang@liu.se

Challenges in the core acquisition process mainly result from the uncertainties of: return volume and timing, and core quality. Core acquisition management is to actively manage these uncertainties though various activities. This presentation aims to provide an overview of core acquisition management issues and then present a core quality classification/refund policies model for improving the acquisition management. Using quality classification methods, refund policies with different numbers of quality classes are studied. Under the assumption of uniformly distributed quality, analytical solutions for these refund policies are derived. Numerical examples indicate that the customers' valuation of cores is an important factor influencing the return rates and the remanufacturer's profit. Refund policies with a small number of quality classes could already bring major advantages. Credit refund policies (without deposits) are included for comparisons.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Special Session 1:

Qualitative Multiple Criteria Decision Making

A QUALITATIVE MULTI-CRITERIA MODEL FOR THE EVALUATION OF ELECTRIC ENERGY PRODUCTION TECHNOLOGIES IN SLOVENIA

Marko Bohanec[1], **Nejc Trdin**[1,2], **Branko Kontić**[1]
[1] Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
E-mail: {marko.bohanec, nejc.trdin, branko.kontic}@ijs.si

Abstract: A methodological approach to the strategic evaluation of electric energy production technologies in Slovenia is presented. The aim of this work is to make a transparent and reproducible identification of reliable, rational, and environmentally sound production of electric energy in Slovenia by 2050. The approach is based on a qualitative multi-criteria modelling method DEX and consists of three stages: (1) assessment of individual technologies for electricity production, (2) assessment of mixtures of technologies, and (3) evaluation of scenarios of shutting-down existing old power plants and constructing the new ones until 2050. Technology alternatives include both conventional and renewable energy sources: coal fired, gas fired, biomass fired, oil fired, nuclear, hydro, wind, and photovoltaic. The results indicate that only mixtures of nuclear, hydro, and gas fired technologies can meet expected energy needs in a sufficiently reliable and rational way.

Keywords: Electric energy production technology, power plants, decision analysis, multi-criteria decision modelling, decision rules, qualitative model, method DEX

1 INTRODUCTION

Electric energy is a strategic resource that plays a vital role in the operation and development of every country. Electric energy production is a complex process, which requires strategic management and careful planning years ahead. The selection of appropriate technologies for electric energy production depends on a number of factors: energy needs of a country, availability of fuel and other natural resources, feasibility, efficiency, effectiveness and rationality of production, environmental impacts, and many more. Not only that these factors are multiple, they are often conflicting and influence the decisions in a variety of ways; thus they have to be carefully assessed individually and against each other.

For this kind of problems, Operations Research provides Multi-Criteria Decision Modelling (MCDM) methods [4, 6] that assess decision alternatives using multiple criteria. Each alternative is first assessed according to each criterion. These individual assessments are then aggregated into an overall evaluation of the alternative, which provides a basis for comparison, ranking and analysis of alternatives, and eventual selection of the best one. MCDM methods are commonly employed in the assessment of electric energy production [8], either in a general setting [10], or considering the specifics of countries, such as Germany [5] or Portugal [9].

In Slovenia, almost 13 TWh of electricity is consumed annually (net figure for the year 2014). The electricity is produced by thermal, hydro, and nuclear power plants in approximately equal shares. After a recent introduction of a controversial and expensive Unit 6 of the coal-fired power plant at Šoštanj (TEŠ6), which is expected to produce up to 3.5 TWh of electricity annually, there are important decisions to be taken for the next decades. Slovenia has one nuclear power plant in Krško, which produces around 5 TWh of electricity annually, and which will be according to plans closed down in 2023. However, there is an option to extend its operation until 2043. Another large power plant, coal-fired unit TEŠ5, will be closed down in 2027. There are plans to finalize, by 2025, two hydro power plants on the lower Sava river, which is the last Slovenian water resource available for hydro power

plants. There are also plans to introduce gas fired plants, and energy production from renewable sources: wind, biomass and sun.

In order to contribute to strategic planning of electrical energy production in Slovenia, a project called OVJE [7] was conducted with the aim to make a transparent and reproducible identification of reliable, rational, and environmentally sound production of electric energy in Slovenia by 2050. Eight electric energy production technologies were considered: hydro, coal, oil, gas, nuclear, biomass, photovoltaic (PV), and wind. Hereafter we present the methodological approach to this sustainability appraisal and summarize the main results.

2 METHODS

The methodological approach is based on a qualitative MCDM method DEX, which is used in three stages, and involves two MCDM models and one simulation model:

1. *Model T*: Evaluation of eight individual electric energy production technologies.
2. *Model M*: Evaluation of mixtures of technologies, considering the shares of individual technologies in the total installed capacity.
3. *Model S*: Simulation of possible implementations of technology mixtures in the period 2014–2050, taking into account various scenarios of shutting down the existing power plants and constructing new ones.

2.1 Qualitative Multi-Attribute Modelling Method DEX

DEX (Decision EXpert) [1] is a *multi-criteria decision modelling* method. As all other MCDM methods, it is aimed at the evaluation and analysis of a set of decision *alternatives* $A = \{a_1, a_2, \dots, a_m\}$. These alternatives are described with a set of variables $X = \{x_1, x_2, \dots, x_n\}$, called *attributes*. Each attribute represents some observed or evaluated property of alternatives, such as "price", "quality", and "efficiency".

DEX is a *hierarchical* method. This means that the attributes X are organized in a *hierarchy*. Observed in the top-down direction, the hierarchy represents a decomposition of the decision problem into sub-problems. The bottom-up direction denotes dependence, so that higher-level attributes depend on the lower-level, more elementary ones. The most elementary attributes, called *basic attributes*, appear as terminal nodes of the hierarchy and represent the basic observable characteristics of alternatives. Higher-level attributes, which depend on one or more lower-level ones, are called *aggregated attributes*; they represent evaluations of alternatives. The topmost nodes (usually, there is only one such node) are called *roots* and represent the final evaluation(s) of alternatives.

Furthermore, DEX is a *qualitative* method. While most of MCDM methods are quantitative and thus use numeric variables, qualitative methods use symbolic ones. In DEX, each attribute $x_i \in X$ has a *value scale* $D_i = \{v_{i1}, v_{i2}, \dots, v_{ik_i}\}$, where each v_{ij} represents some ordinary word, such as "low", "high", "acceptable", "excellent". Scales are usually small, containing 2 to 5 values. Also, scales are usually preferentially ordered so that $v_{i1} \preccurlyeq v_{i2} \preccurlyeq \cdots \preccurlyeq v_{ik_i}$ (here, $a \preccurlyeq b$ denotes weak preference: the value b is preferred equally or more than a). Attributes that have preferentially ordered scales are called *criteria* [4].

Finally, DEX is a *rule-based* method. The bottom-up aggregation of alternatives' values is defined in terms of *decision rules*, which are specified by the decision maker and usually represented in the form of *decision tables*. Suppose that $x_{(0)} \in X$ is some aggregated attribute and that $x_{(1)}, x_{(2)}, \dots, x_{(r)} \in X$ are its immediate descendants in the hierarchy. Then, the aggregation function $x_{(0)} = f_{(0)}(x_{(1)}, x_{(2)}, \dots, x_{(r)})$ is defined with a set of decision rules of the form

if $x_{(1)} = v_{(1)}$ **and** $x_{(2)} = v_{(2)}$ **and** ... **and** $x_{(r)} = v_{(r)}$ **then** $x_{(0)} = v_{(0)}$

Here, $v_{(i)} \in D_{(i)}, i = 0,1, \ldots, r$.

The method DEX is implemented as DEXi [2], freely available software that supports both the development of DEX models and their application for the evaluation and analysis of decision alternatives. DEXi checks the quality of decision rules so that its models, when properly developed, are guaranteed to be *complete* (they provide evaluation results for all possible combinations of basic attributes' values) and *consistent* (defined aggregation functions obey the principle of dominance, i.e., they are monotone with respect to all preferentially ordered basic criteria).

For further information of DEX and DEXi, please refer to [1] and [2], respectively.

(a) Model T

Scales

Attribute	Scale
Technology	unsuit; weak; suit; ***good***; ***exc***
Rationality	inapprop; low; ***med***; ***high***
Contribution to development	low; med; ***high***
Economic	low; med; ***high***
Societal	low; med; ***high***
Economic-Technical advancement	low; med; ***high***
Technical level	low; med; ***high***
Expected development	low; med; ***high***
Economy	low; med; ***high***
Financial aspects	less_suit; suit; ***more_suit***
Energy price	high; med; ***low***
Financing	less_suit; suit; ***more_suit***
Financial sources	uncertain; less_certain; ***certain***
Financial shares	less_suit; suit; ***more_suit***
Long-term liabilities	less_suit; suit; ***more_suit***
Efficiency	low; med; ***high***
Energy ratio	low; med; ***high***
Return period	long; med; ***short***
Independence	low; med; ***high***
Dependence	v_high; high; med; low; ***none***
Land use and pollution	unsuit; less_suit; suit; ***more_suit***
Spatial availability	less_suit; suit; ***more_suit***
Land availability	low; med; ***high***
Energy share provision	low; med; ***high***
Resource protection	weak; present; ***effective***
Water protection	weak; present; ***effective***
Land protection	weak; present; ***effective***
Landscape protection	weak; present; ***effective***
Pollution	high; med; ***low***
Health impact	high; med; ***low***
Air pollution	high; med; ***low***
Greenhouse gases	high; med; ***low***
Other pollutants	high; med; ***low***
Public health status	low; med; ***high***
Contribution to development	low; med; ***high***
Feasibility	low; med; ***high***
Technical feasibility	low; med; ***high***
Technological complexity	less_suit; suit; ***more_suit***
Infrastructure availability	low; med; ***high***
Accessibility	low; med; ***high***
Fuel availability	low; med; ***high***
Fuel accessibility	low; med; ***high***
Economic feasibility	low; med; ***high***
Investment feasibility	low; med; ***high***
Return of investment	less_suit; suit; ***more_suit***
Spatial feasibility	low; med; ***high***
Societal feasibility	low; med; ***high***
Social acceptance	low; med; ***high***
Permitting	no; ***yes***
Spatial suitability	low; med; ***high***
Uncertainties	v_high; high; med; low; ***none***
Technological dependence	v_high; high; med; low; ***none***
Foreign dependence	v_high; high; med; low; ***none***
Construction	high; med; ***low***
Licences	strong_restr; moder_restr; ***no restr***
Operation	high; med; ***low***
Licences	strong_restr; moder_restr; ***no restr***
Contracts	strong_restr; moder_restr; ***no restr***
Special materials	strong_restr; moder_restr; ***no restr***
Weather dependence	high; med; ***low***
Fuel supply dependence	high; med; ***low***
Political stability	no; low; ***high***
Possible changes	neg; no; ***pos***
Possible societal changes	neg; no; ***pos***
Possible world changes	neg; no; ***pos***
Perception of risks	v_high; high; med; low; ***none***

(b) Model M

Scales

Attribute	Scale
Technology mix	unsuit; weak; suit; ***good***; ***exc***
Reasonability	unreas; less_reas; reas; ***desired***
Energy demand coverage	low; med; good; ***high***
Reliability of supply	low; med; high; ***v_high***
Availability	low; med; ***high***
Installed capacity	unsuit; suit; ***exceed***
Energy produced	unsuit; suit; ***exceed***
Base load	low; med; ***high***
Peaks	no; ***yes***
Uncertaintites	v_high; high; med; ***low***
Health impacts	high; med; ***low***
Possible changes	neg; no; ***pos***
Feasibility and rationality	weak; low; med; ***high***
Feasibility	low; med; ***high***
Economy	low; med; ***high***
Long-term appropriateness	low; med; ***high***
Fulfilment of goals and interests	low; med; ***high***
Environmental goals	low; med; ***high***
Low carbon	low; med; ***high***
Rational land use	low; med; ***high***
Nature protection	low; med; ***high***
National interests	low; med; ***high***
Independence	low; med; ***high***
Energy users capabilities	low; med; ***high***
Energy supply to all	low; med; ***high***
Protection of vulnerable groups	low; med; ***high***
Lifetime of supply	short; med; ***long***

(c) Decision rules

DEXi

Decision rules

	Rationality	Feasibility	Uncertainties	Technology
	43%	29%	28%	
1	inapprop	*	*	unsuit
2	<=low	<=med	v_high	unsuit
3	<=*med*	low	v_high	unsuit
4	>=low	low	high:med	weak
5	>=low	*high*	v_high	weak
6	>=*med*	>=med	v_high	weak
7	*high*	low	<=med	weak
8	*high*	*	v_high	weak
9	low:*med*	low	>=low	suit
10	>=low	low	low	suit
11	>=low	>=med	high	suit
12	low	>=med	>=med	*good*
13	low:*med*	med	med:low	*good*
14	>=low	>=med	med	*good*
15	*high*	low	*none*	*good*
16	>=*med*	>=med	*none*	*exc*
17	>=*med*	*high*	>=low	*exc*
18	*high*	>=med	>=low	*exc*

Figure 1: Hierarchical structure and value scales of (a) Model T and (b) Model M, and (c) example of decision rules that aggregate *Rationality*, *Feasibility* and *Uncertainties* into *Technology*

2.2 Model T: Evaluation of Technologies

The DEX model, used in the first stage of appraisal, is called *Model T* ('T' stands for "Technologies"). It is aimed at the evaluation and comparison of individual energy production technologies: $A = \{$Hydro, Coal, Oil, Gas, Nuclear, Biomass, PV, Wind$\}$. Evaluation criteria X are organised in a hierarchy shown in Figure 1(a). The hierarchy contains 36 basic and 28 aggregated attributes. There are two aggregated attributes that appear twice in Figure 1(a), because they affect more than one higher-level attribute: *Licenses* and *Contribution to development*. Figure 1(a) also shows attributes' value scales; all scales are preferentially ordered increasingly in the direction from left to right.

Model T consists of three main sub-trees of criteria: *Rationality*, *Feasibility*, and *Uncertainty*. *Rationality* assesses how much a particular technology contributes to the overall societal development, the economy, and the prudent use of land with low pollution. Each of these aspects is represented by a corresponding attribute and decomposed further. The sub-tree *Land use and pollution*, for instance, specifically addresses *Spatial availability*, *Pollution*, and *Health impacts*. Similarly, the assessment of *Feasibility* takes into account *Technical*, *Economic* and *Spatial feasibility*. *Uncertainty* evaluation comprises *Technological dependence* (in terms of foreign, uncontrollable factors, operation of supplier, and political stability), *Possible changes* in society and in the world, and *Perception of risks* with respect to technical advancement of a technology and trust into safety management system.

Since Model T contains 28 aggregated attributes, there are also 28 corresponding decision tables, which were defined by experts and decision analysts in the OVJE project. Here, we present only the one that corresponds to the root attribute *Technology*: Figure 1(c) shows a condensed form of decision rules that aggregate intermediate assessments of *Rationality*, *Feasibility* and *Uncertainties* into the overall evaluation of *Technology*. The first rule, for instance, says that whenever *Rationality* is inappropriate, then *Technology* is considered unsuitable, regardless on its *Feasibility* and *Uncertainties* (the symbol '*' denotes any value). The last rule defines *Technology* as excellent when its *Rationality* is high, *Feasibility* at least medium and *Uncertainties* low or better (the symbols '>=' and '<=' denote weak preference). The percentages shown in Figure 1(c) represent the importance of each attribute (determined by linear approximation of decision rules, see [2]). As indicated, *Rationality* is more important (43%) than *Feasibility* and *Uncertainties*, which are of similar importance (29% and 28%, respectively).

2.3 Model M: Evaluation of Technology Mixtures

While Model T evaluates individual technologies, *Model M* evaluates technology mixtures. A *technology mixture* is defined as a collection of technologies, considering a specific share of each technology in the total installed capacity. For example, some technology mixture may employ three technologies, nuclear, coal and hydro, with respective relative installed capacity shares of 0.3, 0.6 and 0.1; this mixture is denoted {nuclear/0.3, coal/0.6, hydro/0.1}.

Model M is structured as shown in Figure 1(b). The two top-level attributes, *Reasonability* and *Long-term appropriateness*, measure the certainty of supply by some mixture, and fulfilment of goals and interests: environmental, social, and national. In total, Model M has 15 basic and 12 aggregated attributes.

Models T and M are connected and used in succession. When evaluating mixtures with Model M, some of its basic attributes receive values from Model T: *Health impacts*, *Possible changes*, *Feasibility*, *Economy*, *Low carbon* (determined from *Greenhouse gasses*), *Rational land use* (from *Spatial availability*), *Nature protection* (from *Resource protection*), and *Independence*. The input values of the remaining basic attributes are determined from

scenarios (see section 2.4) for each mixture as a whole. The evaluation of mixtures with Model M takes into account the relative shares of individual technologies and employs an evaluation method based on probabilistic value distributions; see [11] for a detailed description of the method.

2.4 Model S: Simulation of Implementation Scenarios

In contrast with the two Models T and M, which are of multi-attribute type, *Model S* ('S' stands for "Scenarios") is a simulation model. It uses Models T and M, and "runs" them through the years 2014 to 2050. For each year, Model S evaluates technology mixtures that are expected to be in place in Slovenia in that year according to different management scenarios. The following management decisions have been considered:

1. Closing-down of the nuclear power plant (NPP) Krško Unit1 in 2023.
2. Construction of Unit2 at the NPP Krško by 2025.
3. Finalisation of the two hydro power plants on the lower Sava river by 2025.
4. Construction of a gas fired power plant by 2025.
5. Closing-down of Unit5 of the coal fired power plant at Šoštanj in 2027.
6. Construction of the chain of hydro power plants on the mid Sava river by 2035.

Since each of these decisions can be either yes or no, they collectively make $2^6 = 64$ possible scenarios. The simulation of these scenarios is implemented in an on-line decision support system [3].

3 RESULTS

In the first stage, individual electric energy production technologies were evaluated by Model T as shown in Figure 2. In addition to the overall evaluation (second row), Figure 2 displays intermediate evaluation results obtained on two lower levels of the Model T hierarchy. Some evaluation values are presented as intervals, which are due to uncertainties regarding future values of several evaluation criteria. The lower and upper interval bounds correspond to pessimistic and optimistic assessment of evaluation criteria, respectively.

Attribute	Hydro	Coal	Oil	Gas	Nuclear	Bio	PV	Wind	Impor
Technology	suit - ***exc***	**unsuit**	**unsuit**	weak - ***good***	weak - ***exc***	**unsuit**	**unsuit**	**unsuit**	**unsuit**
Rationality	low - ***high***	**inapprop**	**inapprop**	***high***	***high***	**inapprop**	**inapprop** - low	**inapprop**	**inapprop**
Contribution to development	med - ***high***	***high***	med	***high***	***high***	med	**low** - med	**low**	**low**
Economy	med - ***high***	***high***	**low**	med - ***high***	med - ***high***	**low**	**low**	**low**	med
Land use and pollution	**less_suit** - ***more_suit***	**unsuit**	**unsuit**	***more_suit***	***more_suit***	**less_suit**	**unsuit** - ***more_suit***	**unsuit** - **less_suit**	**less_suit**
Feasibility	***high***	***high***	***high***	***high***	**low** - ***high***	**low** - med	**low**	**low**	***high***
Technical feasibility	***high***	***high***	***high***	***high***	***high***	med	med - ***high***	med	med
Economic feasibility	***high***	med	med	med	***high***	**low** - med	**low**	**low**	***high***
Spatial feasibility	***high***	***high***	***high***	***high***	**low** - ***high***	**low** - ***high***	**low** - ***high***	**low** - ***high***	***high***
Uncertainties	high - ***none***	low	**v_high** - low	**v_high** - med	**v_high** - low	low	**v_high**	**v_high**	med
Technological dependence	high - ***none***	low	**v_high** - med	**v_high** - med	**v_high** - low	med	**v_high**	**v_high**	high
Possible changes	***pos***	***pos***	no	no	***pos***	no	no	no	***pos***
Perception of risks	med - ***none***	med - low	***none***	high - med	**v_high** - low	***none***	low	***none***	low

Figure 2: Evaluation results of individual electric energy production technologies with Model T

These results indicate that there are only three technologies of sufficient suitability for Slovenia: Hydro, Gas, and Nuclear. Among these, Hydro is the best. Gas and Nuclear are similar, with Nuclear worse in terms of *Feasibility* and *Perception of risks*, but better in terms of *Economic feasibility* and *Possible changes*. Coal and Oil are unsuitable particularly because of inappropriate *Rationality* due to *Land use and pollution*. All the remaining "green" technologies are unsuitable for a number of reasons, including *Economy*, *Land use*, *Economic feasibility* and *Technological dependence*. See [7] for a more detailed justification of this assessment and its consequences.

Results of simulating the 64 scenarios [7, 3] indicate that only the mixtures that include extension of operation of Unit1 of NPP Krško, construction and operation of Unit2 of NPP

Krško, construction of all planned hydro power plants on the Sava river and construction of the gas fired thermal power plant ensure coverage of energy needs by 2050 in Slovenia. Renewable energy sources – wind and PV – do not constitute a sustainable choice since they are not reliable due to land-use context (almost 40% of the Slovenian territory is under Natura2000 protection regime), and are consequently not capable of meeting a substantial share of energy demands; they may only constitute an option for covering 8% to 15% of energy needs.

4 CONCLUSION

With the aim to contribute to better strategic planning of electrical energy production in Slovenia, this work proposes a systematic, transparent and reproducible sustainability appraisal of technologies and strategic management scenarios. The approach is based on qualitative multi-attribute modelling and simulation, and proceeds in three stages: assessment of (1) individual technologies, (2) technology mixtures and (3) management scenarios in the period 2014–2050. The method is implemented in an on-line decision support system [3].

Evaluation results clearly identify three main technologies that are most suitable for Slovenia: Hydro, Gas, and Nuclear. Only a proper mixture of these technologies is reliable and rational in the context of meeting expected energy needs. Biomass, wind and photovoltaic sources of energy are less sustainable than others and may provide only from 8% to 15% of energy in Slovenia.

References

[1] Bohanec, M., Rajkovič, V., Bratko, I., Zupan, B., Žnidaršič, M. (2013). DEX methodology: Three decades of qualitative multi-attribute modelling. Informatica 37, 49-54.

[2] Bohanec, M. (2014). DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version 4.01. Report IJS DP-11739, Ljubljana: Jožef Stefan Institute.

[3] Electric Energy Needs and Production Through Time. Ljubljana: Jožef Stefan Institute. http://sepo.ijs.si/naloge/OVJE/energetic_scenario_comparative_model/

[4] Figueira, J., Greco, S. and Ehrgott, M. (2005). Multi criteria decision analysis: State of the art surveys. London: Springer.

[5] Hirschberg, S., Dones, R., Heck, T., Burgherr, P., Schenler, W. and Bauer, C. (2004). Sustainability of electricity supply technologies under German conditions: A comparative evaluation, Paul Scherrer Institut, PSI Bericht Nr. 04-15, ISSN 1019-0643.

[6] Ishizaka, A. and Nemery, P. (2013). Multi-criteria decision analysis: Methods and software. Somerset: Wiley.

[7] Kontić, B., Kontić, D., Zagorc, S., Matko, M., Dermol, U., Bohanec, M. and Trdin, N. (2014). Ocena vzdržnosti za razvoj energetike v Sloveniji do leta 2030 s poudarkom na jedrski tehnologiji (Appraisal of Sustainability of Electrical Energy Production in Slovenia with the Emphasis on Nuclear Technology). Report IJS DP-11583, Ljubljana: Jožef Stefan Institute.

[8] Pohekar, S.D. and Ramachandran, M. (2004). Application of multi-criteria decision making to sustainable energy planning – a review. Renewable and Sustainable Energy Reviews, 8, 365-381.

[9] Ribeiro, F., Ferreira, P. and Araújo, M. (2013). Evaluating future scenarios for the power generation sector using a Multi-Criteria Decision Analysis (MCDA) tool: The Portuguese case. Energy 52, 126–136.

[10] Stein, E.W. (2013). A comprehensive multi-criteria model to rank electric energy production technologies. Renewable and Sustainable Energy Reviews 22, 640–654.

[11] Trdin, N. and Bohanec, M. (2014). New Generation Platform for Multi-Criteria Decision Making with Method DEX. DSS 2.0 – Supporting Decision Making with New Technologies, Supplemental Proceedings, In: Phillips-Wren, G., Carlsson, S., Burstein, F., Respício, A., Brézillon, P. (Eds.), IFIP Working Group 8.3 digital publications, ISSN 1647-7103, 12.

DECISION SUPPORT MODELLING FOR ENVIRONMENTALLY SAFE APPLICATION OF PESTICIDES USED IN AGRICULTURE

Marko Debeljak
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
marko.debeljak@ijs.si

Sašo Džeroski
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
saso.dzeroski@ijs.si

Vladimir Kuzmanovski
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
vladimir.kuzmanovski@ijs.si

Jonathan Marks-Perreau
ARVALIS, Institut du végétal
CS 30200 Estrées Mons, France
j.marksperreau@arvalisinstitutduvegetal.fr

Aneta Trajanov
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
aneta.trajanov@ijs.si

Benoit Real
ARVALIS, Institut du végétal
91 720 Boigneville, France
b.real@arvalisinstitutduvegetal.fr

Abstract:
The application of pesticides in agriculture is not always safe for human health and the environment. Despite being officialy approved by authorities in terms of ecological risk, they appear in surface and ground water in concentrations above the official thresholds. To reduce and eliminate water pollution with pesticides, a decision support system (DSS) for ecological risk assessment of pesticide applications and ecological risk management, is proposed. The DSS is built by using the framework proposed by the US Environmental Protection Agency and Multi Criteria Decision Analysis (MCDA) implemented into the DEX (Decision EXpert) integrative methodology was implemented to build qualitative multi-attribute decision models by the DEXi modelling tool. A conceptual solution is demonstrated on an assessment of a proposed crop management plan for winter wheat, where the herbicide Isoproturon is planned to be applied. The DSS identified that the proposed plan is risky, therefore mitigation measures that have to be included in the proposed crop management plan are proposed.

Keywords: pesticides, risk assessment, risk management, mitigation measures, qualitative MCDM, DEX methodology

1 INTRODUCTION

The use of pesticides in agriculture has to be implemented in accordance with safe and environmentally sound agricultural crop management. Their use must be consistent with the European water framework directive [1] and the Directive on the sustainable use of plant protection products [2] in order to provide the most effective protection of surface and ground waters through the implementation of best crop management practices.

Though the crop management uses active substances previously approved for commercial use respecting EU regulations (No 1107/2009) [3] and permitted according to the Commission Implementing Regulation (EU) No 540/2011 [4], they can be still found in surface and ground water in concentrations above the official thresholds. The main reasons for the pollution of waters with pesticides are the inappropriate use and storage of pesticides [5].

Each approved and permitted active substance has passed very rigorous ecological risk assessment during its registration process. The procedure for preregistration risk assessment is described in detail by the respective authorities (e.g., European Food Safety Authority, US Environmental Protection Agency). However, the post-market risk assessment of pesticides used in agriculture is not at that level. To make a progress on this issue, the European Commission (through the environmental program LIFE) and the European plant protection industry association (ECPA) launched the project TOPPS (Train the Operators to Promote best management Practices and Sustainability) [6], which aim is to reduce water pollution due to

the improper use of pesticides. The TOPPS project addresses both point and diffuse sources of water pollution by pesticides and it tries to diagnose the level of the pollution risk and give instructions for mitigation measures that would reduce and prevent the pollution of water with pesticides. Though the end-users (agricultural advisers and farmers) benefitted from the TOPPS project, its first results are very difficult to use at the field level and the end-users are not very flexible in terms of selecting a set of proposed mitigation measures.

To overcome these problems and to address the problem of water pollution with pesticides at the field level, the French agricultural institute ARVALIS – Institut du végétal launched a project EVADIFF (EVAluation of existing models and development of new decision-making tools to prevent DIFFuse pollution caused by plant protection products), whose purpose is to upgrade the approach used in the TOPPS project through a combination of existing expert knowledge collected in the framework of the TOPPS project and experiences that ARVALIS experts obtained from the application of different crop management practices on the reduction and elimination of water pollution with pesticides.

The aim of this paper is to combine and structure domain knowledge of risk assessors and risk managers into a decision support system (DSS) comprised of risk assessment and risk management decision modules for pesticides approved for use in the agriculture. The DSS should be applicable on the field level and it should give its end-users (e.g., farm advisers) flexibility regarding the choice of mitigation measures from the list proposed by the decision support system.

2 POST MARKET RISK ASSESSMENT

To propose the methodological solution for post-market assessment of water pollution with pesticides, we used the approach proposed by the US Environmental Protection Agency (EPA). The methodology evaluates the likelihood that adverse ecological effects may occur or are occurring because of an exposure to one or more stressors [7]. The methodology does not rely only on deterministic descriptions of the studied system using empirical data, but it takes into account also expert knowledge accumulated through empirical systematic observations and management experiences.

To evaluate the potential transfer of applied pesticides to the water and to find the appropriate management solutions to reduce or eliminate the pollution if identified in the prior step, the EPA methodology proposes to combine risk assessment and risk management respectively.

The risk assessment is performed in the context of what techniques one should use to objectively describe and evaluate the pollution risk. The results of risk assessment are primarily for providing information and insight to those who make decisions about how that risk should be managed. The process of combining a risk assessment with decisions on how to address that risk is a central task of ecological risk management. This includes decisions about whether to respond to an assessed level of ecological risk and which of the provided alternatives should be selected. As such, both ecological risk assessment and risk management require combining the results from decision modelling for either diagnosis (assessment) or mitigation purposes.

3 METHODOLOGY

To achieve the research objectives, we used two complementary approaches which we implemented in two methodological modules. The first module deals with the assessment of ecological risk of water pollution with pesticides and the second module addresses risk management, which analyses and compares various alternative mitigation measures in order to prevent water pollution by pesticides used in crop management.

To assess the ecological risk, we have focused first on the determination of the prevailing water pathways in a field, and on the assessment of their flow intensities, time in and duration during the crop growing season. Regarding the water flow types, we focused on surface runoff, drainage runoff and infiltration. For individual water pathways we assessed the risk that they could pollute surface and ground water with the transfer of pesticides. The results of the assessment of ecological risk are used as input data in the second module.

The second module deals with risk management. Its goal is to analyse and compare the various alternative mitigation measures to reduce the pollution risk that is assessed in the first module. The result of the risk management module is a list of mitigation measures that the end-user may use to protect water from pesticides applied in the field for crop management purposes.

To integrate the existing expert knowledge, we used a methodology originally developed for Multi Criteria Decision Analysis (MCDA) to generate multi-attribute decision models (MADM) of risk assessment and risk management. The approach is based on a hierarchical integration of subcomponents (e.g., water pathways, used active substance, applied soil management techniques, application time of pesticides, etc.), forming several hierarchical levels beginning with the integration of basic attributes at the lowest hierarchical level.

In general, MADM are built by a quantitative approach using the numeric values of attributes [8], while we generate MADM using a DEX (Decision EXpert) integrative methodology [9], which is based on attributes with a finite set of qualitative (nominal) values instead of attributes with numerical values. The integrative functions in DEX are adjusted for qualitative variables and therefore represented with if-then rules, which are given in a tabular form compared to the more common weight-based integration functions used in quantitative multi-attribute decision modelling. The DEX methodology enables the construction of a transparent and comprehensive models and it provides mechanisms for presenting aggregation rules in a user friendly way, i.e. in the form of decision trees.

In addition to the mere evaluation of alternatives, the DEX methodology provides what-if examination analysis of alternatives. Both possible applications of the DEX methodology were used in our research. The evaluation of alternatives was used for risk assessment (module one), while the what-if analyses were used for selection of mitigation measures in the risk management module (module 2). The decision models were built with the software modelling tool DEXi, which is based on the DEX methodology. DEXi facilitates the development of qualitative MADM [10] and enables an evaluation and analysis of decision options. This is particularly useful for complex decision-making problems, where an option that satisfies the goals of decision makers has to be selected from a set of possible ones (e.g., mitigation measures).

4 THE EXPERT KNOWLEDGE

The expert knowledge that we used to build qualitative multi-criteria decision models for risk assessment and risk management was obtained from experts involved in the TOPPS project and experts for pesticide use from ARVALIS. The decision models were evaluated on data obtained from the experimental station La Jaillière, which is located in western France and managed by ARVALIS. The data are collected on 11 fields from 1987 on. Each field is described with data about water pathways (duration and water quantity) and the concentrations of active substances in water outflows (total 76 active substances). Beside data related to water outflows, meteorological data and data about applying soil and crop management measures were also collected.

5 RESULTS

The decision models for risk assessment and risk management are the central parts of the DSS of plant protection products approved for use in agriculture. The conceptual diagram (Fig.1) shows the structure of the DSS. The input to the DSS is a proposed crop management plan from which several data are extracted and pre-processed for ecological risk assessment. There are two types of input data: the first type consists of data describing the soil hydrological properties of the assessed field (water pathway, flow/drainage period, soil properties) and the second type consists of data describing the crop management plan for that field (crop, pesticide application time, active substance, dosage).

Because the transfer of pesticides to the surface or ground water is made by water flows, the central focus of the risk assessment module is on assessing the prevailing water pathways in a field. In this study, the surface runoff and infiltration were the two general types of water pathways, but due to local soil hydrological specifics, we divided these two categories into a few subcategories. Beside infiltration, we took into account also drainage outflows from the fields with installed tile drainage system that drains the surplus of soil water from the fields. Regarding the surface runoff category, we made a distinction between runoff by saturation, simple surface runoff and runoff on capping soil.

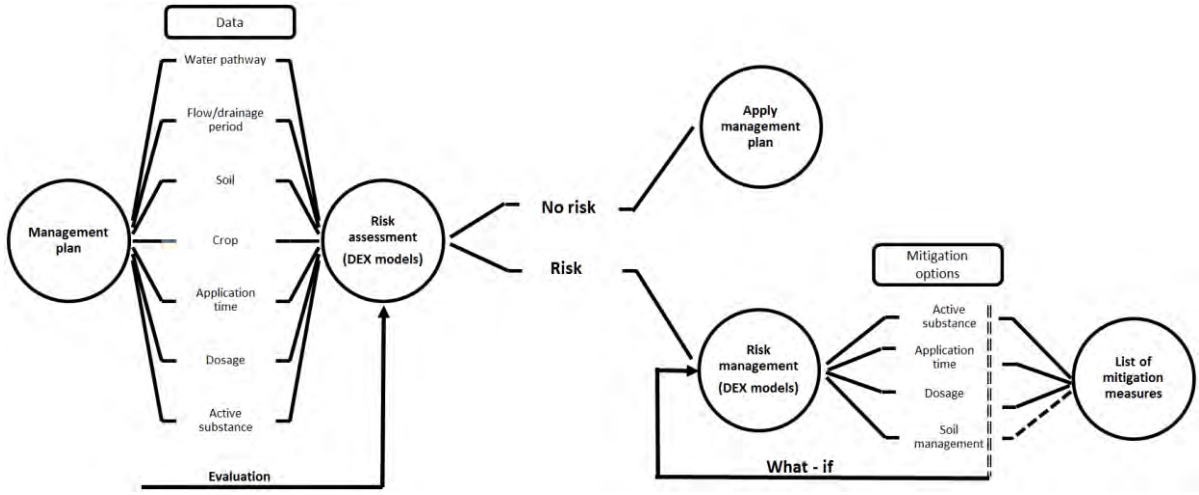


Figure 1: A conceptual diagram of the decision support system for assessing the risk of water pollution by pesticides and for proposing a list of mitigation measures if the risk of pollution exists.

According to the data describing the soil hydrological properties of the assessed field, the DSS first assesses the prevailing water pathway and its intensity. It selects the water pathways with the most intensive flows and in combination with data describing the crop management plan (crop, pesticide, planed application time of pesticide and dosage) makes an assessment of the pollution risk that the prevailing water pathways might cause with the transfer of the pesticides into surface or ground water.

In case the ecological assessment of the proposed crop management plan does not predict any risk for the environment, the management plan can be applied as such. But if the ecological assessment predicts a risk of pollution, the management plan is given as input to the risk management module of the DSS. Its goal is to find which of the planned management measures should be changed in order to avoid the risk of water pollution.

Usually the end-users have technical and financial constraints regarding the selection of mitigation measures. Therefore, they would prefer changes of only a few components of the management plan. In our case, the system can propose mitigation measures with a selection

and combination of four suitable mitigation measures to avoid the risk of water pollution: change of the application time of a pesticide depending on the intensity of water pathways; change of dosage; change of the type of active substance; and change of soil management (tillage, no tillage). The risk management module iteratively searches for a combination of suitable mitigation options that would reduce the level of pollution risk to an acceptable level. In order to give the end-users flexibility in terms of their management preferences, the risk management module proposes a list of several possible solutions from which the end-users can choose the one that best fits their management capacity.

To demonstrate how the DSS works for a particular case, we demonstrate its use in the case of maize production in a field with an installed drainage system (Fig. 2). The proposed management plan described in Fig. 2 is assessed as ecologically risky, therefore, the risk management part proposes a list of mitigation measures that the end-user may apply on his field without any risk of water pollution by pesticides.

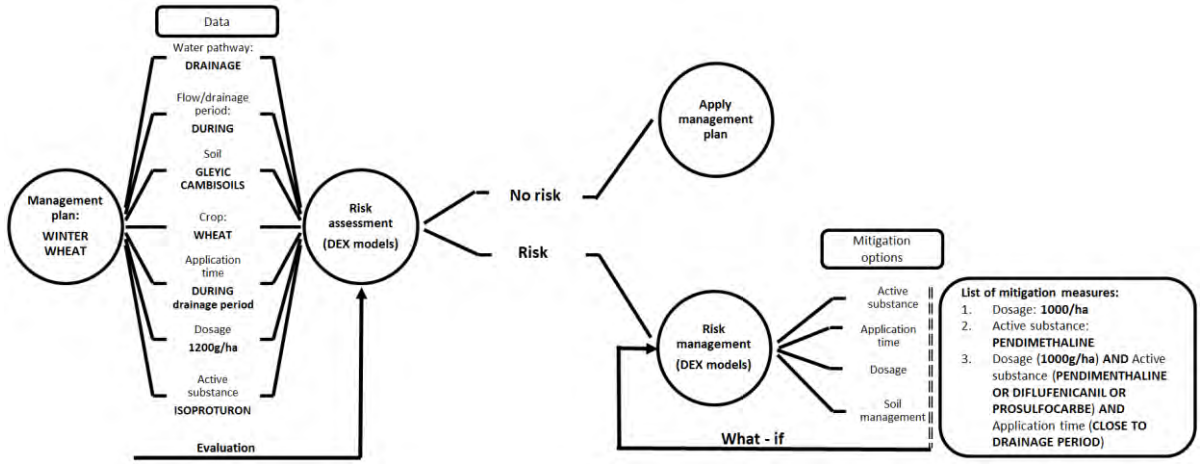


Figure 2: The application of the decision support system to assess the risk of water pollution by isoproturon planned to be applied to winter wheat at a dosage of 1200 g/ha during the winter drainage period. The proposed management plan is assessed as risky, therefore the risk management module proposes a list of mitigation measures.

6 CONCLUSIONS

Since the post market ecological risk assessment of pesticides approved for use in crop management is not as developed as pre-registration assessment, the decision support system presented in this paper makes an important contribution to this very serious environmental issue. The applied MCDA built through the DEX methodology has enabled us to structure the existing expert knowledge according to the approach proposed by the EPA. The applied methodology facilitated the representation of existing domain knowledge about pesticide use and environmental protection crop management measures.

The results have been recognized as very useful because they address different aspects, ranging from the assessment of ecological risk, comparisons of assessed risk under different settings of input data, and what-if analyses of mitigation options that generate a list of mitigation measures for reduction and elimination of assessed pollution risk by pesticides. The results are applicable at the field level and give large flexibility to end-users in terms of their selection of mitigation options.

Since knowledge and practical experiences accumulate through time, the applied methodology enables improvement of the DSS with the latest expert knowledge. The general structure of the DSS presented in Figure 1 can potentially be very widely applicable, given

enough knowledge about local soil properties and soil hydrology is available. The presented DSS could be easily implemented as a web application and put into everyday on-site use by advisors.

Acknowledgement

The authors wish to acknowledge the support of the project EVADIFF (Evaluation of existing models and development of new decision-making tools to prevent diffuse pollution caused by phytopharmaceutical products) financed by ARVALIS – Institut du vegetal and Florence Leprince from the Department of Scientific Affairs of ARVALIS.

References

[1] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. (2000). Official Journal L 327, 22/12/2000, pp. 0001 – 0073.

[2] Directive 2009/128/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for Community action to achieve the sustainable use of pesticides. (2009). Official Journal L 309, 24.11.2009, pp. 71–86.

[3] Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market. (2009). Official Journal L 309, 24.11.2009, pp. 1- 50.

[4] Commission Implementing Regulation (EU) No 540/2011. (2011). Official Journal L 153, 11.6.2011, pp. 1–186.

[5] Reichenberger, S., Bach, M., Skitschak, A., Frede, H. G. (2007). Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness: A review. Science of the Total Environment, 384: 1–35.

[6] Roettele, M. (2008). Strategies to reduce point source losses of PPP to water focus on "behaviour, technique and infrastructure": results and lessons learned from the TOPPS – project. Aspects of Applied Biology, 84: 375-368.

[7] Overview of Risk Assessment in the Pesticide Program. (2015). Environmental Protection Agency, USA. http://www.epa.gov/pesticides/about/overview_risk_assess.htm [16/6/2015]

[8] Mandoza, G.A., Martins H. (2006). Multi-criteria decision analysis in natural resource management: A critical review of methods and new modelling paradigms. Forest Ecology and Management. 230: 1–22.

[9] Bohanec, M., Rajkovic, V. (1990). DEX: An expert system shell for decision support, Sistemica 1: 145–157.

[10] Bohanec, M. (2014). DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version 4.01. IJS Report DP-11739, Jožef Stefan Institute, Ljubljana.

ASSESSMENT OF CLOUD HIGH PERFORMANCE COMPUTING POTENTIAL FOR SMES

Mirjana Kljajić Borštnar
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, Kranj, Slovenia
mirjana.kljajic@fov.uni-mb.si

Tomi Ilijaš
Arctur, d.o.o.
Industrijska cesta 15, Nova Gorica, Slovenia
tomi.ilijas@arctur.si

Andreja Pucihar
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, Kranj, Slovenia
andreja.pucihar@fov.uni-mb.si

Abstract: High Performance Computing (HPC) services offered in cloud are believed to be one of the key competitiveness enablers for companies of all sizes throughout the world. On the EU level several measures, like experiments funding, have been taken to boost adoption among SMEs, particularly manufacturing. However, HPC adoption is in its early stages, and business models are not yet fully explored. In order to support perspective business ideas, and assure transparent and efficient public funds spending, there is a need for assessment of the SMEs potential prior to funding the experiments. Assessment relies on many criteria and stakeholders. Currently, there are some tools developed for selection of the experiments, but they are not complete and rely mostly on individual assessment of designated reviewers. The aim of this paper is to propose a qualitative multi-criteria model for SMEs' cloud HPC potential assessment. Using Decision EXpert methodology (DEX), the model is based on theoretical and practical knowledge, elicited from experts and use cases. The model will be verified on a set of experiments conducted within several EU projects in I4MS initiative.

Keywords: High Performance Computing, Cloud services, Assessment Criteria, Multi-Aattribute Modelling, DEXi

1 INTRODUCTION

High Performance Computing (HPC) refers to computing performance needed for solving complex computing problems that could not easily (or timely) be computed by typical desktop computers. It is generally used for solving large scale problems in science, engineering and business [1]. So far HPC was mainly reserved for the large companies and research institutes, who could afford high costs that are associated with HPC. From the industry perspective the HPC is predominantly used in manufacturing sector with financial sector just behind it [2].

In recent years cloud computing services has reached high adoption rates among companies, also SMEs [3] and so the possibility of hiring HPC services in the cloud became immanent [4]. However, moving HPC to cloud services is the one least exploited, especially among SMEs [5]. The problem lies not merely in high costs but predominantly in the lack of competencies (knowledge, maintenance, proprietary software etc.). One of the promising changes is identified in redefining a business model, which consists not only in hiring HPC services in the cloud, but includes also other services of modelling, maintenance, implementation, and software adaptation. In this way HPC services can become an interesting opportunity for other industries, as well as for the SMEs [6].

Across the EU, there were 21.2 million SMEs (99.8%) in the non-financial business sector in 2013 [6]. The number of manufacturing SMEs, its added value and employment is still

below that of the year 2008 and is anticipated to shrink in 2015. However, it is still a very important sector, employing more than 17 million individuals and generating 21% of SME added value in Europe [6]. It is believed that this group particularly can benefit from adopting new technologies, such as HPC, and become successful on the global market, and a leading sector in EU.

On the global level governments are investing in boosting the adoption of cloud HPC [7, 8]. The European initiative ICT Innovation for Manufacturing SMEs (i4MS) is set to support the European leadership in manufacturing through the adoption of ICT technologies. In fact, Europe's competiveness in this sector depends on its capacity to deliver highly innovative products, where the innovation often originates from advances in ICT [9]. Within i4MS initiative four areas are supported: robotics, HPC cloud based simulation services, laser based applications, and intelligent sensor-based equipment. By developing and supporting new business ideas, particularly the use of HPC services in manufacturing SMEs, the initiative aims to foster the new economic growth and competitiveness.

Several projects within i4MS address adoption of HPC cloud services by selecting experiments for showcasing the best practices, develop, test and demonstrate the use of infrastructure and the business model as a one-stop pay-per-use shop. Experiments include all actors (SME, Innovation centres/clusters, experts, code providers) throughout the value chain of an innovation ecosystem. One of the important propositions of the initiative is the "development of a sustainable business model, which is crucial for the successful adoption of these services" [9]. Therefore, general criteria for selection of an experiment are: demonstration of HPC needs for new product development in manufacturing industry; should be end-user driven, address a real use-case, and demonstrate the use of HPC and high potential to benefit from cloud technology [10].

The problem addressed in this paper is the selection of appropriate experiments to facilitate the early adopters and early majority group in order to boost the competitiveness of the European manufacturing SMEs. There are some tools developed, like questionnaires and selection criteria, mostly for the purpose of open call proposals evaluation, that are not complete and are mostly designed to rank the proposals and decide what to fund or not.

1.1 Related work

There are several studies focusing on identifying factors influencing adoption of cloud services in SMEs, mainly through researching adoption factors business perspective, and technology, and security from both vendors and users perspective [11, 12]. There are scarce studies on the topic of decision support and cloud computing services, predominantly from the viewpoint of web-based decision support systems [13] and decision support on migration to the cloud [14, 15].

To address the problem of assessing the potential of cloud HPC services adoption we propose a qualitative multi-attribute decision model based on DEX methodology. The proposed methodology belongs to Multi-Attribute Decision Making (MADM), which is rooted in the decision theory and utility theory, and well accepted in practice.

Multi-attribute decision modelling is a process of evaluation in which we develop the model that supports the alternative evaluation according to the stated goals and preferences. The model is based on a set of criteria, parameters, variables and factors, recognized in the process of decision-making. MADM is a formal basis for model development, where the basic problem is in integrating individual parameters into a final value. Core of the model is based on the methods of expert knowledge modelling of the expert systems, which support the transparent evaluation and reasoning [16]. These methods, however, are not compensating the human decision-maker, but can contribute to more systematically and

organized decision-making. Supported by such models, the decision maker is stimulated to understand the problem, to reduce the possibility of error or missing important factors [16].

This research contributes to the discussion of how to identify SMEs with best potential, engage them in public funding schemes and stimulate the adoption of cloud HPC services. Based on literature review on HPC and cloud services adoption, and expert interviews, during several iterations, we propose a qualitative multi-attribute decision model, which can support both vendors of cloud HPC services and SMEs in this decision-making process.

2 METHODOLOGY

The proposed methodological approach is rooted in Design Science Research [17, 18], which dates back as early as 1990 [19, 20] and gained recognition in 2004 by a MISQ paper [17]. Its basic philosophy derives from other engineering disciplines – where development of an artefact was common. The main driver is developing an IT artefact (construct, model, prototype, instantiation) that will demonstrate practical relevance, and is fundamentally rooted in problem-solving paradigm. On the other hand, "the practical relevance of the research result should be valued equally with the rigor of the research performed to achieve the result" [21]. Three basic cycles of DSR are defined [17]: 1) Relevance (definition of environment, application domain, problem or challenge), 2) Design (iterative artefact building process and evaluation method), and 3) Rigor cycle (theory grounding and contribution to theory and practice).

The IT artefact developed in this study is a model, based on a qualitative multi-criteria decision modelling methodology [22, 23]. Figure 1 presents the process of DEX modelling in the context of DSR cycles.

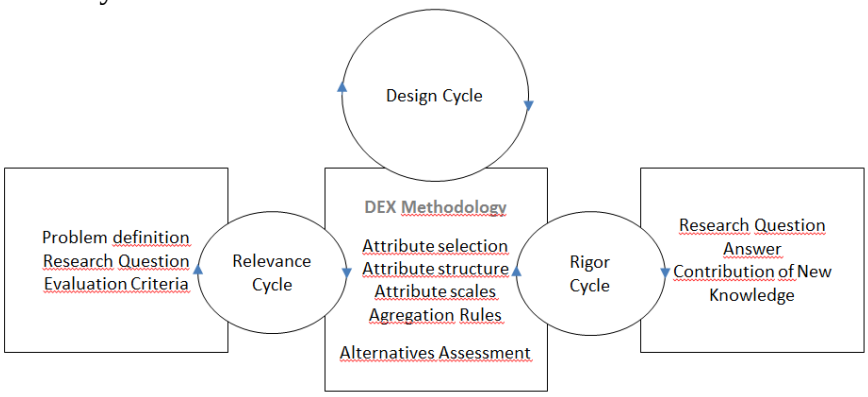


Figure 1: Research design

DEX (Decision EXpert) belongs to a qualitative multi-attribute decision modelling methodology, based on an integration of multicriteria decision modelling with rule-based expert systems [24]. A qualitative multi-attribute model, with which decision alternatives are evaluated and analysed, is developed by a team of decision modelling and domain experts. The model itself represents a decomposition of the decision problem into smaller, less complex sub-problems. The decomposition is represented by a hierarchy of attributes. Attribute scales are qualitative; therefore they are easily understood by the decision-makers. Basic problem represents the aggregation of individual parameter values to a final value (criteria function). Furthermore, the parameter interdependencies, weak determination or ability to measure, and changing influence (weight) make the modelling a complex task. Domain knowledge is modelled by a combination of hierarchical tree of attributes and aggregation based on "if-than" rules. The result is the transparent evaluation, which supports the explanation of the evaluation results and the decision-making process itself [16].

In this paper we model the assessment of SMEs readiness to use the HPC services in cloud as a complex decision-making problem. Our aim, to develop a multi-attribute decision model, is based on the proposition that such a model can be built using DEX methodology and utilized on a set of real-world problems. Two assumptions have to be met: 1) ability to observe and measure criteria in a real-life environment and 2) transparent assessment of SMEs HPC cloud readiness.

3 RESULTS

3.1 Attributes identification

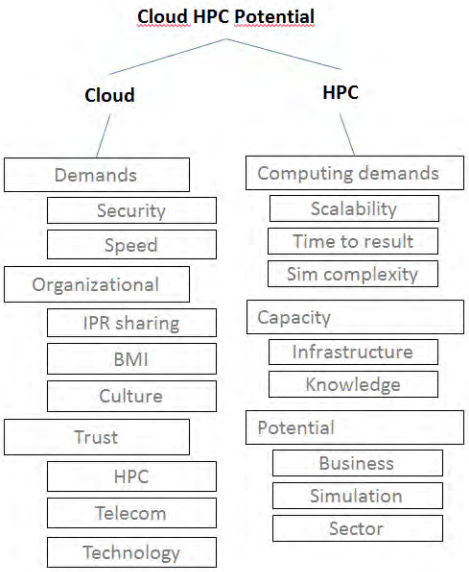
Attributes were defined on the basis of literature review, experiment call for proposals [9, 10], current experiments, 15 experiment proposals, and in a set of group interviews with domain experts (3 rounds of interviews with experts from the fields of HPC, business model innovation, code parallelization). In the first iteration there were 59 attributes identified. In the following iterations the total number of attributes was reduced to 33 basic and 22 aggregate attributes, in total 55 attributes.

3.2 Hierarchical model of decision criteria

From the reviewed documents and interviews we were able to distinct two basic groups of attributes defining the potential of cloud HPC services: Cloud (describing the possibility to use the service in the cloud) and HPC (describing the need for high performance computing). These two groups were further partitioned into subgroups of attributes to the 5th or 6th level Simplified tree of attribute is described in Figure 2.

3.3 Attribute scales

Qualitative attribute scales were defined by domain experts to reflect various phenomenon otherwise difficult to describe numerically (i.e. “Culture”, “BMI”, “trust in HPC provider”). Typically the attribute scales have 3 to 4 values, but the scales of “Cloud” and “HPC” aggregated attributes range from 1 to 4, with final assessment of Cloud HPC potential on a 5 values scale. All scales were ordered from “good” to “bad”.



	Cloud	HPC	Cloud HPC Potential
	47%	53%	
1	not acceptable	<=reassess	No potential
2	not acceptable	>=potential	Low potential
3	reassess	<=reassess	Low potential
4	>=reassess	no potential	Low potential
5	reassess	potential	Medium
6	>=potential	reassess	Medium
7	reassess	great potential	Good potential
8	>=potential	potential	Good potential
9	>=potential	great potential	Great potential

Figure 2: Hierarchical tree of attributes and a decision table

3.4 Aggregation rules

Values of the basic attributes were determined by the experts, whereas the values of the aggregated attributes are derived by “if-than” rules that are easily understood. These rules in

combination with the hierarchical attribute structure provide transparent evaluation and explanation of the evaluation results. A single rule represents a part of domain expert knowledge. DEXi provides an approximation of a linearly weighted sum, but this can be overruled by an expert. This way we can consider non-linearity in the domain knowledge. An example of decision rules, defined by the domain experts, as set in DEXi is presented in Figure 2 on the right.

3.5 Validation of the model

Evaluation of experiments has been conducted in a team of experiment evaluators and proposers. The attribute values were derived mostly from the experiment proposals itself; some information was further elicited from the proposers. In Figure 3 (left) we present two experiments for the purpose of model validation. Results suggest that existing experiments (E1, E2) were evaluated as "acceptable" for the Cloud and "medium" and "reassess" for HPC respectively. Since both experiments were selected for funding, we further analysed E2 (Figure 3). E2* was negotiated in the part, where business model is developed.

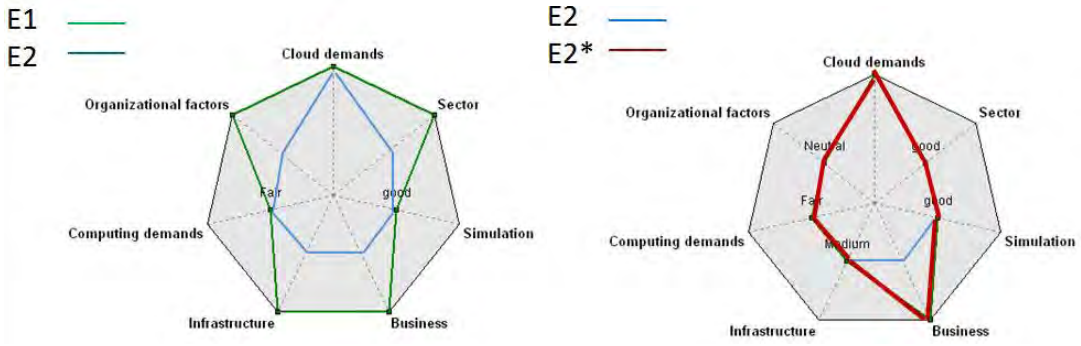


Figure 3: Cloud HPC Potential assessment (left); what if analysis of E2 (right)

4 CONCLUSIONS

HPC offers great potential in new product and services development for the manufacturing SMEs. However, this requires specialized knowledge, infrastructure and software and is as such not available for the SMEs. HPC services offered in the cloud present an important possibility for the SMEs, where other actors in the value chain, such as HPC providers, researchers and other specialists, have an important role in creating new business models.

In order to assess the potential of SMEs, or their proposed experiments, to uptake the cloud HPC services we developed a qualitative multiple-attribute decision model. Together with the domain experts, the attributes were identified, structured in a hierarchical attribute tree, attribute scales were defined and aggregation rules were set. The model was validated by the existing experiments. Based on findings the model will be refined, and further evaluated. Opportunities for using the model in decision and negotiation process were explicated. Preliminary results suggest high usability of the decision model as an assessment tool and the potential to be used in similar set of problems (i.e. business model innovation).

Acknowledgement

This research was funded by Slovenian Research Agency, Grant No. P5-018.

References

[1] Vecchiola, C. Pandey, S. Buyya, R. (2009). High-Performance Cloud Computing: A View of Scientific Applications. 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN), pp.4-16. doi: 10.1109/I-SPAN.2009.150.

[2] Osseyran, A. Giles, M. (2015). Industrial Applications of High-Performance Computing: Best Global Practices. Chapman & Hall/CRC Computational Science.

[3] RightScale 2015: State of the Cloud Report. http://www.rightscale.com/ [Accessed 20/06/2015].

[4] Rehr, J.J. Vila, F.D. Gardner J.P. Svec, L. Prange, M. (2010). Scientific Computing in the Cloud. Vol. (12). No. (3). pp. 34-43. IEEE Computer Society.

[5] Trader, T. (2014). Boosting HPC Access in SME Community. http://www.hpcwire.com/2014/04/18/boosting-hpc-access-sme-community/ [Accessed 20/06/2015].

[6] Muller, P. Gagliardi, D. Caliandro, C. Bohn, N, U. (2014). Demetrius KlitouAnnual Report on European SMEs 2013/2014. European Comission.

[7] Feldman, M. (2011).White House Announces Project to Spur HPC Adoption in US Manufacturing. HPC Wire. http://www.hpcwire.com/2011/03/03/white_house_announces_project_to_spur_hpc_adoption_in_us_manufacturing/ [Accessed 20/06/2015].

[8] Horizon 2020, Work Programme 2014-2015. https://ec.europa.eu/research/participants/portal/doc/call/h2020/common/1617606-part_5_i_ict_v2.0_en.pdf [Accessed 20/06/2015].

[9] ICT Innovation for SMEs. http://i4ms.eu/i4ms/i4ms.php [Accessed 20/06/2015].

[10] 1st Open Call for Application Experiments. (2015). http://www.eu-cloudflow.eu/files/open-calls_first-call_CF-short-form.pdf [Accessed 20/06/2015].

[11] Alshamaila, Y. Papagiannidis, S. Li, F. (2013). Cloud computing adoption by SMEs in the north east of England: A multi-perspective framework. Journal of Enterprise Information Management Vol. (26), No. (39), pp. 250-275. Emerald Group Publishing Ltd. DOI 10.1108/17410391311325225.

[12] Bogataj, K. Pucihar, A. (2013). Business Model Factors Influencing Cloud Computing Adoption: Differences in Opinion. BLED 2013 Proceedings. Paper 3. http://aisel.aisnet.org/bled2013/3 [Accessed 20/06/2015].

[13] Russell, S. Yoon, V. Forgionne, G. (2009) Cloud-Based DSS and Availability Context: The Probability of Successful Decision Outcomes. Designing E-Business Systems. Markets, Services, and Networks. Lecture Notes in Business Information Processing Vol. (22), pp. 96-109.

[14] SWiFT 10-2012: Adopting the Cloud – decision support for cloud computing. (2012).National Standards Authority of Ireland. file:///C:/Users/mirjanak/Downloads/swift10-2012%20(1).pdf [Accessed 20/06/2015].

[15] Khajeh-Hosseini, A. Sommerville, I. Bogaerts, J. Teregowda, P. (2011). Decision Support Tools for Cloud Migration in the Enterprise. Arxiv preprint arXiv:1105.0149. http://arxiv.org/ftp/arxiv/papers/1105/1105.0149.pdf [Accessed 20/06/2015].

[16] Bohanec, M. Rajkovič, V. (1999) Multi-attribute decision modeling: Industrial applications of DEX. Informatica 23, 487-491.

[17] Hevner, A. March, S. Park, J. Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly (28:1), pp. 75-105.

[18] Kuechler, B. Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. European Journal of Information Systems, 17(5), 489-504.

[19] Simon, H. (1996). The Sciences of Artificial, 3rd edn. MIT Press, Cambridge, MA.

[20] Nunamaker, J. Chen, M. Purdin, T, D, M. (1991). Systems development in information systems research, Journal of Management Information Systems 7 (3), pp. 89–106.

[21] Hevner, A. Chatterjee, S. (2010). Design Science Research in Information Systems., Design Research in Information Systems, Integrated Series 9 in Information Systems 22, DOI 10.1007/978-1-4419-5653-8_2.

[22] Keeney, R.L. Raiffa, H. (1976). Decisions with multiple objectives, John Wiley & Sons.

[23] Saaty, T.L. (1980). The analytic hierarchy process, McGraw-Hill.

[24] Alić, I. Siering, M. Bohanec, M. (2013). Hot Stock or Not? A Qualitative Multi-Attribute Model to Detect Financial Market Manipulation. BLED 2013 Proceedings. Paper 1. http://aisel.aisnet.org/bled2013/1 [Accessed 20/06/2015].

EMPIRICAL COMPARISON OF THREE METHODS FOR APPROXIMATING DEX UTILITY FUNCTIONS

Matej Mihelčić
Laboratory for Information Systems, Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
Marko Bohanec
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Abstract: DEX is a qualitative multi-criteria decision analysis method. It provides support to decision makers in evaluating and choosing decision alternatives, using discrete attributes and rule-based utility functions. This work builds upon our previous attempt of approximating DEX utility functions with methods UTA and ACUTA, aimed at improving the sensitivity of qualitative models and providing an interpretation of DEX utility functions. In this work we empirically compare three methods for approximating qualitative DEX utility functions with piecewise-linear marginal utility functions: Direct marginals, UTADIS and Conjoint analysis. The results show that these methods can accurately approximate complete, monotone DEX utility functions.

Keywords: decision support, multi-criteria decision making, utility function, DEX, UTADIS, conjoint analysis, direct marginals method

1 INTRODUCTION

Multi criteria decision analysis (MCDA) [7] deals with solving decision problems involving multiple, possibly conflicting, criteria. It provides a number of methods to create decision models by using information provided by the decision maker. Provided information can be given in various forms, using different representations. Converting representations from one form to another is often highly desirable, as it can bridge the gap between different methodological approaches and enrich the capabilities of individual ones.

At a general level, this study addresses two types of utility function representations, qualitative and quantitative, and investigates how to convert the former to the latter. At a specific level, we compare three methods of approximating DEX utility functions by piece-wise linear marginal utility functions: the Direct marginals method, UTADIS and the Conjoint analysis method. DEX [5] is a qualitative MCDA method, which employs discrete attributes and discrete utility functions defined in a rule-based point-by-point way (see section 2.1). This makes DEX suitable for classifying decision alternatives into discrete classes. The Direct marginals method (section 2.3) establishes marginal utility functions by a projection of a DEX utility function to individual attributes. UTADIS [6] (section 2.4) is a quantitative method that constructs numerical additive utility functions from a provided subset of alternatives and assigns this alternatives to predefined ordered groups. Conjoint analysis [8] (section 2.5) is a method that constructs numerical additive utility functions through determining attribute importance, the appropriate importance levels and the effects of combining different attributes on the measured variable. The three methods were experimentally assessed on a collection of artificially generated complete monotone DEX utility functions.

All three methods are aimed at providing an approximate quantitative representation of a qualitative DEX function. This extends the capabilities of DEX and is useful for several reasons. First, the newly obtained numerical evaluations facilitate an easy ranking and comparison of decision alternatives, especially those that are assigned the same class by DEX. Consequently, the sensitivity of evaluation is increased. Second, the sheer form of numerical functions may

provide additional information about the properties of underlying DEX functions, which is useful in verification, representation and justification of DEX models. In this study, we focus on the accuracy of representation.

There have been several previous attempts to approximate DEX utility functions. A linear approximation method is commonly used in DEX to assess criteria importance [3]. An early method for ranking of alternatives and improving the sensitivity of evaluation called QQ [12] has been proposed in [2]. Recently, extensive research has been carried out to approximate DEX functions with copulas [12]. This paper builds upon our previous work on approximating DEX utility functions by using methods UTA and ACUTA [11]. The methods used in the present study were chosen because they do not have convergence issues when approximating discrete functions as opposed to the methods tried in [11].

2 METHODS

2.1 DEX method

DEX [5] is a qualitative MCDA method for the evaluation and analysis of decision alternatives, and is implemented in the software DEXi [4]. In DEX, all attributes are qualitative and can take values represented by words, such as *low* or *excellent*. Attributes are generally organised in a hierarchy. The evaluation of decision alternatives is carried out by utility functions, which are represented in the form of decision rules.

In the context of this paper, we focus on individual utility functions. For simplicity, we assume that all attributes are ordinal and preferentially ordered, so that a higher ordinal value represents a better preference. In this setting, a DEX utility function f is defined over a set of attributes $\vec{x} = (x_1, x_2, \ldots, x_n)$ so that

$$f : X_1 \times X_2 \times \cdots \times X_n \to Y$$

Here, $X_i, i = 1, 2, \ldots, n$, denote value scales of the corresponding attributes x_i, and Y is the value scale of the output attribute y:

$$X_i = \{1, 2, \ldots, k_i\}, i = 1, 2, \ldots, n \quad \text{and} \quad Y = \{1, 2, \ldots, c\}$$

The function f is represented by a set of decision rules

$$F = \{(\vec{x}, y) | \vec{x} \in X_1 \times X_2 \times \cdots \times X_n, y \in Y, y = f(\vec{x})\}$$

Each rule $(\vec{x}, y) \in F$ defines the value of f for some combination of argument values $\vec{x}$. In this study, we assume that all functions are *complete* (defined for all combinations of argument values) and *monotone* (when argument values increase, the function value increases or remains constant).

2.2 Approximation of DEX utility functions

All methods assessed in this study are aimed at approximation of some DEX utility function f with marginal utility functions $u_i : X_i \to \mathbb{R}, i = 1, 2, \ldots, n$. The functions u_i are assumed to take a piece-wise linear form: the numeric value of $u_i(v)$ is established from f for each $v \in X_i$, while its value for $v \notin X_i$ is linearly interpolated from the closest neighbouring points.

On this basis, f is approximated as a weighted sum of marginal utility functions:

$$u(x) = u(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \omega_i u_i(x_i)$$

Here, $\omega_i \in \mathbb{R}, i = 1, 2, \ldots, n$, are weights of the corresponding attributes, normalised so that $\sum_{i=1}^{n} \omega_i = 1$.

2.3 Direct marginals method

The direct marginals method establishes the marginal utility function $u_i(v)$ as an average value of target attribute y for decision rule $a \in F$, where $x_i(a) = v$. Let $F_{i,v} \subset F$ denote all decision rules where $x_i(a) = v$. Then

$$u_i(v) = \frac{1}{|F_{i,v}|} \sum_{\{a \in F \mid x_i(a)=v\}} y(a), \quad i = 1, 2, \ldots, n, \quad v \in X$$

In the experiments (section 2.6), all functions $u(x)$ were scaled to the $[0, 1]$ interval, therefore importance weights for attributes were computed as a percentage of total utility range covered by the range of a particular attribute.

2.4 UTADIS method

The UTADIS method [6] is an extension of UTA (UTilités Additives) method [9] that enables decision maker to assign alternatives to predefined ordered groups. Thus it is very well suited to our problem of approximating discrete DEX functions, assuming that each DEX decision rule $a \in F$ defines some (hypothetical) decision alternative. UTADIS approximates u_i as:

$$u_i(x_i(a)) = u_i(x_i^J) + \frac{x_i(a) - x_i^J}{x_i^{J+1} - x_i^J}[u_i(x_i^{J+1}) - u_i(x_i^J)]$$

It is assumed that each alternative values are divided to $(\alpha_i - 1)$ equally sized intervals $[g_i^J, g_i^{J+1}]$. The alternatives are assigned to groups by using thresholds t_i: $u(x_j) \geq t_1 \Rightarrow a \in C_1,\ t_2 \leq U(g_j) < t_1 \Rightarrow a \in C_2, \ldots, U(g_j) < t_{c-1} \Rightarrow a \in C_c$.

UTADIS searches for marginal utility functions by solving the linear programming problem $min\ E = \sum_{k=1}^{c} \frac{\sum_{a_j \in C_k} \sigma(a)_j{}^+ + \sigma(a)_j{}^-}{m_k}$, where σ^+, σ^- denote errors after violation of upper/lower bound of a group C_k and m_k denotes a number of alternatives assigned to the group C_k.

2.5 Conjoint analysis method

Conjoint analysis [8] is designed to explain decision maker's preferences. It outputs attribute importance, their interactions and utility functions for each attribute in a decision making problem. The original decision table is transformed in a binary matrix x_b, that encodes the original attribute values by using a fixed number of bits. This matrix is used to compute a matrix of deviation scores $x = x_b - \mathbb{1}\mathbb{1}^\tau x_b(\frac{1}{n})$. The utility value is computed as $b = (x^\tau x)^{-1} \cdot (x^\tau y)$, where y denotes a vector containing deviation scores of the target variable. Attribute importance is obtained by observing the percentage of total utility range covered by the range of a particular attribute.

2.6 Experimental procedure

The goal of experiments was to assess and compare the performance of the three methods – Direct marginals, UTA, and Conjoint analysis – on artificially generated, complete, and monotone DEX utility functions. For this purpose, we generated all monotone functions for spaces with dimensions $3 \times 3 \to 4$, $3 \times 4 \to 3$, $4 \times 4 \to 3$ and $5 \times 6 \to 7$ (The notation $3 \times 3 \to 4$ denotes the space of all utility functions having two three-valued arguments, that map to 4 values). Evaluation was also performed on several randomly generated function sets of different sizes: $3 \times 4 \times 3 \times 5 \to 6$, $4 \times 5 \times 5 \to 6$, $5 \times 6 \to 7$, $6 \times 7 \to 7$, $8 \times 7 \to 7$ containing 1000 functions, and $3 \times 5 \times 3 \times 4 \to 4$ containing 100 functions.

The experimental procedure consisted of predicting the target utility function for all the generated functions by using three selected methods, and computing evaluation scores for each method's resulting utility function. Two measures were used for evaluation: the Area Under the Curve (AUC) and the Root Mean Squared Error (RMSE). Finally, we computed the average of AUC and RMSE with corresponding standard deviation for sets of functions with given dimensions to compare method performance on the whole function set. Since these methods compute utility values in different ranges, all the functions were scaled to the $[0, 1]$ interval.

All experiments were performed in R programming language by using 'MCDA' [10], 'conjoint' [1] and 'pROC' [13] R packages. In addition, we implemented Direct marginals method, the RMSE measure, monotone function generator that generates all monotone functions in some space with given dimensions, and a random monotone function generator that generates a number of random monotone functions in a space with given dimensions.

3 RESULTS

In this section we present results of approximating DEX utility functions with methods Direct marginals, Conjoint analysis and UTADIS. A thorough evaluation can be seen in Table 1.

method	space dimension	num.	avg. AUC	avg. RMSE	succ.
Direct marginals	$3 \times 3 \to 4$	979	0.996 ± 0.015	0.532 ± 0.246	100%
	$3 \times 4 \to 3$	489	0.998 ± 0.011	0.404 ± 0.162	100%
	$4 \times 4 \to 3$	2014	0.995 ± 0.013	0.416 ± 0.135	100%
	$5 \times 6 \to 7$	1000	0.981 ± 0.021	0.981 ± 0.354	100%
	$6 \times 7 \to 7$	1000	0.978 ± 0.021	1.0 ± 0.327	100%
	$8 \times 7 \to 7$	1000	0.975 ± 0.023	0.980 ± 0.308	100%
	$4 \times 5 \times 5 \to 6$	1000	0.945 ± 0.025	1.056 ± 0.245	100%
	$3 \times 4 \times 3 \times 5 \to 6$	1000	0.921 ± 0.027	1.145 ± 0.225	100%
	$3 \times 4 \times 5 \times 3 \times 4 \to 4$	100	0.928 ± 0.018	0.818 ± 0.101	100%
Conjoint analysis	$3 \times 3 \to 4$	979	0.989 ± 0.026	0.564 ± 0.236	100%
	$3 \times 4 \to 3$	489	0.990 ± 0.026	0.416 ± 0.159	100%
	$4 \times 4 \to 3$	1763	0.987 ± 0.025	0.423 ± 0.132	100%
	$5 \times 6 \to 7$	1000	0.971 ± 0.027	1.014 ± 0.329	100%
	$6 \times 7 \to 7$	1000	0.967 ± 0.028	1.023 ± 0.305	100%
	$8 \times 7 \to 7$	1000	0.964 ± 0.029	0.996 ± 0.291	100%
	$4 \times 5 \times 5 \to 6$	1000	0.925 ± 0.033	1.056 ± 0.233	100%
	$3 \times 4 \times 3 \times 5 \to 6$	1000	0.896 ± 0.035	1.142 ± 0.214	100%
	$3 \times 4 \times 5 \times 3 \times 4 \to 4$	100	0.904 ± 0.028	0.808 ± 0.102	100%
UTADIS	$3 \times 3 \to 4$	979	0.970 ± 0.063	0.722 ± 0.293	99.7%
	$3 \times 4 \to 3$	489	0.976 ± 0.064	0.567 ± 0.215	99.6%
	$4 \times 4 \to 3$	1763	0.972 ± 0.069	0.567 ± 0.212	99.9%
	$5 \times 6 \to 7$	1000	0.931 ± 0.065	1.569 ± 0.688	100%
	$6 \times 7 \to 7$	1000	0.924 ± 0.064	1.574 ± 0.646	100%
	$8 \times 7 \to 7$	1000	0.916 ± 0.068	1.545 ± 0.672	100%
	$4 \times 5 \times 5 \to 6$	1000	0.898 ± 0.054	1.299 ± 0.389	100%
	$3 \times 4 \times 3 \times 5 \to 6$	1000	0.880 ± 0.049	1.292 ± 0.312	100%
	$3 \times 4 \times 5 \times 3 \times 4 \to 4$	100	0.911 ± 0.034	0.875 ± 0.151	100%

Table 1: Comparison results for the Direct marginals, Conjoint analysis and UTADIS method on various generated DEX monotone utility functions. For each method and space dimensions, the columns show the number of utility functions (num.), average AUC and RMSE with standard deviation, and the percentage of successfully approximated functions (succ.).

The results from Table 1 show that all three methods can approximate the majority of artificially created complete monotone DEX utility functions; only UTADIS returns errors when faced with trivial functions (containing equal target value for every alternative), which is likely a problem of implementation. The Direct marginals method achieved the best evaluation score on all tested functions in both AUC and RMSE measures and is closely followed by the Conjoint analysis method. UTADIS method has somewhat lower results and higher standard deviation. The results indicate that the AUC value decreases with the increase of function domain dimensions and the cardinality of attribute value set for all tested methods (see Figure 1). Results for the RMSE measure are little less conclusive. The error rises slowly for the Conjoint analysis and Direct marginals method but drops for UTADIS method. AUC increase and RMSE decrease on the last dataset could be caused by a small generated function sample (100 random functions) and the fact that the target attribute could have only 4 different values.

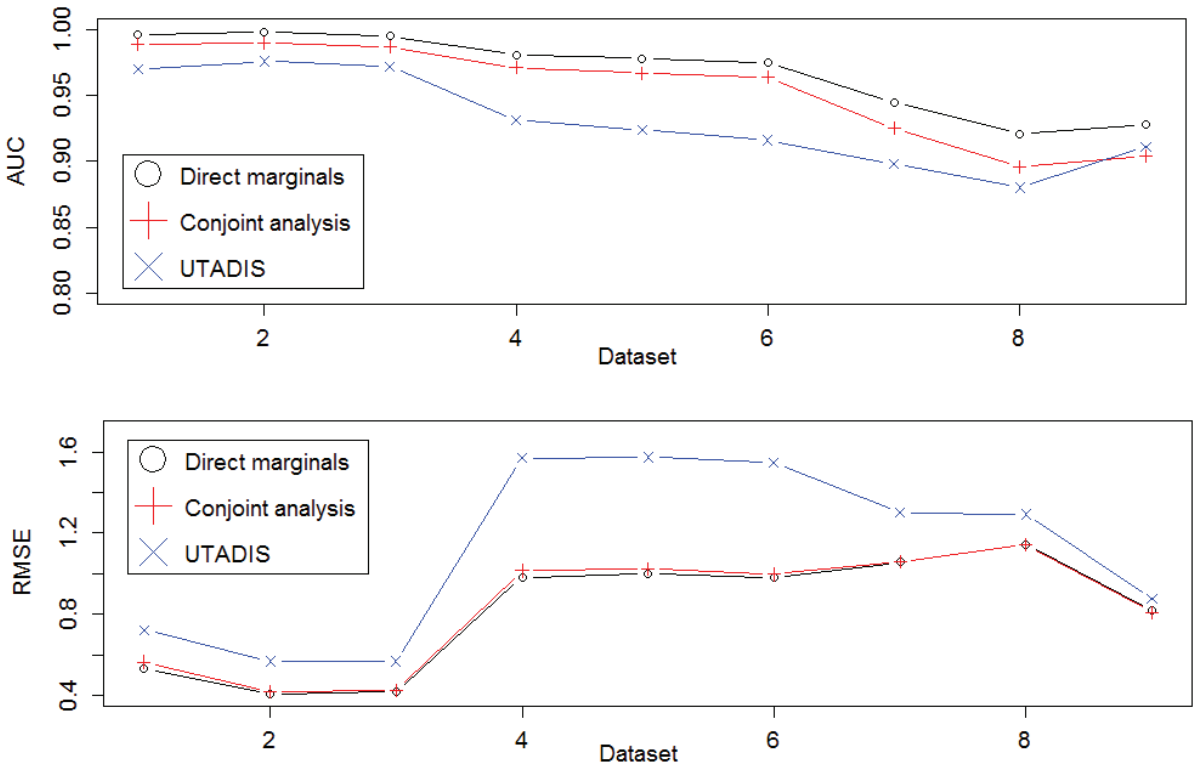


Figure 1: AUC and RMSE comparison for all three methods. Function sets are presented in the same order as in Table 1.

4 CONCLUSION

In this work we presented a new method for approximating monotone DEX utility functions, the Direct marginals method, and compare its performance with two known decision support methods: UTADIS and Conjoint analysis. The methods were evaluated on several sets of randomly generated functions with domains of different dimensions and the resulting utility functions were scaled to the $[0, 1]$ interval, to allow comparative analysis. The overall quality of approximation is assessed by using multi-class AUC and RMSE measures. The Direct marginals method outperformed other approaches on all test functions with respect to the AUC method and on majority of test functions with respect to the RMSE measure. Conjoint analysis follows very closely. All tested methods give fairly good approximations of monotone DEX utility functions and give additional insight into decision makers preferences on attribute level, but also

between different attributes. We believe that such insight might be useful for different decision problems, for instance, product manufacturers to evaluate their products and locate important and interesting features that should be improved or changed to satisfy their customers.

In the future work, we would like to address the problem of approximating incompletely defined DEX functions and DEX functions defined with distribution of classes.

5 Acknowledgements

We would like to thank prof. Lavoslav Čaklović for introducing us to Conjoint analysis method.

References

[1] Bak, A. (2009.), *Statystyczna analiza danych z wykorzystaniem programu R*, Warszawa: Wydawnictwo Naukowe PWN.

[2] Bohanec M., Urh B., Rajkovič V. (1992.), *Evaluating options by combined qualitative and quantitative methods*, Acta Psychologica 80, 67–89.

[3] Bohanec M., Zupan B. (2004.), *A function-decomposition method for development of hierarchical multi-attribute decision models*, Decision Support Systems, 36, 215–233.

[4] Bohanec M.: DEXi: Program for Multi-Attribute Decision Making (2014.), User's Manual, Version 4.01. IJS Report DP-11739, Ljubljana: Jožef Stefan Institute.

[5] Bohanec M., Rajkovič V., Bratko I., Zupan B., Žnidaršič M. (2013.), *DEX methodology: Three decades of qualitative multi-attribute modelling*, Informatica, 37, 49–54.

[6] Devaud, J., M. and G. Groussaud and E. Jacquet-Lagrez (1980.), *UTADIS: Une methode de construction de fonctions d'utilite additives rendant compte de jugements globaux"*, European working group on MCDA, Bochum.

[7] Figueira J.R., Greco S., Ehrgott M. (2005.), *Multiple Criteria Decision Analysis: State of the Art Surveys*, Boston: Springer.

[8] Green, P., E. and Srinivasan, V. (1990.), *Conjoint analysis in marketing: new developments with implications for research and practice*, The Journal of Marketing, 54(4):3-19.

[9] Jacquet-Lagreze, E. and Siskos, J. (1982.), *Assessing a set of additive utility functions for multicriteria decision-making, the UTA method*, European Journal of Operational Research, 10(2):151–164.

[10] Meyer P., Bigare S. (2015.), *Package 'MCDA'*
`http://cran.r-project.org/web/packages/MCDA/MCDA.pdf [13.05.2015]`

[11] Mihelčić, M. and Bohanec, M. (2014.), *Approximating DEX utility functions with methods UTA and ACUTA*, In Proceedings of the 17th International Conference Information Society IS 2014, A, 62-65, Ljubljana.

[12] Mileva-Boshkoska B., Bohanec M. (2012.), *A method for ranking non-linear qualitative decision preferences using copulas*, International Journal of Decision Support System Technology, 4(2), 42–58, 2012.

[13] Robin X. and Turck N. and Hainard A. and Tiberti N. and Lisacek F., Sanchez J.C, and Müller M. (2011.), *pROC: an open-source package for R and S+ to analyze and compare ROC curves*, BMC Bioinformatics, 12:77.

MANAGING DIFFERENT INTERESTS IN GROUP DECISION MAKING

Uroš Rajkovič
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55 a, SI-4000 Kranj, Slovenia
uros.rajkovic@fov.uni-mb.si

Vladislav Rajkovič
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55 a, SI-4000 Kranj, Slovenia
vladislav.rajkovic@fov.uni-mb.si

Abstract: The paper discusses group decision making in the frame of multi criteria decision knowledge modelling. Described are pros and cons of group decision making. Special emphasis is given to the leveraging of different interests and possibilities of formulating a joint decision. Available methods and techniques together with a properly organized group work can make a substantial contribution.

Keywords: decision making, multi criteria, knowledge modelling, groups, conflict of interests

1 INTRODUCTION

There is probably no apparent reason for doubting the old proverb »More heads are better than one« in decision making. Still, it is useful to understand the advantages of participating in a group compared to decisions made by an individual. Clearly, such an approach may also be met by certain problems as implies the proverb that »Too many cooks spoil the broth«.

The purpose of this contribution is to present how people work with one another in a decision group through stages of decision knowledge acquisition, processing and use. In other words it deals with *management of decision knowledge*. Special emphasis is on qualitative multi criteria decision making and its added value [6], [13]. Group decision making assumes participation of different people. It is a process in which two or more people influence one another while the decision is being carried out. Usually, the decision in question will affect those participating or their representatives at some point in the future. Participation is furthermore built around the idea of *different interests* that need to be integrated into a *joint decision* [8], [17], [18].

What is the role of operational research methods and techniques in facilitating decision making? Specifically, what can be expected from the information communication technology (ICT) that lies at our disposal? During the decision making as a socio-technological process we can justly count on the *synergy between a human and technology*. According to Dreyfus [10] neither human nor computer can achieve on their own what they can achieve together. It is by far not enough to be aware of existing methods, techniques and technologies. Group decision making has to be appropriately organized. The individual and the group have to be technologically literate. The aim is to harmonize the work among the members of the decision group by using different evaluation and decision making tools [2], [4], [18].

2 HOW TO MANAGE CONFLICTING INTERESTS?

It is completely natural for people to come to different decisions on the same issues. It is due to differences in preferential knowledge what can be attributed to differences in relations to the decision situation, values, principles, understanding of circumstances, knowledge and lack thereof. A decision regarding a new family car is subject to differences in preferences among parents and children, for instance.

On the basis of preferential knowledge a preferential relation between alternatives is established. This way they are listed according to their desirability – utility. An evaluation model can also be used to assess the degree of desirability of a specific alternative, for example by assigning scores on a scale from 0 to 10. Children and parents assess different cars in a different way.

How to merge different scores in order to reach a single decision? First we should check if different scores are not due to insufficient knowledge regarding goals, alternatives and possibilities. Providing arguments for one's different preferences can be helpful. Afterwards we face the different interests.

We differentiate among two basic approaches that are founded on the distinction whether different interest groups are willing to cooperate or not in search of a righteous decision or choice.

Those groups that do not wish to consult one another and cooperate can implement one of the formal methods, for example voting. Again, every method has its advantages and disadvantages. Nobel laureate Arrow (awarded Nobel Prize in 1972) [1] demonstrated and proved through the *impossibility theorem* that an ideal method cannot and does not exist. Still, this does not preclude us from group decision making altogether but rather encourages us to look for the most appropriate method in a given situation.

If we decide that each interest group assigns to each alternative its own degree of utility and if they are willing to look for a *compromise solution*, a few other approaches are available [14]. Let us take the two already mentioned interest groups, namely parents and children deciding on a new family car. Each group assesses each car that matches a point in a system of coordinates, for example V1 (value 1 assigned by parents) and V2 (value 2 assigned by children) as shown in Figure 1.

It is sensible to deal with only non-inferior alternatives which lie on the bolded line depicted in Figure 1. Cars below this line have clear superior alternatives with a higher score given by one group and same or higher score given by the other group. Being aware of this can save us quite some further work.

The remaining question is, which of the alternatives that do not fall among the inferior ones should be chosen as a final group decision? If we choose the approach of »equal satisfaction«, graphically this means deciding for the intersection of a straight line connecting points where V1=V2 with a bolded line in Figure 1. Our imaginary family would thus look for a car that would be similarly assessed by both parents and children. Harsanyi [12] proposes to choose the alternative that maximizes the sum of individual utilities. It is disputed that what can occur are situations in which some groups sacrifice their interest for the common good. Nash (awarded Nobel Prize in 1994) proposed leveraging of interests by maximizing the product of utilities (individual utilities multiplied) [15]. In other words, we consider not only ourselves but also others. It is in the group's best interest not to allow sacrificial lamb.

Examples of leveraging in Figure 1 depict leveraging of interest based on final utilities (scores). Decision knowledge is expressed only with the final utility value. Still, we lack the understanding of the origin of the different scores. The final score is only a consequence.

2.1 Multi-attribute group decision making

When we try to leverage the origins of different scores and not only the consequence, that is the final score, we can apply the hierarchical multi-attribute models [4], [14], [13]. They are structured, have internally devised parameters and are open. This is why they not only produce final scores but also enable us to »look inside« and see how and why the scores came about. We can address specific parameters, their values and relationships among them. All of the evaluation elements are at our hand.

Figure 2 shows an example of a tree of criteria constructed to evaluate a car. The basic criteria or the tree leaves are the following: Purchase price, Maintenance costs, Number of doors, Number of passengers, Luggage space, and Safety. Introduced are also intermediate aggregate criteria. Intermediate criterion Price is composed of two sub-criteria, namely Purchase price and Maintenance costs, while intermediate criterion Comfort consists of three basic criteria: Number of doors, Number of passengers, and Luggage space. Intermediate criterion Technical characteristics combines subordinate intermediate criterion Comfort and basic criterion Safety. The criteria were assigned discrete value domains, which can be observed in Figure 2. For basic criteria were used either descriptive or numerical values (Number of doors). Meaning of descriptive values, such as for instance high Price or acceptable Safety, also had to be defined simultaneously.

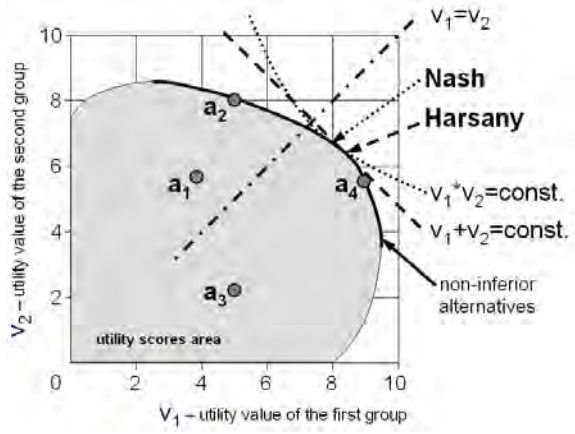


Figure 1: Comparison of alternatives evaluated by two different interest groups

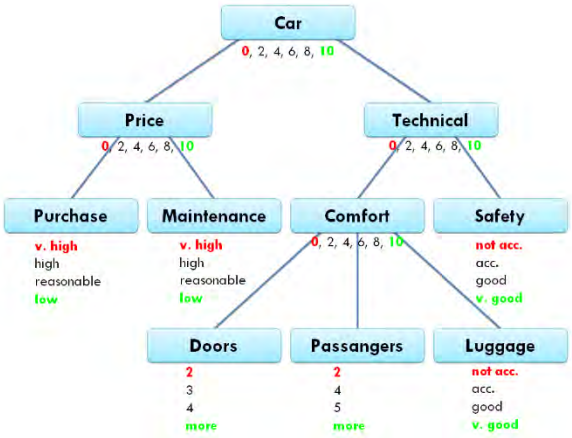


Figure 2: Tree of criteria for car evaluation with value domains

Our experiences [11], [3], [6] show that a unified model structure should be used despite different preferences that may arise due to different interests. Each interest group can however define within this structure its own utility function [16], [13]. The model is then used to evaluate the alternatives for each group separately. Usually, we end up with different scores for the same alternatives. We are not faced with the diversity only when it comes to final scores but can also gain insight into the reasons and origins of the scores provided for specific parameters for each alternative. Instead of leveraging (harmonizing) only the final score, we can investigate at where the differences stem from and what they are like. An explanation helps us realize the key stumbling blocks responsible for disagreements that can serve as a foundation for further interest leveraging among groups.

Figure 3 presents the evaluation score for the alternative Car 3 for parents and children according to all the criteria in the tree structure. Final scores 2 and 4 are inherently different. The difference stems from how the Price and Comfort are perceived. Even though both parents and children consider the Purchase price and Maintenance costs on the same terms, the aggregate score for parents is 6 and for children 8. Let us take a look at the utility functions which are described in table form in Figure 4 and thus establish the source of differences in the scores. Discrepancies are evident in four combinations. Each combination can be interpreted also as a logical rule. For the final score of Car 3 row 14 clearly bears importance. Parents think that low Purchase price and high Maintenance costs yield the value of aggregate score 6, while children think it is 8. Obviously, parents are responsible for financing a car and are far more put off by high Maintenance costs than children. Discussion and negotiations then focus only on the identified differences. Such differences should be discussed and underlying reasons

should be made clear. The aim is to reach a decision that is in favour of the decision maker who carries the burden of the discrepancy, in this case the parents providing the financial means.

Attribute	Parents	Children
Car	2	4
Price	6	8
Purchase	low	
Maintenance	high	
Technical	2	
Comfort	4	2
Doors	3	
Passengers	5	
Luggage	acc.	
Safety	good	

Figure 3: Comparison of different evaluation results for alternative Car 3

	Purchase	Maintenance	Price	
			Parents	Children
			58% : 42%	64% : 36%
1	v. high	v. high	0	
2	v. high	high	0	
3	v. high	reasonable	0	
4	v. high	low	0	
5	high	v. high	0	
6	high	high	2	
7	high	reasonable	2	
8	high	low	6	4
9	reasonable	v. high	0	
10	reasonable	high	2	
11	reasonable	reasonable	6	
12	reasonable	low	8	
13	low	v. high	4	6
14	low	high	6	8
15	low	reasonable	8	10
16	low	low	10	

Figure 4: Utility functions by parents and children for aggregate criterion Price

3 CONCLUSION

Group decision making is more demanding when differences in preferences of team members are present. Argumentation, why a certain decision was made in a particular way and not the other, increases the probability for a good decision or at least diminishes probability of a bad one. Group decision is more easily understood and can be better justified as we focus only on the differences among team members.

A clear and well justified decision is crucial for a sensible leveraging of different interests. Final score of the alternative is a consequence of numerous factors that appear in the evaluation process. Our decision processes can be and need to be transparent all the way from specific criteria (measures), to their aggregation and final score assigned to an alternative.

Various existing approaches, methods and techniques supported with ICT can be applied [5], [7], [9]. Let us make use of them. We should strive for open and clear models in order to make decision knowledge available to everyone affected. When we are deciding on the most suitable alternative, let us not consider only ourselves but also everyone else involved.

References

[1] Arrow, K. J., Sen, A. K. and Suzumura, K. (2002). Handbook of social choice and welfare. Amsterdam: North Holland.

[2] Bazerman, M. H. and Moore, D. (2009). Judgement in managerial decision making. London: Wiley.

[3] Bohanec, M. and Rajkovič, V. (1990). DEX: an expert system shell for decision support. Sistemica, Vol(1): 145-157. URL: http://www-ai.ijs.si/MarkoBohanec/pub/Sistemica90.pdf

[4] Bohanec, M. (2006). Odločanje in modeli (ang. Decision making and models). Ljubljana: DMFA – založništvo.

[5] Bohanec, M., Urh, B. and Rajkovič, V. (1992). Evaluating options by combined qualitative and quantitative methods. Acta Psychologica, Vol(80): 67-89.

[6] Bohanec, M., Rajkovič, V., Bratko, I., Zupan, B., and Žnidaršič, M. (2013). DEX Methodology: Three Decades of Qualitative Multi-attribute Modelling. Informatica, Vol(37): 49-54.

[7] Bohanec, M. (2014). DEXi: Program for Multi-Attribute Decision Making, User's Manual, Version 4.01. Jožef Stefan Institute. IJS Report DP-11739, Ljubljana.

[8] De Haan, A. and de Heer P. (2015). Solving Complex Problems: Professional Group Decision Making Support in Highly Complex Situations. New York: Eleven.

[9] DEXi: A program for qualitative multi-attribute decision modelling (1999-2014): version 4.01. Ljubljana: Jožef Stefan Institute. http://kt.ijs.si/MarkoBohanec/dexi.html (Acc.: 12.5.2015)

[10] Dreyfus, H. L. (1992). What computers still can't do: A critique of artificial reason. Cambridge: The MIT Press.

[11] Efstathiou, J. and Rajkovič, V. (1979). Multi-attribute decision making using a fuzzy heuristic approach. IEEE Transaction on Systems, Man and Cybernetics, Vol(9): 326-333.

[12] Harsanyi, J. (1955). Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. Journal of Political Economy, Vol(63): 309-321.

[13] Ishizaka, A. and Nemery, P. (2013). Multi-Criteria Decision Anlysis: Methods and Software. London: Wiley.

[14] Lu, J., Zhang, G. and Ruan, D. (2007). Multi-objective decision making: Methods, software and applications with fuzzy set techniques. London: Imperial College Press.

[15] Nash, J. (1950). Equilibrium points in n-person games. Proceedings of the National Academy of Sciences, Vol(36): 48-49.

[16] Rajkovič, V., Bohanec, M. and Batagelj, V. (1988). Knowledge engineering techniques for utility identification. Acta Psychologica, Vol(68): 271-286.

[17] Saaty, T. L. and Peniwati, K. (2013). Group Decision Making: Drawing out and Reconciling Differences. Pittsburgh: RWS Publications.

[18] Triantaphyllou, E. (2000). Multi-criteria decision making models: A comparative study. London: Kluwer Academic Publishers.

[19] Zarate, P., Belaud, J. P., Camilleri, G. and Ravat, F. (eds.) (2008). Collaborative decision making: Perspectives and challenges. Amsterdam: IOS Press.

APPLICATION OF MULTI CRITERIA DEX MODEL IN HOP BREEDING

Črtomir Rozman
University of Maribor, Faculty of agriculture and life science
Pivola 11, 2311 Hoče, Slovenia
crt.rozman@um.si

Karmen Pažek
University of Maribor, Faculty of agriculture and life science
Pivola 11, 2311 Hoče, Slovenia
crt.rozman@um.si

Viljem Pavlovič
University of Maribor, Faculty of agriculture and life science
Pivola 11, 2311 Hoče, Slovenia
crt.rozman@um.si

Marko Bohanec
Jožef Stefan Institute, Department of Knowledge Technologies,
Jamova 39, 1000 Ljubljana, Slovenia
marko.bohanec@ijs.si

Martin Pavlovič
University of Maribor, Faculty of agriculture and life science
Pivola 11, 2311 Hoče, Slovenia
crt.rozman@um.si

Abstract: The planning process in agriculture often requires consideration of many conflicting criteria and participation of multiple stakeholders with conflicting interests. The multi criteria decision method DEX is therefore a viable option for decision support in farm management. This study briefly reviews The DEXi-HOP 1.0 model enables an assessment and ranking of individual hop hybrids' and hop varieties' breeding potentials. The model has 18 attributes, hierarchically grouped within four aggregated attributes: Biology, Chemistry, Morphology and Brewing value. Furthermore, utility functions in the model were defined by sets of elementary decision rules through the entire hierarchy for all aggregated attributes. Four Slovenian hop hybrids, A1/54, A2/104, A3/112, A4/122 and a reference hop variety Hallertauer Magnum with target characteristics in plant resistance and brewing value, were used for the model assessment.

Keywords: multi criteria decision making, DEXi, hop breeding

1 INTRODUCTION

Multi criteria decision analysis can be applied when the evaluation involves several variables that cannot be easily transformed into quantitative units, and the assessment process is likely to be influenced by multiple competing criteria. Such situation often emerges in agriculture and the multi criteria analysis for different kind of assessmets systems has been applied in many cases (Pavlovič et al. 2011; Žnidaršič et al. 2008; Bohanec et al. 2008; Mazetto and Bonera 2003; Griffits et al. 2008, Tiwari et al. 2009; Tojnko et al. 2011).

The most common methods like analytical hierarchical process (AHP) and multi attribute utility theory are based on quantitative assessment. For instance AHP has been used for variety assessment before (Rozman et al., 2015; Srđevićet al., 2004). On the contrary, the method DEXi (Bohanec et al. 2000) is based on discrete values of attributes and utility functions in the form of "if…then" decision rules. In particular, some methods, such as DEXi (Bohanec and Rajkovič 1990; Bohanec et al. 2000), facilitate the design of qualitative

(symbolic) decision models. In contrast to conventional quantitative (numeric) models, qualitative models use symbolic variables. These seem to be well-suited for dealing with 'soft' decision problems, that is, less-structured and less-formalized problems that involve a great deal of expert judgment and where qualitative scales can be more informative than quantitative scores. The DEXi method has already been successfully used in numerous real life decision and assessment problems such as for the estimation of tourist farm service quality (Rozman et al. 2009) or assessment of multifunctional contributions of "Streuobst" stands.

The aim of this paper is to present the applications of method DEXi in agriculture on real world agricultural decision problem, namely hop breeding.

2 ASSESSMENT OF NEW HOP CULTIVARS

The hop model (Pavlovič et al. 2011) was developed in order to assess new potential hop hybrids. Within the hop breeding research program carried out at the Slovenian Institute of Hop Research and Brewing, thousands of hop hybrids appeared to be perspective according to research objectives (Cerenak 2006). In this research the data from four different Slovenian hop hybrids A1/54, A2/104, A3/112, A4/122 were compared with a reference German variety Hallertauer Magnum, which had the desired characteristics plant resistance and brewing value. The assessment was carried out by a qualitative multi-attribute model based on the DEX methodology (Bohanec et al. 2000). We first developed the model and then applied it to assess the aforementioned perspective hybrids. The model hierarchy is shown in figure 1.

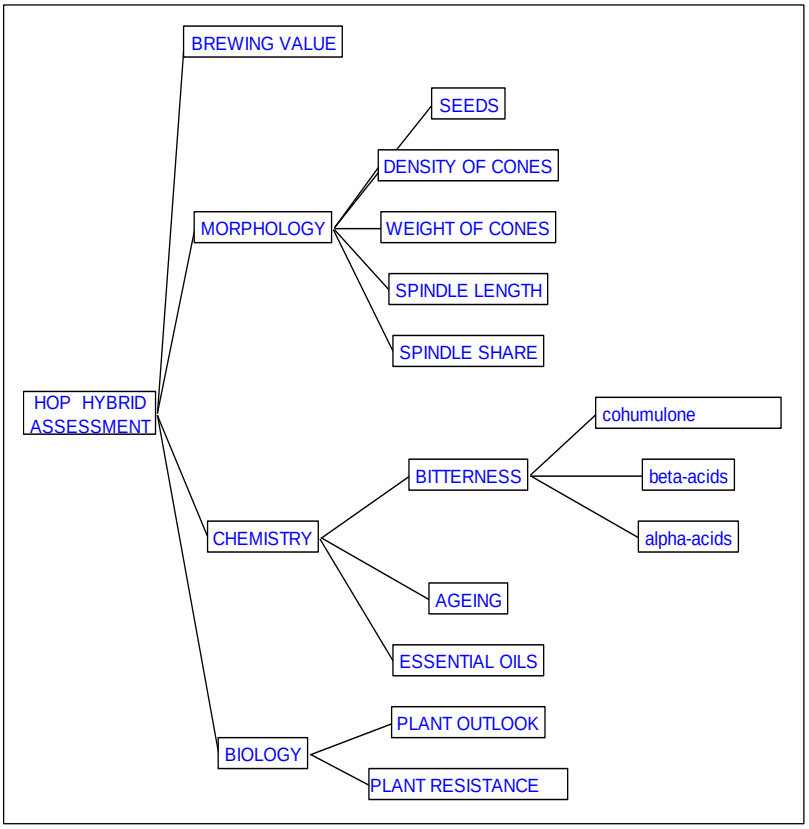


Figure 1: Hop decision model

Among over one thousand of hybrid hop plants analyzed and eliminated stepwise through a selection procedure, the four Slovenian hop hybrids such as A1/54, A2/104, A3/112, A4/122 and a reference variety Hallertauer Magnum were involved into a comparative model assessment. The hop hybrids had been selected through a hop breeding process among sets of seedlings analysed and assessed as highly forthcoming and promising new hop varieties. Numerical data of analyses and measurements of hop cones as well as beer sensory estimation were used to describe hybrids production and brewing quality parameters. They were analyzed and results were additionally discussed. The model enabled a final assessment of hybrids based on defined attributes and decision rules within defined utility functions figures 2).

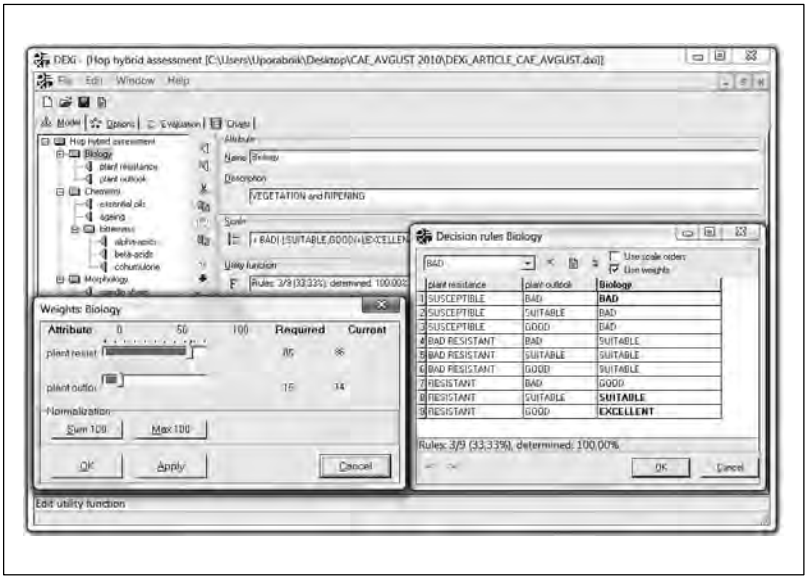
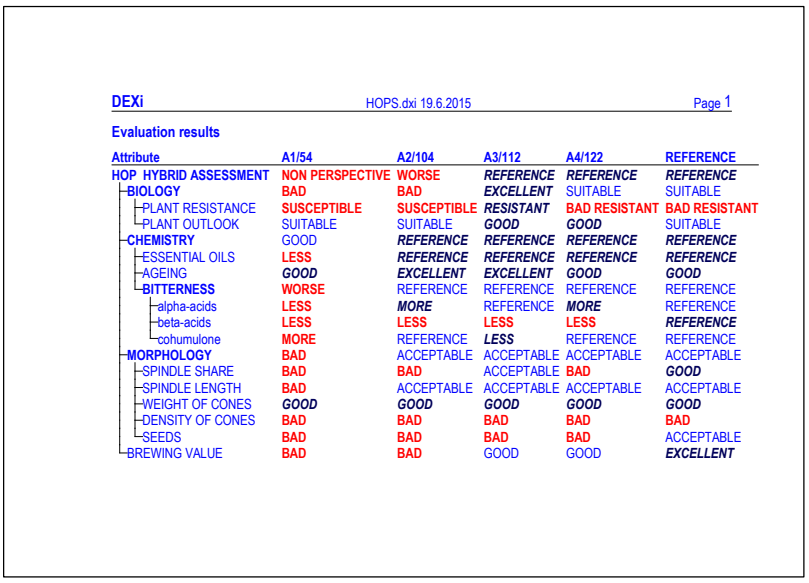


Figure 2: Determination of the DEX-HOP model utility functions for attribute *Biology*.

Based on breeding experiences and the DEXi-HOP 1.0 model results (figure 3), the overall as well as individual (aggregated and derived) attributes assessments were carried out. The results are shown on figures A3/112 and A4/122 reached the overall level of reference and were thus assessed as appropriate for further breeding. On the contrary, A1/54 and A2/104 did not meet expectations in their attributes related to the reference variety. A2/104 was in overall assessed as WORSE, while A1/54 as NON PERSPECTIVE. Therefore, they were considered as hybrids with less breeding potentials. The DEXi model was able to provide additional information on 4 hop hybrids that were initially all considered as perspective by the breeders. We were able to additionally rank them within the group of previously identified hybrids marked as perspective on the basis of breeder's assessment.



DEXi HOPS.dxi 19.6.2015 Page 1

Evaluation results

Attribute	A1/54	A2/104	A3/112	A4/122	REFERENCE
HOP HYBRID ASSESSMENT	NON PERSPECTIVE	WORSE	*REFERENCE*	*REFERENCE*	*REFERENCE*
BIOLOGY	BAD	BAD	*EXCELLENT*	SUITABLE	SUITABLE
PLANT RESISTANCE	SUSCEPTIBLE	SUSCEPTIBLE	*RESISTANT*	BAD RESISTANT	BAD RESISTANT
PLANT OUTLOOK	SUITABLE	SUITABLE	*GOOD*	*GOOD*	SUITABLE
CHEMISTRY	GOOD	*REFERENCE*	*REFERENCE*	*REFERENCE*	*REFERENCE*
ESSENTIAL OILS	LESS	*REFERENCE*	*REFERENCE*	*REFERENCE*	*REFERENCE*
AGEING	*GOOD*	*EXCELLENT*	*EXCELLENT*	*GOOD*	*GOOD*
BITTERNESS	WORSE	REFERENCE	REFERENCE	REFERENCE	REFERENCE
alpha-acids	LESS	*MORE*	REFERENCE	*MORE*	REFERENCE
beta-acids	LESS	LESS	LESS	LESS	*REFERENCE*
cohumulone	MORE	REFERENCE	*LESS*	REFERENCE	REFERENCE
MORPHOLOGY	BAD	ACCEPTABLE	ACCEPTABLE	ACCEPTABLE	ACCEPTABLE
SPINDLE SHARE	BAD	BAD	ACCEPTABLE	BAD	*GOOD*
SPINDLE LENGTH	BAD	ACCEPTABLE	ACCEPTABLE	ACCEPTABLE	ACCEPTABLE
WEIGHT OF CONES	*GOOD*	*GOOD*	*GOOD*	*GOOD*	*GOOD*
DENSITY OF CONES	BAD	BAD	BAD	BAD	BAD
SEEDS	BAD	BAD	BAD	BAD	ACCEPTABLE
BREWING VALUE	BAD	BAD	GOOD	GOOD	*EXCELLENT*

Figure 3: DEXi assessment for all four analyzed hop hybrids and the reference

Different kind of analyses can be conducted using DEXi. For instance figure 4 shows a comparison between reference hybrid (Magnum) and hybrid A1/54 that was assessed as NON PERSPECTIVE.

None of the hybrids was able to achieve the same Brewing value as reference cultivar. The chart on figure 5 shows scatter chart for the attribute Brewing value.

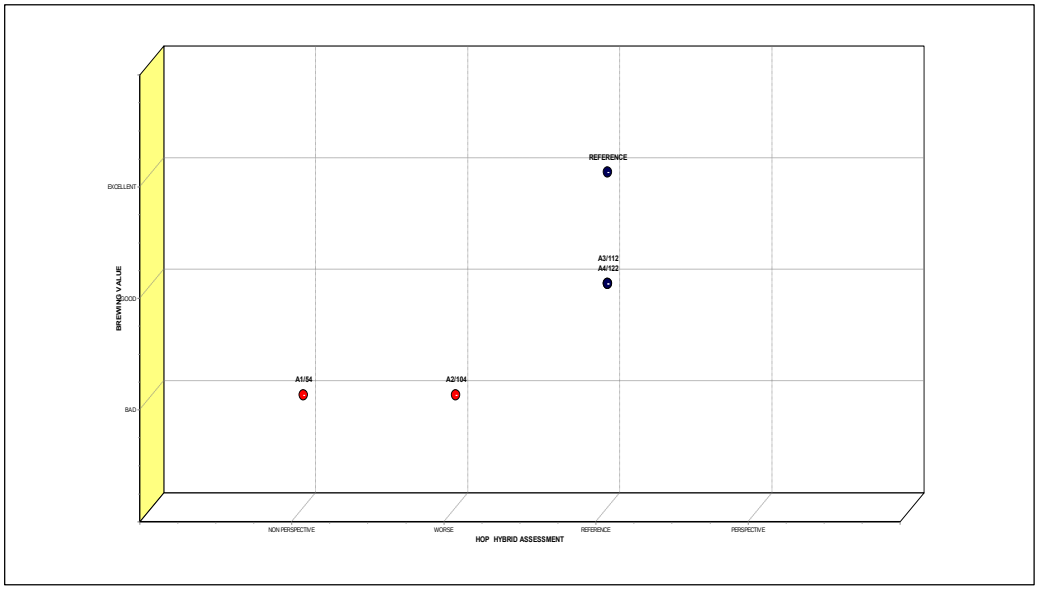


Figure 5: Scatter chart for attribute Brewing value

3 CONCLUSION

In this paper, an attempt was made to present multi-criteria method DEXI, based on qualitative attribute values and utility functions in the form of decision rules, and its possible application in the field of hop breeding.

Despite of the minor deficiencies (such as use of qualitative data only), it was found out that the approach has fulfilled most of the breeders' expectations and revealed considerable advantages in comparison with other approaches. The multi attribute model DEX-HOP 1.0 can therefore be regarded as a useful alternative tool for hop hybrids assessment. We can observe that none of the hybrids is fully equal with the reference cultivar.

This method cannot entirely replace experts, but it can be their additional tool in decision-making, since decisions based on model testing offered much faster results that validate the application of the model for further research. In future, data of new coming hybrids will be added and assessed in comparison to experts' decisions. Furthermore, also new attributes as a response to new goals in hop breeding programs will be included into the model.

Acknowledgement
This research was funded by Slovenia research agency program P4-0022.

References

[1] Bohanec M., 2008. DEXi: Program for multi-attribute decision making, User's manual, Version 3.00. IJS Report DP-9989, Jožef Stefan Institute, Ljubljana, Available at: http://kt.ijs.si/MarkoBohanec/pub/DEXiiManual30r.pdf (10.6.2012)

[2] Bohanec M., Messean A, Scatasta S, Angevin F, Griffiths B, Krogh P.H, Znidarsic M., Dzeroski S., 2008. A qualitative multi-attribute model for economic and ecological assessment of genetically modified crops. Ecol. Model., 215:247-261.

[3] Bohanec M. and Rajkovič V., 1990. DEXi: An expert system shell for decision support. Sistemica, 1(1):45–157.

[4] Bohanec M. and Rajkovič V., 1999. Multi attribute decision modelling: industrial application of DEXi. Informatica, 23:448-491.

[5] Bohanec M., Rajkovič B., Semolič B., Pogačnik A., 1995. Knowledge – based portfolio analysis for project evaluation. Information and Management,;28:293–302.

[6] Bohanec M., Zupan B., Rajkovič V., 2000. Applications of qualitative multi-attribute decision models in health care. Int. J. Med. Inf., 58–59:191–205.

[7] Griffiths B., Hallett P., Bohanec M., Osler G., Ball B.C., 2008. A qualitative multi-attribute model for assessing the impact of management options on soil sustainability. Agriculture and the environment VII: Land management in a changing environment (eds. K. Crighton, R. Audsley), Proceedigs of the SAC and SEPA biennial conference, Edinburgh, 63-67.

[8] Mazzeto F. and Bonera R., 2003. MEACROS: a tool for multi-criteria evaluation of alternative cropping systems. Eur. J. Agr.; 18:379-387.

[9] Pavlovič M., Čerenak A., Pavlovič V., Rozman Č., Pažek K., Bohanec M. 2011. Development of DEXi-HOP multi-attribute decision model for preliminary hop hybrids assessment. Computers and Electronics in Agriculture;75:181-189.

[10] Rozman, Č., Hühner, M., Kolenko, M., Tojnko, S., Unuk, T., & Pažek, K. (2015). Apple Variety Assessment with Analytical Hierarchy Process. Erwerbs-Obstbau, 57(2), 97-104.

[11] Rozman Č., Potočnik M., Pažek K., Borec A., Majkovič D., Bohanec M. 2009.A Multi-criteria assesment of tourist farm service quality. Tour. Manage.;30:629-637.

[12] Srdjevic, B., Srdjevic, Z., & Kolarov, V. (2004, October). Group evaluation of walnut cultivars as a multicriterion decision-making process. In *CIGR International Conference, Beijing, China* (pp. 11-14).

[13] Tojnko S., Rozman Č., Unuk T., Pažek K., Pamič S., 2011. A Qualitative multi-attribute model for the multifunctional assessment of "Streuobst Stands" in NE Slovenia. Erwerbs-obstbau. 53(4):157-166.

[14] Žnidaršič M., Bohanec M., Zupan B. 2008. Modelling impacts of cropping systems: Demands and solutions for DEXI methodology. Eur. J. Oper. Res.; 189:594-608.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Special Session 2:

Inventory Research

FLOATING POINTS OF A COLD SUPPLY CHAIN IN AN ENVIRONMENT OF THE CHANGING ECONOMIC GROWTH

David Bogataj and Marija Bogataj
MEDIFAS-CERRISK, Mednarodni prehod 6 Vrtojba, SI-5290 Šempeter pri Gorici ; and Faculty of Organizational Studies Novo mesto, SI-8000, Slovenia
dbogataj@actuary.si; Marija.Bogataj@guest.arnes.si

Abstract: In global economy nearly one third of all food produced is lost or wasted. It is estimated that majority of these loss which could be managed better is a result of long time delays in global supply chains. The urban population in 2014 accounted for 54% of the total global population and continues to grow, as reported by World health organization. As we face the problem of feeding over 9 billion people by the middle of this century, more than half of them living in the cities, being fare away from the growing agricultural products, the methods and technologies are needed for better postharvest loss prevention as a critical global food security and sustainability issue for growing cities. To meet expected demands, yields will have to more than double, while we have the constant natural resources. Therefore we need better understand supply chains of perishables in global economic environment. In global cold supply chains (CSC) when more intelligent technologies are introduced in transportation modes, like ships, to improve postharvest loss prevention, supply chains have to consider the location of sensors in a chain as floating points of the SC graph. This paper presents the possibility how to measure the impact of these technologies on the Net Present Value of the activities in CSC in case that transportation lines are long and the detection of deterioration by sensors in the transportation node could influence lead time. In the case of longer transportation time also the changes in economy can influence the final gains of a chain. The evaluation of these new technologies is proposed through the Net Present Value approach of extended Material Requirements Planning (MRP) models, where a volatile economic growth is assumed.

Keywords: Supply Chain; time delay; perishability; Cold Supply Chain; MRP; EMRP; Postharvest Loss Prevention

1 INTRODUCTION

As already mentioned in our paper (Kovačić et al, 2014) decision makers in food logistics have not many appropriate tools to best decide about CSC logistics and how to control CSCs. More distant locations for growing the food and their transportation in multistage logistics systems often require very long lead times for transportation, warehousing and quarantines for quality control testing, which also take time. Therefore we suggested a model for multistage production and distribution of perishable goods when we need to consider many interactions of different factors. Their joint impacts on the added value of a global supply chain can be better evaluated by using Extended Material Requirement Planning models as previously developed by Bogataj et al. (2011) and later by Bogataj and Grubbström (2012, 2013). Material Requirements Planning (MRP) is a quite well known process modelling among managing production processes, covering both production planning and inventory management (Orlicky 1975). It is well established in practice since most of the multi-level production systems are operated using MRP's obvious advantages. The strong technical background of the MRP makes it also an exceptionally good basis for deeper scientific research, for which the term MRP Theory has been established (Grubbström 2007). For the purpose of the scientific observation, structures from the Bill of Materials (BOM) can be conveniently captured within a pair of input and output real matrices **H** and **G** (Grubbström and Bogataj eds. 1997) of dimension nxm where i=1,2,…n different group of items are processed on the activity level j=1,2,…,m where the items i from the child node i (often we are taking the same notation for items and their activity cell, where

they are produced) are sent to activity cell *j*, where the item *j* is produced. To these structures, lead times have been assigned using the Laplace transform theorems in well-developed MRP Theory. This allows us to evaluate cash flows with the use of the Net Present Value (NPV) calculation also in food industry (Kovačić et al, 2012, 2014). A detailed review of the MRP Theory and its background can be found in Grubbström and Tang (2000). Recently, MRP Theory has been recognized as a very useful method for studies of entire supply chains, covering not only production but also distribution, consumption and recycling processes (Grubbström et al. 2007). When transportation delays expressed by transportation matrix are included such systems can conveniently be scientifically researched using the so-called Extended MRP (EMRP) Theory (Bogataj and Grubbström 2013, Bogataj et al 2013). Detailed structures of input and output matrices H and G of such a complex system of food and other supply chains in repeated feedback logistics can be found in Kovačić and Bogataj (2011, 2015) and Kovačić et al (2014).

According to the basic MRP theory developed by Grubbström the *j*-th process is run on activity level (node) P_j, the volume of required inputs of item *i* is $h_{ij}P_j$ and the volume of produced (transformed, or uploaded, unloaded, conserved, detected as changed in quality) outputs of item *k* is $g_{kj}P_j$. The total of all inputs may then be collected into the column vector **HP**, and the total of all outputs into the column vector **GP**, from which the net production is determined as (**G - H**)**P**. In general **P** (and thereby net production) will be a time-varying vector-valued function. In MRP systems, *lead times* are essential ingredients and could be studied simultaneously using Laplace Transforms methodology. The lead time of a process is the time in advance of completion that the requirements are requested. If $P_j(t)$ is the rate of items *j* planned to be completed at time *t*, then the quantity $h_{ij}P_j(t)$ of items *i* need to be available for production (assembly, packaging…) the lead time τ_j in advance of time *t*, i.e. at time $(t-\tau_j)$. The volume $h_{ij}P_j$ of item *i*, previously having been part of *available inventory*, at time $(t-\tau_j)$ was ear-marked for the specific production or manipulation $P_j(t)$ and thereby moved into *work-in-process or logistic process* (*activities in general*). At time *t*, when this activity is completed, the identity of the items type *i* disappear and the newly produced items appear instead. While the item i is assumed to be located previously at the location i it will be available for activity j at the location j before the activity $P_j(t)$ starts and it will need τ_{ij} to arrive there.

2 A NEW APPROACH WITH FLOATING POINTS IN GRAPH OF MRP

Now we shall introduce sensors attached to the items with known perishability β_{ij} in the flow between i *and* j as the activity j' located on the link between two physical nodes and if the change in perishability appears having higher intensity β_{ij}^* the sensor there reports on these changes in the time delay $\tau_{ij}-\delta_{ij}$ before coming to the child node . If there are no perturbations, than the value of δ_{ij} is zero ($\delta_{ij}=0$). In order to incorporate the lead times of transportation and manipulations in the nodes for the processes, we transform the relevant time functions into Laplace transforms in the frequency domain (see Bogataj and Grubbström, 2012, 2013). For the case of simplicity we shall consider a *linear logistic* system, for which the components of activity *j* need to be in place τ_j time units before completion (packaging, sorting…) and sent from parent node i to j having a floating point of sensor in j' for additional time delay τ_{ij} in advance. We shall denote by $\boldsymbol{\beta}$ the matrix of

dynamics of rotting and decay from each parent node to each child node in the graph of the logistic chain by $\boldsymbol{\beta}^*$ as the matrix of perturbed dynamics. We shall also denote the matrix of distances between a floating point and its destination node in case of perturbation by matrix $\boldsymbol{\delta}$ having coefficients in the row belonging to the index of the floating point and in the column of the index of the child node of each floating point. $\boldsymbol{\delta}$ is zero matrix if there are no perturbations in the system. Let us denote by $\tilde{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta})$ the matrix of CSC graph which include floating points between each parent and child node of activity cells

$$\tilde{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta}) =$$

$$= \begin{bmatrix} 0 & 0 & \ldots & \ldots & \ldots & 0 \\ h_{21}e^{s\beta^*_{21}\delta_{21}} & 0 & 0 & 0 & \ldots\ldots & 0 \\ h_{31}e^{s\beta_{31}\tau_{31}} & h_{32}e^{s\beta_{32}(\tau_{31}-\delta_{21})} & 0 & 0 & \ldots\ldots\ldots & 0 \\ & & h_{43}e^{s\beta^*_{43}\delta_{43}} & 0 & \ldots\ldots & 0 \\ & & h_{53}e^{s\beta_{53}\delta_{53}} & h_{54}e^{\beta_{54}s(\tau_{53}-\delta_{43})} & \vdots & \vdots\qquad\vdots \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots\ldots & \ldots\vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots\ h_{n-1,n-2}e^{s\beta^*_{n-1,n-2}\delta_{n-1,n-2}} & 0 \\ 0 & 0 & 0 & 0 & \ldots\ h_{n,n-2}e^{s\beta_{n,n-2}\delta_{n,n-2}} & h_{n,n-1}e^{s\beta_{n,n-1}(\tau_{n,n-2}-\delta_{n-1,n-2})} \end{bmatrix} \tag{1}$$

and lead times in activity cells by

$$\boldsymbol{\tau} = \begin{bmatrix} e^{s\tau_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{s\tau_n} \end{bmatrix}. \tag{2}$$

Applying the time translation theorem like in Grubbström (1998) the input requirements as transforms will be here similarly expressed:

$$\tilde{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta}) \begin{bmatrix} e^{s\tau_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{s\tau_n} \end{bmatrix} \tilde{\mathbf{P}}(s) = \breve{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta})\tilde{\mathbf{P}}(s) \tag{3}$$

where $\breve{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta})$ is the perturbed *generalized transportation-logistics activity input matrix* capturing the volumes of requirements as well as their advanced timing, and $\breve{\mathbf{H}}(s,\boldsymbol{\beta},[0])$ is its unperturbed case. The vector $\mathbf{u}(s,\boldsymbol{\beta},[0]) = \breve{\mathbf{H}}(s,\boldsymbol{\beta},[0])\tilde{\mathbf{P}}(s)$ describes in a compact way all component volumes that need to be in place for the logistic plan $\tilde{\mathbf{P}}(s)$ to be possible and

perturbations described by $\mathbf{u}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta}) = \breve{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta})\tilde{\mathbf{P}}(s)$ need some feedback control to get a solution close to the required one. If a component of the net production vector is negative, there is a need either for taking this amount from available inventory or for importing this amount into the system and, if it is positive, it may be delivered (sold) to the environment of the system or added to available inventory.

Let $\tilde{\mathbf{F}}(s)$ be the vector of deliveries from the system and $\tilde{\mathbf{F}}^*(s)$ its perturbed case. Given the plan $\tilde{\mathbf{P}}(s)$, available inventory $\tilde{\mathbf{R}}(s)$ will develop according to

$$\tilde{\mathbf{R}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta}) = \frac{\mathbf{R}(0) + \left\{\mathbf{I} - \breve{\mathbf{H}}(s,\boldsymbol{\beta}^*,\boldsymbol{\delta})\right\}\tilde{\mathbf{P}}(s) - \tilde{\mathbf{F}}^*(s)}{s}, \tag{4}$$

where R(0) collects initial available inventory levels. The division by s represents a time integration of the flows represented by the other terms. Also here we can consider cyclical processes, repeating themselves in constant time intervals , j = 1, 2, … , m, we may write the plan in the same way as in equation (12) of the article of Bogataj and Grubbström (2013).

3 THE ECONOMIC RELATIONSHIPS OF THE MODEL

3.1 The economic relationships at constant GDP

Items are assumed to have unit economic values, which could be different in different nodes. If in the case that the cargo becomes suddenly highly exposed to the risk of decay and the smart devices recognise that, the system could report to near city, which hosts the child node, and smart city can organize the transactions for such cargo locally at lower but acceptable prices or, in the worst case, the city can organize the disposal of rotten goods. Therefore, we shall write the unperturbed price vector **p** being a row vector:

$$\mathbf{p} = \left[p_1, p_2, ..., p_n\right] \tag{5}$$

Usually at floating points it is $p_2 = p_4 = ... = p_{n-1} = 0$, but in case of perturbations the values could be lower than p_1 or even negative if immediately a disposal is needed. Therefore, there exists a perturbed vector of lower or even negative prices:

$$\mathbf{p}^* = \left\lfloor p_1^*, p_2^*, ..., p_n^* \right\rfloor \tag{6}$$

Following the procedure of Bogataj and Grubbström (2012, 2013) we can write the NPV of an unperturbed CSC. In the case of a stable economy and constant GDP, NPV is:

$$\mathrm{NPV} = \mathbf{p}\left(\mathbf{I} - \breve{\mathbf{H}}(\rho)\right)\tilde{\mathbf{P}}\left(\rho\right) - \mathbf{K}\tilde{\mathbf{v}}(\rho) \tag{7}$$

In case of perturbations in perishability we can write

$$\mathrm{NPV}^* = \mathbf{p}^*\left(\mathbf{I} - \breve{\mathbf{H}}(\rho,\boldsymbol{\beta}^*,\boldsymbol{\delta})\right)\tilde{\mathbf{P}}_1\left(\rho\right) - \mathbf{K}_1\tilde{\mathbf{v}}_1(\rho) \tag{8}$$

Here **K** are ordering costs appearing at each node, also in floating points (could be 0), collected into the row vector

$$\mathbf{K} = \left[K_1, K_2, ..., K_n\right] \ , \ \tilde{\mathbf{v}}(s) = \begin{bmatrix} \tilde{v}_1(s) \\ \vdots \\ \tilde{v}_m(s) \end{bmatrix} \tag{9}$$

and ρ is continuous interest rate. The formulas (7) – (9) are valid when GDP does not change in the time horizon.

3.2 NPV in changing economic growth

We shall introduce the general economic growth into models (4) and therefore in its criterion function (7) and (8). When we consider the economic growth $e^{\omega \cdot t}$ which may be obtained by increasing productivity in the long term, we can write the following equation:

$$\mathrm{NPV}(\omega) = \mathbf{p}(\mathbf{I} - \mathbf{H}(\rho - \omega))\tilde{\mathbf{P}}(\rho - \omega) - \mathbf{K}\mathbf{v}(\rho - \omega) \tag{10}$$

and in case of long-term perturbations in perishability we have to consider

$$\mathrm{NPV}^*(\omega) = \mathbf{p}^*(\mathbf{I} - \mathbf{H}(\rho - \omega, \boldsymbol{\beta}^*, \boldsymbol{\delta}))\tilde{\mathbf{P}}_1(\rho - \omega) - \mathbf{K}_1\mathbf{v}_1(\rho - \omega) \tag{11}$$

At a given CSC the values $\mathrm{NPV}^*(\omega)$ and $\mathrm{NPV}(\omega)$ depends on time delays in the detection of deterioration and on proper forecasting of the development of deterioration, which all depend on the quality of the smart devices and good cooperation with cities in the neighborhood of the nodes in the supply chain. Therefore, supply chains and cities involved in cooperative supply chain need smart devices to control, decide optimally and communicate with agglomerations in the surroundings of CSC to sell quickly the products which are still of acceptable quality, and to order the goods which do not reach the final destination according to the plan earlier, to mitigate the consequences of the risk realization regarding postharvest loss prevention and the adequate supply of goods to growing urban population.

4 CONCLUSION

The urban population in 2014 accounted for 54% of the total global population and continues to grow, as reported by WHO, http://www.who.int). It is estimated that by 2017, even in less developed countries, a majority of people will be living in urban areas. As we face the problem of feeding over 9 billion people by the middle of this century, more than half of them living in the cities (in Europe even 80% of Europeans will live in urban areas), being fare away from the growing agricultural products, the methods and technologies are needed for better postharvest loss prevention. To meet expected demands, yields will have to more than double, while we have the constant natural resources. Therefore we need better understand supply chains of perishables in global economic environment and cooperation of cities to act as buyer of cargo which is exposed to rick if staying in transport. In global CSCs intelligent technologies are introduced in transportation modes, to improve postharvest loss prevention. Here we presented how extended MRP model can be used as a support for planning and control the perishability, when sensors on the transportation roads are consider as floating point and economic growth or decline is predicted. Using NPV approach as presented here also very robust perturbations can be evaluated.

References

[1] Bogataj, M., Grubbström, R.W. (2012). On the Representation of Timing for Different Structures within MRP Theory. International Journal of Production Economic, 140 (2), 749–755.

[2] Bogataj, M., Grubbström, R.W. (2013). Transportation delays in reverse logistics. International Journal of Production Economics, 143 (2), 395–402.

[3] Bogataj, M., Grubbström, R.W., Bogataj, L. (2011). Efficient location of industrial activity cells in a global supply chain. International Journal of Production Economics, 133 (1), 243–250.

[4] Bogataj, D., Vodopivec, R., Bogataj, M. The extended MRP model for the evaluation and financing of superannuation schemes in a supply chain. *Technological and economic development of economy*, 2013, vol. 19, suppl. 1, str. S119-S133.

[5] Bogataj M., Bogataj L., and Vodopivec R. 2005. Stability of perishable goods in cold logistic chains. International journal of production economics, 93/94, 345–356.

[6] de-la-Fuente Ma. V. and L. Ros, What to Do with Rejected Loads in the Agri-Food Chain, ENTERprise Information Systems Communications in Computer and Information Science, vol. 109, pp. 251-260, Oct 2010

[7] Grubbström , R. W. (1967). On the application of the Laplace transform to certain economic problems, Manag. Sci., 13: 558-567

[8] Grubbström , R. W. (1998a). Stochastic properties of a production-inventory process with planned production using transform methodology. Int. J.Prod. Econ., 45, pp. 407–419

[9] Grubbström , R. W. (1998b) A net present value approach to safety stocks in a planned production, Int. J.Prod. Econ., 56-57: 213-229

[10] Grubbström , R. W. and L. Bogataj, (Eds.), (1997). Input–Output Analysis and Laplace Transforms in Material Requirements Planning, Storlien, 1997. FPP, Portorož, 1998.

[11] Grubbström, R.W. (2007). Transform Methodology Applied to Some Inventory Problems. Zeitschrift für Betriebswirtschaft, 77(3), 297–324.

[12] Grubbström, R.W., Bogataj, M., Bogataj, L., (2007). A compact representation of distribution and reverse logistics in the value chain. Ekonomska fakulteta, Ljubljana.

[13] Grubbström, R.W., Tang, O., (2000). An Overview of Input-Output Analysis Applied to Production-Inventory Systems. Economic Systems Review, 12, 3-25.

[14] Kovačić, D., Hontoria, E., Ros McDonnell, L., Bogataj, M. (2014). Location and lead-time perturbations in multi-level assembly systems of perishable goods in Spanish baby food logistics. Central European journal of operations research, Vol. 22, pp. 1-17. http://link.springer.com/article/10.1007/s10100-014-0372-5,

[15] Kovačić, D., Bogataj, M., 2013. Reverse logistics facility location using cyclical model of extended MRP theory. Central European Journal of Operations Research, 21(1), 41–57.

[16] Kovačić, D., Hontoria, E., Bogataj, M., Ros, L., 2012. Application of the Extended MRP Theory to a baby food company. In: Z. Babić (ed.), Croatian operational research review (14th International conference on operational research, Trogir, Croatia, September 26 - 28, 2012), 3, 41–51.

[17] Orlicky, J.A., 1975. Material Requirements Planning. McGraw-Hill, New York.

INVESTMENTS IN SMART NANO-CONTROL SYSTEMS IN COLD SUPPLY CHAINS

David Bogataj[1], Marija Bogataj[1] , Damjana Drobne[2] , Lorenzo Ros-McDonnell[3] , Rebeka Rudolf[4] and Domen Hudoklin[2]

[1]MEDIFAS-CERRISK Mednarodni prehod 6 Vrtojba, SI-5290 Šempeter pri Gorici ; dbogataj@actuary.si; Marija.Bogataj@guest.arnes.si

[2]University of Ljubljana, SI-1000 Ljubljana; Damjana.Drobne@bf.uni-lj.si; domen.hudoklin@fe.uni-lj.si

[3]Technical University Cartagena, ETSII, C/Dr. Fleming, s/n, 30202, Cartagena, Spain; Lorenzo.ros@upct.es

[4]University of Maribor, Slomškov trg 15, 2000,Slovenia rebeka.rudolf@uni-mb.si

Abstract: Fruit and Vegetable Supply Chain Management (F&V SCM) requires more intelligent technology to improve postharvest loss prevention; however, the current technology has to be improved to reduce the lead times and perishability dynamics in an intelligent F&V supply chain. This can be achieved by developing a more accurate environmental monitoring using advanced measurement techniques and by introducing new gold nanoparticle sensors (Au-NP-sensors) which would allow direct sensing of rotting or decaying. This paper presents the possibility how to measure the impact of these technologies on the management of fruit and vegetable supply chains, i.e. from packaging, warehousing to fruit and vegetable distribution, and thus improving the supply chain safety by tracking and tracing cargo and its environment. The evaluation of these new technologies is proposed through the Net Present Value approach of extended MRP models.

Keywords: supply chain; nanosensors; environment monitoring; time delay; perishability; MRP

1 INTRODUCTION

The fruit and vegetable supply chain (F&V SC) encompasses all the activities involved in the sourcing and procurement of fruit and vegetables to the final customer. The subject of this paper is planning, management and control of such systems, concerning F&V that enter a supply system, being transferred and transformed into finished goods at the final consumer. To enable safe transfer through a distribution system, automation can improve F&V SC visibility and efficiency and thus yield higher added value. Contemporary technological solutions include wireless sensor networks (WSNs), which have great impact on the added values in F&V SCM. A WSN is a wireless network consisting of small, spatially allocated devices in a supply network where many scattered sensors are controlling environmental and physical conditions where the cargo is exposed to perishability. They measure humidity, temperature, vibration and pressure, and how a cargo is transported, changing location and quality. As Yonzon et al. mentioned (2005), through the use of smart devices possessing multiple micro-sensors deployed in large numbers over wide areas covered by supply networks, an unprecedented capability exists for monitoring, tracking, and controlling F&V packages, pallets or individual items as cargo in transportation or warehousing.

Nanotechnology is the creation and use of materials or devices at 1 to 100 nanometer (nm) scale. At these dimensions, materials exhibit different physical properties and behaviours not observed at the microscopic level. These effects at the nanoscale have been used in nano-devices with new functionalities stemming from unique nanoscale characteristics. Surprisingly, in the agriculture and food sector the application of nanotechnology-based products is relatively recent and unrevealed compared with their use in other fields. At present, two major applications related to nanotechnology are being expected, which are food nanosensing and food nanostructured ingredients. In controlling environment conditions, nanosensors could measure physico-mechanical properties. Although sensors have a long history, the realm of nanosensors is relatively new (Lima and Ramakrishna 2006, Gruere, 2012). In general, nanosensors convey information about nanoparticles to the macro-world like

supply chains. They can be attached to containers, pallets and individual items in cargo to function as an active transport tracking devices. As such they can actively monitor the fruit and vegetable transportation and warehousing process, control and verify the handling conditions of cargo, like temperature or humidity of fresh foods and trigger feedback actions. Also, such devices can detect damage of items due to sudden shocks, pressures or changes of temperature and other activities which could cause more rapid perishability of fruit or vegetables. Therefore, we can expect that the abilities of nanosensors to detect and monitor the changes in the environmental (ambiental) conditions can improve the quality of F&V logistics substantially (Lu and Bowles, 2013), which in turn lead to a higher net present value of a supply chain.

The **aim** of this article is to describe the potential for developing Au nanosensors, and consider their applicability in the postharvest loss prevention in F&V SCM. We shall also show how the precise monitoring and earlier detection of changes in cargo can influence the manipulation with cargo and reduce the perishability dynamics, influencing an increase of the Net Present Value of logistics in SC by introduction of floating point in MRP Theory.

2 AU NANOPARTICLES IN NANOSENSORS

Nanostructure-engineered chemical sensors could be applied to detect gas and vapour in storage containers or environments. The sensing could be based on sensor resistance that is decreased in the presence of the gases or volatile molecules to be detected. Due to the large surface area, low surface energy barrier and high thermal and mechanical stability, nanostructured sensors potentially can offer higher sensitivity, lower power consumption and better robustness than the state-of-the-art systems, which make them more attractive for a variety of applications. For our needs, i.e. regarding the control of supply chains, gold nanoparticles (Au-NPs) are very promising particles for nanosensors. The unique properties of gold nanoparticles have already stimulated the increasing interest in their application in different areas. Due to their excellent characteristics like biocompatibility, conductivity, catalytic properties, high surface-to-volume ratio and high-density, Au-NPs have attracted a lot of attention in designing (bio)nanosensors Several methods have been described in the literature for the synthesis of Au-NPs of various sizes and shapes. One limitation in using nanomaterial-based products is in the limited possibility to be synthesized in sufficient amounts and pure, i.e. industrial synthesis. The major challenge, and the reason why beneficial properties of Au-NPs have not been introduced to nanosensors up to now, is the high-cost and large batch-to-batch variations in Au-NP production by standard procedures, indicating that commercially available methods for the production of Au-NPs are not sustainable for their large-scale application. The key objective of the cooperation of our research group with the industrial partner is in the production / testing of Au-NPs by a conceptually proven Ultrasonic Spray Pyrolysis (USP) positioned in a clean room in Zlatarna Celje d.d. (Rudolf et al. 2012, 2014). The authors from the University of Ljubljana together with the Zlatarna Celje researchers are already testing the biocompatibility of the Au-NPs produced by the industrial partner Zlatarna Celje, and the laboratory experiments have proved that they are able to improve a biosensor under specific conditions. An important advantage of using gold and not other NPs in nanosensors is the possibility to functionalize the Au NPs with a broad variety of biomolecules, which is important for standardization and mass production of nanosensors as part of smart devices for SC control of perishable goods in general and F&V in particular. Depending on the intended use of the synthesized nanoparticles, they can be collected in an electrostatic field or suspended in a desired medium. USP as a nanoparticle production method is a relatively inexpensive and quite a versatile technique for fine metallic, oxidic and composite nanoparticles. New activities were introduced together with the Laboratory of

Metrology and Quality of UL-FE-LMK, which is a holder of a primary dew-point generator that is now routinely used as a reference for high-end humidity measurements in industry, based on its evaluation and comparison (Heinonen et al, 2012, Hudoklin et al, 2008, Evers et al, 2007).

3 METHODOLOGICA APPROACH EXTENDED MRP MODELS OF F&V SC WITH EMBEDDED NANOSENSORS

Food losses can be forecast through a combination of Au-nanosensors, humidity and temperature sensors embedded in the smart device, developed and installed in trucks and warehouses. Internet of Things (IoT) enables intelligent control and management of F&V SC. IoT might allow us to enhance F&V SC, data and things via devices and sensors. Fruits and vegetables are handled and transferred between farmers and suppliers, distribution centres, the retailer and the final customer. In flexible supply chains, the destination nodes can change rapidly. These many nodes in the supply chain require agile and informed F&V SC. Key to fruit and vegetable in-transit visibility are cloud-based GIS, including GPS, smart measurement devices linked to different sensors, such as bionanosensors, temperature and humidity sensors, which provide the identity, location, quality of cargo, ambient and other tracking information. These are the backbone of the IoT as it relates to the supply chain. With the data gathered by smart technologies, a detailed visibility of cargo is provided all the way from the fruit and vegetable producer to the final user. Data gathered from GIS and smart devices based on sensor technologies allow supply chain professionals to automate shipping and delivery by exactly predicting the dynamics; they can monitor the perishability of cargo and ambient control which impact the quality of fruit and vegetable cargo in-transit. IoT brings all of these tools together by putting a smart measurement device in a trailer and combines them in an integrated device, enabling the transfer of important data onto the cloud, with the devices being able to identify the cargo regarding position, quality, temperature and humidity conditions, as well as traffic and the driving pattern influencing lead time in a supply chain. This allows the stakeholders to make efficient decisions that influence the sustainability of F&V SC. By enabling devices to communicate as required, IoT can help supply chain professionals: (a) ensure temperature and humidity stable and adaptive; (b) reduce postharvest loss; (c) save on fuel costs by optimizing the routes regarding traffic and the perishability of cargo, and create fleet efficiency; (d) monitor and optimally replenish the inventory in warehouses, and (e) enable user insight in the quality of food. In general, this technology can reduce time delays in decision-making and actions in the supply chain management. A prototype of strategic storage nodes can be designed in a collaborative network as an innovation, which in the case of rejected loads in the F&V SC will permit some treatments and a re-introduction of these loads in the product value chain, as described in the paper by de-la-Fuente and Ros (2010). We can conclude that the F&V SC involves the transportation, warehousing, packaging, loading and unloading of temperature sensitive products along a SC under thermal and humidity control to protect the integrity of these shipments. The question arises, till when and how the cargo handling can be improved in order to mitigate the postharvest loss. What should be the conditions and restrictions of a dynamic system management so that the final consumer's need for healthy food will be met? Are the static limitations of the perturbations at each moment as good as dynamic restrictions? To overview the chain flows and to achieve the optimal control in the procedures of cooling and improvements, the input–output approach studied by Grubbström (1996, 1998), Bogataj et al. (2005), Bogataj and Bogataj (2007) and the recent papers of Bogataj and Grubbström (2012) should give the basic approach. They developed the models which help to evaluate the influence of SC parameters and their perturbations on the chain performances, especially on the net present value of the supply chain operations, named the extended material requirements

planning (EMRP) model, which uses the input–output approach in packaging, warehousing and transportation formalization of a supply chain. For a certain time unit we can write:

$$S(t) = (I - H(t,\tau))P(t) - F(t) \tag{1}$$

Here we denote: the inventory per time unit is S, the net output from the system per time unit is $(I-H)P$, where P is the intensity of activities in nodes, the internal demand per time unit is HP, and the delivery per time unit is F. There exist lead times between the activity cells in the matrix $H(t,\tau)$ which could be perturbed and controlled, and such events could be better studied after the transformation of this presentation of flows in frequency domain. Let us denote the perturbed matrix with $H(\tau^*)$. The results of F&V supply chain management (F&V SCM) are finally expressed and evaluated as the net present value (NPV) of the final delivery reduced by the costs of production, distribution, additional cooling and deterioration of goods. How more precise measurement and feedback control increases the NPV of all activities of a supply chain can be expressed through the following formalization:

The F&V in a supply chain is transported altogether through n stages (k = 1; 2; ... ; n), where the set of stages includes all activity cells (warehouses, loading and unloading points including packaging activities and others) and also all artificial nodes, one at each branch between two nodes where sensors are located. The artificial node is located between a parent node and a child node in a random distance from the parent node. It presents the location where the perishability dynamics is changed, detected and reported through the smart device. Having in mind this artificial supply chain points of unknown time delay τ_k, the results of F&V supply chain management (F&V SCM) are finally expressed and valuated as the NPV of the final delivery reduced by the costs of production, distribution, additional cooling and deterioration of goods. How more precise measurement and feedback control increase the Net Present Value of all activities of a supply chain can be expressed through the following formalization, where at any of n stages (k = 1; 2; … ; n), the development of the unperturbed state of the system is described by a set of first-order linear differential-delay equations (DDE):

$$\frac{d\vec{x}_k(t)}{dt} = A_k \vec{x}_k(t) + B_k \vec{x}_k(t-\tau_k) \ , \qquad k = 1,2,...,n \tag{2}$$

with the initial condition: $\vec{x}_{k0}(\theta) = \vec{\varphi}(\theta)$, $\qquad -\tau_k \le \theta \le 0$ (3)

Here we have $\vec{x}_k(t) \in \Re^2$, where $\vec{x}_{1,k}(t)$, is the quantity of good products at time t at stage k and $x_{2,k}$ presents the deteriorated quantity of products. The matrix $A_k^* = A_k + \Delta A_k \in L(\Re^2, \Re^2)$ describes the intensity of deterioration of goods at stage k of the warehousing, packaging or distribution part of the supply chain, where ΔA_k is assumed to be detected by Au nano-biosensors. Here $L(\Re^2, \Re^2)$ denotes the real Banach space of all continuous linear maps: $\Lambda : \Re^m \to \Re^m$. The matrix $B_k^* = B_k + \Delta B_k \in L(\Re^2, \Re^2)$ presents the conservation effect which is activated with delay τ_k, which could be perturbed and therefore written as τ_k^* and depends on the state of the items in the system and the quality of sensors as well as the total feedback procedure which can have an additional time delay. Combining the basic input–output Eq. (1), developed for dynamic consideration of MRP in Grubbström (1998), and Bogataj, Grubbström (2012), with Eq. (2), the following procedures as explained in Bogataj et al. (2005) can be written, where the initial value at stage k is assumed to be equal to the final value at stage $k-1$: $\vec{x}_{k0}(\theta) = \vec{x}_{k-1f}(\theta)$, $-\tau_k \le \theta \le 0$

$$\begin{bmatrix}\dot{\vec{x}}_p(t)\\ \dot{\vec{x}}_1(t)\\ \dot{\vec{x}}_2(t)\\ \cdot\\ \cdot\\ \dot{\vec{x}}_n(t)\end{bmatrix}=\left[\begin{array}{cccc}0&0&\ldots&0\\ \hline 0&A_1^*&\ldots&0\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ 0&0&\ldots&A_n^*\end{array}\right]\begin{bmatrix}\vec{x}_p(t)\\ \vec{x}_1(t)\\ \vec{x}_2(t)\\ \cdot\\ \cdot\\ \vec{x}_n(t)\end{bmatrix}+\left[\begin{array}{cccc}I-H(\vec{\tau}^*)&0&\ldots&0\\ \hline 0&0&\ldots&0\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ 0&0&\ldots&0\end{array}\right]\begin{bmatrix}P(t)\\ 0\\ 0\\ \cdot\\ \cdot\\ 0\end{bmatrix}-\begin{bmatrix}F(t)\\ 0\\ 0\\ \cdot\\ \cdot\\ 0\end{bmatrix}+\left[\begin{array}{cccc}0&0&\ldots&0\\ \hline 0&B_1^*&\ldots&0\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ \cdot&\cdot&\ldots&\cdot\\ 0&0&\ldots&B_n^*\end{array}\right]\begin{bmatrix}0\\ \vec{x}_1(t-\tau_1^*)\\ \vec{x}_1(t-\tau_2^*)\\ \cdot\\ \cdot\\ \vec{x}_n(t-\tau_n^*)\end{bmatrix} \tag{4}$$

At any stage the total quantity of items is the sum of deteriorated or higher exposed to deterioration quantity (depending on the criteria defined in advance) and the good quantity of items

$$\left[\begin{array}{cccccc:cccccc}-1&0&0&\cdot&\cdot&0&1&1&0&0&\cdot&0\\ 0&-1&0&\cdot&\cdot&0&0&1&1&0&\cdot&0\\ 0&0&-1&\cdot&\cdot&0&0&0&1&1&\cdot&0\\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\ 0&0&0&\cdot&\cdot&-1&0&0&0&0&1&1\end{array}\right]\begin{bmatrix}\vec{x}_p(t)\\ \vec{x}_{11}(t)\\ \vec{x}_{12}(t)\\ \cdot\\ \cdot\\ \vec{x}_{n2}(t)\end{bmatrix}=0 \tag{5}$$

DDE (4) has the following compact presentation (4a, 5a) and conditions (5b,c):

$$\dot{\hat{X}}(t)=\hat{A}^*\hat{X}(t)+\left(\left(I-H(\tau^*)\right)\hat{P}(t)+\hat{B}^*(t)\hat{X}(t,\tau_1^*,\tau_2^*,\ldots,\tau_n^*)-\hat{F}(t)\right. \tag{4a}$$

$$\hat{G}\hat{X}(t)=0;t\in\left[0,t_f\right] \tag{5a}$$

$$initial:\hat{X}_0(\theta)=\hat{\Phi}(\theta);\quad \max_k\left\{\tau_k^*\right\}\le\theta<0 \tag{5b}$$

$$sequential:\vec{x}_{k0}(\theta)=\vec{x}_{k-1f}(\theta);\quad -\tau_k\le\theta\le 0 \tag{5c}$$

Let us write the dynamics (4.a) as $\hat{S}(t)=\dot{\hat{X}}(t)$ and control as explained in (4): $\hat{U}(t)=\widehat{(B+\Delta B^*)}\hat{\tilde{X}}(t,\vec{\tau}^*)$. When we consider the net present value of revenues achieved when the good part of products is sold, reduced by the costs of cooling and other costs in the supply chain, distributed on the time horizon, when the interest rate of costs and revenues is substantial, the approach in the Laplace transformed space gives more straightforward solutions than the solutions in time domain. From the basic knowledge about the Laplace transformed expressions:

$$\begin{aligned}&\pounds(\hat{X}(t)=\tilde{\hat{X}}(s)=\int_0^\infty e^{-st}\hat{X}(t)dt;\qquad \pounds\left\{\frac{d}{dt}\hat{X}(t)\right\}=s\tilde{\hat{X}}(s)-\hat{X}(0)\\ &\pounds\left\{\int_0^t\hat{S}(\alpha)d\alpha\right\}=\tilde{\hat{S}}(s)/s;\qquad \pounds\left\{\hat{X}(t,\vec{\tau})\right\}=\tilde{\hat{\tau}}(s)\cdot\pounds\left\{\int_0^t\hat{S}(\alpha)d\alpha\right\}=\tilde{\hat{\tau}}(s)\frac{\tilde{\hat{S}}(s)}{s}\end{aligned} \tag{6}$$

Similar as explained in Bogataj et al. (2005), also for the case of floating each second node Laplace transformed Eq. (4) in this space has the similar expression, while in case some floating nodes coincide with the child node on the branches of the graph the lead time there is zero. We can write $\tilde{\hat{S}}(s)=\left[I-\frac{\hat{A}+\hat{B}\tilde{\hat{\tau}}^*}{s}\right]\cdot\left[\hat{X}(0)+\left(\widehat{I-H(s,\tau^*)}\right)\tilde{\hat{P}}(s)-\tilde{\hat{F}}(s)\right]$ (7)

Using the Laplace transform of extended Eq. (6) in the same way as the transform of the basic Eq. (4), the perturbations can be studied in frequency space, where the influence of delays can be presented explicitly. The appropriate criterion function could be the net present value of the system described by (7):

$$J = NPV = \int_0^\infty e^{-st}\left\{\left(-\hat{C}_A\left(\widehat{A+\Delta A}\right)-\hat{C}_B\left(\widehat{B+\Delta B}\right)\right)\hat{S}(t)-\hat{C}_F\hat{M}\left(\left|\hat{S}(t)-\hat{D}(t,\beta)\right|_-\right)+\hat{C}_F^*\hat{D}(t,\beta)\right\}dt \tag{8}$$

$\hat{C}_A$ in (8) is the vector of deterioration costs evaluated by smart devices, $\hat{C}_B$ is the vector of cooling costs, also evaluated by smart devices, $\hat{C}_F^*$ is delivery revenue, $\hat{C}_F$ is the cost of shortage, and M is the matrix having 1 in the rows $3k-1$ and column $3k-1$ and 0 otherwise, where k are the stages where demand appears or cargo is deteriorated so that it should not be transported further if we follow the optimal policy. In the case where demand appears only at the end of the supply chain, 1 is placed only at $(3n-1,3n-1)$. $\hat{D}(t,\beta)$ is the demand which will not be satisfied by probability β. For Eq. (7) above we can write:

$$\tilde{\tilde{\tau}}(s)=\left[\begin{array}{cccc|cccccccc}
1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0\\
0 & 1 & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots\\
\vdots & \ddots & \ddots & 0 & \vdots & \vdots & \cdots & 0 & 0 & 0 & 0 & 0\\
0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0\\
\hline
0 & & 0 & 0 & e^{-s\tau_1} & 0 & 0 & & \cdots & & & 0\\
0 & & & & 0 & e^{-s\tau_1} & 0 & & \cdots & & & \vdots\\
 & & & & 0 & 0 & e^{-s\tau_2} & & & & & \vdots\\
 & & 0 & & \vdots & \ddots & & \ddots & & & & 0\\
 & & & & & & \ddots & & & \ddots & & \\
 & & & & 0 & \cdots & 0 & & \ddots & \ddots & & 0\\
 & & & & 0 & & & & & & e^{-s\tau_n} & 0\\
0 & 0 & & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 & 0 & e^{-s\tau_n}
\end{array}\right]_{3nx3n} \tag{9}$$

and $\widetilde{\Delta\tau}$ is the diagonal matrix, having values $e^{-s\Delta\tau_i}$ instead of $e^{-s\tau_i}$ on the diag. The control U has to be determined and registered in cloud which appears with a delay, so that the deterioration process will stay stable. It follows:

$$U=\left(\widehat{B+\Delta B}\right)\tilde{\tilde{\tau}}\,\widetilde{\Delta\tau}\frac{\tilde{\tilde{S}}(s)}{s}=\tilde{\tilde{S}}(s)-\left(\widehat{A+\Delta A}\right)\frac{\tilde{\tilde{S}}(s)}{s}-\hat{X}(0)+\left(\widehat{I-H(\tau^*)}\right)\tilde{\tilde{P}}(s)-\tilde{\tilde{F}}(s) \tag{10}$$

The optimization problem $\max(NPV)$ of (8) under constraints (10) and (5a–5c) can be solved, using one of the known mathematical programming approaches. In case of the constant $\hat{P}(t)=\hat{P}$ and $\hat{F}(t)=F$, the parametric linear programming approach gives us the proper evaluation of perturbations and the control needed.

5 CONCLUSION

Finally, we can conclude that F&V SC is thus a science, a technology and a process. It is a science since it requires the understanding of the chemical and biological processes linked to perishability and the systems theory which allows for developing a theoretical framework for control of systems with perturbed time-lags. Secondly, it is a technology developed in engineering since it relies on the physical means to assure appropriate temperature conditions along the F&V SC and, thirdly, it is also a process since a series of tasks must be performed to prepare, store, transport, monitor temperature and humidity sensitive cargo as well as to give a proper feedback control. Therefore, we have to break the silos of separated knowledge to build an interdisciplinary and multidisciplinary science of postharvest loss prevention. Considering the sensors as floating activity cells as nodes in a graph of such a system, an extended MRP theory can be applied to determine the optimal feedback control.

References

[1] Bogataj, D., Bogataj, M. 2007.Measuring the supply chain risk and vulnerability in frequency space. International journal of production economics, Vol. 108 (1-2), pp. 291-301.

[2] Bogataj M., Grubbström R. W. 2012. On the representation of timing for different structures within MRP theory. International journal of production economics, 140, 2, 749–755.

[3] Bogataj M., Bogataj L., and Vodopivec R. 2005. Stability of perishable goods in cold logistic chains. International journal of production economics, 93/94, 345–356.

[4] de-la-Fuente, Ma. V. and Ros, L. 2010. What to Do with Rejected Loads in the Agri-Food Chain. ENTERprise Information Systems Communications in Computer and Information Science. 109, 251–260.

[5] Evers L., Havinga P. J. M., Kuper J., Lijding M. E. M., Meratnia N. 2007. Sensor Scheme: Supply Chain Management Automation using Wireless Sensor Networks. ETFA. IEEE Conference on, Issue Date: 25–28 Sept. 2007.

[6] Gruere G. P. 2012. Implications of nanotechnology growth in food and agriculture in OECD countries. Food Policy 37, 191–198.

[7] Grubbström, R.W., 1996. Stochastic properties of a production-inventory process with planned production using transform methodology. International Journal of Production Economics 45, pp. 407–419

[8] Grubbström, R. W. 1998. A net present value approach to safety stocks in a planned production, *International Journal of Production Economics*: 56–57: 213–229.

[9] Heinonen M., Bojkovski J., Hudoklin D. et al. 2012. Investigation of the equivalence of national dew-point temperature realiza-tions in the -50 °C to +20 °C range. International journal of thermophysics, 33 (8/9).

[10] Hudoklin D., Bojkovski J., Nielsen. J, Drnovšek J. 2008. Design and validation of a new primary standard for calibration of the top-end humidity sensors. Measurement: 41 (9), 950-959.

[11] Lima T.-C. and Ramakrishna S. 2006. A Conceptual Review of Nanosensors. Z. Naturforsch. 61a, 402–412.

[12] Lu J. and Bowles M. 2013. How Will Nanotechnology Affect Agricultural Supply Chains? International Food and Agribusiness Management Review, 2013, vol. 16, issue 2, pp. 24–42.

[13] Rudolf, R., Friedrich, B., Stopic, S., Anžel, I., Tomic, S., Colic, M. Cytotoxicity of gold nanoparticles prepared by ultrasonic spray pyrolysis. Journal of biomaterials applications, ISSN 0885-3282, Jan. 2012, 26, 5, 595–611.

[14] Rudolf, R., Tomic, S., Anžel, I., Zupancic Hartner, T., Colic, M. 2014. Micro-structure and biocompatibility of gold-lanthanum strips. Gold bulletin, ISSN 2190-7579, 47, 4, 263–273.

[15] Yonzon C. R., Stuart D. A., Zhang X., McFarland A. D., Haynes C. L., Van Duyne R. P. 2005. Towards advanced chemical and biological nanosensors—An overview. Talanta 67, 438–448.

SOME ASPECTS ON THE DISTRIBUTION OF THE NET PRESENT VALUE OF A STOCHASTIC CASH FLOW GENERATED BY A COMPOUND RENEWAL PROCESS

Robert W Grubbström FVR RI, FLO K
Linköping Institute of Technology, SE-58183 Linköping, Sweden
Mediterranean Institute for Advanced Studies, Šempeter pri Gorici, Slovenia
E-mail: robert@grubbstrom.com

Abstract: A cash flow is generated by discrete payments along a continuous time axis. Each payment has a stochastic amplitude and two consecutive payments are separated in time by a stochastic interval. The amplitudes are assumed to be stochastically independent and taken from the same given distribution, and similarly, the intervals are stochastically independent and taken from a second distribution.

This paper analyses properties of the distribution of the Net Present Value (NPV) of the cash flow, given a continuous interest rate used for discounting. Whereas the moments of the NPV distribution are easily obtained from the properties of the amplitude and interval distributions, the NPV distribution is not easily found even in the simplest cases. Accepting the axioms of standard risk-preference theory, alternative cash flows belonging to the class treated may be compared coherently by determining their certainty equivalent NPV value.

Keywords: Renewal process, stochastic cash flow, net present value, NPV, risk preference function.

1 INTRODUCTION

Although the closed form formula for the Net Present Value (NPV) of a deterministic sequence of constant payments has been known for more than 200 years,

$$\mathrm{NPV} = \sum_{i=1}^{n} a\left(1+r\right)^{-i} = a \cdot \frac{1-\left(1+r\right)^{-n}}{r}, \tag{1}$$

where a is the amount of each discrete payment, n the number of payments, and r the simple interest rate corresponding to the length of the intervals between consecutive payments, its stochastic extension has only been touched upon in the literature for about 70 years. The corresponding NPV of a constant continuous payment flow of b monetary units per time unit follows the similar formula

$$\mathrm{NPV} = \int_0^T b e^{-\rho t} dt = b \cdot \frac{1-e^{-\rho T}}{\rho}, \tag{2}$$

where T is the duration of the stream and ρ the continuous interest rate. The simple interest rate and the continuous interest rate are related by

$$r = e^{\rho\tau} - 1, \ \rho = \frac{\ln\left(1+r\right)}{\tau}, \tag{3}$$

where τ is the time interval on which r is based.

In this paper we study properties of the distribution of the NPV of a sequence of n discrete stochastic payments $\mathbf{A}_n, \mathbf{A}_{n-1}, \ldots, \mathbf{A}_1$ separated by stochastic time intervals $\boldsymbol{\tau}_n, \boldsymbol{\tau}_{n-1}, \ldots, \boldsymbol{\tau}_1$. The reverse numbering is chosen for algebraic purposes in the next section. The first payment in time $\mathbf{A}_n$ thus takes place at time $\boldsymbol{\tau}_n$, the second $\mathbf{A}_{n-1}$ at $\boldsymbol{\tau}_n + \boldsymbol{\tau}_{n-1}$, and so on, and the last nth payment $\mathbf{A}_1$ at $\sum_{i=1}^{n} \boldsymbol{\tau}_i$. The payments are assumed to be stochastically independent and taken

from the same distribution $F_{\mathbf{A}}(A)$, and similarly for the time intervals having the common distribution $F_{\boldsymbol{\tau}}(\tau)$, These two sets of distributions are assumed to be stochastically independent. Although *n* is given, we may view the sequence of payments as a compound renewal process. We thus attempt to study properties of the distribution of a stochastic NPV:

$$\mathrm{NPV} = \sum_{i=n}^{1} \mathbf{A}_i e^{-\rho \sum_{j=n}^{i} \boldsymbol{\tau}_j} = \sum_{i=1}^{n} \mathbf{A}_i e^{-\rho \sum_{j=i}^{n} \boldsymbol{\tau}_j} . \tag{4}$$

We make use of the Laplace transform as a moment-generating function for non-negative probability density functions (pdfs)

$$\mu_m = (-1)^m \lim_{s \to 0} \frac{d^m \tilde{f}(s)}{ds^m},\ m = 0,1,\ldots, \tag{5}$$

where μ_m is the *m*th moment (about zero), $\tilde{f}(s)$ is the Laplace transform $\tilde{f}(s) = \pounds\{f(x)\} = \int_{x=0}^{\infty} f(x) e^{-sx} dx$ of a probability density $f(x)$ and *s* the compex Laplace frequency. The moments and central moments (moments about the mean) written $\mu'_m,\ m = 0,1,\ldots$ are related by

$$\mu'_m = \sum_{j=0}^{m} \binom{m}{j} (-\mu)^{m-j} \mu_j ,\ \mu_m = \sum_{j=0}^{m} \binom{m}{j} \mu^{m-j} \mu'_j , \tag{6}$$

where $\mu = \mu_1$ is the mean and μ'_2 the variance (σ^2). We always have $\mu'_0 = \mu_0 = 1$ and $\mu'_1 = 0$.

Among the selected references below are listed papers related to the problem treated here and to methodology applied. In this current paper, space is not offered for discussing and comparing with the various contributions.

2 THE MOMENTS OF THE DISTRIBUTION OF THE NPV

Let $\mathbf{Y}_n$ be the NPV with *n* payments according to (4), then

$$\mathbf{Y}_n = \sum_{i=1}^{n} \mathbf{A}_i e^{-\rho \sum_{j=n}^{i} \boldsymbol{\tau}_j} = e^{-\rho \boldsymbol{\tau}_n} (\mathbf{A}_n + \mathbf{Y}_{n-1}) = \mathbf{X}_n (\mathbf{A}_n + \mathbf{Y}_{n-1}), \tag{7}$$

where $\mathbf{X}_n = e^{-\rho \boldsymbol{\tau}_n}$ is the discount factor. The three expressions in the right-hand member are obviously independent. Therefore the *m*th moment of $\mathbf{Y}_n$ may be written

$$\mu_{\mathbf{Y}_n,m} = \mathrm{E}\left[(\mathbf{Y}_n)^m \right] = \mathrm{E}\left[\left(\mathbf{X}_n (\mathbf{A}_n + \mathbf{Y}_{n-1}) \right)^m \right] =$$

$$= \mathrm{E}\left[(\mathbf{X}_n)^m \right] \mathrm{E}\left[\sum_{j=0}^{m} \binom{m}{j} (\mathbf{A}_n)^{m-j} (\mathbf{Y}_{n-1})^j \right] = \mu_{\mathbf{X},m} \sum_{j=0}^{m} \binom{m}{j} \mu_{\mathbf{A},m-j} \mu_{\mathbf{Y}_{n-1},j} . \tag{8}$$

Since all $\mathbf{X}_n$ and all $\mathbf{A}_n$ have the same respective distribution, we have dropped the subscripts of **X** and **A** in the right-hand member. Eq.(8) may thus be written in matrix-vector form as

$$\boldsymbol{\mu}_{\mathbf{Y}_n} = \Phi_{\mathbf{X}} \Psi_{\mathbf{A}} \boldsymbol{\mu}_{\mathbf{Y}_{n-1}} , \tag{9}$$

or,

$$\begin{bmatrix} \mu_{\mathbf{Y}_n,0} \\ \mu_{\mathbf{Y}_n,1} \\ \mu_{\mathbf{Y}_n,2} \\ \vdots \\ \mu_{\mathbf{Y}_n,m} \end{bmatrix} = \begin{bmatrix} \mu_{\mathbf{X},0} & & & & \\ & \mu_{\mathbf{X},1} & & & \\ & & \mu_{\mathbf{X},2} & & \\ & & & \ddots & \\ & & & & \mu_{\mathbf{X},m} \end{bmatrix} \begin{bmatrix} \binom{0}{0}\mu_{\mathbf{A},0} & & & & \\ \binom{1}{0}\mu_{\mathbf{A},1} & \binom{1}{1}\mu_{\mathbf{A},0} & & & \\ \binom{2}{0}\mu_{\mathbf{A},2} & \binom{2}{1}\mu_{\mathbf{A},1} & \binom{2}{2}\mu_{\mathbf{A},0} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \binom{m}{0}\mu_{\mathbf{A},m} & \binom{m}{1}\mu_{\mathbf{A},m-1} & \binom{m}{2}\mu_{\mathbf{A},m-2} & \cdots & \binom{m}{m}\mu_{\mathbf{A},0} \end{bmatrix} \begin{bmatrix} \mu_{\mathbf{Y}_{n-1},0} \\ \mu_{\mathbf{Y}_{n-1},1} \\ \mu_{\mathbf{Y}_{n-1},2} \\ \vdots \\ \mu_{\mathbf{Y}_{n-1},m} \end{bmatrix}, \quad (10)$$

where $\boldsymbol{\mu}_{\mathbf{Y}_n}$ and $\boldsymbol{\mu}_{\mathbf{Y}_{n-1}}$ are (m+1)-dimensional column vectors with elements being the moments $\mu_{Y_n,j}$, $j = 0,1,...m$ (similarly for $\boldsymbol{\mu}_{\mathbf{Y}_{n-1}}$), $\Phi_{\mathbf{X}}$ a diagonal matrix with $\mu_{\mathbf{X},j}$ as its (j+1)st diagonal element, and $\Psi_{\mathbf{A}}$ a triangular matrix with non-zero elements only on and below its main diagonal and with $\binom{m}{j}\mu_{\mathbf{A},m-j}$ as its element in row ($m - j$ + 1) and column (j + 1). For n = 1 (only one payment), we have $\mathbf{Y}_1$=$\mathbf{X}_1$. Since $\Phi_{\mathbf{X}}$ and $\Psi_{\mathbf{A}}$ are constant matrices (independent of n), (9) may be written

$$\boldsymbol{\mu}_{\mathbf{Y}_n} = \left(\Phi_{\mathbf{X}}\Psi_{\mathbf{A}}\right)^{n-1}\boldsymbol{\mu}_{\mathbf{X}}, \quad (11)$$

Eqs (8)–(11) therefore provide a relatively simple formula for obtaining all moments of the NPV distribution for every $n > 0$, when the moments of the payments $\mathbf{A}$ and the moments of the individual discount factor $\mathbf{X} = e^{-\rho\boldsymbol{\tau}}$ are known (where the subscript of $\boldsymbol{\tau}$ has been omitted). For the very special case of all $\mathbf{A}_i$=1, $\Psi_{\mathbf{A}}$ becomes Pascal's triangle.

In the stationary limit of $n \to \infty$, we would have $\boldsymbol{\mu}_{\mathbf{Y}_n} = \boldsymbol{\mu}_{\mathbf{Y}_{n-1}}$, so the moments then form an Eigen-vector of the matrix $\mathbf{I} - \Phi_{\mathbf{X}}\Psi_{\mathbf{A}}$, where $\mathbf{I}$ is the identity matrix. Solving (10) for the stationary $\mu_{\mathbf{Y},i}$, i.e. an infinite sequence of payments, the first few elements are the following:

$$\begin{cases} \mu_{\mathbf{Y},0} = 1, \\ \mu_{\mathbf{Y},1} = \dfrac{\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} \cdot \mu_{\mathbf{A}}, \\ \mu_{\mathbf{Y},2} = \dfrac{\mu_{\mathbf{X},2}}{1-\mu_{\mathbf{X},2}} \cdot \left(2\mu_{\mathbf{A}}^2 \dfrac{\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} + \mu_{\mathbf{A},2} \right), \\ \mu_{\mathbf{Y},3} = \dfrac{\mu_{\mathbf{X},3}}{1-\mu_{\mathbf{X},3}} \cdot \left(3\mu_{\mathbf{A},2}\mu_{\mathbf{A}} \cdot \dfrac{\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} + 3\mu_{\mathbf{A}} \dfrac{\mu_{\mathbf{X},2}}{1-\mu_{\mathbf{X},2}} \cdot \left(2\mu_{\mathbf{A}}^2 \dfrac{\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} + \mu_{\mathbf{A},2} \right) + \mu_{\mathbf{A},3} \right). \end{cases} \quad (12)$$

A closed-form general expression is probably relatively easily developed.

Using (6), the variance of the infinite sequence in terms of the means and variances of the payments and discount factor will be:

$$\sigma_{\mathbf{Y}}^2 = \sigma_{\mathbf{A}}^2 \frac{\sigma_{\mathbf{X}}^2 + \mu_{\mathbf{X}}^2}{\left[1-\sigma_{\mathbf{X}}^2 - \mu_{\mathbf{X}}^2\right]} + \mu_{\mathbf{A}}^2 \left(\frac{\sigma_{\mathbf{X}}^2 + \mu_{\mathbf{X}}^2}{\left[1-\sigma_{\mathbf{X}}^2 - \mu_{\mathbf{X}}^2\right]} \left(\frac{1+\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} \right) - \left(\frac{\mu_{\mathbf{X}}}{1-\mu_{\mathbf{X}}} \right)^2 \right). \quad (13)$$

This expression is more complex than one would expect.

3 THE DISTRIBUTION OF THE DISCOUNT FACTOR

The distribution of the individual discount factor $\mathbf{X} = e^{-\rho\boldsymbol{\tau}}$ is a consequence straight from the distribution of the time interval $\boldsymbol{\tau}$.

The probability that $\mathbf{X}$ is on or below x, $x \geq 0$, i.e. $\Pr\{\mathbf{X} \leq x\} = \Pr\{e^{-\rho\boldsymbol{\tau}} \leq x\} = F_{\mathbf{X}}(x)$, which is the distribution function of $\mathbf{X}$. So if the pdf of $\boldsymbol{\tau}$, $f_{\boldsymbol{\tau}}(\tau)$, is continuous, we have

$$F_{\mathbf{X}}(x) = \begin{cases} 1 - F_{\boldsymbol{\tau}}\left((-\ln x)/\rho\right), & x \leq 1, \\ 1, & x > 1, \end{cases}, \quad f_{\mathbf{X}}(x) = \begin{cases} f_{\boldsymbol{\tau}}\left((-\ln x)/\rho\right)/(\rho x), & x \leq 1, \\ 0, & x > 1. \end{cases} \tag{14}$$

Three examples of distributions of $\boldsymbol{\tau}$ are listed in Table 1 together with the corresponding $f_{\mathbf{X}}(x)$ and moments.

Table 1: Three examples of time interval distributions and consequences for the pdf of the discount factor and its moments. The symbol δ refers to Dirac's delta function.

	Exponential (Poisson)	***Rectangular***	***Deterministic***
$f_{\boldsymbol{\tau}}(\tau)$	$\lambda e^{-\lambda\tau},\ \tau \geq 0$	$\left(\hat{\tau}_2 - \hat{\tau}_1\right)^{-1},\ 0 \leq \hat{\tau}_1 \leq \tau \leq \hat{\tau}_2$	$\delta(\tau - \hat{\tau}),\ \hat{\tau} \geq 0$
$\tilde{f}_{\boldsymbol{\tau}}(s)$	$\lambda/(\lambda + s)$	$\left(e^{-s\hat{\tau}_2} - e^{-s\hat{\tau}_1}\right)/\left(s(\hat{\tau}_2 - \hat{\tau}_1)\right)$	$e^{-s\hat{\tau}},\ \hat{\tau} \geq 0$
$f_{\mathbf{X}}(x)$	$(\lambda/\rho)x^{(\lambda/\rho)-1},\ x \leq 1$	$\left(\rho(\hat{\tau}_2 - \hat{\tau}_1)x\right)^{-1},\ e^{-\rho\hat{\tau}_2} \leq x \leq e^{-\rho\hat{\tau}_1}$	$\delta\left(x - e^{-\rho\hat{\tau}}\right)$
$\mu_{\mathbf{X},m}$	$\lambda/(\lambda + m\rho)$	$\left(e^{-m\rho\hat{\tau}_1} - e^{-m\rho\hat{\tau}_2}\right)/\left(m\rho(\hat{\tau}_2 - \hat{\tau}_1)\right)$	$e^{-m\rho\hat{\tau}}$
$\mu_{\mathbf{X}} = \mu_{\mathbf{X},1}$	$\lambda/(\lambda + \rho)$	$\left(e^{-\rho\hat{\tau}_1} - e^{-\rho\hat{\tau}_2}\right)/\left(\rho(\hat{\tau}_2 - \hat{\tau}_1)\right)$	$e^{-\rho\hat{\tau}}$
$\sigma_{\mathbf{X}}^2 = \mu'_{\mathbf{X},2}$	$\dfrac{(\lambda/\rho)}{((\lambda/\rho)+2)((\lambda/\rho)+1)^2}$	$\dfrac{\left(\left(e^{-2\rho\hat{\tau}_1} - e^{-2\rho\hat{\tau}_2}\right)(\rho(\hat{\tau}_2 - \hat{\tau}_1)) - 2\left(e^{-\rho\hat{\tau}_1} - e^{-\rho\hat{\tau}_2}\right)^2\right)}{2\rho^2(\hat{\tau}_2 - \hat{\tau}_1)^2}$	0

Alternatively, one may find all moments directly from the transform $\tilde{f}_{\boldsymbol{\tau}}(s)$:

$$\mathrm{E}\left[\left(e^{-\rho\boldsymbol{\tau}}\right)^m\right] = \mathrm{E}\left[e^{-m\rho\boldsymbol{\tau}}\right] = \int_{\tau=0}^{\infty} f_{\boldsymbol{\tau}}(\tau) e^{-m\rho\tau} d\tau = \tilde{f}_{\boldsymbol{\tau}}(m\rho). \tag{15}$$

4 THE DISTRIBUTION OF THE NPV

Knowing all moments of a distribution is close to knowing the distribution itself. A Maclaurin expansion of the Laplace transform of the pdf $f_{\mathbf{Y}_n}(y)$ yields

$$\tilde{f}_{\mathbf{Y}_n}(s) = \int_{y=0}^{\infty} f_{\mathbf{Y}_n}(y) e^{-sy} dy = \sum_{i=0}^{\infty} \int_{y=0}^{\infty} f_{\mathbf{Y}_n}(y) \frac{(-sy)^i}{i!} dy =$$

$$= \sum_{i=0}^{\infty} \frac{(-s)^i}{i!} \int_{y=0}^{\infty} f_{\mathbf{Y}_n}(y) y^i dy = \sum_{i=0}^{\infty} \frac{\mu_{\mathbf{Y}_n,i} (-1)^i s^i}{i!}. \tag{16}$$

So the ith Maclaurin coefficient of $\tilde{f}_{\mathbf{Y}_n}(s)$ is the ith moment $\mu_{\mathbf{Y}_n,i}$ multiplied by $(-1)^i / i!$ proving Eq.(5), and for a wide class of distributions the moments will define the distribution uniquely [3, p.230]. Here (and below), we have limited our treatment to cases when $\mathbf{Y}_n$ is non-negative, although generally the $\mathbf{A}_i$ might be of either sign. Generalisations in this respect are left to future research.

From (7), knowing that $\mathbf{Y}_1 = \mathbf{A}_1\mathbf{X}_1$, we have

$$\mathbf{Y}_n = \sum_{j=1}^{n} \mathbf{A}_j \prod_{i=1}^{j} \mathbf{X}_i \ , \tag{17}$$

so the general form of the distribution of $\mathbf{Y}_n$ would be given by

$$F_{\mathbf{Y}_n}(y) = \Pr\left\{ \sum_{j=1}^{n} \mathbf{A}_j \prod_{i=1}^{j} \mathbf{X}_i \le y \right\} = \int\int\cdots\int_{\sum_{j=1}^{n} a_j \prod_{i=1}^{j} x_i \le y} \prod_{j=1}^{n} f_{\mathbf{A}_j}(a_j) \prod_{j=1}^{j} f_{X_i}(x_i) \prod_{l=1}^{n} da_l \prod_{k=1}^{n} dx_k \ . \tag{18}$$

5 RISK PREFERENCE CONSIDERATIONS

In risk preference theory (decision theory), one considers a basic lottery ticket with one major positive and one major negative outcome, say $\underline{Z}$ and $\overline{Z}$. The probability for receiving $\overline{Z}$ is π and for receiving $\underline{Z}$ therefore $(1-\pi)$. A rational person prefers to own a ticket with a higher value of π and has the ability to find a value of π making the ticket equally preferable to any given other certain outcome *Z*, with $\underline{Z} \le Z \le \overline{Z}$. So equating the preference of a ticket π with any relevant value of *Z*, a continuous and monotonically increasing risk preference (utility) function $\pi(Z)$ is defined. If $\pi(Z)$ is strictly concave in a certain region of *Z*, this defines risk aversion, in a strictly convex region the person is a risk taker, and in a linear region the person is risk neutral. A typical assumption is that $\pi(Z)$ is concave for lower values of *Z* and convex for higher.

We assume a risky project with several stochastic outcomes $z_1, z_2, ..., z_N$, the probabilities of which are known to be $p_1, p_2, ..., p_N$. Owning this project would be equally preferred to receiving lottery tickets $\pi(z_1), \pi(z_2), ..., \pi(z_N)$ with the same probabilities, in which case the overall probability for receiving $\overline{Z}$ would be $\sum_{i=1}^{n} p_i \pi(z_i)$ and for receiving $\underline{Z}$ would be $1 - \sum_{i=1}^{n} p_i \pi(z_i)$. But this corresponds to a certainty equivalent outcome $\pi^{-1}\left(\sum_{i=1}^{n} p_i \pi(z_i) \right)$, where π^{-1} is the inverse of the risk preference function. The Arrow-Pratt measure of absolute risk aversion [1, 14] is defined as $\gamma(z) = -\dfrac{d^2\pi(z)}{dz^2} \Big/ \dfrac{d\pi(z)}{dz}$. A person with constant absolute risk aversion will thus have the risk preference function $\pi(z) = \beta - e^{\alpha - \gamma z} / \gamma$, where α and β are constants of integration and the inverse risk preference function $\pi^{-1}(p) = (\alpha - \ln \gamma(\beta - p))/\gamma$, for any $0 \le p \le 1$.

Applying this reasoning to the NPV of a sequence of stochastic payments the certainty equivalent of $\mathbf{Y}_n$ as defined above will be:

$$\text{NPV}_{\text{certainty equivalent}} = \pi^{-1}\left(\int_{y=-\infty}^{\infty} f_{\mathbf{Y}_n}(y)\pi(y)\,dy \right). \tag{19}$$

Comparing certainty equivalents for different stochastic payment sequences will thus provide a coherent procedure for their evaluation.

By way of example, assuming a simple risk preference function with constant absolute risk aversion $\pi(z) = 1 - e^{-\gamma z}$, i.e. $\alpha = \ln\gamma$ and $\beta = 1$, and that *y* in (19) is non-negative, we have

$\pi^{-1}(p) = \left(-\ln(1-p)\right)/\gamma$ and $\int_{y=0}^{\infty} f_{\mathbf{Y}_n}(y)\pi(y)dy = \int_{y=0}^{\infty} f_{\mathbf{Y}_n}(y)\left(1-e^{-\gamma y}\right)dy = 1-\tilde{f}_{\mathbf{Y}_n}(\gamma)$.

Thus

$$\mathrm{NPV}_{\text{certainty equivalent}} = -\frac{\ln\left(\tilde{f}_{\mathbf{Y}_n}(\gamma)\right)}{\gamma}. \tag{20}$$

In (16) the transform $\tilde{f}_{\mathbf{Y}_n}(s)$ is given by its moments, so the right-hand member of (20) is evaluated exchanging *s* for γ in (16) and taking the moments from (8)–(11). Eq. (20) is easily shown to hold for any risk preference function with constant absolute risk aversion.

References (selected)

[1] Arrow, K. J. (1965). Aspects of the Theory of Risk Bearing. The Theory of Risk Aversion. Helsinki: Yrjo Jahnssonin Saatio.

[2] Barnes, J.W., Zinn, C.D. and Eldred S., (1978). A Methodology for obtaining the probability density function of the present worth of probabilistic cash flow profiles. AIIE Transactions, 10(3), 226-236.

[3] Feller, W., (1966). An Introduction to Probability theory and its Applications, Vol. II. New York: John Wiley and Sons.

[4] Grubbström, R.W. (1967). On the Application of the Laplace Transform to Certain Economic Problems. Management Science, 13(7), 558-567.

[5] Grubbström, R.W. (1980). A principle for determining the correct capital costs of inventory and work-in-progress. International Journal of Production Research, 18(2), 259-271.

[6] Grubbström, R.W. (2010). The Newsboy problem when customer demand is a compound renewal process. European Journal of Operational Research, 203(1), 134-142.

[7] Grubbström, R.W. and Kingsman, B.G. (2004). Ordering and Inventory Policies for Step Changes in the Unit Item Cost: A Discounted Cash Flow Approach. Management Science, 50(2), 253-267.

[8] Grubbström, R.W. and Tang, O. (2006). The moments and central moments of a compound distribution. European Journal of Operational Research, 170(1), 106-119.

[9] Hillier, F.S. (1963). The Derivation of Probabilistic Information for the Evaluation of Risky Investments, Management Science, 9(3), 443-457.

[10] Hillier, F. S. (1965). Supplement to the Derivation of Probabilistic Information for the Evaluation of Risky Investments, Management Science 11(3), 485-487.

[11] Jang, J.-W. (2004). Martingale Approach for Moments of Discounted Aggregate Claims. The Journal of Risk and Insurance, 71(2), 201-211.

[12] Perrakis, S. and Henin, C. (1974). The Evaluation of Risky Investments with Random Timing of Cash Returns, Management Science, 21(1), 79-86.

[13] Perrakis, S. and Sahin, I. (1976). On risky investments with random timing of cash returns and fixed planning horizon. Management Science, 22(7), 799-809.

[14] Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. Econometrica, 32(1–2), 122–136.

[15] Reisman, A. and Rao, A.K. (1972). Stochastic Cash Flow Formulae Under Conditions of Inflation, The Engineering Economist, 18(1), 49-69.

NEAR-OPTIMALITY OF MYOPIC POLICY FOR THE DUAL SOURCING INVENTORY PROBLEM WITH STOCHASTIC CAPACITATED SUPPLIER

Marko Jakšič
Faculty of Economics, University of Ljubljana, Kardeljeva ploščad 17, Ljubljana, Slovenia
marko.jaksic@ef.uni-lj.si

Abstract: We model a periodic review inventory system with a stationary stochastic demand, in which a manufacturer is procuring a component from two available supply sources. The faster supply source is assumed to be stochastic capacitated with immediate delivery, while the slower supply source is modelled as uncapacitated with a longer fixed lead time. The objective of a manufacturer is to choose how the order should be split between the two supply sources in each period, where the slower supply source is used to compensate for the supply capacity unavailability of the faster supply source. We derive the optimal dynamic programming formulation that minimizes the total expected inventory holding and backorder costs over a finite planning horizon. We show that the optimal policy is relatively complex, therefore we study the myopic policy, which can be characterized as the two-level base-stock policy. We derive some additional properties of the myopic policy and show that it provides a near-perfect estimate of the optimal costs. More specifically, we shows that the accuracy is lower when both the demand uncertainty and the supply capacity uncertainty are high, as expected.

Keywords: inventory, dual sourcing, stochastic models, myopic policy, uncertain capacity.

1. INTRODUCTION

In the age of agile supply chains the two main determinants of the customer service level are the speed of replenishment and its reliability. To guarantee the customer satisfaction the companies are seeking for a supply base that would enable them to pursue these two goals. It is often the case, that a supplier might offer fast delivery while its reliability will suffer occasionally. This has forced the companies to search for alternative supply channels, through which they would improve the supply process reliability, where often more reliable supply comes with the price, either in higher purchasing costs per unit of a product or longer replenishment lead time.

A number of procurement options nowadays is growing, particularly due to increased production outsourcing, but these may differ substantially in terms of flexibility, availability and reliability. While the companies can now rely on a spot market for flexibility and possibly favourable market price, the risk of insufficient product availability is too high and it needs to be compensated through utilization of a more reliable supply channel. To hedge this risk companies can rely to long term contractual agreements with their alternative suppliers, which can guarantee a reliable delivery.

In this paper we address the problem of a manufacturer procuring a component used for production of a finished product, where the two sourcing options are considered: a fast capacitated supplier and a reliable supplier with the longer lead time. When the manufacturer anticipates the supply shortage at the faster supplier, the decision has to be made to what extent the slower supplier should be utilized. In each period the manufacturer has to decide about the quantities to procure from the two supply channels. Due to the production delay both decisions are made before the demand for the finished product is realized, thus the decision maker is faced with both, the uncertain replenishment from the faster supplier and the uncertainty of demand.

We proceed with a review of the relevant literature. The supply process of the faster supplier is modelled as in [2, 8, 5, 6] and [7], where the random supply/production capacity determines

a random upper bound on the supply availability in each period. For a finite horizon stationary inventory model it holds that the optimal policy is of base-stock type, where the optimal base-stock level is increased to account for possible, albeit uncertain, capacity shortfalls in future periods.

The search for the best dual-sourcing strategy in the literature revolves around the dilemma of when to use a faster and necessarily more expensive supplier to compensate for a slow response by a cheaper supplier. Several papers discuss the setting in which the lead times of the two suppliers differ by a fixed number of periods [4, 1, 9]. More specifically, [4] shows the optimality of the two-level base-stock policy for the case where the two suppliers' lead times differ by one period (so-called consecutive lead times). The policy instructs that first the order with the fast supply source is placed so that the inventory position is raised to the first base-stock level, and then the slow supply source is used to raise the inventory position to the second base-stock level. For general, nonconsecutive, lead times, [9] found the optimal policy to be quite complex and lacking structure.

The dual sourcing papers that assume either one or both suppliers are capacitated all limit themselves to having fixed capacity limits on an individual order or on cumulative order with the supplier [3, 12, 10, 11]. They generally show that the two-level base-stock policy is optimal when leadtimes differ for one period. While the modeling perspective of the papers referred to above is similar to ours, they all still primarily focus on exploring the trade-off between the price difference and responsiveness. To our knowledge we are the first to study the trade-off between stochastic supply capacity availability and responsiveness.

The remainder of the paper is organized as follows. We present the model formulation and derive the optimal cost formulation in Section 2. In Section 3, the myopic dual sourcing policy is introduced, the structure of the policy is given and the analysis of the differences between the optimal and myopic policy is presented. We summarize our findings in Section 4.

2. MODEL FORMULATION

In this section, we give the notation and the model description. The faster, zero-lead-time, supply source is stochastic capacitated where the supply capacity is exogenous to the manufacturer and the actual capacity realization is only revealed upon replenishment. The slower supply source is modeled as uncapacitated with a fixed one period lead time. The demand and supply capacity of the faster supply source are assumed to be stochastic non-stationary with known distributions in each time period, although independent from period to period. In each period, the customer places an order with either an unreliable, or a reliable supply mode, or both.

Presuming that unmet demand is fully backordered, the goal is to find the optimal policy that would minimize the inventory holding costs and backorder costs over finite planning horizon T. We intentionally do not consider any product unit price difference and fixed ordering costs as we are chiefly interested in studying the trade-off between the capacity uncertainty associated with ordering from a faster supply source and the delay in the replenishment from a slower source. The notation used throughout the paper is summarized in Table 1 and some is introduced when needed.

We assume the following sequence of events. (1) At the start of the period, the manager reviews the inventory position before ordering x_t, where $x_t = \tilde{x}_t + v_{t-1}$ is the sum of the on-hand stock $\tilde{x}_t$ and the order v_{t-1} to a slower supply source made in the previous period. (2) Order z_t to a faster supply source and order v_t to a slower supply source are placed. For the purpose of the subsequent analysis, we define two inventory positions after the order placement. First, after placing order z_t the inventory position is raised to y_t, $y_t = x_t + z_t$, and order v_t is placed to w_t, $w_t = x_t + z_t + v_t$. Observe that it makes no difference in which sequence the orders are actually placed as long as both are placed before the current period's capacity of

Table 1: Summary of the notation.

T	:	number of periods in the finite planning horizon
c_h	:	inventory holding cost per unit per period
c_b	:	backorder cost per unit per period
α	:	discount factor ($0 \leq \alpha \leq 1$)
$\tilde{x}_t$	:	on-hand inventory before ordering in period t
x_t	:	inventory position before ordering in period t
y_t	:	inventory position after ordering from a faster capacitated supply source in period t
w_t	:	inventory position after ordering from a slower uncapacitated supply source in period t
z_t	:	order placed with the faster supply source in period t
v_t	:	order placed with the slower supply source in period t
d_t, D_t	:	actual realization and random variable denoting demand in period t
q_t, Q_t	:	actual realization and random variable denoting the available supply capacity of the faster capacitated supply source in period t

the fast supply source $q-t$ and demand d_t are revealed. (3) The order with the slower supply source from the previous period v_{t-1} and the current period's order z_t are replenished. The inventory position can now be corrected according to the actual supply capacity realization $w_t - (z_t - q_t)^+ = x_t + min(z_t, q_t) + v_t$, where $(z_t - q_t)^+ = \max(z_t - q_t, 0)$. (4) At the end of the period, demand d_t is observed and satisfied through on-hand inventory; otherwise, it is backordered. Inventory holding and backorder costs are incurred based on the end-of-period on-hand inventory, $\tilde{x}_{t+1} = y_t - (z_t - q_t)^+ - d_t$. Correspondingly, the expected single-period cost function is defined as $C_t(y_t, z_t) = \alpha E_{Q_t,D_t}\tilde{C}_t(\tilde{x}_{t+1}) = \alpha E_{Q_t,D_t}\tilde{C}_t(y_t - (z_t - Q_t)^+ - D_t)$, where $\tilde{C}_t(\tilde{x}_{t+1}) = c_h(\tilde{x}_{t+1})^+ + c_b(-\tilde{x}_{t+1})^+$. The minimal discounted expected cost function that optimizes the cost over a finite planning horizon T from period t onward, starting in the initial state x_t, can be written as:

$$f_t(x_t) = \min_{x_t \leq y_t \leq w_t} \left\{ C_t(y_t, z_t) + \alpha E_{Q_t,D_t} f_{t+1}(w_t - (z_t - Q_t)^+ - D_t) \right\}, \text{ for } 1 \leq t \leq T \qquad (1)$$

and the ending condition is defined as $f_{T+1}(\cdot) \equiv 0$.

3. NEAR-OPTIMAL MYOPIC POLICY

In this section, we review the characteristics of the optimal policy, where the order with the faster supply source depends on the inventory position before ordering, while the order with the slower supply source is placed up to a state-dependent base-stock level. We continue by introducing the myopic policy where the optimal orders are the solutions to the extended single period problem, and show the properties of the two myopic base-stock levels. We conduct the numerical analysis to show that the costs of the myopic policy provide a very accurate estimate of the optimal costs.

As single-period costs C_t in period t are not influenced by order v_t, we can rewrite (1) in the following way:

$$f_t(x_t) = \min_{x_t \leq y_t} \left\{ C_t(y_t, z_t) + \min_{y_t \leq w_t} \alpha E_{Q_t,D_t} f_{t+1}(w_t - (z_t - Q_t)^+ - D_t) \right\}, \text{ for } 1 \leq t \leq T, \quad (2)$$

which now enables us to introduce auxiliary cost functions $J_t(y_t, z_t)$ and $H_t(w_t, z_t)$:

$$J_t(y_t, z_t) = C_t(y_t, z_t) + \min_{y_t \leq w_t} \alpha E_{Q_t,D_t} f_{t+1}(w_t - (z_t - Q_t)^+ - D_t), \text{ for } 1 \leq t \leq T \qquad (3)$$

$$H_t(w_t, z_t) = \alpha E_{Q_t,D_t} f_{t+1}(w_t - (z_t - Q_t)^+ - D_t), \text{ for } 1 \leq t \leq T \qquad (4)$$

In the literature review, we refer to a series of papers studying the dual sourcing inventory problem with consecutive lead times. The two-level base-stock policy characterizes the structure of the optimal policy in all of them, both in the case of uncapacitated supply sources and in the case where one or both supply source exhibit the fixed capacity limit. However, by studying the convexity properties of the cost functions given in (2), (3) and (4), it can be shown that these are not convex in general. The optimal inventory position after ordering with the faster supplier is not independent of the inventory position before ordering x_t, and therefore cannot be characterized as the optimal base-stock level.

Next, we introduce the notion of the myopic policy, which optimizes the so-called extended single period problem in every period. As the model under consideration assumes different lead times for the faster and the slower supplier, the extended period covers the time interval in which both ordering decisions are made that directly influence the single period costs C_t in period t. It effectively starts by placing an order v_{t-1} with the slower supplier in period $t-1$, followed by placing an order z_t with the faster supplier in period t. The ability to cover the demand D_t at the end of period t depends on the replenishment of the two orders and the realization of the demand. Observe also that with the order v_t we cannot influence the costs in period t.

The myopic equivalent to the optimal cost function as given in (2) can therefore be written in the following way:

$$f_t^M(x_t) = \min_{x_t \leq y_t} C_t(y_t, z_t) + \min_{y_t \leq w_t} \alpha E_{Q_t, D_t} f_{t+1}(w_t - (z_t - Q_t)^+ - D_t), \text{ for } 1 \leq t \leq T, \qquad (5)$$

where the search for the optimal y_t resorts to minimizing the single period cost function C_t.

In the following proposition, we show that the myopic policy can be characterized as the two-level base-stock policy, where the optimal order $\hat{z}_t$ is placed up to the first base-stock level $\hat{y}^M$, and the optimal order $\hat{v}_t$ is placed up to the second, state-dependent, base-stock level $\hat{w}(\hat{z})$. For clarity reasons we omit the subscript t in parts of the following text.

Proposition 3.1 *Under stationary stochastic demand and supply capacity, the following results hold for all t:*

1. *$f_t^M(x_t)$ is convex in x_t.*
2. *The optimal myopic policy is the two-level base-stock policy with the constant base-stock levels $\hat{y}^M$ and $\hat{w}^M(\hat{z})$, where:*
3. *$\hat{y}^M = \hat{y_t}^M$ is the solution to $\min_{x_t \leq y_t} C_t(y_t, z_t)$: $\hat{y}^M = G^{-1}\left(\frac{c_b}{c_b + c_h}\right)$, and*
4. *$\hat{w}^M(\hat{z}) = \hat{w_{t-1}}^M(\hat{z}_{t-1})$ is the solution to $\min_{y_{t-1} \leq w_{t-1}} \alpha E_{Q_{t-1}, D_{t-1}} \min_{x_t \leq y_t} C_t(y_t, z_t)$.*

Proof. We prove Part 1 by regular inductive argument on t. It is clear that Part 1 holds for $T+1$ due to $f_{T+1}(\cdot) \equiv 0$. Suppose that Part 1 also holds for $t+1$. We need to show that it also holds for t. We need to show that both parts of (5) are convex. Note that the single-period cost function $C_t(y_t, z_t)$ is convex in y_t and quasiconvex in z_t for any t as shown in [2]. Partially differentiating C_t with respect to z_t and setting it to zero proves Part 3. The second partial derivative of C_t with respect to y_t evaluated at the optimal $\hat{y}^M$ is nonnegative, which proves the convexity. The second part of (5) holds due to the fact that taking expectations over Q_t and D_t preserves the convexity of f_{t+1}, and evaluating the second partial derivative of $H_t(w_t, z_t)$ with respect to w_t and applying the first-order optimality condition for w_t, shows that minimization over w_t also preserves convexity. Part 4 comes directly from the definition of the myopic dual sourcing policy.

To prove Part 2, we need to show that the myopic policy behaves as a two-level base-stock policy. By differentiating the first-order optimality condition for $H_t(w_t, z_t)$ with respect to y_t,

we see that $d\hat{w}_t(z_t)/dy_t = 0$ satisfies the equation, which is a definition of the base-stock policy. Similarly, differentiating the first-order optimality condition for $J_t(y_t, z_t)$ with respect to x_t, we see that $d\hat{z}_t/dx_t = -1$ needs to hold, which proves Part 2. □

For the capacitated single supplier model, [2] show that while $\hat{y}^M$ optimizes a single period problem, it is far from optimal in a multiperiod setting. However, we show that in the dual sourcing model under consideration the appropriate combination of the two myopic base-stock levels provides a very good substitute for the optimal base-stock levels. In Table 2, we present the optimal and the myopic inventory positions. Numbers in bold are used to denote cases, where the myopic inventory positions and orders differ from the optimal ones. The results for $\hat{y}(x)$ confirm that $\hat{y}$ is a function of x, therefore the optimal order $\hat{z}$ is not placed in the manner of a base-stock policy. We see that with increasing x, $\hat{y}(x)$ is increasing, approaching the myopic $\hat{y}^M$. For lower x, the optimal policy suggests that it is not optimal to constantly increase $\hat{z}$ by ordering up $\hat{y}^M$. This can be attributed to the increased uncertainty about the replenishment of $\hat{z}$. To compensate for this, the decision maker should rely more on the order to the slower supplier by increasing $\hat{v}$ above $\hat{v}^M$.

Table 2: The optimal and the myopic inventory positions and orders.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
$\hat{y}(x)$	14	14	14	14	15	15	15	15	15	15	15	15	15	16	16	16	16							
$\hat{y}^M$	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	16	16	16	16							
$\hat{z}$	14	13	12	11	11	10	9	8	7	6	5	4	3	3	2	1	0							
$\hat{z}^M$	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	3	2	1	0							
$\hat{w}(\hat{z})$	29	28	27	27	27	26	25	25	25	24	24	24	23	23	23	23	23	23	23	23	23	23	23	23
$\hat{w}^M(\hat{z}^M)$	**31**	**30**	**29**	**28**	27	**27**	**26**	25	25	**25**	24	24	**24**	23	23	23	23	23	23	23	23	23	23	23
$\hat{v}(\hat{z})$	15	14	13	13	12	11	10	10	10	9	9	9	8	7	7	7	7	6	5	4	3	2	1	0
$\hat{v}^M(\hat{z}^M)$	15	14	13	**12**	**11**	11	10	**9**	**9**	9	**8**	**8**	8	7	7	7	7	6	5	4	3	2	1	0

Despite the fact that the myopic base-stock levels generally differ from the optimal inventory positions, we now proceed to show that the costs of the myopic policy provide a nearly perfect estimate of the optimal costs. In Figure 1, we provide the histogram of the relative differences in costs of the optimal and myopic policy. The accuracy of the myopic policy is tested on 300 scenarios, where in almost 60% of the scenarios the costs of the myopic policy are equal to the optimal costs. Only in 4 cases, myopic costs differed for around 0.3%, which is also the highest cost difference observed. A careful study of suboptimal scenarios has not revealed a clear pattern that would point out the characteristics of these scenarios. In general, the accuracy is lower when both the demand uncertainty and the supply capacity uncertainty are high, which is not unexpected.

4. CONCLUSIONS

In this paper, we analyze a sourcing situation in which both a faster stochastic capacitated supply source and a reliable slower supply source may be used. Accordingly, we model a finite horizon dual-sourcing inventory model with a stochastic capacitated, zero-lead-time supplier and an uncapacitated supplier with the lead time of one period. We derive the dynamic programming formulation for the optimal inventory holding and backorder costs, where we show that the structure of the optimal inventory policy is relatively complex as the optimal cost function is not convex in general. Thus, the optimal inventory position after ordering with the faster supplier is a function of the inventory position before ordering and thus the policy does not behave as a base-stock policy. We study the myopic policy, where the optimal inventory position after ordering with the faster supplier corresponds to the solution to a single

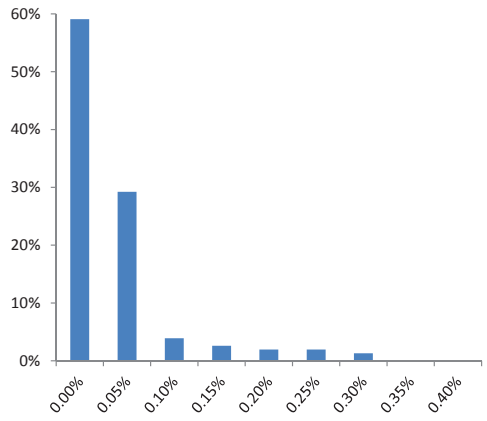


Figure 1: The relative difference in costs between the optimal and the myopic policy.

period uncapacitated inventory problem. We show that by carefully balancing the two base-stock levels, the costs of the myopic policy provide a very good approximation for the optimal costs.

References

[1] Bulinskaya, E. (1964). Some results concerning optimal inventory policies. Theory Prob. Appli., Vol(9): 502–507.

[2] Ciarallo, F.W., Akella, R. and Morton, T.E. (1994). A periodic review, production planning model with uncertain capacity and uncertain demand - optimality of extended myopic policies. Management Science, Vol(40): 320–332.

[3] Daniel, K. (1963). A delivery-lag inventory model with emergency. Multistage inventory models and techniwues. Stanford University Press, Stanford, CA.

[4] Fukuda, Y. (1964). Optimal policies for the inventory problem with negotiable leadtime. Management Science, Vol(10(4)): 690–708.

[5] Iida, T. (2002). A non-stationary periodic review production-inventory model with uncertain production capacity and uncertain demand. European Journal of Operational Research, Vol(140): 670–683.

[6] Jakšič, M., Fransoo, J., Tan, T., de Kok, A.G. and Rusjan, B. (2011). Inventory management with advance capacity information. Naval Research Logistics, Vol(58): 355–369.

[7] Jakšič, M. and Fransoo, J. (2015). Optimal inventory management with supply backordering. International Journal of Production Economics, Vol(58): 254–264.

[8] Khang, D.B. and Fujiwara, O. (2000). Optimality of myopic ordering policies for inventory model with stochastic supply. Operations Research, Vol(48): 355–369.

[9] Whittemore, A.S. and Saunders, S.C. (1977). Optimal inventory under stochastic demand with two supply options. SIAM Journal on Applied Mathematics, Vol(32(2)): 293–305.

[10] Xu, N. (2011). Myopic optimal policy for a multi-period, two-delivery-lead-times, stochastic inventory problem with minimum cumulative commitment and capacity. International Journal of Production Economics, Vol(133): 719–727.

[11] Yang, J. and Qi, X. (2005). A production-inventory system with Markovian capacity and outsourcing option. Operations Research, Vol(53): 328–349.

[12] Yazlali, O. and Erhun, F. (2009). Dual-supply inventory problem with capacity limits on order sizes and unrestricted ordering costs. IIE Transactions, Vol(41(8)): 716–729.

INVENTORY ROUTING AND CONTROL MODELS FOR THE DISTRIBUTION OF INDUSTRIAL GASES: A MULTIPERIOD DETERMINISTIC AND A DYNAMICAL PREDICTIVE APPROACH

Athanassios Nikolakopoulos
National Technical University of Athens, Department of Chemical Engineering
Heroon Polytechneiou 9, 15780, Greece
nikolako@mail.ntua.gr

Ioannis Ganas
Technological Educational Institute of Epirus
Department of Accounting and Finance, Psathaki, 481 00 Preveza, Greece
ganas@teiep.gr

Abstract: This work proposes a multi-period deterministic and a dynamical predictive model as new decision support tools for Inventory in process Routing and Control in the distribution of industrial gases via truck deliveries. The first is a mixed integer linear model and describes the problem of the optimal selection of the customer replenishment schedule giving priority to the minimization of the fleet size, and backhauling of empty gas tanks. The dynamic version is a mixed integer model with an exact linearization of the non-linear objective function terms, suitable for a predictive approach of the inventory control problem.

Keywords: Inventory Routing, Industrial Gas Distribution, Mathematical Modelling, Inventory Control, Mixed Integer Programming, Model Predictive Control

1 INTRODUCTION

The complete life cycle of a product includes scheduling, production planning, manufacturing, inventory management and distribution planning. The latter two prove to be of particular importance due to the economic benefits they offer to the overall process. The integration of forward and reverse flows of products and materials offers additional profit. Reverse logistics deal with the management of return flows of products and materials in commercial and industrial processes and offer economic and environmental benefits. In Vendor managed resupply the supplier manages the inventory replenishment of the customers. Using inventory routing [3], vendors can offer attractive discounts by saving on distribution cost and are able to better coordinate deliveries to different customers. Supply chain managers also wish to balance profit maximization with risk minimization. One way to minimize risk is to hold some buffer stock so that the supply chain can respond to rush orders or demand spikes. This work first presents a multi-period deterministic model for the distribution of industrial gas with small tanks that captures all attributes of reverse logistics and inventory routing. The single period version of the deterministic model of has been presented in [4, 5] and it contained continuous variables for the truck loads. In this work the problem is expanded to multiple periods scheduling, where the truck loads are represented by integer variables. Also more effective sub-tour elimination constraints replace the ones in the single period models [4, 5], by demanding smooth flow of commodities between nodes and consistency with the vehicle capacity. The system is modelled in Section 2 as a multi-period Inventory Routing problem with Simultaneous Pick-up and Delivery (IRPSPD) to account for the reverse logistics of empty gas tanks and multiple period scheduling. Section 3 describes the discrete time dynamical version of the model, which is utilized by the MPC strategy proposed for solving the inventory control problem, where future demands are considered as the disturbance to the system. The dynamical model balances profit maximization versus risk minimization. Section 4 illustrates the methodology using an example of stabilizing an inventory routing system via

the proposed predictive strategy. The paper concludes with a discussion of the results and directions for future research.

2 DEFINITION OF THE MULTI-PERIOD IRPSPD

In the multi-period *IRPSPD* a homogeneous fleet V with $K_{avl} = |V|$ available vehicles of equal capacity Q must be serve a set N of n customers, $N = \{1, 2\ldots, n\}$. The planning horizon considers T periods, $t = 1, \ldots, T$. There is a demand $d_{i,t}$ for every customer $i \in N$. Each customer i maintains its own inventory up to capacity C_i and incurs inventory carrying cost of h_i per period per unit. The complete directed graph induced by the customers is $G = (N^+, A)$, where $N^+ = N \cup \{0, n+1\}$ is the set of vertices. The depot is represented by vertex 0 and $n + 1$. A is the set of arcs connecting any pair of vertices, while no arc terminates in vertex 0, and no arc originates from vertex $n+1$. A travelling cost $c_{i,j}$ is assigned to each arc $(i, j) \in A$ *and* $c_{\max} = \max\left(c_{i,j}\right)$. Zero delivery demands and pickup demands are assigned to nodes $\{0, n+1\}$, that is, $d_0 = d_{n+1} = p_0 = p_{n+1} = 0$. The primary objective is the minimization of the required number of vehicles. The secondary objective is to minimize the overall transportation and inventory carrying cost incurred over a specific planning horizon. Each customer is serviced at most once in each period t, ($p_i, d_i \le Q$, $\forall\, i \in N$), no split delivery is permitted, capacity constraints must be satisfied, and no backlogging is permitted.

2.1 Mixed Integer Programming formulation for the multi-period IRPSPD

The next four sets of variables are considered in the problem:

$$z_{k,i,,j,t} = \begin{cases} 1 \text{ if vehicle } k \text{ drives from } i \text{ to } j \text{ in period } t \\ 0 \text{ otherwise} \end{cases}, \quad (i,j) \in \text{A}, k \in \text{V and } t = 1, ..., T$$

$Y_{k,i,j,t}$ and $P_{k,i,j,t}$ are the vehicle loads of full and empty tanks on that trip respectively, and $I_{i,t}$ and $EI_{i,t}$ are the inventory of full and empty tanks at customer i, at the end of period t respectively. The *IRPSPD* can then be formally described as the following multi-period commodity network flow model with capacity constraints:

$$Min \quad \sum_{t=1}^{T} \left\{ 10^{a^*} \left(\sum_{k \in V} \sum_{(0,i) \in A} c_{aux_{0,i,t}} z_{k,0,i,t} \right) + \sum_{k \in V} \sum_{(i,j) \in A} c_{i,j} z_{k,i,j,t} + \sum_{i \in N} h_i I_{i,t} \right\} \tag{1}$$

s.t.

$$\sum_{\substack{j=0 \\ j \neq i}}^{N} z_{k,i,j,t} \le 1, \qquad \forall i \in N,\ k \in V \text{ and } t = 1,...,T \tag{2}$$

$$\sum_{\substack{m=0 \\ m \neq i}}^{N \cup \{0\}} z_{k,i,m,t} - \sum_{\substack{l=0 \\ l \neq i}}^{N \cup \{n+1\}} z_{k,l,i,t} = 0 \qquad \forall k \in V,\ \forall i \in N,\ t = 1,...,T \tag{3}$$

$$Y_{k,i,j,t} - z_{k,i,j,t} Q \le 0 \qquad \forall k \in V,\ \forall i, j \in N, i \neq j,\ t = 1,...,T \tag{4}$$

$$\sum_{\substack{l=0 \\ l \neq i}}^{N \cup \{0\}} Y_{k,l,i,t} - \sum_{\substack{m=0 \\ m \neq i}}^{N \cup \{n+1\}} Y_{k,i,m,t} \ge 0 \qquad \forall k \in V,\ \forall i \in N,\ t = 1,...,T \tag{5}$$

$$P_{k,i,j,t} - z_{k,i,j,t} Q \le 0 \qquad \forall k \in V,\ \forall i, j \in N, i \neq j,\ t = 1,...,T \tag{6}$$

$$\sum_{\substack{m=0\\m\neq i}}^{N\cup\{n+1\}} P_{k,i,m,t} - \sum_{\substack{l=0\\l\neq i}}^{N\cup\{0\}} P_{k,l,i,t} \geq 0, \qquad \forall k \in V,\ \forall i \in N,\ t = 1,...,T \tag{7}$$

$$I_{i,t-1} - I_{i,t} + \sum_{k=1}^{V}\left(\sum_{\substack{l=0\\l\neq i}}^{N} Y_{k,l,i,t} - \sum_{\substack{m=0\\m\neq i}}^{N} Y_{k,i,m,t} \right) = d_{i,t} \qquad \forall i \in N,\ t = 1,...,T \tag{8}$$

$$EI_{i,t} = EI_{i,t-1} - \sum_{k=1}^{V}\left(\sum_{\substack{m=0\\m\neq i}}^{N} P_{k,i,m,t} - \sum_{\substack{l=0\\l\neq i}}^{N} P_{k,l,i,t} \right) + d_{i,t} \qquad \forall i \in N,\ t = 1,...,T \tag{9}$$

$$Y_{k,i,j,t} + P_{k,i,j,t} \leq Q, \qquad \forall i \in N^{+},\ \forall k \in V,\ t = 1,...,T \tag{10}$$

$$I_{i,t} + EI_{i,t} \leq C_i, \qquad \forall i \in N,\ t = 1,...,T \tag{11}$$

$$Y_{k,i,j,t} \geq 0,\ I_{i,t} \geq 0,\ P_{k,i,j,t} \geq 0 \quad \forall i, j \in N,\ \forall k \in V,\ t = 1,...,T \tag{12}$$

$$Y_{k,i,\{n+1\},t} = 0, \qquad \forall i \in N,\ \forall k \in V,\ t = 1,...,T \tag{13}$$

$$P_{k,\{0\},j,t} = 0, \qquad \forall j \in N,\ \forall k \in V,\ t = 1,...,T \tag{14}$$

$$z_{k,i,j,t} \in \{0,1\},\ Y_{k,i,j,t}, P_{k,i,j,t}, I_{i,t}, EI_{i,t} \in Z, \quad \forall (i,j) \in A,\ \forall k \in V,\ t = 1,...,T \tag{15}$$

The objective function (1) minimizes the number of vehicles, the inventory and the transportation costs, giving priority to the attainment of the first goal, where:

$$a^{*} = \arg\min_{a} \left\{ \min\left(10^{a\text{-}1}\right),\ 10^{a\text{-}1} > C_{\max} \right\},\ a \in Z^{+}, C_{\max} = (K_{\text{av}} + n)c_{\max} \tag{16}$$

Constraints set (2) states that a vehicle will visit a location no more than once in a time period, and constraints set (3) ensure route continuity. Constraints (4, 6) ensure that the amount transported between two locations will always be zero whenever there is no vehicle moving between these locations, and that the amount transported is less than or equal to the vehicle's capacity. Constraint sets (5, 7) along with the other elements of the model ensure that efficient solutions will not contain sub-tours. Constraint sets (8, 9) are the balances of full and empty tanks at customer *i* at the end of period *t.* According to the capacity constraint set (10), when a vehicle leaves a customer, the total load must not exceed the capacity *Q*. Constraint set (11) ensures that the amount of tanks (empty and full) at a customer cannot exceed its' capacity to hold inventory C_i.

3 PREDICTIVE INVENTORY CONTROL APPROACH

In this work, the MPC strategy described in [2] is adjusted to the inventory routing and control problem. As shown in Figure 1 the future outputs for the prediction horizon *Np*, are predicted at each instant *t* using the process model (2-15). The set of future control signals is calculated by optimizing a *balanced* set-point tracking criterion (quadratic function) and the economic cost criterion of equation (1). The control signal $u(t|t)$ (deliveries) is sent to the process whilst the future calculated deliveries are rejected, because at the next sampling instant $y(t+1)$ is already known and step 1 is repeated with this new value and all the sequences are brought up to date. Thus the $u(t+1|t+1)$ is calculated using the receding horizon concept. The basic structure of the adopted strategy is shown in Figure 2. Model (2-15) is used to predict the future inventories, based on past and current values and on the proposed optimal future delivery actions. These actions are calculated by the optimizer taking into account the constraints and the cost function (where the future tracking and the economic cost error are balanced). We use a state space representation of the system:

$$x^{+} = Ax + Bu - B_d d \tag{17}$$

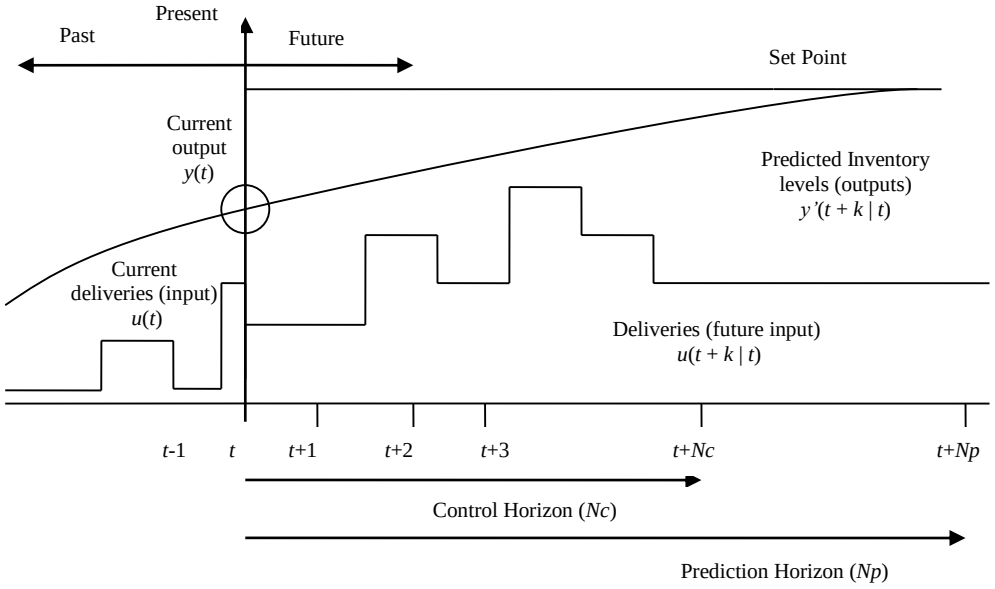


Figure 1: The Model Predictive Control strategy

in which $x \in Z^N$ is the system state i.e. the inventory at customers $\left[I_1 I_2 ... I_N\right]'$; $\left(x_{i,t} = I_{i,t}\right)$, $u \in Z^N$ is the manipulated input, i.e. the delivered quantities $\left[u_1 u_2 ... u_N\right]'$, where $u_i = \sum_{k=1}^{V}\left(\sum_{\substack{l=0 \\ l\neq i}}^{N} Y_{k,l,i} - \sum_{\substack{m=0 \\ m\neq i}}^{N} Y_{k,i,m}\right)$ and $d \in R^N$ is the disturbance to the system i.e. the demand of customers $\left[d_1 d_2 ... d_N\right]'$. The outputs are the states themselves $y \equiv x$. For the nominal disturbance (demand) d_s, assuming stability of the system [1], we adopt the model of the online terminal constraint MPC problem (18). With $g^c\left(z(t)\right) = \sum_{k\in V}\sum_{(i,j)\in A} c_{i,j} z_{k,i,j,t} + \sum_{i\in N} h_i I_i, t$ we formalize the transportation and inventory cost, while the minimization of the number of vehicles is not considered now as a primary objective, and $g^{tr}\left(x(t)\right) = \frac{1}{2}\left(x_t - x_{t,s}\right)' P \left(x_t - x_{j,s}\right)$ is the tracking cost, with matrix P for the cost of tracking the safety stock x_s. Coefficients a and $(1 - a)$, where $a \in [0, 1]$, are relative weightings assigned to the economic and the tracking costs respectively.

$$\left.\begin{aligned} &\min_{\mathbf{u}} \quad \sum_{t=0}^{Np-1}\left[a g^c\left(x(t)\right) + (1-a) g^{tr}\left(x(t)\right)\right] \\ &s.t. \quad x(t+1) = Ax(t) + Bu(t) - B_d d(t), \;\; t \in \{0,...,Np-1\} \end{aligned}\right\} \tag{18}$$

A, B and B_d derive from (8) and with respect to (2-15) for period $t = 1,..., Np$. The non-linear term $\frac{1}{2}\left(x_t - x_{t,s}\right)' P\left(x_t - x_{j,s}\right)$, can be replaced by $\frac{1}{2}\sum_{i=1}^{N}\left[diag(P)_i \left(x_{i,t}{}^2 - 2x_{i,t}x_{i,j,s} + x_{i,j,s}{}^2\right)\right]$ since P is diagonal. By substituting $q_{1,t} = x_t^{\,2}$ $q_{2,t} = x_t$ it becomes

$$
\left.
\begin{aligned}
& g^{tr}\left(x(t)\right) = \frac{1}{2}\sum_{j=1}^{N}\left[diag(P_{j,t})\left(q_{1,j,t} - 2q_{2,j,t}x_{j,s} + x_{j,s}{}^{2}\right)\right] \\
\text{also let} \quad & q_{1,t} = y_{1,t} + 4y_{2,t} + \ldots + 400y_{20,t} \quad \text{and} \\
& q_{2,t} = y_{1,t} + 2y_{2,t} + \ldots + 20y_{20,t} \quad (x_t \le 20) \\
& \sum_{k=1}^{20} y_{k,t} \le 1, \;\; y_{k,t} \in \{0,1\}
\end{aligned}
\right\} \qquad (19)
$$

Model (18) with the linearization constraints of (19) is a mixed integer linear model, which can be solved to optimality within reasonable computational times for moderate instances.

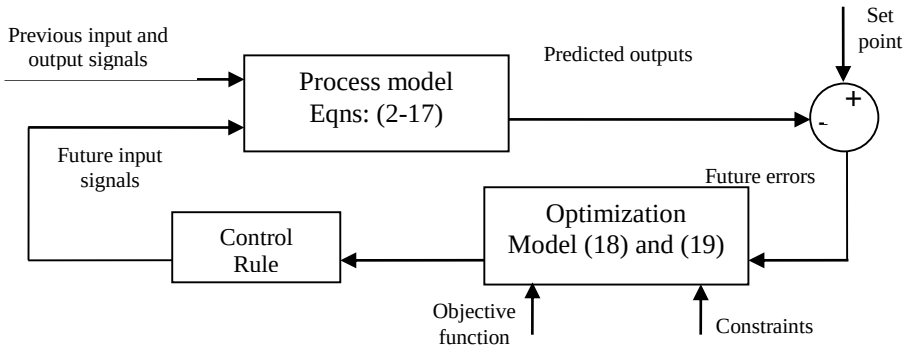


Figure 2: Structure of the Model Predictive strategy for the Inventory Control problem

4 AN ILLUSTRATIVE EXAMPLE

To illustrate the methodology a distribution system of 1 supplier, 3 customers, and 3 vehicles is considered. The vehicles' capacity is Q = 20. The nominal demand of the customers is d_s = [10 7 13]. The initial inventories of full and empty tanks are I_0 = [2 15 20] and P_0 = [11 5 2] respectively and the unit inventory cost is h = 1. The scheduling horizon spans 12 time periods. The unit transportation costs of node pairs are: $c_{0,1}$ = 15, $c_{0,2}$ = 18, $c_{0,3}$ = 22, $c_{1,2}$ = 32, $c_{1,3}$ = 14, $c_{2,3}$ = 34 and the cost table is symmetric.

The optimal schedule for the multi-period steady state case at the nominal demand is presented in Table 1. Rows and columns correspond to customer nodes and periods respectively.

Table 1: Multi-period steady state schedule

	Periods					
Customers	***1***	***2***	***3***	***4***	***5***	***6***
1	(3, 0, 8, 10)	(-, -, 0, 0)	(3, 0, 20, 10)	(-, -, 0, 0)	(3, 0, 13, 3)	(1, 0, 20, 13)
2	(-, -, 0, 8)	(-, -, 0, 1)	(2, 0, 20, 14)	(-, -, 0, 7)	(-, -, 0, 0)	(2, 0, 14, 7)
3	(-, -, 0, 7)	(2, 0, 19, 13)	(-, -, 0, 0)	(3, 0, 20, 7)	(1, 0, 19, 13)	(-, -, 0, 0)
	Periods					
	7	***8***	***9***	***10***	***11***	***12***
1	(1, 0, 7, 10)	(-, -, 0, 0)	(1, 0, 13, 3)	(1, 3, 7, 0)	(3, 0, 10, 0)	(1, 0, 20, 0)
2	(-, -, 0, 0)	(2, 0, 14, 7)	(-, -, 0, 0)	(2, 0, 8, 1)	(2, 0, 20, 14)	(-, -, 0, 7)
3	(1, 1, 13, 0)	(3, 0, 19, 6)	(1, 1, 7, 0)	(1, 0, 13, 0)	(1, 0, 19, 6)	(2, 0, 20, 13)

The quadruplet entries of the table represent: (vehicle, previous node, delivery, and inventory at the end of each period).The total cost is 967 units and it is consisted of 777 units for transportation and 190 units for inventory. For the disturbance predictions of the dynamic system, we make a demand forecast ***d***. We assume that we have perfect demand information for one period, thus we select $Nc = 1$, ($Np = 12$). For the remainder of the horizon, the demand forecast is set to the nominal demand d_s.

The Np-horizon dynamic case

The targeted safety stock is the solution of $x(t+1) = Ax(t) + Bu(t) - B_d d(t)$ for one period delivery delay: x_s = [10 7 13]. The targeting cost of the safety stock is a diagonal matrix P, where $P_{i,j} = 2$ for $i = j$. We assume that the customer demand is normally distributed around the nominal demand. The variance of the demand is 0.5. Table 2 contains the actual demand. Figure 3, presents the response of the inventories with α varying from 0 (i.e. strategy for shortage risk aversion) to $\alpha = 1$ (i.e. opt for economic criterion). Results show the safety stock is reached and maintained promptly in the first case. For example, for a = 0.2, Figure 3 shows that the response of the closed-loop system, quickly settles to the safety steady state. Table 3 reports the distribution of costs with varying *a*, which shows that balanced decisions can be made based on the change of a single parameter (*a*) and the desired inventory tracking can be matched with the best possible routing schedules.

Table 2: Actual demand for 12 periods

	Periods											
Cust	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***
1	9	9	10	11	10	10	9	9	10	11	10	10
2	7	8	6	7	7	8	7	8	6	7	7	8
3	14	13	12	13	12	14	14	13	12	13	12	14

Table 3: Cost distribution with varying *a*

a	0	0.2	0.3	0.4	0.5	0.8	0.9	0.95	1
Total cost	1651	1606	1527	1451	1338	1246	1145	1053	1044
Transportation cost	1291	1248	1173	1099	992	917	886	851	814
Inventory cost	360	358	354	352	346	329	259	202	230
Tracking cost	0	6	26	86	210	577	1005	1560	1510

Conclusions

The paper presents a linear deterministic model for the multi-period IRPSPD and a dynamical predictive inventory control strategy for the industrial gas distribution business. It is shown that by using the proposed MPC strategy, the supply chain can be stabilized. The method offers a useful tool for producing balanced decisions between costs and tracking of the safety inventory. Future research may consider the presence of time windows for the customer service, which is a realistic approach to the actual problem. Also the development of custom algorithms for the solution of instances of higher cardinality will be required since the problem is NP-hard.

Acknowledgement

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds of the Operational Program "Education and Lifelong Learning"

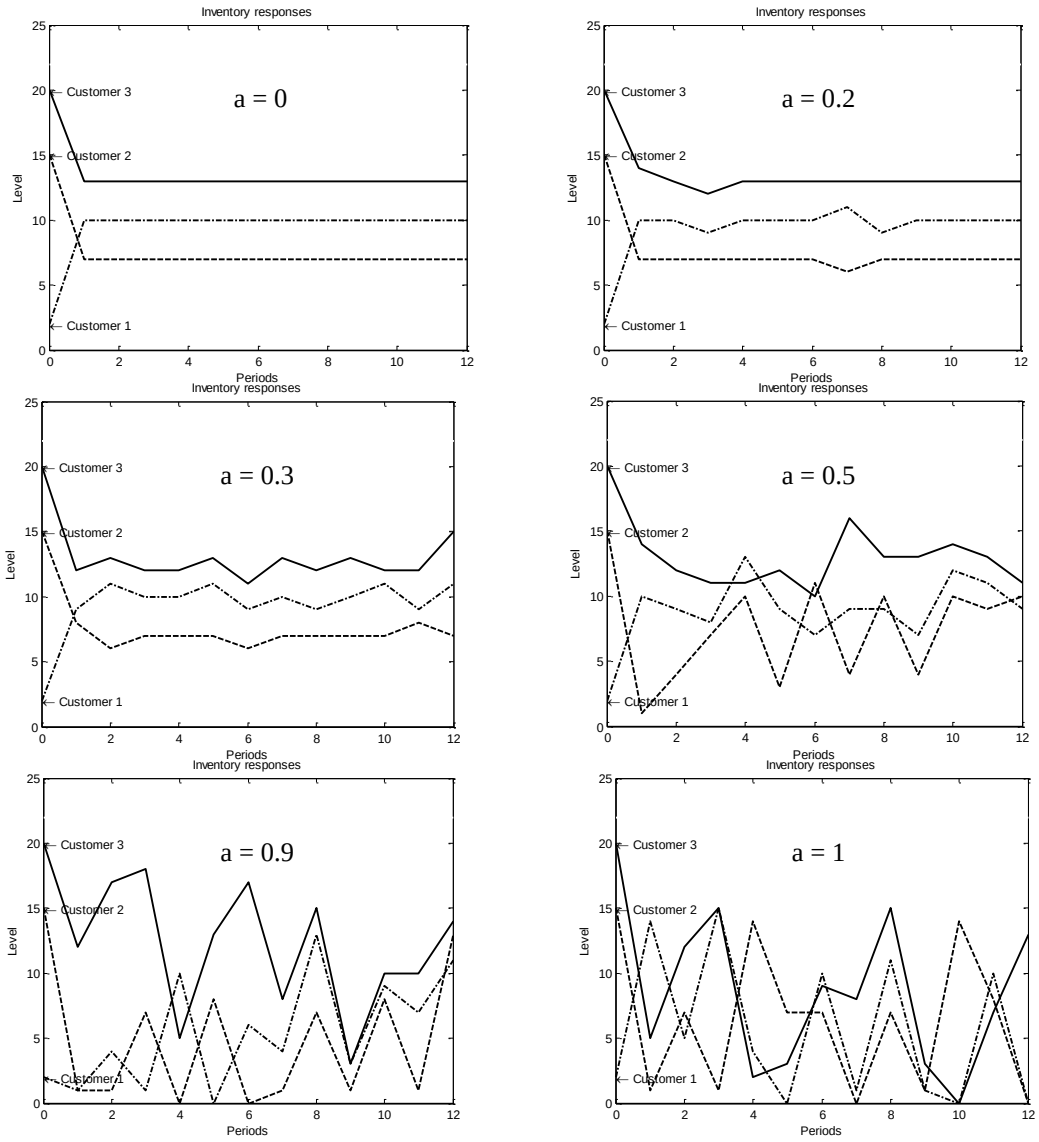


Figure 3: Inventory responses for a = 0, 0.2, 0.3, 0.5, 0.9 and 1

of the National Strategic Reference Framework (NSRF) - Research Funding Program ARCHIMEDES III, "Investing in knowledge society through the European Social Fund".

References

[1] Amrit, R., Rawlings, J.B., Angeli, D. (2011) Economic optimization using model predictive control with a terminal cost. Annual Reviews in Control, 35 (2): 178-186.

[2] Camacho E.F. and Bordons C. (1999), Model Predictive Control, Springer

[3] Kleywegt, A.J., Nori, V.S., Savelsbergh, M.W.P. (2002) The stochastic inventory routing problem with direct deliveries. Transportation Science, 36 (1): 94-118.

[4] Nikolakopoulos, A. and Sarimveis, H. (2006). A Heuristic Approach to the Vehicle Routing Problem with Time Windows and Simultaneous Pick-up and Delivery. Proceedings of the Odysseus 2006, Third International Workshop on Freight Transportation and Logistics, p.263-266.

[5] Nikolakopoulos, A. (2015) A Metaheuristic Reconstruction Algorithm for Solving Bi-level Vehicle Routing Problems with Backhauls for Army Rapid Fielding, Springer Operations Research/ Computer Science Interfaces Series, 56: 141-157.

Managing different demand classes in a spare parts inventory: a practical dynamic allocation strategy

Erica Pastore
Politecnico di Torino, c.so Duca degli Abruzzi 24, Torino, ITALY, erica.pastore@polito.it,
Arianna Alfieri
Politecnico di Torino, c.so Duca degli Abruzzi 24, Torino, ITALY, arianna.alfieri@polito.it,
Giulio Zotteri
Politecnico di Torino, c.so Duca degli Abruzzi 24, Torino, ITALY, giulio.zotteri@polito.it

Abstract: When different demand classes must be served from the same inventory, stock is allocated following a rationing policy. Under rationing, low priority demand is filled as long as the inventory is above a critical level. When inventory decreases below the critical level, parts are reserved for future high priority demand. In this paper, we develop a dynamic rationing policy that changes the critical level as the next delivery is approaching.
Simulations have been run to compare the effect of dynamic versus static critical levels on service level. Results show that the dynamic critical levels lead to a higher service level.

Keywords: inventory management, allocation strategies, rationing policies, dynamic critical levels

1 INTRODUCTION

Inventory systems often face various classes of orders (for a given item) with different priorities. For instance, in an automotive spare parts warehouse, the same item can be requested to repair a vehicle out of road (VOR order) or as regular restocking order from the retailer. In such a case, the company may want not to fill a demand with a low priority and reserve stock for potential VOR future demand.
In the literature, this problem is known as *inventory rationing problem*. Inventory rationing deals with finding a critical level such that, when inventory-on-hand falls below it, low priority orders are back-ordered and parts are reserved for future high priority demand. Critical levels can be constant (*static inventory rationing*) or can change in time (*dynamic inventory rationing*). The majority of the literature contributions analyses systems with static inventory rationing. In particular, optimal critical levels are computed minimizing expected cost [2, 3, 1] or fixing expected service level for each demand class [10, 9]. A comprehensive and accurate review on static critical levels literature can be found in [11] and [7].
The literature about dynamic rationing is relatively scarce, although one of the first papers dates back to 1968 [12]. Differently from static approaches, the main idea of the dynamic ones is to link critical levels to time. For example, the lead time is divided in subintervals and the critical levels are constant over each subinterval [6]. Instead, [8] introduces the concept of *rationing trigger times*, i.e. critical times after which low priority demand is no longer satisfied. Critical levels are computed on the basis of the residual lead time before the next shipment in [11, 4], while [5] uses the stock-out probability of future high priority demand in the residual lead time to set the levels.
This paper presents a heuristic rule to determine dynamic critical levels. The idea is that, as the expected number of higher priority orders decrease as the next replenishment is approaching, also the critical level must be reduced as well. This is the same idea pursued in [11] and [4]. However, while [11] and [4] device an objective function whose minimization leads to the critical levels, we determine their values as if they were the safety stock for the residual high priority demand before the next replenishment. This is achieved by setting a constraint on the service level for high priority demand. Our heuristic has been tested on a European automotive spare parts inventory and compared with a static rationing policy. The results show that it is

possible to achieve higher service level allocating stock by using the dynamic critical levels. The remainder of the paper is structured as follows. Our model is discussed in section 2. Section 3 presents the results of the case study and section 4 concludes the paper.

2 DYNAMIC INVENTORY RATIONING MODEL

2.1 System Description

We analyse a single-echelon single-warehouse system, characterized by periodic-review policy, two demand classes and backlogging. Demand is stochastic and each demand class is an independent random variable $D_{i,t}, i = H, L$. At time t the total demand is $D_t = D_{H,t} + D_{L,t}$, where $D_{H,t}$ has higher priority than $D_{L,t}$. This assumption could model different service level requirements or different penalty cost for the two classes. We assume the supplier's lead time (LT) is constant and deterministic.
We control inventory by using an order-up-to S policy, i.e., every τ periods an order is placed to bring the inventory position up to the value S. Hence, the order issued at time t has to cover demand up to the delivery of the next order, over the so-called *out of control period* ($LT + \tau$). The value S is computed as:

$$S = F_{LT+\tau} + z_\alpha \sigma_{LT+\tau}, \tag{1}$$

where $F_{LT+\tau}$ is the forecasted demand during the out of control period, z_α the α-quantile of the demand distribution and $\sigma_{LT+\tau}$ a measure of the demand uncertainty. The term $z_\alpha \sigma_{LT+\tau}$ represents the so-called *safety stock*. Some allocation strategy must be used to choose whether (and in case, to which extent) low priority demand should be filled or backlogged. In this situation, it might be appropriate not to fill some low priority orders and to reserve stock for potential high priority future demand; hence, critical levels must be set to decide how much stock to reserve for high priority demand.

2.2 A Dynamic critical level

As introduced in section 1, the purpose of this paper is to devise a heuristic procedure to set dynamic critical levels. The logic behind the procedure is that, as the next replenishment is approaching, the expected number of high priority orders that inventory currently in the warehouse needs to meet decreases. Thus, also the amount of stock reserved for high priority demand can be reduced (Figure 1).

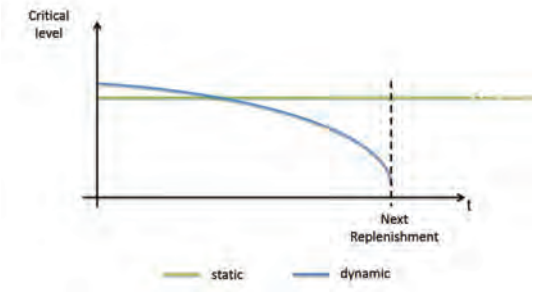


Figure 1: Static vs Dynamic critical levels

Just after a replenishment, the dynamic critical level reserves more inventory to cover the high priority demand over all the period until the next replenishment. However, this occurs at a time when the amount of total available inventory is large, thus there is a negligible effect on the fulfilment of low priority orders. On the other hand, just before the next replenishment, the stock kept for high priority orders can decrease and this decrease frees part of inventory

that can be used to satisfy low priority demand. The decrease of the critical level could result in a small reduction of the service level for high priority orders meanwhile increasing that for low priority ones, with a possible positive balance.
Following the idea presented above, any time the system faces a new replenishment, critical levels are updated. Let T denote the length of the period between two successive replenishments and t the current day, then $t = 1, \cdots, T$. Let CL_t be the critical level at day t and CL_0 the initial critical level, calculated at the time the first replenishment enters the system (i.e., the warehouse). The initial critical level CL_0 represents the amount of stock the system needs to guarantee a target service level for high priority demand within period T. For this reason, CL_0 can be considered as a safety stock for high priority demand. According to our heuristic, CL_0 is computed as follows:

$$CL_0 = z_{H,\alpha}\sigma_{H,T}, \tag{2}$$

where α is the target service level for high priority demand satisfaction, $z_{H,\alpha}$ the α-quantile of high priority demand distribution and $\sigma_{H,T}$ a measure of its uncertainty within period T. Assuming that demand fluctuations are independent within T, the demand variability $\sigma_{H,T}$ can be computed as $\sigma_{H,T} = \sigma_H\sqrt{T}$.

At day t, the critical level CL_t must guarantee a target service level for high priority demand within period $T - t$, thus it must be updated to a lower value $CL_t < CL_0$. Using the idea of equation (2), CL_t must be updated as follows:

$$CL_t = z_{H,\alpha}\sigma_{H,T-t}. \tag{3}$$

Assuming independent demand over time, it is easy to show that CL_t is proportional to CL_0 and, in particular,

$$CL_t = CL_0\sqrt{\frac{T-t}{T}}. \tag{4}$$

As stated in equation (4), CL_t decreases as the next replenishment is approaching, in order to free stock for low priority demand.

3 AN APPLICATION TO A REAL SYSTEM

The heuristic described in section 2 has been tested in a real context to improve the performance of a European automotive spare parts central depot. In such a context, the same item can be requested to repair a vehicle out of road (VOR order) or as regular restocking order from dealers. The cost linked to a VOR backorder is higher than the one associated to regular orders. For this reason, the company may decide to have different rules to allocate its inventory-on-hand among orders with different priority, especially when the inventory level is down to a few units. In this situation, it might be appropriate not to fill a demand with lower penalty costs and reserve stock for potential VOR future demand.
The warehouse plans orders to suppliers according to the order-up-to S policy with monthly frequency. For the sake of simplicity, suppliers are assumed to have deterministic lead time LT (perfect reliability). Demand is represented by the orders arriving from dealers that must be shipped within the end of the day. Unfulfilled demand is backlogged. The sequence of events in each day is the following. First, the replenishment order placed LT days ago arrives and dealers' demand (for both high and low classes) is collected. Then, available stock is allocated to orders. Initially, both high and low priority backorders (in case there is any) are fulfilled; then, the inventory-on-hand is used to satisfy high priority demand. Finally, if the remaining stock is above the critical level, low priority orders are served until the critical level is reached.

The company was interested in improving the performance of their warehouse in terms of increasing orders fulfilment for both demand classes. Thus, the performance measure we used

was the fill rate for each class i, i.e.,

$$SL_i = \frac{Fulfilled\ class\ i\ demand\ quantity}{Total\ class\ i\ demand\ quantity} \qquad i = H, L. \tag{5}$$

We compared the developed heuristic with the *as is* static inventory rationing policy of the company.

3.1 *As is* static rationing policy

The critical levels CL_m^{static} used by the company are computed at the beginning of each month m as a percentage UP_m of the monthly safety stock SS_m:

$$CL_m^{static} = SS_m UP_m. \tag{6}$$

The percentage UP_m corresponds to the smoothed fraction of high priority demand $D_{H,m-1}$ over total demand D_{m-1} in the previous month, i.e.,

$$UP_m = \beta \frac{D_{H,m-1}}{D_{m-1}} + (1-\beta)\, UP_{m-1}, \tag{7}$$

where β is the smoothing coefficient of the high priority demand percentage and it is set to $\beta = 0.2$. SS_m is the safety stock calculated in month m, i.e., the same used in equation (1):

$$SS_m = z_\alpha \sigma_{LT+\tau}. \tag{8}$$

The main difference between this policy and the one represented by equations (2) - (4) is that the dynamic critical level changes everyday, decreasing with the approaching of a new replenishment, while the static one is monthly updated and stays constant within the month. In the following, the dynamic critical level is indicated by CL_t^{dyn}.

3.2 Numerical study

We tested dynamic critical levels on 35 fast moving items over a time horizon of 3 years. Time-bucket is a single day. The inventory planning process has monthly frequency, i.e., 30 days. Suppliers deliver at the beginning of the month.
We adopted a 2-months moving average as demand forecasting technique and the Root Mean Squared Error (RMSE) as measure of demand uncertainty, i.e.,

$$\sigma_t = \frac{\sum_{k=1}^{t} (D_k - F_k)^2}{t}. \tag{9}$$

The company inventory behaviour has been simulated under the two different allocation strategies. The demand pattern for both systems is the one observed by the company from dealers during the three years time horizon. Instead, the replenishment ordering pattern has been generated using two different target service level. The **low-SL** pattern has been generated using $z_\alpha = 2$ in equation (1), while the **high-SL** one with $z_\alpha = 5$.
A first experiment has been the comparison between the two policies. For each ordering pattern, the same high priority service level SL_H has been fixed for static and dynamic rationing and low priority service level SL_L has been compared. The results are shown in Table 1.
It can be noticed that, with the **low-SL** pattern, the two policies guarantee the same performance: giving $SL_H = 93\%$, low priority orders are fulfilled with $SL_L = 86\%$. On the other hand, when a **high-SL** ordering pattern is considered, the dynamic policy leads to higher performance on low priority demand. Indeed, with equal $SL_H = 97\%$, SL_L can be increased by 6% in the dynamic case.

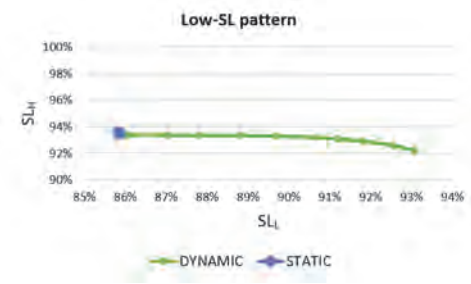
$z_\alpha = 2$	SL_H	SL_L
STATIC	93%	86%
DYNAMIC	93%	86%

$z_\alpha = 5$	SL_H	SL_L
STATIC	97%	89%
DYNAMIC	97%	95%

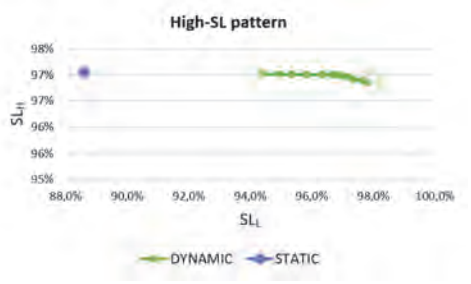
Table 1: High and low priority service level

As second experiment, we changed the high priority target service level for the dynamic policy. In fact, setting different values to $z_{H,\alpha}$ in equations (2) - (3), managers can choose different performance for high priority demand, which obviously impacts on SL_L. On the contrary, with the static rationing policy used by the company, there is no freedom to choose the most appropriate trade-off between SL_H and SL_L.

Figure 2 presents the trade-off curve for high and low priority service levels achievable with the



(a) High and low priority Service Level trade-off with **low-SL** ordering pattern



(b) High and low priority Service Level trade-off with **high-SL** ordering pattern

Figure 2: High and low priority Service Level trade-off

dynamic critical levels for both the ordering patterns. Focusing on the **low-SL** pattern (Figure 2a), if for example the inventory managers choose $SL_H = 92\%$, then low priority service level turns out to be equal to $SL_L = 93\%$. Having seen in Table 1 (on the left) that $SL_H = 93\%$ implies $SL_L = 86\%$, it means that, reducing SL_H by 1%, SL_L can be improved by 7%.
Considering the **high-SL** pattern (Figure 2b), we obtain a similar behaviour. For example, for $SL_H = 96.5\%$, we get $SL_L = 98\%$, i.e., comparing with results in Table 1 (on the right), a decrease on SL_H of 0.5% leads to an increase of 3% on SL_L.

Determining the best policy is, however, not a simple issue to deal with. The optimal trade-off between SL_H and SL_L should take into account the costs, either direct or indirect, the company faces for each class backorder.

4 CONCLUSION

In this paper, we devised a heuristic approach to dynamically set critical levels for the inventory rationing problem. The critical levels set with the developed rule decrease while the next replenishment is approaching, as the expected number of high priority orders to meet does.

We have tested our model in a real context to increase the performance of a European automotive spare parts inventory. We have compared the proposed heuristic with the static rationing rule used by the company. The results show that, with the dynamic policy it is possible to control the trade-off between the performance for high and low priority orders by setting the most appropriate high priority service level.

Future research will follow two directions. The first direction deals with the inclusion of backorder costs in the computation of critical levels. In fact, different priority demands usually lead to different backorder costs for the company and the critical levels must account for this

difference. The second direction, instead, will have to link the critical level setting not only to the next replenishment date but also to its size, as discussed in [13].

References

[1] Dekker, R., Hill, R., Kleijn, M., and Teunter, R. (2002). On the (s- 1, s) lost sales inventory model with priority demand classes. Naval Research Logistics (NRL), 49(6):593–610.

[2] Ha, A.Y. (1997). Inventory rationing in a make-to-stock production system with several demand classes and lost sales. Management Science, 43(8):1093–1103.

[3] Ha, A.Y. (1997). Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. Naval Research Logistics (NRL), 44(5):457–472.

[4] Haynsworth, H.C. and Price, B.A. (1989). A model for use in the rationing of inventory during lead time. Naval Research Logistics (NRL), 36(4):491–506.

[5] Hung, Y.F. and Hsiao, J.Y. (2013). Inventory rationing decision models during replenishment lead time. International Journal of Production Economics, 144(1):290–300.

[6] Melchiors, P. (2003). Restricted time-remembering policies for the inventory rationing problem. International Journal of Production Economics, 81:461–468.

[7] Möllering, K.T. and Thonemann, U.W. (2008). An optimal critical level policy for inventory systems with two demand classes. Naval Research Logistics (NRL), 55(7):632–642.

[8] Moon, I. and Kang, S. (1998). Rationing policies for some inventory systems. Journal of the Operational Research Society, pages 509–518.

[9] Nahmias, S. and Demmy, W.S. (1981). Operating characteristics of an inventory system with rationing. Management Science, 27(11):1236–1245.

[10] Tempelmeier, H. (2006). Supply chain inventory optimization with two customer classes in discrete time. European Journal of Operational Research, 174(1):600–621.

[11] Teunter, R.H. and Haneveld, W.K.K. (2008). Dynamic inventory rationing strategies for inventory systems with two demand classes, poisson demand and backordering. European Journal of Operational Research, 190(1):156–178.

[12] Topkis, D.M. (1968). Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. Management Science, 15(3):160–176.

[13] Wang, Y., Zhang, S.H., and Sun, L. (2013). Anticipated rationing policy for two demand classes under service level constraints. Computers & Industrial Engineering, 65(2):331–340.

MODELLING THE CRUDE OIL STORAGE MANAGEMENT

Lorenzo Ros-McDonnell
Mediterranean Institute for Advances Studies (MEDIFAS)
Mednarodni prehod 6, Vrtojba, Sempeter pri Gorici, Slovenia
lorenzo.ros@upct.es

Maria Victoria de-la-Fuente Aragon
Technical University of Cartagena, Research Group Industrial Engineering and Management
S/ Dr. Fleming, s/n; 30202 Cartagena, Spain
marivi.fuente@upct.es

Abstract: Stock management in an oil refinery is a complex operation, and it is necessary to consider a large number of variables and factors. In this paper the researchers make approach to this data, and we present he results obtained through a developed application, which is directly connected to the operative system of the refinery, in order to plan the unloading operations when the arrival of a tanker, an a better management of the stock in the storage tanks.

Keywords: Modelization, simulation, crude oil, refinery, stock management.

1 INTRODUCTION

The main activity of a refinery is to meet the demand for refined products (mostly fuels) in quantity, quality and time, according to market needs. To meet this goal, refineries will adapt their production schemes for a set of processes; in order to get the process of refining crude oil becomes more suitable to achieve this technical objective.

The supply of raw materials needed to develop the refining activity is performed by tankers, in the case of refineries near the seaport, or by pipeline, in inland refineries. These are the two basic means by which Spain get crude oil from producing countries.

Crude oil storage is done in specially designed tanks. The importance of this operation is critical for refining, it must properly operate with the storage capacity of these facilities, in order to enable supply of charge to distillation units, (in which will occur the first separation of the components that constitute the petroleum products), to fulfill the safety stock (to avoid unwanted refinery stops), and to optimize the storage capacity, ensuring the raw material supply and a proper stock rotation, to avoid the freezing of oil for long periods time.

Therefore, the refinery tries to meet the market needs thorough a comprehensive planning and a proper stock management. In this paper, we present he results obtained through the developed application, in a spread sheet, which is directly connected to the operative system of the refinery, will plan the unloading operations when the arrival of a tanker.

Now, the operator in charge of the system management does this process manually, with possible errors associated to the human factor. With the new application, we aim to automate this process, facilitating planning and minimizing further operator mistaken.

1.1 Overview of planning and scheduling in oil refineries

The short-term oil refinery-scheduling problem is one of the most challenging problems in operational research due to the complexity of the refinery operations and the corresponding process models (Pinto et al, 2000).

Traditionally, long-term and plant-wide planning problems in the Petrochemical Industry have been mainly addressed by mathematical programming techniques (Bodington and Baker, 1990). Pure Linear Programming (LP) methods have been used for long-term planning, but are not applicable to short-term scheduling and on-line optimisation.

In fact, short-term refinery scheduling combines a set of problems that have been mainly addressed in literature as cost/profit optimisation problems, such as the allocation of crude oil to tanks, the problem of determining the flows from several input tanks, etc. Paolucci et al. (2002) proposed a simulation based decision support system for allocating crude oil supply to port and refinery tanks. Pinto et al. (2000) present the results of the application of a mixed integer optimisation model in a similar real world problem. In their model, time is represented by variable length time slots, which correspond to crude oil receiving operations (vessels unloading) as well as to periods between two receiving tasks. Shah (1996) proposed a mathematical programming approach, in which, tanks may store only one crude type and each CDU runs exactly one crude type from one tank at a time.

In general, the availability of LP-based commercial software for refinery production planning has allowed the development of general, long-term (Paolucci et al., 2002) production plans of the whole refinery (Pinto et al., 2000). On the other hand, the refinery short-term scheduling processes (requiring fast decisions) are affected by stochastic disturbances (Paolucci et al., 2002).

In this complex scenario, commercial tools for short-term production scheduling are few and these do not allow for a rigorous representation of plant particularities. The optimisation based formulations for short term scheduling in literature are very few. They also show that there are still a number of issues to address in order to be applicable (Simons, 1995; Pinto et al., 2000; Paolucci et al., 2002; Chryssolouris et al., 2005):

(a) the difficulty of controlling the computational time frame,
(b) the difficulty of imposing directly the time structure in the model, and
(c) the optimisation criteria, which are usually expressed in cost or profit terms, while units operation and the overall plant objectives may not always be directly represented with cost/profit terms.

2 THE APPROACH

To model the crude oil stock management system of the refinery (see fig. 1), we should identify all elements involved in this process, which includes the receipt of oil from production centres and transported by tankers, and a its later storage in tanks for final refine operations.

2.1 Elements involved in the crude oil storage system operation

- Tanker (tankship): It allows the transport of raw materials from the wells to the refinery, and domestic or international distribution of the finished products from refineries to consumption centres.
- Storage Tanks (T_i): there are 20 floating roof tanks with a usable storage capacity of 92,477 m^3/tank (total usable volume of 1,849,540 m^3)
- Pipeline: allows pumping crude to facilities located within the country. It has a large capacity (12,000-25,000 ton/day), the most safe and respectful environment for transferring hydrocarbon mean. In this study it has been considered a variable pumping stream S= [500 - 1000] m^3/h.
- Distillation unit (CDU): where the first stage of refining crude oil occurs. In the present work, the distillation unit loads crude from the storage tanks fed from the oil terminal, which receives the discharge of the tanker. Usually it works with a variable load stream S = [500-750] m3/h.
- Strategic storage - C_i - (CORES): Compulsory volume of crude oil, as strategic stock, with the aim of ensuring supply to the whole country for a reasonable period of time to deal with a potential supply crisis. The strategic quantity of oil, according to

Spanish Law 34/1998 and European Directive 2009/119/EC, will be at least equal to 90 days of the country consumption. This situation forces the refinery having immobilized 5 tanks from the set of 20 tanks, with a approximate capacity of 500,000 m^3

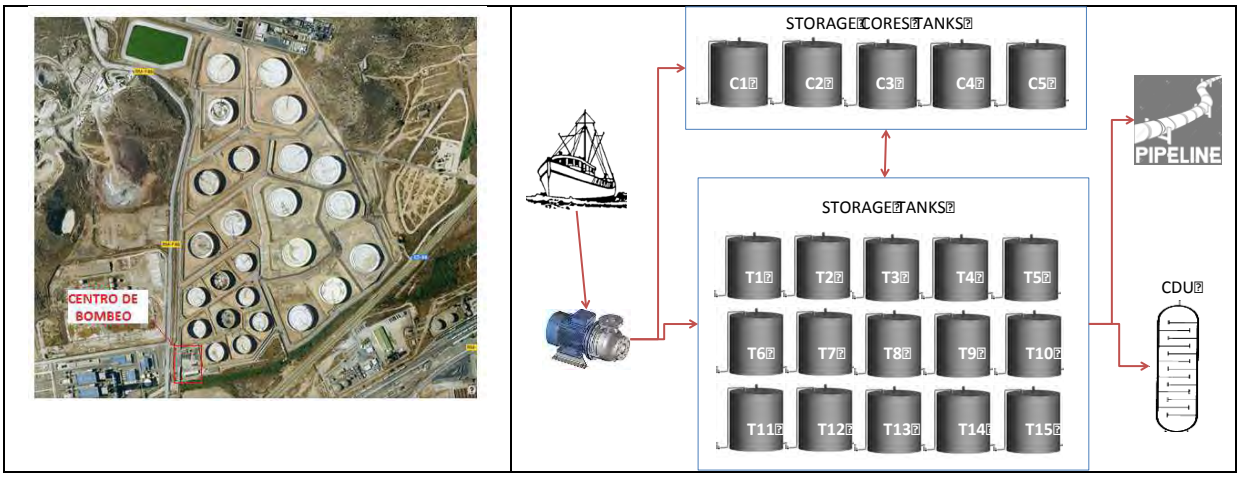


Figure 1: Air photography and diagram of the crude oil storage system in a Spanish refinery.

2.2 The system operation

The storage tank filling operations and their operative conditions are performed as follows: Once the tanker ship docks in the port, it is inspected to verify that the load is in good condition, operators proceed to connect the tanker to the unloading lines through which the crude oil will be loaded to the storage tanks.

The unloading rate can range from 6,000 to 10,000 m3/h. (For calculations in this work it has been considered an average stream S = 8,000 m3/h). The tanker usually ship unloads continuously, filling one tank after another one, of those who are scheduled to receive the crude oil volume.

After the unloading operative is finished, a new inspection rechecks the ship's tanks and storage tanks to calculate the unloaded volume. Once the unloading process is closed, the tanker is disconnected from the pipeline, and leaves to another production centre to be reloaded and restart the cycle.

2.3 Modelling the system. Decision-making process.

Once the elements of the crude oil storage system have been defined (section 2.1), and the operation mode of this system has been detailed (section 2.2), for modelling its operation it is necessary to define the steps of the decision-making process, together with the set of rules, constraints and boundary conditions that make possible the loading/unloading of the storage tanks (T_i).

Then, we describe the steps facing the arrival of a tanker to the port for delivering their crude oil volume. These steps are:

Step 1. Scheduled arrival of a tanker into port. Unloading planning.
Parameters to consider: date and time of arrival, volume to unload, stream and estimated time of operation.

Step 2. Determine initial situation of the storage tanks. Determine inventory levels.

Parameters to consider: used volume per tank, empty volume per tank.

- Only T_i will be possible nominated tanks, because C_i tanks have the immobilized stock.
- T_i with operation in process (supplying to the distillation unit, the pipeline, or another tank – T_i) will not be selected for the unloading process.

Step 3. Define unloading sequence for loading the storage tanks.
Selection criteria tanks: Total empty volume of selected tanks should be similar to the volume of the tanker load. (Establish possible combinations and select those feasible).
Definition of loading sequence: from highest to lowest empty volume tanks.

Step 4. Estimate need to transfer from Ti tanks to crude distillation unit (CDU).
Parameters to consider: stream. Estimated time for unloading nominated tanks.

Step 5. Estimate need to transfer from Ti tanks to pipeline.
Parameters to consider: stream. Estimated time for unloading nominated tanks.

Relevant issues to consider:

- Safety Stock: Volume which keeps the distillation unit operative for almost 45 days. (approx. 810.000m3), reaching a volume level near this amount, the refinery should be planned future purchases of crude oil.
- Data necessary for the decision making process will be provided by the refinery operation system.
- The refinery has established a group of crudes for purchase, so that there are no problems regarding their possible blending, and in order to be suitable for further processing.
- All T_i tanks are eligible for loading and supplying to CDU and Pipeline, and the movement of crude among them is allowed.
- C_i tanks are immobilized (strategic stock) and will not participate in the loading/unloading operations. Only once a year, the movement of crude among C_i and T_i is made for maintaining them and to avoid problems with raw material in C_i.

3 SIMULATING THE STORAGE SYSTEM MODELLED. AN IMPROVEMENT PROPOSAL.

The operator in charge of the storage system will start the decision-making process whenever a tanker arrival is scheduled by the refinery, needed supply to the pipeline or CDU, or every certain period of time (set by the refinery). Depending on the stock level in each T_i tank of the system, the operator will schedule the operations to be performed on T_i tanks.

In the presented study case in the paper, it is scheduled a tanker unloading at this moment (because it is right now docked at the port): Tanker GENMAR SPYRIDON, ready to download 94,000 m^3 of crude oil type "Castilla".

Then, the initial conditions of the storage system is fixed as shown in Table 1. After that, we have analyzed the evolution of the stock in the set of T_i and C_i tanks, results obtained from the simulation of the modelled system. We have established appropriate time sequences, in this case has set a time-window of 24h (that includes the full unloading of the tanker - approx. 12h). As shown in table 2 and figure 2.

Table 1: input data necessary for the study of the storage Ti tanks levels. Data provided by the refinery

Tank	*Used Volume (m3)*	*Empty Volume (m3)*	*Operation*	*Pumping rate S(m3/h)*	*Loading / Unloading time (hours)*
C1	92,206.34	254.86	No		
C2	91,992.97	468.23	No		
C3	92,396.00	65.20	No		
C4	91,803.30	657.90	No		
C5	92,449.35	11.85	No		
T1	67,141.06	25,320.14	No		
T2	717.17	91,744.03	Loading crude oil	8320	11.03 h
T3	25,646.13	66,815.07	No		
T4	91,868.50	529.70	No		
T5	83,120.25	9,340.95	Nominated for loading crude oil		
T6	37,506.06	54,955.14	No		
T7	41,761.64	50,699.56	Pumping to CDU	450	92.80 h
T8	8,179.26	84,281.94	Pumping to CDU	250	32.72 h
T9	80,749.45	11,711.75	No		
T10	31,910.97	6,550.23	Pumping to T15	650	49.09 h
T11	18,450.75	7,410.45	Pumping to pipeline	670	27.54 h
T12	50,272.81	42,188.39	Nominated for pumping to pipeline		
T13	2,032,96	90,428.24	Pumping to pipeline	230	8.84 h
T14	67,745.61	24,715.59	Nominated for pumping to pipeline		
T15	717.17	82,800.19	Pumping from T10	650	127.38 h

Table 2: Results obtained from the simulation of crude oil storage until the arrival of a new oil tanker

	***Total** Used Volume (m3)*	***Used volume in CORES tanks***	***Total used volume in normal storage tanks***	***Total empty volume***
t	1,069,432.2	460,847.958	616,763.62	745,438.79
t+1	1,076,152.32	460,847.958	615,304.36	738,718.79
t+2	1,082,872.318	460,847.958	622,024.36	731,998.79
t+3	1,089,592.318	460,847.958	628,744.36	725,278.79
…	…		…	…
…	…	…	…	…
t+23	1,223,992.318	460,847.958	763,144.36	590,878.79
t+24	1,230,712.318	460,847.958	769,864.36	584,158.79

From the results shown in Figure 2, can be noted the following issues:

The total used volume increases as a result of the crude oil unloaded from the tanker, which has increased the total stock of the refinery. Actually, it increases the used volume of Ti tanks, because the used volume of Ci tanks remains constant, due to its definition as a strategic reserve of crude oil in the country.

As a real improvement for the modelled system, and in the way of optimizing the loading/ unloading process of tanks, we propose a division of the set of 20 tanks into three groups:

- Group 1: CORES (Ci) tanks for the storage of strategic reserves.
- Group 2: storage tanks for pumping to pipeline, P1- P7 (seven tanks)
- Group 3: storage tanks for loading the distillation unit (CDU), T1-T8 (eight tanks).

Always keeping the operative conditions previously defined, that is, all storage tanks can receive load from tankers, and crude oil movements are permitted among all storage tanks.

This proposal seems to reduce the operational flexibility of the system (facilities), because each tank will se assigned to exclusive uses, but the benefits, which will be achieved in the short-medium term of the system operation, are:

- Simplicity in facility circuits, with greater clarity in the system operation.
- Reduction of set-up times, operation and maintenance.
- -Higher control and safer facilities (system).

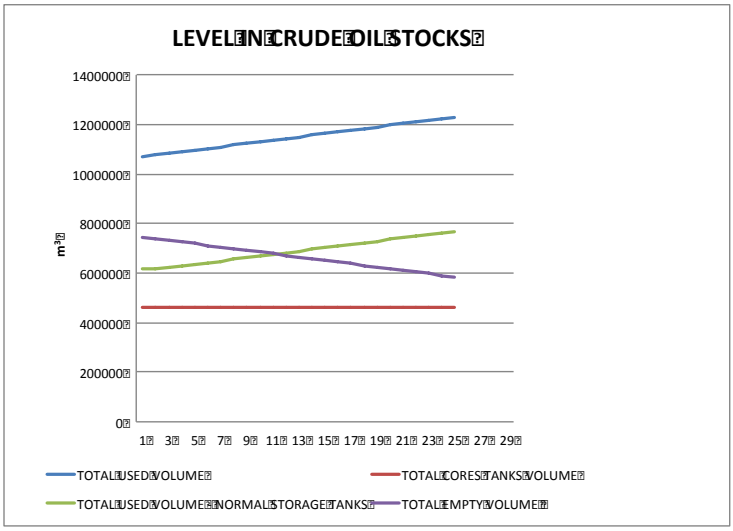


Figure 2: Simulated evolution of crude oil stocks in the Spanish refinery.

References

[1] Bodington, C.E., Baker, T.E., 1990. A history of mathematical programming in the petroleum industry. Interfaces 20, 117–127.

[2] Chryssolouris, G.; Papakostas, N. and Mourtzis, D. (2005) Refinery short-term scheduling with tank farm, inventory and distillation management: An integrated simulation-based approach. European Journal of Operational Research, 166: 812-827.

[3] Paolucci, M.; Sacile, R. and Boccalatte, A. (2002). Allocating oil supply to port and refinery tanks: a simulation-based decision support system. Decision Support Systems, 33:39-54.

[4] Pinto, J.M.; Joly, M. and Moro, L.F.L. (2000). Planning and scheduling models for refinery operations. Computers and Chemical Engineering, 24: 2259-2276.

[5] Shah, N., 1996. Mathematical programming techniques for crude oil scheduling. Computers and Chemical Engineering 20, S1227–S1232.

[6] Simons, R., February 1995. Planning and scheduling BP_s oil refineries—new and old approaches, MP in Action—The Newsletter of Mathematical Programming in Industry and Commerce. Available from <http://www.eudoxus.com/ mpac9502.pdf>.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Special Session 3:

Metaheuristic Optimization

SUITABILITY OF MASA ALGORITHM FOR TRAVELING THIEF PROBLEM

Petra Drevenšek
Faculty of computer and information science, University of Ljubljana, 1000 Ljubljana, Slovenia
Vida Vukašinović
Gregor Papa
Peter Korošec
Computer Systems department, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

Abstract: Real-world problems usually consist of different subproblems which are noticed to be highly dependent between each other. The aim of traveling thief problem introduction was to combine two well studied interdependent problems in order to imitate real-world conditions. In this paper we study the suitability of the MASA algorithm for solving traveling thief problem. The emphasis was on studying the influence of the number of items on the quality performance of the MASA.

Keywords: MASA, treveling thief problem, traveling salesman problem, knapsack problem

1 Introduction

The transportation scheduling, which includes vehicle routing and items pick-up plan, is an important real-world optimization problem. In general, large amounts of goods have to be transported from one place to another as quickly as possible, with the limited number of available vehicles, limited vehicle capacitances, and variable item values. This VRP-based combinatorial optimization problem has been already solved by many approaches; like genetic algorithm (GA) [1, 2] and ant-colony optimization [3, 4].

Most ot the real-world problems posses the following characteristics:

- Combination - they are actually constructed of two or more combined sub-problems;
- Interdependence - the solutions of the sub-problems are interdependent, since the solution of one sub-problem influences the quality of the solution of other sub-problem(s).

Due to the above-mentioned interdependence solving one sub-problem to optimality, without considering the other sub-problem, is not optimal in general. To have in general more optimal approach in solving the transportation scheduling/routing problem we should model it with the travelling thief problem (TTP). The TTP is a novel optimization problem proposed by Bonyadi et al. [5] consisting of two well-known interdependent subproblems. The first is the traveling salesman problem (TSP) [8] of finding Hamiltonian cycle with the smallest cost and the second is the knapsack problem (KP) [9], a problem of picking items with a weight and a value to fill a fixed capacity knapsack while maximizing the overall value. The aim of the TTP is to maximize the value of picked items and minimize the time needed to visit all the cities.

The TTP was proposed with the aim of making a problem more similar to the real-world problems - these usually consist of several interconnected subproblems. Two main approaches to tackle such problems are as follows. In the first approach (atomistic), the problem is divided into more subproblems where each of them is solved separately, while in the second holistic approach all subproblems are solved together as one complex problem. Mei et al. [6] found that the combination of TSP and KP in TTP can lead to much higher computational complexity of the search. To reduce the computational complexity, they proposed to divide the entire search into two stages. In the first stage they perform TSP search in order to find the shortest

Hamiltonian cycle (tour) while in the second stage they perform KP search in order to find the best picking plan under the given tour. In this paper we study the efficiency and suitability of multilevel ant stigmergy algorithm for holistic approach to TTP problem solving.

2 Formal Definition of Traveling Thief Problem

The formal definition of traveling thief problem is as follows. There are n cities for which a distance $d_{i,j}$ between any pair of cities i, j is known. In each city i there are m_i items and each item k has assigned weight w_{ik} and value p_{ik} . The idea of TTP is to visit all cities exactly once, at each visited city decide which items to pick, and end the path in the first visited city. In order to carry picked items, the thief has a rented knapsack with capacity W. The goal of the problem is to maximize the value of picked items and minimize the time needed to travel. The time needed to travel from city i to city j is defined as $t_{ij} = \frac{d_{i,j}}{v_i}$ where v_i is the speed of travel at city i. Speed v_i depends on the current weight of the knapsack W_i, i.e. the weight of items picked in the tour up until the city i is visited. It is calculated as $v_i = v_{max} - W_i v$, where the constant v is defined as $\frac{v_{max}-v_{min}}{W}$ and v_{max}, v_{min} are maximum, minimum possible speeds, respectively. To enhance the interdependence of subproblems, rent R is defined as the fee the thief must pay for each travel time unit. Let $x = (x_1, \ldots, x_i, \ldots, x_n)$, $x_i \in \{1, \ldots, n\}$ be a tour and let $\delta_{ik} = 1$ if item k is picked in the city i and 0 otherwise. Then, the objective function of TTP is defined as

$$Z = \sum_{i=1}^{n} \sum_{k=1}^{m_i} p_{ik}\delta_{ik} - R\left(\frac{d_{x_n,x_1}}{v_n} + \sum_{i=1}^{n-1} \frac{d_{x_i,x_{i+1}}}{v_{x_i}}\right)$$

and is interpreted as the sum of values of all picked items in all cities, minus the travel time multiplied by rent.

3 MASA Algorithm

The Multilevel Ant Stigmergy Algorithm (MASA) is an algorithm for solving discrete multi-parameter problems [10] on graphs, where a search space is a connected, directed, non-weighted, acyclic, rooted and ordered graph. The MASA is based on an ant stigmergy - a form of indirect communication between ants. The algorithm exploits the idea of ants making the pheromone traces in nature, its deposition on paths and its evaporation with time. Ants prefer to choose paths with more pheromones. On that way, they direct their search toward more promising areas. At the beginning of the search, the same amount of pheromone is on all paths in the graph, resulting in ants choosing paths more or less randomly. After each iteration the paths richer with pheromones found are awarded with more pheromones than the rest of them.

The main idea of the MASA is in multilevel approach, which consist of two parts: coarsening of the graph into multilevel graph by merging vertices together in multiple steps and refinement of previously coarsened multilevel graph with optimization being done on every level. An i-th level of coarsened multilevel graph consists of a set of merged vertices after i-th step and edges between them.

In order to solve Traveling thief problem by the MASA algorithm, some basic modifications of the MASA were needed. The characteristics of the modified MASA are as follows. The first level of a coarsened multilevel graph is a graph with vertices as cities and edges as paths between them. In each step the process of coarsening pairs vertices in such a way that it is constructed a half smaller level, where each vertex is a group of two vertices from the previous level. In the case of odd number of vertices in the level, one newly created vertex consist of only one vertex from the previous level. The process is repeated until the level with just a few vertices is reached. Furthermore, in each level l, $1 < l \leq L$, there is an edge between two

vertices u and v if in level $l-1$ there exist at least one edge connecting a vertex from the set of vertices, which are in the next step merged into u, and a vertex from the set of vertices, which are in the next step merged into v.

Algorithm 1: Pseudocode of MASA algorithm

```
input  : Graph G = (V, E) where V is a set of vertices representing cities and E is a set
         of edges that represent connections among cities.
output: The (near) optimal tour on all vertices of graph G with optimized picking plan.
1 Initialization of multilevel graph M with L levels, where G is its first level;
2 Coarsening of multilevel graph M;
  for each level l of M do
      for each pair P of closest vertices in l do
          combine P into a new vertex at level (l + 1);
          remove nodes in P from consideration for new pairs;
      end
  end
3 Optimization;
  for each level l of M do
      while there are still improvements in best path of level l do
          All ants search a path on level l;
          Ants deposit normalized amount of pheromone on edges in such a way, that
          better paths get more pheromones;
          Additionally deposit extra pheromones on edges of best paths of level l (Elitism);
          Evaporation of pheromones;
          if l == 0 then
              All ants search the (sub)optimal picking plan for their paths;
          end
      end
      Refinement of level l;
  end
```

After the multilevel graph is constructed, the optimization phase is started. The optimization starts at the coarsest level L. At each level l, $1 < l \leq L$, the ant colony optimization is used to find the best path and then the graph is refined by transferring pheromones from the current level l to the level $l-1$. At each level l with n_l vertices, i_l iterations are allowed and N_l ants are created. Each ant is appointed a starting vertex and the next visited vertex is chosen based on the probability rule. In the MASA algorithm, the probability of choosing vertex k after visited vertex j in a given level l is:

$$P(j,k) = \frac{F_{j,k}^{\alpha} \cdot d_{j,k}^{\beta}}{\sum_{i=1}^{n_l} (F_{j,i}^{\alpha} \cdot d_{j,i}^{\beta})},$$

where $0 < \alpha < \infty$, $\beta < \infty$, and $F_{j,k}$, $d_{j,k}$ are the amount of pheromone on edge between vertices j and k and a distance between vertices j and k, respectively.

After i_l iterations, ant colony returns the best path found on the level l. Refinement phase is then started by putting pheromone on level $l-1$ in the following manner. Let by jk denote an edge from vertex k to j. Further, let us assume jk in an edge on the shortest path in level l. In the coarsening part of the algorithm vertices k_1 and k_2 from level $l-1$ are merged into the vertex k in level l, and similarly, vertices j_1 and j_2 from level $l-1$ are merged in vertex j in level l. When putting amount of pheromone in level $l-1$, each edge of this path is processed

by putting some amount of pheromone on all edges between the sub-vertices of vertex. After the optimization and refinement are completed on all levels, the best path on the first level is our solution.

The picking plan, i.e. the set of items we picked, is being constructed and constantly evaluated by the ants. In the beginning of each level, a table T with m rows and n columns is constructed. Here, m represents the number of all items. The value of i-th row, j-th column denotes how good, in percents, is the objective function if the item i is picked j-th in order (this means that the city that contains item i is visited j-th). Every time an ant chooses next city, it also chooses which items to pick in that city. At the start, ants choose more or less randomly, untill table T has been revised a few times. Then, they choose items according to the table.

4 Performance Evaluation

4.1 Experimental Environment

The computer platform used to perform the experiments is based on an AMD Opteron ™ 2.2-GHz processor, 16 GB of RAM, and the Microsoft® Windows® 8.1 operating system. The algorithms are implemented in Sun Java 1.7.

4.2 Benchmark Set

The algorithms were tested on 5 instances of the benchmark set for TTP (see Table 1). The benchmark instances were chosen such that the balance between subproblems TSP and KP is kept, and that the influence of the number of items is tested.

Tabela 1: Description of the benchmark.

	cities	items
TB1	10	18
TB2	10	36
TB3	10	72
TB4	10	144
TB5	20	18

4.3 Parameter Settings

The parameter settings were selected by conducting experimental runs and observing the behaviour of the algorithm. The following control parameters were used:

- the number of ants at level l, N_l is $n_l/2$, where n_l is the number of nodes at level l
- the number of allowed iterations without improvement $i_l = n_l^2/2$. Absolute maximum of allowed iterations without improvement is 100, even if i_l exceeds it.
- parameters weighting the influence of pheromone trails α, and distance on ant's choice of the next node β equals $\alpha = 1$, and $\beta = 0$
- the evaporation of pheromone trails $evap = 0.99$
- the minimum pheromone amount on edges $p_{min} = 1$ and the maximum pheromone amount on edges $p_{max} = 20$

- the number of best solutions kept at every iteration and every level $n_{top} = 3$

4.4 Results

The authors of comprehensive benchmark set for the TTP [7] created two heuristics, which solve each TTP subproblem separately. A simple heuristic (SH) which construct a solution by processing and picking the items that maximize the objective value according to a given tour and a simple $(1 + 1)$ evolutionary algorithm (EA), where in contrast to SH no domain knowledge is used. The results of MASA are compared with the results of EA.

Tabela 2: The optimization results for MASA and EA.

	MASA				EA				
	best	mean	worst	std		best	mean	worst	std
TB1	39.57	16.83	6.04	12.06	28.30	28.30	28.30	0.00	
TB2	90.58	65.11	48.46	14.49	80.68	80.68	80.68	0.00	
TB3	108.71	101.08	84.05	10.22	130.52	130.52	130.52	0.00	
TB4	158.25	118.09	107.64	19.55	151.97	151.08	151.86	0.88	
TB5	-47.52	-71.62	-95.15	17.78	-178.50	-178.50	-178.50	0.00	

The best, mean and the worst objective values for the MASA and the EA algorithms are compared in Table 2. The results show that the MASA is capable of solving these small instances of the TTP. In all five cases the MASA outperforms the EA. However, it is also obvious that with the growing number of items the difference between the MASA and EA is decreasing.

5 Conclusion and Future Work

In this paper we studied suitability of the MASA algorithm for Traveling thief problem. We compared the MASA with the EA on several small instances of the problem. It was shown that the increasing number of items decreases the performance of the MASA.

Besides improving the performance of the MASA for larger instances of the problem, we should also consider the computational improvements in the sense of parallelizing the MASA.

In the future, we also want to tackle the problem in the way it was proposed, i.e. with holistic approach for solving TTP, where interdependencies of TSP and KP are taken into account.

Literatura

[1] Salhi, S.; Petch, R. 2007. A GA based heuristic for the vehicle routing problem with multiple trips, Journal of Mathematical Modelling and Algorithms, 6(4): 591-613.

[2] Saadatseresht, M.; Mansourian, A.; Taleai, M. 2009. Evacuation planning using multiobjective evolutionary optimization approach, European Journal of Operational Research, 198(1): 305-314.

[3] Rizzoli, A.E.; Montemanni, R.; Lucibello, E.; Gambardella, L.M. 2007. Ant colony optimization for real-world vehicle routing problems, Swarm Intelligence, 1(2): 135-151.

[4] Lee, C.Y.; Lee, Z.J.; Lin, S.W.; Ying, K.C. 2010. An enhanced ant colony optimization (EACO) applied to capacitated vehicle routing problem, Applied Intelligence, 32(1): 88-95.

[5] M. R. Bonyadi, Z. Michalewicz, and L. Barone. The travelling thief problem: The first step in the transition from theoretical problems to realistic problems. In IEEE Congress on Evolutionary Computation, pp. 1037–1044. IEEE, 2013.

[6] Y. Mei, X. Li, and X. Yao. Improving efficiency of heuristics for the large scale traveling thief problem. Simulated Evolution and Learning, Lecture Notes in Computer Science, 8886: 631-643, 2014.

[7] S. Polyakovskiy, M. R. Bonyadi, M. Wagner, Z. Michalewicz, and F. Neumann. A comprehensive benchmark set and heuristic for the traveling thief problem. In Proc. Genetic and evolutionary computation conference, GECCO 2014, pp. 477-484, 2014.

[8] S. Lin and B. W. Kernighan. An effective heuristic algorithm for the traveling-salesman problem. Oper. Res., Vol. 21, pp. 498-516, 1973.

[9] H. Pirkul. A heuristic solution procedure for the multiconstraint zero one knapsack problem. Na. Res. Log. (NRL), Vol. 34, pp. 161-172, 1987.

[10] P. Korošec and J. Šilc. (2006). The multilevel ant stigmergy algorithm for numerical optimization. Lect. Notes Comp. Sc., Vol. 3911, pp. 92-99, 2006.

MULTIOBJECTIVE OPTIMIZATION OF PROCESS PARAMETERS IN STEEL PRODUCTION

Bogdan Filipič, Miha Mlakar, Tea Tušar
Jožef Stefan Institute, Department of Intelligent Systems
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
E-mail: bogdan.filipic@ijs.si, miha.mlakar@ijs.si, tea.tusar@ijs.si

Abstract: The paper presents a multiobjective optimization approach to process parameter optimization in continuous casting of steel, which is the most widely used steel production process. The optimization task is to find parameter values such that the target values of the empirical metallurgical optimization criteria are approached as closely as possible, since this in turn results in high quality of the cast steel. The problem is being solved with a multiobjective evolutionary algorithm coupled with a numerical simulator of the casting process. The resulting trade-off solutions are visualized to support decision-making about the preferred solutions.

Keywords: continuous casting of steel, process parameters, product quality, multiobjective optimization, Pareto dominance, DEMO algorithm

1 INTRODUCTION

Contemporary material production strongly depends on numerical methods and computer support. Numerical simulators are a prerequisite for performing computer experiments and enable insight into process evolution. Moreover, coupled with efficient optimization algorithms, they make it possible to automate process parameter optimization, improve material properties, increase productivity and reduce production costs.

An example of a material production process to which modern computational approaches are being intensively applied is continuous casting of steel. Here molten steel is cooled and shaped into various semifinished products. To produce high-quality steel, it is crucial to properly control the metal flow and heat extraction during the process execution. They depend on several process parameters, such as the casting speed and coolant flows. However, finding the optimal values of process parameters is hard because of several obstacles. Above all, the number of possible parameter settings grows exponentially with the number of considered parameters, the criteria are conflicting, and on-site parameter tuning is infeasible as it may be expensive and even dangerous. The simulator-optimizer coupling is a reasonable alternative to deal with the problem.

In the past, a common way of solving optimization problems with multiple objectives was to aggregate the objectives into a single cost function and solve the simplified problem with a suitable single-objective optimization method. Nowadays it is becoming an increasingly popular practice to address such problems in their original multiobjective form. For this purpose, population-based metaheuristics capable of finding sets of trade-off solutions in a single run, such as evolutionary algorithms and particle swarm optimizers, are typically used.

In this paper we report on multiobjective optimization of process parameters on a steel casting machine where the task is to find parameter settings that maximize the quality of the cast steel given the empirically defined quality indicators. An optimization environment consisting of a numerical process simulator and an evolutionary-algorithm-based optimizer, equipped with result visualization capability, was developed and installed at a steel plant where it was evaluated in continuous casting of a selected steel grade. The paper introduces the related work, presents the optimization problem, reports on the experimental setup and the obtained results, and, in the conclusion, summarizes the study and suggests future work.

2 RELATED WORK

Evolutionary multiobjective optimization algorithms (EMOAs) are widely used in industry to optimize various devices and processes. A comprehensive literature survey on the applications of EMOAs in materials science and engineering is presented in [1]. Here we focus on the optimization problems related to continuous steel casting process. The purpose of optimization is to find parameter settings (controls) that result in improved steel quality, reduced defects, minimized bulgings, or optimize specific parameters, such as the lubrication index and peak friction. If the constraints describing the technological requirements result in an empty set of feasible controls, the optimization problem is usually reformulated into finding a control that violates the constraints as little as possible.

To obtain high-quality cast steel, various algorithms and approaches are applied to different (sub)sets of process parameters. In [5], the authors used multiobjective ant-colony system to optimize the billet surface temperatures and the length of the liquid core. Both objectives were calculated as a difference between the actual and the target values. Additional examples of optimizing the casting process performance are presented in [9] and [10] where a genetic algorithm employing a knowledge base of operational parameters is used.

In [3], two variants of an evolutionary algorithm (generational and steady state) and the downhill-simplex method were applied to significantly improve the manual settings of coolant flows. The work continued in [4] where the core length and temperature deviations in the casting process were optimized with an EMOA called DEMO [8]. On a similar problem exhaustive search and DEMO were evaluated with various discretization steps for parameter settings [6]. The results, analyzed in view of effectiveness and efficiency, showed that the most suitable way to solve such optimization problems is to apply a stochastic optimization approach on the finest reasonable discretization.

3 OPTIMIZATION PROBLEM

Continuous casting of steel is a complex metallurgical process that starts with molten steel being transported from an electric furnace and poured into the ladle and further led through the tundish which serves as a buffer for the liquid steel. The material flow continues into the mold. Cooling water flowing through the channels in the walls of the mold extracts heat from steel and initiates its solidification. Liquid steel with a thin solid shell, called the strand, exits the base of the mold and enters the cooling chamber where it is supported by water-cooled rollers and sprayed with water from the wreath and spray cooling systems. Heat extraction and solidification continue, and at the exit from the casting machine solidified steel is cut into billets of the desired length.

The quality of the cast steel depends on appropriate process control, specifically, on proper tuning of the process parameters. According to the empirical knowledge from steel production, the crucial process parameters include the casting speed, the change of the mold coolant temperature, and the coolant flows in the wreath and spray systems. On the other hand, the three key indicators of the process suitability and, consequently, the expected steel quality, are the length of the liquid core in the strand, known as the metallurgical length, the thickness of the solid shell at the mold exit, and the strand surface temperature at the unbending point.

Based on these observations, we formulated a multiobjective optimization problem, involving input variables (process parameters), output variables (quality indicators), and the desired output values. Both input and output variables are also provided with their boundary constraints. The task is to find the input variable settings resulting in the values of output variables as close as possible to the desired values.

Formally, given N input variables and M output variables, feasible solutions are the ones that satisfy the boundary constraints for each input variable x_i, $x_i^{\min} \leq x_i \leq x_i^{\max}$, $i = 1, \ldots, N$,

Table 1: Input variables, their boundary constraints and discretization steps.

Variable	Lower bound	Upper bound	Discretization
Casting speed [m/min]	1.0	1.2	0.01
Change of the mold coolant temperature [°C]	5	9	1
Wreath system coolant flow [l/min]	20	40	5
Spray system coolant flow [l/min]	40	70	5

Table 2: Output variables, their bounds and desired values.

Variable	Lower bound	Upper bound	Desired value
Metallurgical length [m]	10	12	10.5
Shell thickness [mm]	11	17	14
Surface temperature [°C]	1100	1140	1120

and each output variable y_j, $y_j^{\min} \leq y_j \leq y_j^{\max}$, $j = 1, \ldots, M$. Provided that an output variable is feasible, the corresponding optimization criterion (objective) $f_j \in [0, 1]$ is defined as

$$f_j = \frac{|y_j - y_j^*|}{y_j^{\max} - y_j^{\min}}, \tag{1}$$

where y_j^* is the desired vaue of y_j. The goal of otimization is to find the values of input variables that minimize the objectives.

4 EXPERIMENTAL SETUP

The above problem formulation was applied to the continuous casting process conducted at a specific steel plant where the considered steel grade was 70MnVS4. The optimization problem was approached using an integrated simulator-optimizer software environment named VizEMO-Steel [11]. As a simulator a numerical model of the steel casting process [12] was used, and the optimization procedure was the Differential Evolution for Multiobjective Optimization (DEMO) algorithm [8].

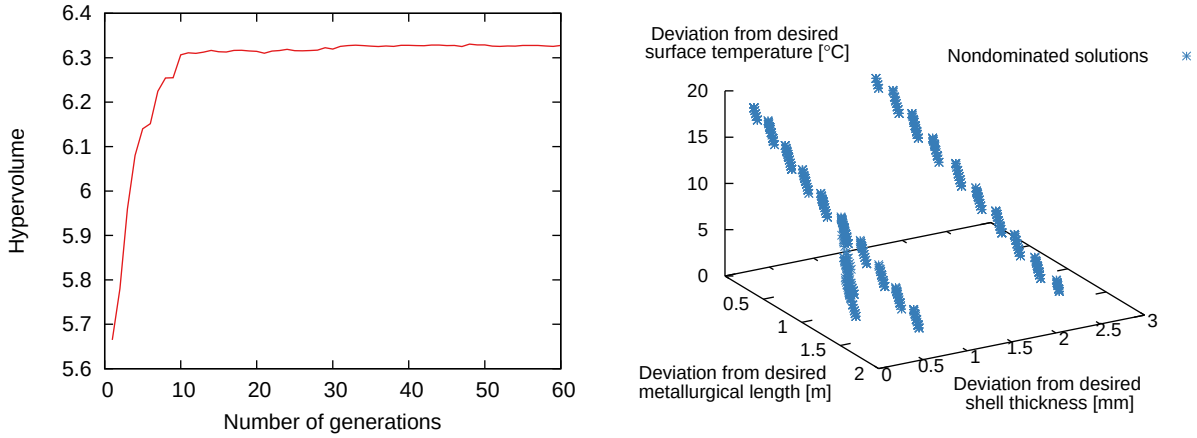
Given the values of input variables together with their boundary constraints and discretization steps (as listed in Table 1), the simulator numerically evaluates the casting process and returns the values of output variables. They are shown in Table 2 together with their lower and upper bounds and desired values as provided by engineers at the plant. Each output value is checked for satisfying the boundary constraints and, if feasible, mapped into a corresponding objective value according to Eq. 1.

DEMO is a population-based algorithm designed for numerical multiobjective optimization. It assumes candidate solutions are encoded as real-valued vectors and creates new solutions from the existing ones using vector addition and scalar multiplication. After creation of a candidate, the candidate and its parent are compared using the Pareto dominance relation. If the candidate dominates the parent, it replaces the parent in the current population. If the parent dominates the candidate, the candidate is discarded. Otherwise, when the candidate and its parent are incomparable, the candidate is added to the population. After constructing candidates for each parent individual in the population, the population size possibly exceeds the predefined value. In this case, the population is truncated to the original size using the nondominated sorting procedure and the crowding distance metric known from NSGA-II [2].

The algorithm parameter values in this study were as follows: population size 50, number of solution evaluations 3000, scaling factor 0.5, and crossover probability 0.3.

5 RESULTS

This section presents and discusses the results of the performed optimization experiment. Figure 1(a) shows the value of the hypervolume indicator over 60 generations of the algorithm execution. We can see that the algorithm is able to converge rather quickly to a hypervolume value that is close to the best value achieved. After generation 30 only minor improvements in the hypervolume indicator can be observed. This implies that in order to save some time without significantly deteriorating the results, we could have stopped the algorithm earlier. Note that the entire run (3000 sequential solution evaluations) took approximately 6.5 days on a 1.9-GHz Intel Xeon server with 32 GB RAM.



(a) algorithm performance (b) nondominated solutions

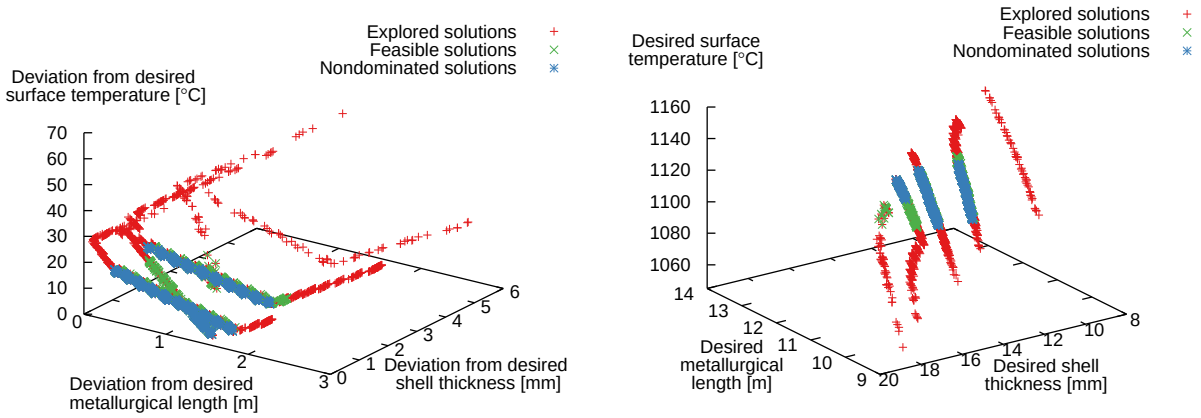
Figure 1: Final results: (a) algorithm performance in terms of the hypervolume indicator and (b) resulting nondominated solutions in the objective space.

Of the 3000 explored solutions, 2090 were feasible and 282 mutually nondominated. The latter are presented in Figure 1(b). The nondominated solutions form three distinct 'strips' in the objective space, which is caused by the discretization of input variables. Figure 2(a) reveals that the shorter strip is actually longer, but only its small part consists of nondominated solutions. Intrigued by the fact that these strips of solutions cross each other in the objective space, we visualized all solutions also in the output variable space. The resulting plot can be seen in Figure 2(b). The solutions are again positioned in strips, but they appear more parallel, which better fits the physics behind the steel casting process. Clearly, the shape of the nondominated front depends on the calculation of objectives from the output variables.

Finally, Figure 3 presents all nondominated solutions in the parallel coordinates plot, which connects the values of the input and output variables for each solution. By interacting with this plot, it is easy to see that the three strips containing nondominated solutions correspond to three different values of the input variable 'Change of the water temperature in the mold'. The parallel coordinates plot is of great help to the decision maker when (s)he needs to select the preferred solutions out of all nondominated ones.

6 CONCLUSION

Quality standards in the steel industry are becoming increasingly stringent, and in continuous casting of steel optimization of process parameters is crucial for achieving high product quality. At contemporary steel plants this optimization is carried out through virtual experimentation involving numerical simulators of the production process and advanced optimization techniques. In this study, the traditional single-objective treatment of the problem that involves multiple



(a) objective space

(b) output variable space

Figure 2: All explored, feasible and nondominated solutions in the (a) objective and (b) output variable space.

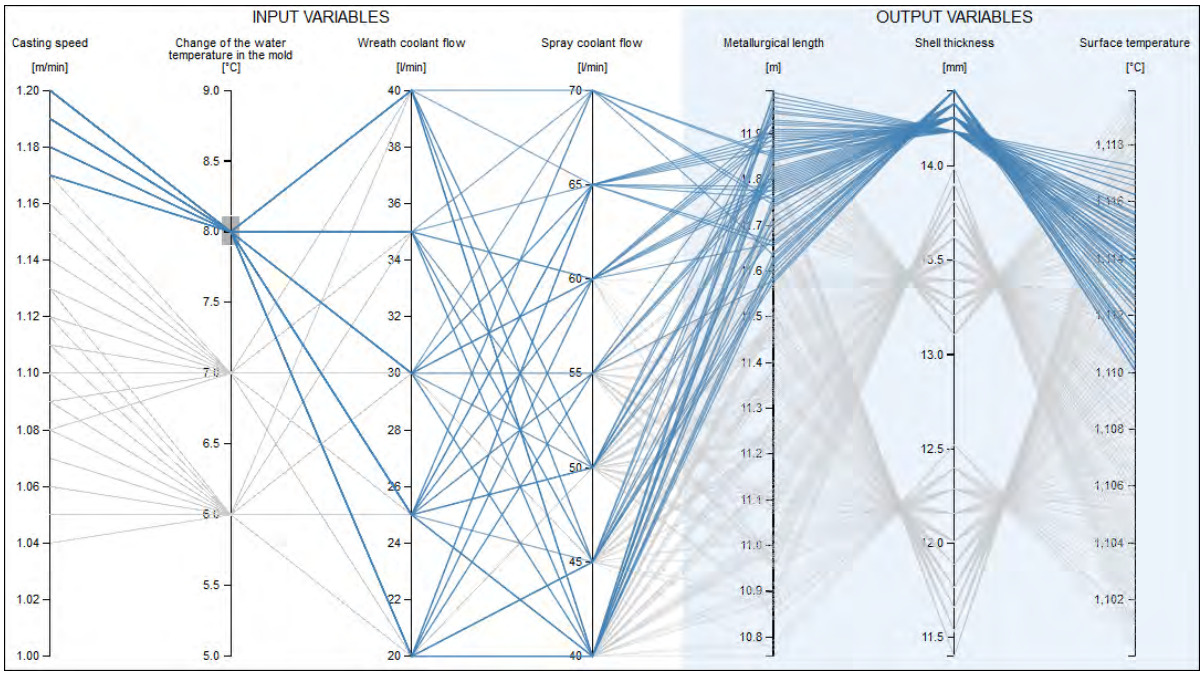


Figure 3: Screenshot of the parallel coordinates plot from VizEMO-Steel showing the connection between input and output variables for the nondominated solutions. Emphasized (blue) solutions correspond to value 8.0 of the input variable 'Change of the water temperature in the mold'.

criteria was replaced with multiobjective optimization as performed by an iterative population-based technique. Specifically, the DEMO algorithm coupled with a numerical process simulator was deployed in tuning critical process parameters with respect to three indicators of the product quality. The study assumes steady-state process conditions, where the optimization results are mainly intended to analyze the process and evaluate the casting machine performance, and not control the process itself.

The simulator-optimizer environment was equipped with a visualization tool and experimentally installed at a steel plant. As illustrated in this paper for a specific steel grade, the resulting approximation sets of Pareto optimal fronts offer an informative insight into process properties and, when appropriately visualized, support decision making about the final parameter setting to be applied. The decision depends on the user preferences that may change from

one order to another.

As a key direction for future work, the presented optimization environment will be extended by involving a surrogate model to reliably approximate the process at a significantly lower computational cost than the currently used numerical simulator. This will increase the efficiency of the optimization procedure, which is an imperative for its regular deployment in practice.

Acknowledgement

The work presented in this paper was carried out under research program P2-0209 and research project L2-3651, both funded by the Slovenian Research Agency. The authors are grateful to Robert Vertnik and Božidar Šarler for making available a numerical simulator of the casting process and the related data.

References

[1] C. A. C. Coello and R. L. Becerra. Evolutionary multiobjective optimization in materials science and engineering. *Materials and Manufacturing Processes*, 24(2):119–129, 2009.

[2] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[3] B. Filipič and T. Robič. A comparative study of coolant flow optimization on a steel casting machine. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2004*, volume 1, pages 569–573, 2004.

[4] B. Filipič, T. Tušar, and E. Laitinen. Preliminary numerical experiments in multiobjective optimization of a metallurgical production process. *Informatica*, 2(31):233–240, 2007.

[5] Z. Ji and Z. Xie. Multi-objective optimization of continuous casting billet based on ant colony system algorithm. In *Proceedings of the Pacific-Asia Workshop on Computational Intelligence and Industrial Application, PACIIA '08*, volume 1, pages 262–266, 2008.

[6] M. Mlakar, T. Tušar, and B. Filipič. Discrete vs. continuous multiobjective optimization of continuous casting of steel. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation, GECCO '12*, pages 587–590, 2012.

[7] T. Robič and B. Filipič. In search for an efficient parameter tuning method for steel casting. In *Proceedings of the International Conference on Bioinspired Optimization Methods and their Applications, BIOMA 2004*, pages 83–94. Jožef Stefan Institute, Ljubljana, 2004.

[8] T. Robič and B. Filipič. DEMO: Differential evolution for multiobjective optimization. In *Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization, EMO 2005*, volume 3410, pages 520–533, 2005.

[9] C. Santos, J. Spim, and A. Garcia. Mathematical modeling and optimization strategies (genetic algorithm and knowledge base) applied to the continuous casting of steel. *Engineering Applications of Artificial Intelligence*, 16(56):511–527, 2003.

[10] C. A. Santos, N. Cheung, A. Garcia, and J. A. Spim. Application of a solidification mathematical model and a genetic algorithm in the optimization of strand thermal profile along the continuous casting of steel. *Materials and Manufacturing Processes*, 20(3):421–434, 2005.

[11] T. Tušar, G. Žgajnar, M. Mlakar, and B. Filipič. VizEMO-Steel v1.3: A program for optimization of continuous steel casting and visualization of results. Technical Report IJS-DP 11808. Jožef Stefan Institute, Ljubljana, 2015 (in Slovene).

[12] R. Vertnik and B. Šarler. Simulation of continuous casting of steel by a meshless technique. *International Journal of Cast Metals Research*, 22(1–4):311–313, 2009.

FACTORY TRANSPORT COSTS OPTIMISATION WITH A NOVEL HEURISTIC ALGORITHM

Tadej Kanduč
Faculty of Information Studies, Laboratory of Decision Support Systems
Ulica talcev 3, Novo mesto, Slovenia
tadej.kanduc@fis.unm.si

Blaž Rodič
Faculty of Information Studies, Laboratory of Decision Support Systems
Ulica talcev 3, Novo mesto, Slovenia
blaz.rodic@fis.unm.si

Abstract: This contribution describes the novel algorithm developed within a factory floor layout optimisation project. The goal of optimisation was the reduction of costs of product transport between machines by reducing the product travel distance and consequently the utilization of human and transport resources. Limiting factors are the costs of relocating the machines to the optimal positions. We have analysed the current state of the manufacturing system by developing a detailed discrete event simulation (DES) model. The DES model was then used to verify and compare optimisation methods and algorithms. We have developed a novel heuristic algorithm, based on force-directed graph drawing algorithms which has produced significantly improved floor layouts.

Keywords: Layout optimisation, heuristics, discrete event simulation, force-directed graphs

1 INTRODUCION

This contribution describes the novel algorithm developed within an internal logistics optimisation project in a special furniture manufacturing company. The algorithm was developed and tested using a discrete event simulation (DES) model of the factory. Our goal was to investigate how the layout of machines on the factory floor affects the efficiency and costs of manufacturing processes. In this paper we investigate the problem by minimising two dependant criteria. The considered criteria are the total costs of one-time machine relocation and the labour costs in transport of products between the machines as a result of changes in machine relocation. To be able to solve the optimisation, we have developed a novel heuristic method that is based on force-directed graph drawing algorithms. The optimisation method significantly reduces total transport costs on a longer time period.

The structure of the paper is as follows: In this section we highlight the optimisation problem and our developed optimisation method, and give a review of literature that investigates related problems. In Section 2 we describe the optimisation problem and the resulting novel layout optimisation method. Section 3 contains results of the optimisation project and discussion of the project results.

1.1 Problem presentation

The client company has been manufacturing furniture for more than half a century. During that time, customer demands changed while size of orders and quality requirements grew. New machines were added to the factory as needed and placed within available floor space. Machine placements were determined by experience of foremen and typically never changed. No systematic analysis and optimisation of factory floor machine layout has been done by the company.

Our task in the project was to develop a better factory floor machine layout that would fit the current production needs and order projections for the next time period. Approximately 120

relevant machines are located in the factory. The company catalogue contains more than 30,000 different products. Every product is manufactured according to the prescribed bill of materials. The technical procedure data include lists of suitable machines for each operation, machine setup and machine operation duration times.

The costs of manufacturing can be reduced by decreasing the need for labour in the transport of products between machines through better machine placement, i.e. factory layout. However, relocation of a machine is a difficult and costly measure and disrupts the manufacturing process. Therefore it makes economic sense to move a machine only if the relocation will considerably reduce product travel distance and consequently the need for labour. Namely, relocating the machines is associated with additional expenses: moving the machine m_i costs g_i amount of currency. Presumably, good candidates for relocation are machines with high product flows.

Let $\tilde{f}_{ij}$ be flow of products between a pair of machines m_i, m_j, $i = 1,2, \dots, N$. Flow $\tilde{f}_{ij}$ represents the total amount of volume of products that was directly transported between these two machines (the data are obtained from discrete simulation model). From the product flow $\tilde{f}_{ij}$ we straightforwardly compute cost flow f_{ij}, i.e., cost to move all products between the two machines for distance of 1 m. The distances between the machines are yet not known and are obtained from the optimisation of the factory floor layout. To reduce the overall costs we need to solve the following optimisation problem

$$\min_{\{\boldsymbol{x}_1, \mathbf{x}_2, \dots, \boldsymbol{x}_N\}} \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{N} f_{ij} \cdot d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} g_i \right),$$

where $\boldsymbol{x}_i$ represents position of machine m_i and d is a distance functional.

We developed a novel method that is based on force-directed graph drawing algorithms. These algorithms, also referred to as 'spring embedders', are methods for visual representation of graphs (networks) [1]. The basic idea of these methods is to apply attractive and repulsive forces between the nodes of the graph. As forces are applied to the graph, the nodes start to move toward a configuration that has a local minimum of overall energy.

The developed method allows arbitrary granularity of locations, arbitrary facility sizes, custom criteria optimisation function. In our method, nodes of the graph represent facilities and edges represent overall costs of products for the corresponding machines. By adding attractive forces to edges and repulsive forces between the nearby nodes, the overall system tends to minimise energy similar to the sought functional $\sum_{\substack{i,j=1 \\ i \neq j}}^{N} f_{ij} \cdot d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} g_i$.

1.2 Previous research (review of literature)

Simulation is commonly used for the evaluation of scenarios [8][4][12]. Several papers deal with factory layout optimisation, with paper [2] stating that multi-product enterprises requires a new generation of factory layouts that are flexible, modular, and easy to reconfigure. Debevec et al. [3] describe a new method (PoVEIR) aimed at optimisation of manufacturing processes in SMEs that is based on a DES simulation model. Layout optimisation problem is identified as a hard Combinatorial Optimization Problem, prompting the use of heuristic methods such as evolutionary optimisation and simulated annealing [11]. A novel particle swarm optimization method is proposed by [6] for intelligent design of an unconstrained layout in flexible manufacturing systems.

Factory layout design optimisation is further discussed in [9][10][5]. Authors [9] propose a new facility layout design model to optimise material handling costs. Sources [10] and [5]

propose genetic algorithm based solutions to respond to the changes in product design, mix and volume in a continuously evolving work environment.

Class of force-directed methods are one of the most commonly used methods for graph drawing (see [1][7] and references therein).

2 OPTIMISATION PROBLEM

Factory floor is described as a region Ω in the plane $\mathbb{R}^2$. We simplify the problem by restricting Ω to the rectangular shape. Let us enumerate the machines by m_i, $i = 1,2,\dots,N$. Position of machine m_i is described by point $\boldsymbol{x}_i := (x_i, y_i) \in \mathbb{R}^2$.

Each machine takes certain amount of space on the floor. This can be conveniently described by a metric rectangular-like ball $B_{r_i}(\boldsymbol{x}_i)$ with radius r_i and centre $\boldsymbol{x}_i$ in ∞-norm L_∞,

$$B_{r_i}(\boldsymbol{x}_i) := \left\{(x,y) \in \mathbb{R}^2 : \|(x,y), \boldsymbol{x}_i\|_\infty := \max\{|x - x_i|, |y - y_i|\} < r_i \right\}. \quad (1)$$

For every pair of machines m_i and m_j, $i, j = 1,2,\dots,N$, we obtain a flow of volume of products $f_{ij} \geq 0$ as a result of the simulation of the manufacturing processes.

Manhattan distance $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between the pair of machines m_i and m_j is defined as

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) := \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_1 = |x_i - x_j| + |y_i - y_j|. \quad (2)$$

Relocation costs for machine m_i from original position $\boldsymbol{x}_i'$ to $\boldsymbol{x}_i$ is defined as a value $g_i > 0$ if $d(\boldsymbol{x}_i, \boldsymbol{x}_i') \neq 0$ and zero cost otherwise. Costs g_i were obtained from company planners who consider several criteria such as switching on and off the machine, construction work, resupplying costs, cleaning, etc.

The optimisation problem of minimising the total distance is described as

$$\min_{\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}} \left(\sum_{i,j=1,\ i \neq j}^{N} f_{ij} \cdot d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} g_i(\boldsymbol{x}_i, \boldsymbol{x}_i') \right), \quad (3)$$

where positions $\boldsymbol{x}_i$ must satisfy the non-intersecting conditions

$$B_{r_i}(\boldsymbol{x}_i) \cap B_{r_j}(\boldsymbol{x}_j) = \emptyset \quad (4)$$

for every $i \neq j$ and machines must stay within prescribed region Ω:

$$B_{r_i}(\boldsymbol{x}_i) \subset \Omega \quad (5)$$

for every $i = 1,2,\dots,N$.

2.1 The algorithm

In this section we present a heuristic optimisation algorithm for solving (3) by assigning positions $\boldsymbol{x}_i$ to machines m_i that is based on force-directed graph drawing methods. Graph G, representing a new machine layout, is constructed gradually during simulation. The machines are placed randomly and then move according to the applied forces in the system and converge to configuration with a local minimum of the overall energy. After the systems converges or the maximum number of iterations is reached, the simulation ends.

Graph $G = (V, E)$ consists of sets of nodes $V = \{v_i\}_i$ and edges $E = \{e_{ij}\}_{ij}$. Every machine m_i is presented as a node v_i. Edges e_{ij} represent direct mutual transactions of products between m_i and m_j.

To every node v_i we prescribe the corresponding repelling force F_{ij} to all other nodes v_j,

$$F_{ij} := H_{ij}\left(\|\boldsymbol{x}_j - \boldsymbol{x}_i\|_\infty\right) \cdot \frac{\boldsymbol{x}_j - \boldsymbol{x}_i}{\|\boldsymbol{x}_j - \boldsymbol{x}_i\|_\infty}, \tag{6}$$

where H_{ij} is a positive monotonically decreasing function. Repulsive forces keep the nodes away from each other since we want sufficient space between the machines. For our model we chose

$$H_{ij}(r) = \begin{cases} c_{\mathrm{r}} \cdot (r_i - r)^2, & r \le r_i \\ 0, & r > r_i \end{cases} \tag{7}$$

and c_{r}, r_i are distance influence parameters.
For every pair of neighbouring nodes v_i, v_j we define a weighted edge e_{ij} with weight $w(e_{ij}) = f_{ij}$. Attractive force between the nodes is defined as

$$G_{ij} := -f_{ij} \cdot (\boldsymbol{x}_j - \boldsymbol{x}_i). \tag{8}$$

Attractive forces move the nodes with large edge weights closer to each other so that the expression $f_{ij} \cdot \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1$, which represents a part of the overall product distance, has a small value.

Relocation costs have a discontinuous jump from 0 to g_i. Forces I_i representing relocation costs need to be modeled in a continuous way,

$$I_i := \begin{cases} g_i \cdot dist(\boldsymbol{x}_i, \boldsymbol{x}_i')/d_0, & dist(\boldsymbol{x}_i, \boldsymbol{x}_i') < d_0 \\ \mathrm{g_i}, & dist(\boldsymbol{x}_i, \boldsymbol{x}_i') \ge d_0 \end{cases}, \tag{9}$$

otherwise the simulation would start to tremble. In our case, the threshold parameter d_0 is set to 5 m.

To keep the nodes inside the prescribed location Ω, we also need to define forces that pull the nodes back to the interior if they are outside the prescribed region Ω,

$$J_i := \begin{cases} 0, & \boldsymbol{x}_i \in \Omega \\ \mathrm{dist}(\boldsymbol{x}_i, \Omega), & \boldsymbol{x}_i \notin \Omega \end{cases} \tag{10}$$

and dist is a functional measuring the distance between the objects.

The simulation starts with region Ω empty. Nodes and edges are added to the graph G gradually during simulation. New position of nodes are computed every iteration as

$$\boldsymbol{x}_i = \boldsymbol{x}_i + \delta \cdot \left(\sum_{v_j \in B_i} F_{ij} + \sum_{v_j \in N_G(v_i)} G_{ij} + I_i + J_i \right), \tag{11}$$

$i = 1,2,\dots,|V|$, where δ is constant parameter, B_i is as a set of vertices in the neighbourhood of v_i and $N_G(v_i)$ is a set of nodes adjacent to v_i.

In the final stage of the optimisation, forces I_i are set to zero if $\mathrm{dist}(\boldsymbol{x}_i, \Omega) \ge d_0$ (the machine should be moved and it is better to be closer to other machines, defined by forces G_{ij}). If $\mathrm{dist}(\boldsymbol{x}_i, \Omega) < d_0$ the machine m_i is returned to original position $\boldsymbol{x}_i'$ and fixed there.

3 RESULTS

For our furniture factory floor layout we tested the algorithm for three observed time intervals: one year, three years and ten years. The distances between the machines were optimised by the algorithm to obtain the lowest overall costs. The results of optimisation are presented in Table 1. As expected, for longer time periods it is better to move more machines and relative cost savings are larger (1 %, 5% and 27 %, respectively).

Table 1: Overall costs and number of moved machines for different factory floor layouts.

Time interval	Layout	Cost	Number of moved machines
1 year	current	49,580 €	0
1 year	optimal	49,000 €	9
3 years	current	137,800 €	0
3 years	optimal	131,500 €	28
10 years	current	459,500 €	0
10 years	optimal	336,400 €	54

Comparison of original and optimised factory floor can be observed on Fig. 1 and Fig. 2, respectively.

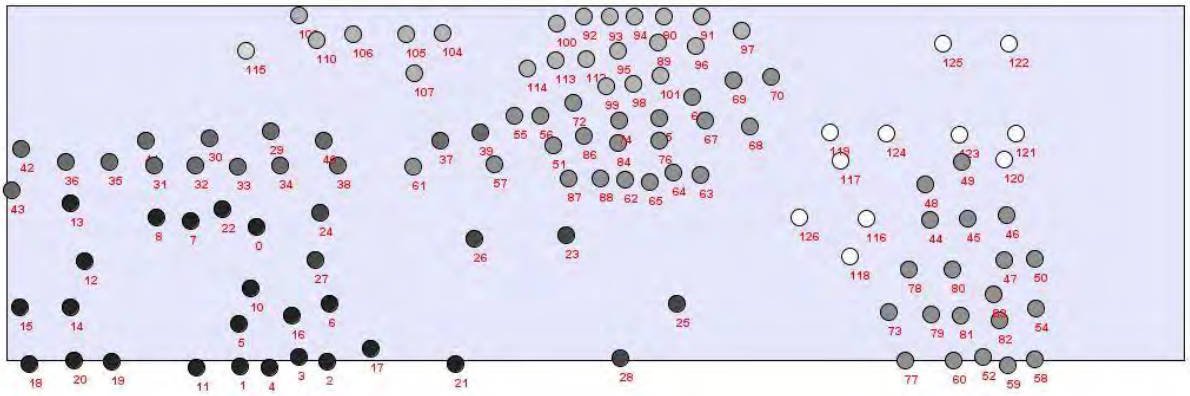


Figure 1: Original machine layout.

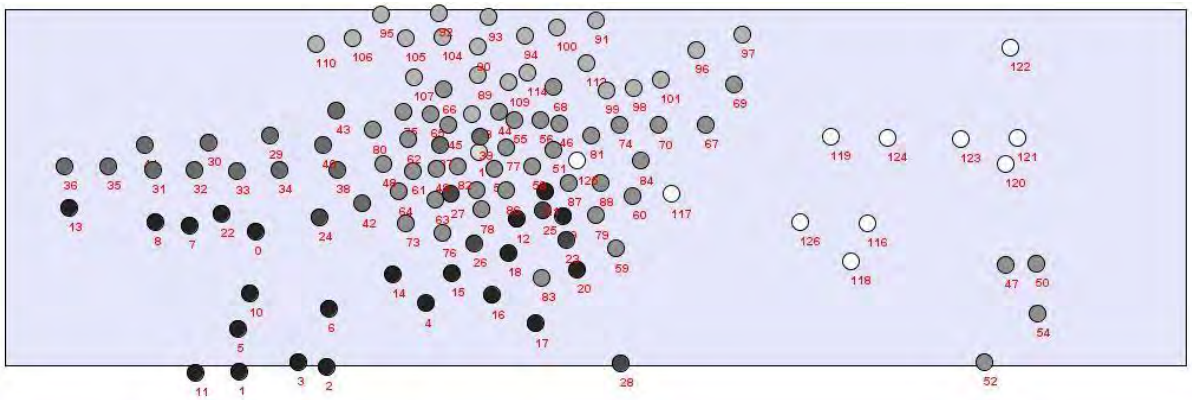


Figure 2: Machine layout after optimisation for 10 years period optimisation.

4 CONCLUSIONS

The described force-directed graph drawing optimisation heuristic method has been used to generate several new factory layouts which have been validated by the factory planners. The optimisation method allows the company to significantly reduce the product transport costs by decreasing the number of workers performing in-factory transport of products. The reduction considers the costs of machine relocation, i.e. only the most relevant machines are moved. Considerable cost reduction can be reached on a larger time scale (e.g. ten years). The factory layouts will be further refined and implemented in the course of ongoing micro-logistic optimisation project.

Acknowledgement
Work supported by Creative Core FISNM-3330-13-500033 `Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

References

[1] Bannister, M. J.; Eppstein, D.; Goodrich, M. T.; Trott, L. (2013). Force-directed graph drawing using social gravity and scaling, *Lecture Notes in Computer Science*, Vol. 7704, 414–425
[2] Benjaafar, S.; Heragu, S. S.; Irani, S. A. (2002). Next generation factory layouts: Research challenges and recent progress, *Interfaces*, Vol. 32, No. 6, 58–76
[3] Debevec, M.; Simic, M.; Herakovic, N. (2014). Virtual factory as an advanced approach for production process optimization, *International Journal of Simulation Modelling*, Vol. 13, No. 1, 66–78
[4] Edis, R. S.; Kahraman, B.; Araz, O. U.; Özfirat, M. K. (2011). A facility layout problem in a marble factory via simulation, *Mathematical and Computational Applications*, Vol. 16, No. 1, 97–104
[5] Enea, M.; Galante, G.; Panascia, E. (2005). The facility layout problem approached using a fuzzy model and a genetic search, *Journal of Intelligent Manufacturing*, Vol. 16, No. 3, 303–316
[6] Ficko, M.; Brezovnik, S.; Klancnik, S.; Balic, J.; Brezocnik, M.; Pahole, I. (2010). Intelligent design of an unconstrained layout for a flexible manufacturing system, *Neurocomputing*, Vol. 73, No. 4, 639–647
[7] Fruchterman, T.; Reingold, E. (1991). Graph drawing by force-directed placement, *Software-Practice and Experience*, Vol. 21, No. 11, 1129–1164
[8] Kljajić, M.; Bernik, I.; Škraba, A. (2000). Simulation approach to decision assessment in enterprises, *Simulation*, Vol. 75, No. 4, 199–210
[9] Krishnan, K. K.; Cheraghi, S. H.; Nayak, C. N. (2008). Facility layout design for multiple production scenarios in a dynamic environment, *International Journal of Industrial and Systems Engineering*, Vol. 3, No. 2, 105–133
[10] Kochhar, J. S.; Heragu, S. S. (1999). Facility layout design in a changing environment, *International Journal of Production Research*, Vol. 37, No. 11, 2429–2446
[11] Sadrzadeh, A. (2012). A genetic algorithm with the heuristic procedure to solve the multi-line layout problem, *Computers and Industrial Engineering*, Vol. 62, No. 4, 1055–1064
[12] Tearwattanarattikal, P.; Namphacharoen, S.; Chamrasporn, C. (2008). Using promodel as a simulation tools to assist plant layout design and planning: Case study plastic packaging factory, *Songklanakarin Journal of Science and Technology*, Vol. 30, No. 1, 117–123

MOBILE TEAMS PROBLEM

Martin Pečar
Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia, m.pecar@ijs.si

Abstract: Recently a paper about the Travelling Thief Problem has sparked some interest in combining already considered problems; it combines Travelling Salesman Problem and Knapsack Problem. This paper introduces the Mobile Teams Problem, which is a combination of Vehicle Routing Problem and Job Scheduling Problem. The Mobile Teams Problem describes real-life routing and scheduling optimization problems that many companies face on a daily basis.

Keywords: Vehicle Routing Problem, Job Scheduling Problem, Travelling Thief Problem, Mobile Teams Problem

1 Motivation

One of the reasons to introduce Travelling Thief Problem (TTP) [2] was the need to define optimization problems that are closer to reality. In many industries, e.g. logistics, we encounter optimization problems, which are much more complex than usual problems in textbooks. For example: when vehicle routes are constructed loading constraints (dimensions of various items, some items may be crushed when stacked, etc.) have to be taken into account.

One of the ways to describe real-life problems is by compositing various already defined problems. The TTP was defined as composition of two very well known problems: the Travelling Salesman Problem and the Knapsack Problem. The TTP has two variants which differ in the way how they are composed: TTP_1 defines a rent which has to be paid for the use of the knapsack, while TTP_2 takes decreasing value of the picked items into account. It is interesting to examine the properties of various compositions – how the types of composed problems and the way they are composed influences the resulting problem. Both subproblems of TTP are NP-hard. All such problems are difficult to solve (no algorithm of polynomial complexity for finding optimal solutions is known), and are also related to each other (there exists a polynomial transformation between each pair of them). However, some problems are more closely related than others – the transformation can be more or less 'natural'. TTP is composed of two quite different problems – what about composing two that are more similar? The Vehicle Routing Problem (VRP) and the Job Shop (Scheduling) Problem (JSP) can (almost) be translated into each other in a rather natural way – vehicles translate into machines, locations into jobs [1]. Each solution proposes the sets and sequence of locations (jobs) for the vehicles (machines).

We are about to define the Mobile Teams Problem (MTP) which is about routing and scheduling vehicles (and workers) so that requested jobs in specific locations are executed. There are many real-life problems, which can be described as MTP - dispatching and scheduling of construction (building sites, repairing) and installation (cable TV and internet) teams. For instance, in the case of construction business, a company may have several construction sites and several vehicles with different functionalities (digging, concrete mixers, transport of material). Sequences of jobs have to be performed at construction sites (e.g. a hole has to be dug, then concrete has to be poured in the hole for the foundation, then the bricks have to be delivered, walls have to be built and finally the roof has to be tiled). In the cable TV installation case, the tasks that need to be done are: drilling holes, inserting cables, connecting cables to set-top boxes and finalizing the installation.

2 Definitions

Let us refresh the definitions of the subproblems.

JSP [6]

- $M = M_1, \cdots M_m$ is a set of machines;
- $J = J_1, \cdots J_n$ is a set of jobs;
- $O = O_1, \cdots O_N$ is a set of operations, belonging to jobs;

We will also use the following notation: J_{O_i} is the job to which operation O_i belongs, M_{O_i} is the machine on which operation O_i is to be processed, t_{O_i} is the start time for operation O_i and p_{O_i} is the processing time for operation O_i. C_{max} is the makespan (the time from the beginning to completion of all operations). On O a binary relation $\rightarrow$ is defined that represents precedence constraints between operations of the same job. If $O_i \rightarrow O_j$, then $J_{O_i} = J_{O_j}$ and there is no O_k satisfying $O_i \rightarrow O_k$ or $O_k \rightarrow O_j$ (operation O_i is the predecessor of operation O_j). Thus, if $O_i \rightarrow O_j$, then $M_i \neq M_j$ by the JSP specifications.

The problem of optimal job shop scheduling is to find a starting time t_{O_i} for each operation $O_i \in O$ such that

$$\max_{O_i \in O}(t_{O_i} + p_{O_i}) = C_{max}$$

is minimised subject to:

$$\forall O_i \in O : t_{O_i} \geq 0 \tag{1}$$

$$\forall O_i, O_j \in O, O_i \rightarrow O_j : t_{O_j} \geq t_{O_i} + p_{O_i} \tag{2}$$

$$\forall O_i, O_j \in O, O_i \neq O_j, M_{O_i} = M_{O_j} : (t_{O_j} \geq t_{O_i} + p_{O_i}) \vee (t_{O_i} \geq t_{O_j} + p_{O_j}) \tag{3}$$

The conditions 2 express precedence constraints which represent technological link-up of operations within the same task. The conditions 3 express machine capacity constraints, i.e. each machine can process at most one operation at a time.

VRP [3]

There is a set of n 'station points' $P_i (i = 1, 2, \cdots, n)$ to which deliveries are made from point P_o , called the 'terminal point'. A 'Distance Matrix' $[D] = [d_{ij}]$ is given which specifies the distance $d_{ij} = d_{ji}$ between every pair of points $(i, j = 0, 1, \cdots, n)$. A 'Delivery Vector' $(Q) = (q_i)$ is given which specifies the amount q_i to be delivered to every point $P_i (i = 1, 2...n)$. The truck capacity is C, where $C > \max q_i$. If $x_{ij} = x_{ji} = 1$ is interpreted to mean that points P_i and P_j are paired $(i, j = 0, 1, \cdots, n)$ and if $x_{ij} = x_{ji} = 0$ means that the points are not paired, one obtains the condition

$$\sum_{j=0}^{n} x_{ij} = 1 (i = 1, 2, \cdots, n)$$

since every point P_i is either connected with P_o or at most one other point P_j. Furthermore, by definition, $x_{ii} = 0$ for every $i = 0, 1, \cdots, n$. The problem is to find those values of x_{ij} which make the total distance

$$D = \sum_{i,j=o}^{n} d_{ij} x_{ij}$$

a minimum under the conditions specified above.

2.1 Mobile Teams Problem (MTP)

Let us define MTP as composition of VRP and JSP. We have:

- n locations $L = \{l_1, \cdots, l_n\}$,
- m jobs $J = \{j_1, \cdots, j_m\}$,

- m_2 operation $O = \{O_1, \cdots, O_{m_2}\}$, which are parts of jobs,
- k vehicles $V = \{v_1, \cdots, v_k\}$,
- l skills $S = \{s_1, \cdots, s_l\}$.

We need to visit all locations, where we need to do some jobs. Each location has at least one operation (part of a job, so $m_2 \geq n$) which also requires a specific skill to be executed. Each driver / vehicle has some skills, which are required for certain operations.

The VRP part is:

- a matrix $T = [t_{ij}]$ shows the travelling time between all pairs of locations.

The JSP part is:

- each operation O_i has specific duration d_i;
- precedence relations $\rightarrow$ to be fulfilled ($O_i \rightarrow O_j$ means that O_i has to be finished before O_j starts).

Additional data, connecting both parts together, are:

- each operation has to be performed at specific locations (O_i at l_l);
- each vehicle (driver) has some of the skills $v_i^S = \{s_p, \cdots, s_r\}$;
- each operation O_i requires certain skill (s_s) to be finished;

Let us also use notation O_l^s meaning this operation has to be performed on location l and requires skill s. $j = \{O_1 \rightarrow \cdots \rightarrow O_i\}$ means that job j consists of sequence of several operations.

The solution to our problem is a schedule, consisting of two types of tasks: travelling between locations ($t(l_1, l_2)$) and performing an operation at a location ($O(O_1, l_1)$). Just the sequence of locations/performed operations is not enough, since we need to make sure the precedence relations are respected - therefore time component needs to be included. We will assume that each task is finished as soon as possible, but there may be some waiting time between tasks. Also, the schedule should tell us the departure and arrival time back to the original location (depot).

Optimization goal of VRP is usually the travelled distance (or overall costs), while the JSP usually optimizes makespan - all jobs should be completed as fast as possible. We are interested in a goal which would be suitable for the composed problem. The goal of the optimization may be minimization of a scalar value (makespan, total working time, etc.) or many Pareto-non-comparable solutions, based on several objectives (number of used vehicles, time, distance, overall costs, etc.). Very often it is really difficult to specify a single criterion for the quality of the solution, so multi-objective optimization can be very useful.

3 Approaches for solving

The VRP and the JSP are NP-hard problems. The MTP problem is composition of both of them. We can reduce MTP to one of its subproblems by specifying some trivial values for the other subproblem; therefore the MTP is (at least) NP-hard, so we need to use heuristic approaches for solving it.

These kind of optimization problems are usually encountered in large companies (e.g. Telcos) which have fleets of several hundreds of vehicles. On the other hand, chains of operations that need to be performed, are typically short. This is why real-life problems are usually solved

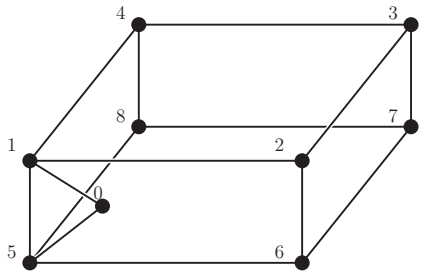


Figure 1: Graph representation of the network, used in the example

with VRP solvers, using additional constraints. However, if jobs have long chains of operations, it is much more difficult to find a good solution. Industry often solves this by breaking jobs into shorter operation chains, and assigning larger crew of workers (with many skills) to a vehicle, thus decreasing the need to schedule the workers. However, there is often a lot of redundant resources (workers) in such cases.

The article [1] discusses the effectiveness of different solvers for different types of problems: if the problem is more similar to pure JSP, then scheduling solvers are more effective, and if it is closer to VRP, then routing solvers are more effective. The 'prevailing' component of the problem probably influences which heuristic will be most effective. The main 'difference' between JSP and VRP is the precedence relation. So - if there are many precedence relations between the jobs, we expect the scheduling approach to be more efficient; if not, routing solvers will probably do the job. However, this still has to be experimentally tested and investigated further.

4 Example

A practical example (Figure 1) of the MTP is presented: we have 8 locations l_1 to l_8 where some operations need to be executed and a starting location l_0; we have a vehicle and driver v_1 with skill s_1 and v_2 with skill s_2; there are several jobs: $j_1 = O_1^1 \to O_5^2$, $j_2 = O_2^1 \to O_6^2$, $j_3 = O_3^2 \to O_7^1$ and $j_4 = O_4^2 \to O_8^1$. Let $d_i = 1$ for all operations, and $T_{12} = T_{23} = T_{34} = T_{41} = T_{56} = T_{67} = T_{78} = T_{85} = 2$ and $T_{15} = T_{26} = T_{37} = T_{48} = T_{01} = T_{05} = 1$.

An optimal solution in this example is represented with a schedule:
v_1: $(l_0, 0), (t(l_0, l_1), 0), (O(O_1^1), 1), (t(l_1, l_2), 2), (O(O_2^1), 4), (t(l_2, l_7), 5), (O(O_7^1), 10),$
$(t(l_7, l_8), 11), (O(O_8^1), 13), (t(l_8, l_0), 14), (l_0, 17)$
v_2: $(l_0, 0), (t(l_0, l_5), 0), (O(O_5^2), 2), (t(l_5, l_4), 3), (O(O_4^2), 6), (t(l_4, l_3), 7), (O(O_3^2), 9),$
$(t(l_3, l_6), 10), (O(O_6^2), 13), (t(l_6, l_0), 14), (l_0, 17)$.
If we take makespan as our optimization goal, we have reached value 17 with this solution.
If we would optimize just a subproblem, a solution for JSP may be
v_1: $O_1^1, O_2^1, O_7^1, O_8^1$
v_2: $O_4^2, O_3^2, O_6^2, O_5^2$
or solution for VRP
v_1: $l_0, l_1, l_2, l_3, l_4, l_1, l_0$
v_2: $l_0, l_5, l_6, l_7, l_8, l_5, l_0$.
However, both these (sub)solutions would construct a solution which is worse that the one presented above. This example shows that optimal solutions for the subproblems don't form an optimal solution for the composed problem, similar as in the TTP.

5 Conclusions and further work

A new problem, MTP, which composes VRP and JSP, is introduced. This problem has many practical applications - routing and scheduling teams in construction, maintenance, installation business. Some approaches for solving the problem have been suggested, but empirical tests still need to be performed. Additional constraints can be added to the simple formulation presented above in order to introduce time windows for specific jobs, different starting points for different vehicles, and other constraints, originating from real-life problems.

It would be interesting to determine what kind of composition of subproblems ensures or forbids optimality of composed problem solutions obtained from optimal solutions of subproblems.

Another important question is: what can we say about the complexity of composed problems? It seems obvious that a composed problem is at least as difficult as any of its subproblems – we can reduce the composed problem to the subproblem. However, the composed problem may become much more difficult. An example is Constrained Shortest Path Problem. If we are looking for a shortest path, or a path with cost under a predefined budget, we know effective algorithms which can find the solution in polynomial time. However, when we combine these two problems, we get Constrained Shortest Path Problem, which is NP-complete. Can we determine what kind of composition will result in specific properties of the resulting problem?

6 Acknowledgements

Many thanks to Tea Tušar, who pointed out TTP to me, and other colleagues for their support and suggestions.

References

[1] Beck, J. C., Prosser P. and Selensky, E. (2003). Vehicle Routing and Job Shop Scheduling: What's the difference?. Proc. of the 13th International Conference on Automated Planning and Scheduling, 267–276.

[2] Bonyadi, M.R., Michalewicz, Z., Barone, L (2013). The travelling thief problem: The first step in the transition from theoretical problems to realistic problems. Evolutionary Computation (CEC), 2013 IEEE Congress on, 1037–1044.

[3] Dantzig, G. B., Ramser, J. H.. (1959). The Truck Dispatching Problem. Management Science, 6 (1): 80-91.

[4] Kouki, Z., Fayech Chaar, B., Hammadi, S., Ksouri, M. (2007). Analogies between Flexible Job Shop Scheduling and Vehicle Routing problems. Industrial Engineering and Engineering Management, 2007 IEEE International Conference on, 880–884.

[5] Selensky, E. (2001). On mutual reformulation of shop scheduling and vehicle routing. Proceedings of the 20th UK PLANSIG.

[6] Šeda, M. (2007). Mathematical Models of Flow Shop and Job Shop Scheduling Problems. World Academy of Science, Engineering and Technology, 1.31: 122–127.

HYBRID METAHEURISTIC ALGORITHM FOR A NEW SCHEDULING PROBLEM IN HYBRID MANUFACTURING ENVIRONMENT

Omer Faruk Yilmaz and Mehmet Bulent Durmusoglu
Istanbul Technical University/Industrial Engineering Department
34367, Macka, Istanbul / TURKEY
{ofyilmaz, durmusoglum}@itu.edu.tr

Abstract: This study presents a hybrid genetic and simulated annealing algorithm (HGSA) to solve scheduling problem in hybrid manufacturing environment. The hybrid manufacturing scheduling problem (HMSP) is a strongly NP-hard combinatorial optimization problem. Nature of this problem is appropriate to handle via evolutionary algorithms (EAs). Genetic algorithm (GA) is mostly known effective population based metaheuristic (P-metaheuristic) of this category and has a powerful diversification property. On the other hand, simulated annealing (SA) is one of the most using single solution based metaheuristic (S-metaheuristic) and has a powerful convergence property. Taking all above mentioned point into consideration, hybrid GA&SA algorithm is developed to handle HMSP problem in this research. A comparison was made between GA and HGSA in terms of solving same problem. Simulation results show that the HGSA has more rapid convergence speed and better searching ability to find an appropriate solution in a reasonable amount of time.

Keywords: Hybrid Manufacturing Systems, Metaheuristic Algorithms, Genetic algorithm, Simulated Annealing Algorithm, Scheduling

1 INTRODUCTION

Hybrid manufacturing systems (HMS) which show typical properties of real manufacturing environment, are consist of manufacturing or assembly cells and functional layout (generally job-shop type). Because of the similarity with real manufacturing environments, scheduling applications of HMS should be handled by practitioners and academicians. The HMSP is not one of the most well-known production planning problem but a typical NP-hard combinatorial optimization problem with strong engineering background [6]. There are some problems which can be seen as other version of this problem such as flexible job-shop scheduling problem (FJSP). The FJSP has been studied in terms of different objectives in literature. Makespan, average flow time and weighted flow time are just some of these objectives. The most common objective of FJSP is makespan. A lot of real-world scheduling problems should be handled as multi-objective because of their nature. Metaheuristic especially evolutionary metaheuristics such as genetic algorithm are very well to obtain good result for multi-objective flow-shop scheduling problems. There are two important terms for multi-objective problems. One of them is diversity preservation and the other one convergence to optimal solution. Genetic algorithm can be combined with local search algorithm to assure these two important property. Especially non-dominated sorting genetic algorithm II (NSGA-II) is very effective with local search to obtain good results in terms of convergence and diversity propoerties (Interested readers may refer to Deb [3].

Like in the multi-objective problems, genetic algorithm is a very efficient metaheuristic for single objective scheduling problems. Genetic algorithm has some strengths and weakness like simulated annealing, but these two metaheuristic algorithms complement each other. GA has strong diversification property due to population and SA has strong neighbourhood search property. Taking all above mentioned point into consideration, hybrid GA&SA algorithm is developed to handle HMSP problem in this research.

The paper develops as follows: The next section describes the basic problem background. Section 3 gives basic structure of GA, SA and proposed GA&SA hybrid algorithm. Section 4

proposes the algorithm. Section 5 describes the experiment performance evaluation of classification algorithm and section 6 shows conclusion of the paper.

2 PROBLEM DESCRIPTION

A novel approach is handled in this study for scheduling problem in hybrid manufacturing environment. There is not any relationship to other studies in terms of novelty, but there can be some similarities in terms of scheduling structure of other type manufacturing systems such as job-shop, flow-shop etc.

The HMSP consists of scheduling N jobs on K machines. For this problem, the following assumptions are made: (1) Each job i is processed at most on one cell/machine at any time, (2) Every cell/machine k can process only one job i at a time, (3) No pre-emption is allowed, i.e. the processing of a job i on a machine k cannot be interrupted, (4) All jobs are available for processing at time zero, (5) The set-up times of the jobs are sequence dependent (6) The machines are available.

In this study, main purpose is to find a sequence σ that minimizes the average flow time [1]. There is set of i,j (i,j=1,…,N) job to be processed in a set of k,l (k,l=1,…,K) cell/machine in the same or different order. The processing time of the jobs on cells/machines are not fixed and dependent on cycle time. The objective is to find a sequence for the jobs to minimize average flow time.

Parameters:

$last_{i,k}$ = If last operation of job i is assigned to cell/machine k then 1, otherwise 0

$d_{i,l,k}$ = If operation of job i is assigned cell/machine l just before cell/machine k then 1, otherwise 0;

$cycmin_{i,k}$ = Minimum cycle time of job i on cell/machine k

$cycmax_{i,k}$ = Maximum cycle time of job i on cell/machine k

q_i = Batch size of job i

Variables:

$p_{i,k}$: Processing time of job i on cell/machine k

$c_{i,k}$: Completion time of job i on cell/machine k

$setupstart_{i,k}$: Starting time of setup of job i on cell/machine k

$setupfinish_{i,k}$: Completion time of setup of job i on cell/machine k

Decision Variables:

$cyc_{i,k}$: Cycle time of job i on cell/machine k

$b_{i,j,k}$: If job i is assigned just before job j on cell/machine k then 1, otherwise 0;

Mathematical model can be expressed as follow:

$$\text{Minimize } f(x) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} last_{i,k} * c_{i,k}}{N} \tag{1}$$

$$c_{i,k} - p_{i,k} \geq \sum_{j=1}^{N} b_{i,j,k} * setupfinish_{i,k} \qquad \forall i,k \tag{2}$$

$$c_{i,k} - p_{i,k} \geq \sum_{l=1}^{K} c_{i,l} * d_{i,l,k} \qquad \forall i,k \tag{3}$$

$$setupfinish_{i,k} = setupstart_{i,k} + \sum_{j=1}^{N} b_{i,j,k} * s_{i,j,k} \qquad \forall i,k \tag{4}$$

$$setupstart_{i,k} \geq \sum_{j=1}^{N} b_{i,j,k} * c_{j,k} \qquad \forall i,k \tag{5}$$

$$cycmin_{i,k} \leq cyc_{i,k} \leq cycmax_{i,k} \qquad \forall i,k \tag{6}$$

$$p_{i,k} = cyc_{i,k} * q_i \qquad \forall i,k \tag{7}$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} p_{i,k} \geq T \qquad \forall i,k \tag{8}$$

Equation 1 shows average flow time of jobs and equations 2-8 are structural constraints for this problem type.

Interested readers may refer to Baker [2] for sequencing and scheduling problems for other type manufacturing systems.

3 PROPOSED HYBRID GENETIC AND SIMULATED ANNEALING ALGORITHM

In this section, we describe the proposed hybrid genetic simulated annealing algorithm for the HMSP. Firstly, we describe basic structure of GA and SA algorithm.

3.1 Genetic Algorithm

GA was developed by Holland [4]. This is an evolutionary P-metaheuristic algorithm which generates solutions using by natural evolution such as selection, crossover, and mutation. Selection mechanism is executed to generate a population using one of the selection method such as tournament selection, roulette wheel etc.. In each population, all solutions are evaluated according to a fitness function and are ranked. Then selected individual of this population are used for crossover to form new off-spring population. In order to allow diversity of search, mutation is used on some individual of population and the GA cycle is repeated [8]. The GA operators are shown in figure 1 as a black text.

3.2 Simulated Annealing Algorithm

SA was developed by Kirkpatrick [5]. This is a probabilistic metaheuristic algorithm for finding a convergence to the global optimum of a given function in a large search space. It was inspired from annealing in metallurgy where temperature controlled cooling of materials is carried out to form crystals with minimum defects and consequently with minimum energy state. Accepting worse solutions allows for a more extensive search for the optimal solution [7]. The newly generated solution is accepted or rejected based on an equation. The SA operators are shown in figure 1 as a red text.

3.3 Hybrid GA&SA

The algorithm uses GA and SA operators, which makes the search-for-optimal-solution process much more effective and efficient. The proposed hybrid GA&SA algorithm is described as follow.

Step 1 Initial population is randomly generated and temperature is initialized

Step 2 After the fitness function evaluations, roulette wheel selection, two point crossover, and polynomial mutation operators are applied to obtain an offspring population

Step 3 Neighbourhood strategies: to conduct a better localized search, mutation and crossover are employed to generate a new offspring population. If new individual is better

than old one then accept the new individual and is put in the place of old one. If new individual is not better than the old one then calculate accepting probability of new individual, If this probability is greater than random number than accept new individual for old offspring population

Step 4 Temperature is decreased and if ending temperature or maximum iteration number is reached then algorithm is ended

The algorithm is applied to an example and compared with GA and SA using same example for HMSP.

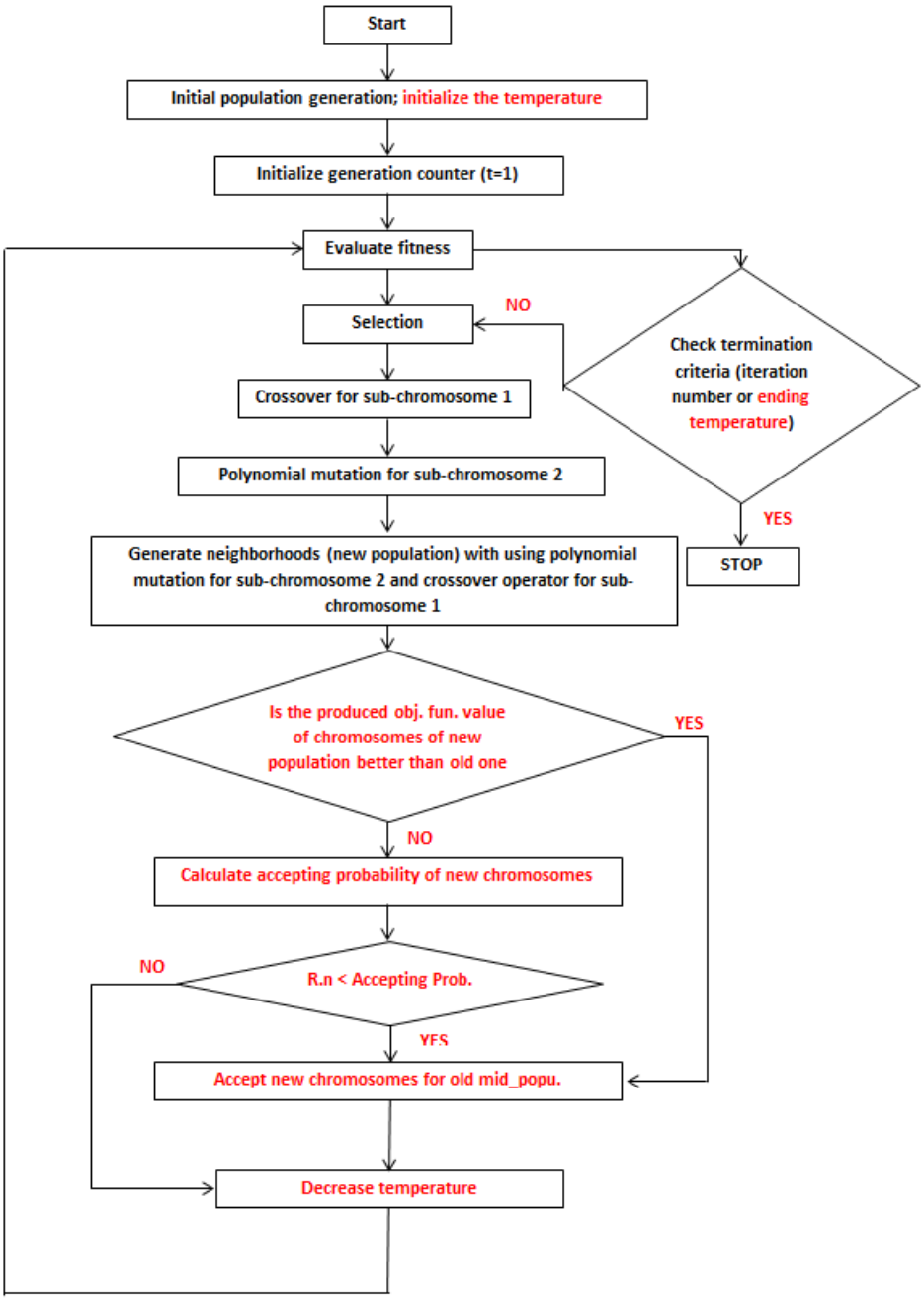


Figure 1: Flow of hybrid of genetic and simulated annealing algorithm

3.4 Solution Encoding

In this study, each solution is defined by a chromosome containing two sub chromosomes in real mode with (number of cells + number of machines)*2 genes (Figure 2). Each sub-chromosome consists of 6 (number of cells + number of machines) sections: for the sub-

chromosome 1, the first 3 sections represent the sequence of products in cells, and the last 3 sections represent the sequence of products in functional layout's machines. For the sub-chromosome 2, the first 3 sections represent the cycle times for cells, and the last 3 sections represent cycle times functional layout's machines. Same chromosomes representation are used for both GA and SA algorithms.

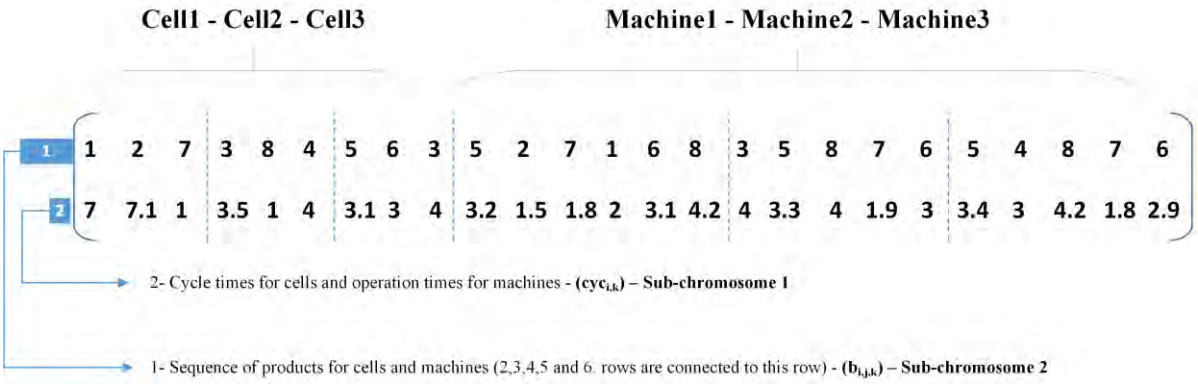


Figure 2: Chromosome representation

4 HYBRID GA&SA and GA APPLICATIONS

8 jobs are to be processed in 6 cells/machines. The operation sequence of products are given in table 1. The cycle times $cyc_{i,k}$ of each job i at each cell/machine k are given in table 2. There are lower and upper limit for cycle time in cells. Cycle times are decision variables for cells. Sequence dependent setup times of products on machines/cells are given table 3. Sequence dependent setup times are symmetric. Each product is not processed in each cell/machine. Because of that reason, some values are not shown in table 3. As an example, the setup time of $b_{1,3,k}$ is 20 minutes for both all cells and machines which product assigned. Because of the symmetric assumption of setup times, the setup time of $b_{3,1,k}$ is also 20 minutes. Time units of the values in table 2 and 3 are in minutes. The batch sizes of products are given in table 4. This problem has been solved using GA and hybrid GA&SA algorithm and results have been compared each other just for 10 runs to show obvious difference between algorithms in Table 5. The parameters which are shown in hybrid GA&SA and GA are indicated in table 6.

Table 1: Operation sequence of products in HMS

Products		**Operations**			
		1	2	3	4
Cell and Machine Order	1	Cell 1	Machine 1	-	-
	2	Cell 1	Machine 1	-	-
	3	Cell 2	Machine 1	Machine 2	
	4	Cell 3	Machine 3	-	-
	5	Cell 3	Machine 1	Machine 2	Machine 3
	6	Machine 1	Machine 2	Machine 3	Cell 3
	7	Machine 1	Machine 2	Machine 3	Cell 1
	8	Machine 1	Machine 2	Machine 3	Cell 2

Table 2 **:** Cycle times for jobs on cells and machines

		Cells			**Machines**		
		1	**2**	**3**	**4**	**5**	**6**
Products	**1**	7/10	-	-	2	-	-
	2	7.1/12	-	-	1.5	-	-
	3	-	8/15	-	4	3.8	-
	4	-	-	10/12	-	-	3
	5	-	-	9.1/14.5	3.2	3.3	3.4
	6	-	-	9.2/10.2	3.1	3	2.9
	7	8/9.6	-	-	1.8	1.9	1.8
	8	-	7.5/14	-	4.2	4	4.2

Table 3: Sequence dependent setup times

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	-	5	20	-	20	5	20	15
2	5	-	20	-	20	5	20	10
3	20	20	-	-	20	20	10	10
4	-	-	-	-	10	10	20	10
5	20	20	20	10	-	10	20	10
6	5	5	20	10	10	-	10	15
7	20	20	10	20	20	10	-	20
8	15	10	10	10	10	15	20	-

Table 4: Batch size of products

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Batch Sizes	100	50	75	100	50	100	75	50

Table 5: Statistical values comparison for algorithms

	Genetic Algorithm Results	**Simulated Annealing Algorithm Results**	**Hybrid Genetic and Simulated Annealing Algorithm Results**
Max.	2999,4	3204	2479
Min.	2964,6	3170	2444,8
Std. Dev.	11,4	9,6	10,7
Mean	2983,1	3186	2458,9
Ave. CPU Time	126,8	88,6	189

The hybrid GA&SA algorithm is used diversification ability GA and local search ability of SA algorithm and has a fast convergence speed to an acceptable point. The results in table 5 show maximum values, minimum values, means, standard deviations and average CPU times (second) of algorithm. Because of each algorithm has an iteration number constraint, these differences can be shown just for ten iteration. Figure 3 shows visualized comparison among them.

Table 6: Parameter for GA and GA&SA algorithms

	GA	**GA&SA**
Population Size	150	200
Crossover Probability	0.9	0.9
Mutation Probability	0.15	0.2
Neighbourhood Rate	-	0.75
Initial Temperature	-	5000
Cool Rate	-	0.75
Ending Temperature	-	10

The parameters combination which are used in GA&SA and GA are shown in table 6. The results of algorithms are shown in figure 3. The results show average of ten run for each algorithm. It can be seen that GA&SA algorithm is the best algorithm among the other algorithms in terms of average flow time. Since the GA&SA algorithm reaches the steady state, both algorithms are terminated at 350 iterations.

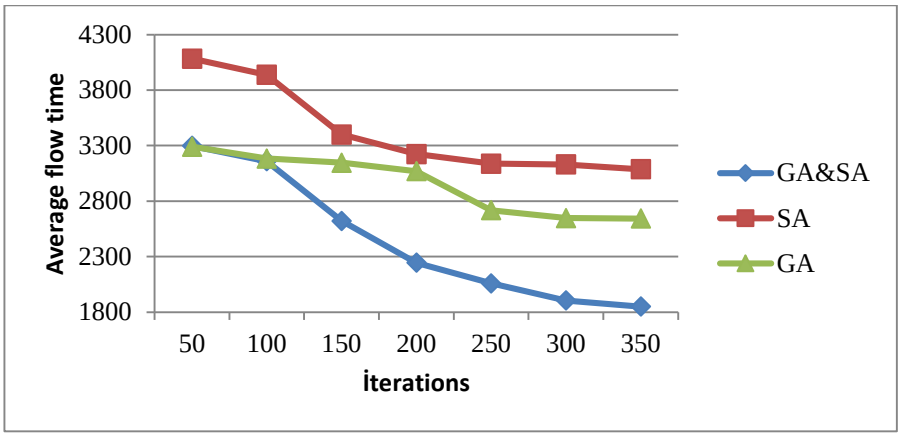


Figure 3: Average flow time for GA, SA and GA&SA algorithm

It can be observed from figure 3 that it is not easy for the single GA to obtain the good results, while the hybrid approach performs better.

5 CONCLUSIONS

In HMSP, with the big number of jobs and machines, the problem's complexity increases. In this paper, the HMSP has been modelled and a hybrid genetic simulated annealing algorithm has been proposed to reach good results. This genetic simulated annealing algorithm has better optimization performance and robustness than single GA and single SA.
Further improvement will be achieved by introducing hybrid GA&SA algorithm to the parameter optimization.

References

[1] Aglan, C., and Durmusoglu, M.B. (2014). Lot-splitting approach of a hybrid manufacturing system under CONWIP production control: a mathematical model. International Journal of Production Research, DOI: 10.1080/00207543.2014.957873.
[2] Baker, K. (1974). Introduction to sequencing and scheduling. Wiley.
[3] Deb, K., and Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Transactions Evolutionary Computation, 6 (2): 182–197.
[4] Holland, J. (1975). Adaptation in natural and artificial systems. Ann Arbor: University of Michigan press.
[5] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220(4598): 671–680.
[6] Satoglu, S.I., Durmusoglu, M.B. and Ertay, T. (2010). A mathematical model and a heuristic approach for design of the hybrid manufacturing systems to facilitate one-piece flow. International Journal of Production Research, 48:17: 5195-5220.
[7] Shokouhifar, M. and Jalali, A. (2015). An evolutionary-based methodology for symbolic simplification of analog circuits using genetic algorithm and simulated annealing. Expert Systems with Applications, 42: 1189–1201.
[8] Zameer, A., Mirza, S.,M., and Mirza, N.,M. (2014). Core loading pattern optimization of a typical two-loop 300 MWe PWR using Simulated Annealing (SA), novel crossover Genetic Algorithms (GA) and hybrid GA(SA) schemes. Annals of Nuclear Energy, 65: 122–131.

METAHEURISTIC BASED AUTOMATED EVALUATION METHOD FOR PUBLIC INVESTMENT PROJECTS

Tulug Figen Yilmaz
METU Computer Engineering Department, Ankara, Turkey
Pinar Karagoz
METU Computer Engineering Department, Ankara, Turkey

Abstract: The public investment projects are essential to improve the state economy and life of people. Since the resources are limited, developing automated project performance evaluation systems are required. In this work, we propose a metaheuristic algorithm based approach for automated project evaluation. Project evaluation involves selecting the attributes effective for evaluation, determining the weights of the selected attributes and determining cut-off values for project evaluation. In the attribute selection part of this model, the effect of supervised learning based models are elaborated. For determining attribute weights and cutoff values, we build a Genetic Algorithm (GA) based solution. The proposed system is tested on the evaluation process of 493 World Bank projects. The results are compared with the classical project evaluation technique using the Analytical Hierarchy Process (AHP) and classification based models. Experimental results show that the proposed approach provides considerable improvement for project evaluation accuracy.

Keywords: project evaluation system, metaheuristic computation, supervised learning

1 INTRODUCTION

Public investment is capital expenditure from government resources on physical infrastructure and soft infrastructure with a productive use that extends beyond a year [8]. In spite of high need for public investment projects, resources are limited. Therefore, there has been increased requirement to support the accountability and efficiency of the projects. In general, to be able to provide efficiency and effective use of the resources, project performance evaluation systems are needed [9, 6].

Project evaluation system basically involves three phases: selecting performance evaluation indexes, finding the weight of evaluation indexes, and building complete evaluation model. The first phase involves deciding as to which attributes are used to evaluate a project. After that, weights of these indexes are determined and assessment methods are applied to carry out evaluation result. Conventionally, these steps are either manually conducted or performed through Analytical Hierarchy Process (AHP) based methods, which involve manual intervention of domain experts.

In this study, we propose an automated project evaluation approach that is based on a metaheuristic algorithm. Metaheuristic algorithms are proven to provide successful results in several optimization and decision problems [2, 13]. Within this work, in order to determine performance evaluation indexes, data mining based feature selection techniques are elaborated. Afterwards, weights of these indexes are determined by utilizing a Genetic Algorithms (GA) based solution.

The basic contributions of this work can be summarized as follows: Although there are several metaheuristic based automated solutions for other domains such as student performance evaluation or inventory evaluation, for public investment project evaluation problem, automated solutions have not been proposed. In this work, we propose a novel metaheuristic algorithm based solution for this problem. Although we use classification based solution as a baseline, since classification based models for project evaluation has not been used and

evaluated before, it is a novel solution developed within this work, as well. We have conducted experimental evaluations to analyze the performance of the proposed approach on the constructed data set in comparison to AHP-based and classification based solutions.

The paper is organized as follows. Section 2 includes a review on the related studies in the literature. In Section 3, the proposed work is described. In Section 4, experiment results are given. In Section 5, conclusion is presented.

2 RELATED WORK

In this section, we summarize related studies on metaheuristic computation and data mining based approaches used in project evaluation problem.

Ai-ling et al. [1] conducted a research on evaluation of engineering projects in the bid system. According to their design, the problem is encoded using GA structure. AHP is used in construction and calculation of fitness function. According to AHP assessment, weight values of bidders are obtained.

Landa-Torres et al. [7] presented a hybrid algorithm for evaluating the internationalization success of a company, based on past data. The algorithm they proposed is composed of Harmony Search (HS) and Extreme Learning Machine (ELM) algorithms. While HS is responsible for forming feature groups, ELM is used for obtaining the objective function for each of such groups. We are inspired by this methodology in feature selection part of our study, however, we use a different encoding procedure and a different algorithm.

Shana and Venkatachalam [11] built a prediction model to estimate the course grade of students by analyzing key performance indicators. In the first part, feature selection techniques are applied to identify the key performance indicators that affect the success of a student in the course. Then, to construct a model, various decision tree classification algorithms are used.

Huping and Chunhua [3] studied to mine a set of general local government performance evaluation indicators out of a massive set of data. In their study, data mining is used for designing evaluation indicators and finding the weights of these indicators. Firstly, Radial Basis Function (RBF) neural network is used as soft clustering technique. Then, hard clustering is done by Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm to mine the local government performance indicators.

3 GENETIC ALGORITHMS BASED PROJECT EVALUATION METHOD

In this section, the general architecture of the proposed project evaluation method is presented and the developed technique is described in detail. The main phases of our system are data gathering, attribute selection, finding the weights of the selected attributes and evaluation. The contribution of this work basically lies in the attribute selection and determining the weights of the attributes by metaheuristic algorithms. In the rest of this section, we describe these phases.

We construct our data set from World Bank projects [10]. In total, 493 projects' data are collected. From the project completion reports, project outcome information, performance ratings, and various attributes are collected manually. Project outcome information has four class labels, which are High Satisfactory (HighS), Satisfactory (S), Unsatisfactory (U), Highly Unsatisfactory (HighU). The collected data includes performance ratings for 25 attributes. Most of the attributes of the project data are ordinal values as in the class label values (HighS, S, U, HighU). As the first step, we converted these ratings into numeric values in [0,1] interval. For example, HighS, S, U, HighU values are mapped to 1.00,0.70,0.30,0.00, respectively. As the next preprocessing step, project IDs are removed from the data. Afterwards, we handled the missing values in the data.

For the attribute selection phase, we used attribute selection feature of WEKA data mining tool. [1] After checking the performance of several attribute selection algorithms, out of 25 attributes, we have selected 7 attributes, which were common in all selections.

In order to model the problem of finding weights of selected attributes and cut-off values, we devised a chromosome structure that is composed of k attribute weights (where k=7 for selected 7 attributes) and 3 cut-off values. Since project outcome information in our dataset has four class labels, chromosome structure has 3 cut-off values to separate the classes. Note that total weight of the attributes is equal to 1, and cut-off values cp_1, cp_2, cp_3 should be set such that $cp_1 < cp_2 < cp_3$.

While using GA for our problem domain, the employed fitness function, crossover and mutation operations are as described below.

Fitness Function: How well the chromosome evaluates the training data set is used as the fitness value of the chromosome. In fitness calculation, we consider the similarity between the predicted class label to the real class label under linear ordering. The fitness of chromosome c is calculated as given in Equation 1.

$$fitness(c) = \frac{1}{n_t} \sum_{p=1}^{n_t} a_p \tag{1}$$

Here, n_t is the number of instances in the training set and a_p is defined as given in Equation 2.

$$a_p = \begin{pmatrix} 1, & \text{if evaluation(c, p) = realValue(p)} \\ 0.4, & \text{if | evaluation(c, p) - realValue(p) | = 0.3} \\ 0.1, & \text{if | evaluation(c, p) - realValue(p) | = 0.7} \\ 0, & \text{otherwise} \end{pmatrix} \tag{2}$$

In Equation 2, $realValue(p)$ is the actual outcome of project p and $evaluation(c, p)$ is the evaluated outcome of project p for a given chromosome c. In addition to exact match, the similarity between the predicted class label and real label is considered, as well. As an example of this condition, if $realValue(p)$ is $HighS$ and $evaluation(c, p)$ is S, then a_p is 0.4. Similarly, if $realValue(p)$ is S and $evaluation(c, p)$ is $HighU$, then a_p is 0.1. The evaluation of a project p for given chromosome c is performed as given in Equation 3.

$$evaluation(c, p) = \begin{pmatrix} \text{HighS}, & \text{if } cp_3 \leq ws(c, p) \\ \text{S}, & \text{if } cp_2 \leq ws(c, p) < cp_3 \\ \text{U}, & \text{if } cp_1 \leq ws(c, p) < cp_2 \\ \text{HighU}, & \text{otherwise} \end{pmatrix} \tag{3}$$

where $ws(c, p)$ is weighted sum of project p for a given chromosome c. The equation for weighted sum is as shown in Equation 4.

$$ws(c, p) = \sum_{i=1}^{k} w_i \frac{p_i - min_i}{max_i - min_i} \tag{4}$$

In this equation, p_i is the rating value of project p for attribute i. max_i and min_i stand for the maximum and minimum rating values of attribute i among all projects in the data set.

Crossover: Uniform crossover is applied on the randomly selected pairs of chromosomes with probability p_c. In order to fulfill problem specifications, crossover operation is applied under control. This operation satisfies the problem constraint so that sum of attribute weights in one chromosome is 1 and value order of cut-off points in one chromosome is maintained.

[1] http://www.cs.waikato.ac.nz/ml/weka/. In WEKA, ReliefFAttributeEval attribute evaluator and Ranker search method; CfsSubsetEval attribute evaluator and GreedyStepwise search method; CfsSubsetEval attribute evaluator and BestFirst search method are used.

Table 1: GA based Algorithm Accuracy Results

	for DS_1	**for DS_2**
Average	**90.83%**	**89.05%**
St. Dev.	0.93	0.89
Overall Average	**89.94%**	

Table 2: GA Algorithm Accuracy Results without Attribute Selection (on data Set DS_1)

	for DS_1
Average (without Attribute Selection)	90.30%
St. Dev. (without Attribute Selection)	1.16
Average (with Attribute Selection)	**90.83%**
St. Dev. (with Attribute Selection)	1.93

Mutation: In the mutation operator, attribute values in genes are set to 0 or 1 with equal probability. This operator is applied on the chromosomes with probability p_m. After mutation operator is applied, chromosome is normalized in order to preserve the specification that the sum of attributes' weights in a single chromosome is 1.

Elitism, which refers to the mechanism that the fittest chromosome always survive to the next generation, is employed in the method.

4 EXPERIMENTAL EVALUATION

Projects are evaluated by using the selected attributes, attribute weights and cut-off values learned by the algorithms explained in the previous section. Firstly, weighted sum value for a given project p is calculated according to Equation 4. The best chromosome is used in the calculation of weighted sum value for a given project p. After weighted sum value for given project p is calculated, it is evaluated according to Equation 5. In this equation, used cut-off values are as in the best chromosome.

$$evaluation_{eval}(p) = \begin{pmatrix} \text{HighS}, & \text{if } cp_3 \leq ws_{eval}(p) \\ \text{S}, & \text{if } cp_2 \leq ws_{eval}(p) < cp_3 \\ \text{U}, & \text{if } cp_1 \leq ws_{eval}(p) < cp_2 \\ \text{HighU}, & \text{otherwise} \end{pmatrix} \quad (5)$$

Experiments are conducted on the project data set collected from World Bank projects that contains 493 samples. For the experiments, 66% of the data set is used for training, and the remaining 34% of the data set is used as the test data. Training data set contains 325 projects, whereas test data set includes 168 project. We applied this partitioning randomly two times, constructing two data sets, DS_1 and DS_2.

As given in Figure 1, GA based solution has about 90% accuracy on average for both DS_1 and DS_2. We repeat the experiment for DS_1 without attribute filtering (i.e., under all attributes obtained from the projects). The results given in Figure 2 reveal that the attribute selection has positive effect for accuracy.

Currently, AHP is the most commonly used method in project evaluation. Therefore, we compare accuracy performance of the proposed meta-heuristic based approaches with that of AHP. As AHP requires expert opinion, we consulted two project evaluation experts for attribute weights. According to expert opinion, weights of selected 7 attributes are determined. Total of 493 projects are evaluated according to these weights and weighted sum method. 254 projects are estimated correctly and true estimate percentage of this technique is 51.5%.

Table 3: Classification based Project Evaluation Accuracy Results

Classifier	DS_1	*10-fold cross valid.*	*Average*
SMO	86.9048%	89.4523%	88.1786%
NaiveBayes	88.6905%	89.0467%	88.8686%
NaiveBayesSimple	88.6905%	89.0467%	88.8686%
RBFNetwork	77.3810%	81.7444%	79.5627%
SimpleLogistic	88.0952%	89.2495%	88.6724%
MultilayerPerceptron	82.1429%	86.6126%	84.3778%
lazy-IB1	85.1190%	84.7870%	84.9530%
MultiClassClassifier	82.7381%	88.4381%	85.5881%
DecisionTable	82.1429%	88.6410%	85.3920%
BFTree	84.5238%	86.6126%	85.5682%
ZeroR	75.5952%	78.9047%	77.2500%
OneR	89.2857%	88.4381%	88.8619%
Ibk	83.3333%	87.6268%	85.4801%
RandomForest	88.0952%	89.4523%	88.7738%
REPTree	89.2857%	88.4381%	88.8619%
SimpleCart	89.2857%	88.0325%	88.6591%

As the second baseline comparison, we used classification algorithms for constructing a model. We used WEKA [12] data mining tool for this purpose and performed experiments with different classifiers, SMO, NaiveBayes, NaiveBayesSimple, RBFNetwork, SimpleLogistic, MultilayerPerceptron, lazy-IB1, MultiClassClassifier, DecisionTable, BFTree, ZeroR, OneR, Ibk, RandomForest, REPTree and SimpleCart, which are commonly used in the literature. In this experiment, we used the same set of 7 features that were used in the previous experiments. The experiments are conducted on data set DS_1 and on the original data set under 10-fold cross validation. The results are presented in Table 3. As shown in the table, all the accuracy results by the classification models are below 90%, whereas GA achieved 90.83% on the same data set.

5 CONCLUSION

In this study we focus on the problem of determining weights of the attributes and the cut-off points automatically for project evaluation, which is currently done in non-automated way by domain experts. We proposed a metaheuristic based algorithms for the problem. In addition, we modeled the problem as a classification problem and analyzed the accuracy performance.

The experimental analysis shows that using metaheuristic algorithms and data mining methods for evaluation of projects is feasible and brings improvement in comparison to the project evaluation by conventional AHP-based approach considerably.

We also elaborated on using a classification model for the problem, and used this as one of the baselines. However, classification based approach for project evaluation problem is also new and is a contribution of this work.

References

[1] Ai-ling,D., Zhao-hui, X. (2009). Research of engineering project evaluation based on genetic algorithm, International Conference on Information Engineering and Computer Science (ICIECS), 1 - 4.

[2] Guan, X., Wang, Y., Tao, L. (2009). Machining scheme selection of digital manufacturing based on genetic algorithm and AHP. Journal of Intelligent Manufacturing: 661–669.

[3] Huping, S., Chunhua, H. (2012). Designing china's local government performance evaluation indicators based on data mining: An exploratory study in four municipal governments in Jiangsu. Fifth International Conference on Business Intelligence and Financial Engineering, 486-494.

[4] Imoto, S., Yabuuchi, Y., Watada, J. (2008). Fuzzy regression model of R & D project evaluation. Applied Soft Computing, 8(3):1266 - 1273.

[5] Iniestra, J. G., Gutierrez, J. G. (2009). Multicriteria decisions on interdependent infrastructure transportation projects using an evolutionary-based framework. Applied Soft Computing, 9(2): 512 - 526.

[6] Khalili-Damghani, K., Sadi-Nezhad, S.(2013). A hybrid fuzzy multiple criteria group decision making approach for sustainable project selection. Applied Soft Computing, 13(1): 339 - 352.

[7] Landa-Torres, I.,Ortiz-Garcia, E. G.,Salcedo-Sanz, S., Segovia-Vargas, M. J., Gil-Lopez, S., Miranda, M., Leiva-Murillo, J. M., Ser, J. D. (2012). Evaluating the internationalization success of companies through a hybrid grouping harmony search-extreme learning machine approach. IEEE Journal of Selected Topics in Signal Processing, 388 - 398.

[8] OECD, Draft OECD principles on effective public investment - a shared responsibility across levels of government (2013).

[9] Taylan, O., Bafail, A. O., Abdulaal, R. M., Kabli, M. R. (2014). Construction projects selection and risk assessment by fuzzy AHP and fuzzy TOPSIS methodologies. Applied Soft Computing, 1(0):105 - 116.

[10] The World Bank open knowledge repository, https://openknowledge.worldbank.org/, Accessed: 2014-05-15.

[11] Shana, J., Venkatachalam, T. (2011). Identifying key performance indicators and predicting the result from student data. International Journal of Computer Applications,45–48.

[12] Weka 3: Data mining software in Java, http://www.cs.waikato.ac.nz/ml/weka/, Accessed: 2015-05-16.

[13] Yu, J., Sun, B. (2010). Research of education evaluation information mining technology based on analytical hierarchy process (AHP) and genetic algorithm (GA). (2010). 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 289 -292.

A HYBRID METAHEURISTIC FOR JOB-SHOP SCHEDULING WITH MACHINE AND SEQUENCE-DEPENDENT SETUP TIMES

Hugo Zupan
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 4, 1000 Ljubljana, Slovenia
E-mail: hugo.zupan@fs.uni-lj.si

Niko Herakovič
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 4, 1000 Ljubljana, Slovenia
E-mail: niko.herakovic@fs.uni-lj.si

Janez Žerovnik
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 4, 1000 Ljubljana, Slovenia
E-mail: janez.zerovnik@fs.uni-lj.si

Abstract: A local search heuristics based on "remove and reinsert" neighbourhood for the job-shop scheduling problem with setup times is proposed. Preliminary experimental results are encouraging.

Keywords: Job-shop scheduling, Remove and Reinsert algorithm, iterative improvement, discrete event simulation, optimization, priority rules

1 INTRODUCTION

Over the past few decades, a great number of studies have been made on job-shop scheduling problem (JSSP). JSSP can be regarded as a scheduling problem and it is one of the most challenging combinatorial optimization problems [Pin]. It is of both theoretical and practical interest, c.f. it is highly popular in production industry [HaoLin]. For conventional JSSP, it is usually assumed that all time parameters are known exactly and in deterministic values. An instance of JSSP can be described as follows: we have a set of *n* jobs that need to be operated on a set of *m* machines [Sul]. Each job has its own processing route; that is, jobs visit machines in different orders. Each job may need to be performed only on a fraction of *m* machines, not all of them. The task is to determine a processing order of all jobs on each machine that minimizes the total flow time.

Another usual assumption is that each job can be processed by at most one machine at a time and each machine can process at most one job at a time. When the process of an operation starts, it cannot be interrupted before the completion; that is, the jobs are non-preemptive. The jobs are independent; that is, there are no precedence constraints among the jobs and they can be operated in any sequence. The jobs are available for their process at time 0. There is unlimited buffer between machines for semi-finished jobs; meaning that if a job needs a machine that is occupied, it waits indefinitely until it becomes available. There is no machine breakdown (i.e. machines are continuously available) [NadFat].

Setup times of machines are typically sequence dependent (or SDST), that is, the magnitude of setup strongly depends on both current and immediately processed jobs on a given machine. For example, this may occur in a painting operation, where different initial paint colours require different levels of cleaning when being followed by other paint colours. We also assume that setup is non-anticipatory, meaning that the setup can only begin as soon as the job and the machine are both available [NadFat].

The JSSP is known to be an NP-hard optimization problem [OmbVen]. Therefore, application of metaheuristics for the JSSP is justified when looking for optimal or near

optimal solutions in reasonable amount of time. This paper proposes such an algorithm, based on Remove and Reinsert algorithm (or RaR) [BreZer, PesSch].

It is well known [EksRah] that using discrete event simulation or virtual factory is very effective tool for "what-if" scenarios, for every type of production system. In our case we have transformed real production system with all the features and limitations into virtual factory. The idea is that the metaheuristic proposes an initial and iteratively improved schedules of orders while the discrete event simulation performs "what-if" scenario for each proposed schedule thus providing the quality measure of the schedule. This process is repeated until the metaheuristic can no longer provide better schedule.

The rest of this short contribution is organized as follows. In the next section, the metaheuristics RaR is outlined and its operation is illustrated with an example in Section 3. Results of the first experiments are given in Section 4. Concluding remarks are in Section 5.

2 THE METAHEURISTIC RaR

The proposed metaheuristic is called RaR algorithm. The idea (algorithms were given various names) was successfully applied to the probabilistic traveling salesman problem (PTSP) [Zer95], the asymmetric traveling salesman problem (ATSP) [BreZer] and to the classical resource-constrained project scheduling problem (RCPSP) [PesSch]. The basic idea of the heuristics is very simple, and this may be a reason for good results. It may be rather surprising that a simple heuristics outperforms much more complicated metaheuristics such as are for example the genetic algorithms, but we believe that this phenomena is not that unexpected, see [Zer14] and the references there.

RaR can be regarded as a hybrid metaheuristic that consists of two phases: generating an initial solution, and iterative improvement. In the first phase, a solution is usually constructed by first solving a subproblem optimally and then inserting the remaining jobs, not unlike the well known heuristics arbitrary insertion for TSP does [RosSte]. The iterative improvement phase, rougly speaking, removes some of the jobs from the current schedule, and reinserts them in back into the schedule in arbitrary order. While the the job-shop scheduling with sequence-dependent setup times on one machine is known to be equivalent to the traveling salesman problem (TSP), the complexity of general JSSP-SDST, in particular the influence of each order to the overall load of the machines, motivates several modifications. Adaptations and modifications used for successful application to the problem studied here are explained in some detail below.

Here, we assume that the initial solution is given (or it is just a random sequence or orders) and focus on the iterative improvement phase.

In the inner repeat loop, a subset of m orders is selected, and an optimal permutation of these m orders is found by exhaustive search. It should be noted that, contrary to RaR applied to some other problems, the orders which are not selected are not removed, only their relative position is frozen. After the optimal permutation of the selected m orders is found, these orders are, one by one, removed and reinserted into the solution into the best position, keeping all other relative positions of orders fixed. The loop is repeated until there is no improvement, and then the position of selected m orders is changed. First, the m orders at positions $1...m$ are chosen, then the positions $2...m+1$, and finally the positions $n-m+1...n$ are regarded.

Algorithm RaR

$S \leftarrow$ initial Schedule (N_1, N_2, …, N_n)

repeat

$S'' \leftarrow S$

```
w = 1 (position from which the permutation starts)
x = 0
repeat
  repeat
    S' ← S
    1. Choose (w … w+m-1) orders from S and start with
       permutation on initial state w
    2. Check every solution for every combination from
       permutation table
    3. Choose the best solution and place it in the S
    4. If permutation found better solution than S' then x = 0
    begin (if x = 0)
      P ← Optimum of first step (S)
      Y = m + w (m – number of orders processed with
      permutation)
      while Y < n (number of all orders) do
        P' = P
        q = 1
        repeat
          P'' ← Insert the order form place Y on place q
          q = q + 1
        until q = Y
        Find the best solution (P'')
        P ← Choose (P'')
        Y = Y + 1
      endwhile
    end
    S ← Choose (P)
  until S' /= S
  w = w + 1
  x = 1
until w > n – m
until S'' /= S
```

3 PROBLEM INSTANCE

The heuristics has been tested on a realistic example. To the best of our knowledge, there are no benchmark instances in the literature, we plan to generate a dataset of several instances for a more extensive experiment, results to be reported in the full paper.

We have a production process with 9 work places where operations are carried out. The plan is to produce 10 orders. For each order, the technology procedure and the sequence of operations are known (see Table 1). Raw material goes from material storage and the finished products are stored in the products storage. The order is always moving from machine one towards the machine nine. Orders can be transported between machines only one at a time. If there are several orders at the same operation, then the orders are awaiting a free operation in the buffer of needs located in front of the operation. For each operation there is information about setup time for the machine and operational time for the order. The setup time is sequence dependent on both current and immediately processed order.

We briefly outline the operation of the algorithm on the instance. The first step is using the permutation table. The idea is that from the whole instance, which has n orders, we take m orders ($m < 7$; we use 5) and optimize them (by finding the best schedule of where the chosen m orders are permuted and the other n-m orders' relative positions do not change) using all

possible combinations, which are enshrined in the permutation table. We continue by considering the orders that were fixed in the first step. One by one, these orders are removed from the schedule and reinserted, this time without changing relative positions of all other orders.

Table 1: The matrix of the type and sequence of operations.

	Machine (i)								
Order (j)	**Mach. 1**	**Mach. 2**	**Mach. 3**	**Mach. 4**	**Mach. 5**	**Mach. 6**	**Mach. 7**	**Mach. 8**	**Mach. 9**
N1	X	X	X	X	X				X
N2	X				X	X	X	X	X
N3	X				X	X	X	X	X
N4	X	X	X	X	X				X
N5	X				X	X	X	X	X
N6	X				X				X
N7	X				X				X
N8	X	X	X	X	X				X
N9	X				X				X
N10	X	X	X	X	X	X	X	X	X

From the layout of production the weighted directed graph was made (see Fig. 1), nodes represents machines and weight ($k=a,b,...,l$) represents transport time between machines.

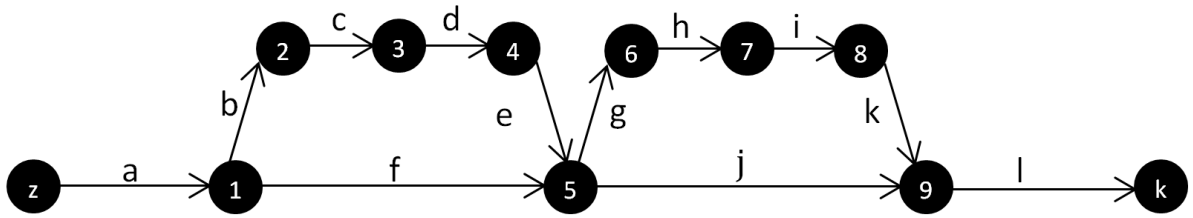


Figure 1: Weighted directed graph of production process.

In our case we have initial schedule of 10 orders (**N1, N2, N3, N4, N5**, N6, N7, N8, N9, N10). Then we choose the first m (5 in our case) orders and find their permutation with minimal cost. (If there are more than one best permutation, the first is taken. In our case, the optimum after first step is: (**N4, N1, N3, N5, N2**, *N6, N7, N8, N9, N10*).

In the second step we start with inserting the other orders into the optimum solution so far. (In particular, here we choose the order *N6* and insert it before order *N4,* before order *N1, ...,* before order *N2.* Thus we have 6 candidate solutions – the one before inserting and for all 5 inserted possibilities. We choose the combination, that gives us the best solution, in this case (**N4, N1, N3,** *N6,* **N5, N2**, N7, N8, N9, N10) and continue by reinserting the next element – order *N7*. We insert order *N7* into 6 possible positions, before orders *N4, N1*, …, *N2* and take the best solution which becomes the initial state before reinserting order *N8*. We do this until we reach the last insertion of the last order (in our case order *N10*). The best solution of this second step is (**N4, N1, N7, N3, N8, N6, N10, N5, N2, N9**).

When we finish the step 2 we repeat the algorithm by using the best solution we got so far.

The algorithm repeats these two steps until it can no longer find an improvement.

After finding a schedule for which no improvement was found by selection the first m orders, the procedure repeats by selecting the orders on positions 2, …,m+1, until no improvement is possible. Then the selection shifts to 3, …,m+2, and so on, until the last selection of orders on positions n-m+1, …,n.

4 COMPUTATIONAL RESULTS

The algorithm was tested on two realistic instances. For the JSSP we had 10 orders and 100 orders. In both cases we compared the algorithm with the genetic algorithm, which is already installed in the Siemens programming environment Plant Simulation ([Siemens]). The results are as follows:

Characteristics of the genetic algorithm (GA):

- Instance with 10 orders: 50 generations: size of generation: 100.
- Instance with 100 orders: 500 generations; size of generation: 100.

Table 2: The results of GA and RaR algorithm.

Algorithm	Total time	… for finding best solution	Quality of the best solution
Instance with 10 orders			
GA	1 min	33 sec	1 day 17h and 8 min
RaR	20 sec	8 sec	1 day 17h and 8 min
Instance with 100 orders			
GA	4h 5 min 22 sec	4h 5 min 22 sec	9 days 23h 45 min
RaR	29 min 30 sec	22 min 41 sec	9 days 13h 28 min

From the results we see that RaR algorithm finds a very good solution in a relatively short time compared to GA. The great advantage of RaR algorithm is that it is not necessary to store large amounts of data, since the algorithm works sequentially, and in almost every step takes only best solution and discards the others.

According to the tests that have been carried out (even those that are not described in here), we can say that RaR algorithm works very well and in quick time gives good solutions.

5 CONCLUSION

This paper proposes Remove and Reinsert heuristics for the job-shop scheduling problem with the sequence dependent setup time of machines. It minimizes the expected average flow time within a reasonable amount of calculation time.

For executions of "what-if" scenarios of the initial schedules, the discrete event simulation (DES) software – Technomatix Plant Simulation was used. A comparison of RaR algorithm with Genetic Algorithm which is built-in module in Technomatix Plant Simulation software. showed that RaR algorithm finds better optimal solution in shorter time compared to Genetic Algorithm.

Motivated by the promising results outlined here, we have also executed "what-if" scenarios for different priority rules inside the production processes. The results showed that by changing priority rules of processing the orders on the machines, we get even shorter flow time of all orders. Details will be given in the full paper.

In our future work, further experiments will be conducted to shorten the calculation time of getting the optimal solutions from the RaR algorithm. Next step would also be to find the optimal sequence of priority rules, or to find the best priority rule with the help of some algorithm.

Finally, as the RaR heuristics performs remarkably well on the JSSP with setup times it may be worth considering the same idea on the other versions of the JSSP problem. In particular because there is a library of benchmark instances [AdaBal, FisTho, Law] which enables a more reliable evaluation of the algorithms.

Acknowledgement
The research has been co-financed by the Operational Programme for Human Resources Development (KROP 13), Priority Axis: Promoting entrepreneurship and adaptability, Main type of Activity 1.1.: Experts and researchers for competitive enterprises.
The third author was supported in part by ARRS, the research agency of Slovenia.

References

[AdaBal] Adams, J., Balas, E. and Zawack, D. (1988). The shifting bottleneck procedure for job shop scheduling, Management Science 34: 391-401.

[BreZer] Janez, B. and Janez, Z. (1999). An approximation algorithm for the asymmetric traveling salesman problem. Ricerca Operativa, 28: 59-67.

[EksRah] Eskandari, H., Rahaee, M. A., Memarpour, M., Hasannayebi, E. and Malek, S.A. (2013). Evaluation of different berthing scenarios in Shahid Rajaee container terminal using discrete-event simulation, Simulation Conference (WSC), 2013 Winter.

[FisTho] Fisher, H and Thompson, G.L. (1963). Probabilistic learning combinations of local job-shop scheduling rules, J.F. Muth, G.L. Thompson (eds.), Industrial Scheduling, Prentice Hall, Englewood Cliffs, New Jersey, 225-251.

[HaoLin] Hao, X., Lin, L., Gen, M. and Ohno, K. (2013). Effective Estimation of Distribution Algorithm for Stochastic Job Shop Scheduling Problem. Procedia Computer Science, 20: 102–107.

[Law] Lawrence, S. (1984). Resource constrained project scheduling: an experimental investigation of heuristic scheduling techniques (Supplement), Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

[NadFat] Naderi, B., Fatemi Ghomi, S. M. T. and Aminnayeri, M. (2010). A high performing metaheuristic for job shop scheduling with sequence-dependent setup times. Applied Soft Computing, Vol(10): 703–710.

[OmbVen] Ombuki, B. and Ventresca, M. (2004). Local search genetic algorithms for the job shop scheduling problem. Applied Intelligence, 21: 99–109.

[PesSch] Pesek, I., Schaerf A. and Zerovnik J. (2007). Hybrid local search techniques for the resource-constrained project scheduling problem. Lect Notes Comput Sci 4771:57–68

[Pin] Pindeo; M. L., (2012). Scheduling: Theory, Algorithms, and Systems, Springer New York Dordrecht Heidelberg London, Springer Science+Business Media, LLC.

[RosSte] Rosenkrantz, D. J., Stearns, R. R. and Lewis P. M. (1977). An Analysis of Several Heuristics for the Traveling Salesman Problem, SIAM J. Comput. 6: 563-581.

[Simens] Tecnomatix Plant Simulation, Siemens PLM Software. http://www.emplant.de/english/fact%20sheet%20plant%20simulation.pdf [accessed on 18/06/2015]

[Sul] Sule, D.R. (1997). Industrial Scheduling, USA: PWS Publishing Company.

[Zer95] Zerovnik, J. (1995). A Heuristics for the Probabilistic Traveling Salesman Problem. Proceedings of the International Symposium on Operational research 1995 (SOR'95), (V.Rupnik, M.Bogataj, eds.), Slovenian Society Informatika, Ljubljana 1995, 165-172.

[Zer14] Zerovnik, J. (2014). Heuristics for NP-hard optimization problems : simpler is better!?. V: IPAVEC, Vesna Mia (ur.), KRAMBERGER, Tomaz (ur.). Pre-conference proceedings of the 11th International Conference on Logistics & Sustainable Transport 2014, Celje, Slovenia, 19-21 June 2014. Celje: Faculty of Logistics.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Special Session 4:

Big Data

ON APPLYING MATHEMATICAL MODELS OF FREQUENCY ASSIGNMENT TO WI-FI THROUGHPUT OPTIMIZATION

Tanja Gologranc[1,4] **Janja Jerebic**[1], **Jaka Kranjc**[2], **Borut Lužar**[2], **Luka Mali**[3], **Janez Povh**[2], **Drago Bokal**[1,4]

Abstract: Frequency assignment has been the motivation behind a large corpus of mathematical research, but the technological progress in telecommunications introduced complexity that makes unclear how to apply existing abstract mathematical models. New insight of the technology is required to close the gap between mathematical optimization and relevant telecommunications problems.

In this contribution, we present a taxonomy of realistic problem instances related to configuring Wi-Fi routers and demonstrate the need for understanding technological details of the problem when developing mathematical models for realistic use cases. We point out that several minor things like the unit in which interference between Wi-Fi routers is measured (dBm or mW) have influence on the optimal configurations of network.

At the end of the paper we provide an instance with a large sample of real Wi-Fi routers and demonstrate that even the simplest, one access point optimization yields significant improvements.

Keywords: frequency assignment, maximum cut, Wi-Fi network, throughput, interference.

1 INTRODUCTION

There will be more than 30 billion connected computers, devices and their parts by 2020, giving rise to the paradigm of Internet of Things, IoT [2]. The demand for throughput is rising and Wi-Fi became de facto technology for data offload from 3G and is standardized as part of the EPC (Evolved Packet Core) in 4G cellular networks. In 2017 alone, over 2.4 billion new Wi-Fi enabled devices will ship, for an installed base of close to almost 20 billions that same year [9]. This means that every human being in 2017 will on average own three Wi-Fi enabled devices, and there will be one Wi-Fi hotspot for every 20 people by 2018 [5].

While cellular radio networks were the key source of frequency assignment technological problems in the past, and their configuration was optimized in several aspects (installing a GSM antenna was a year-long project, while installing a LTE antenna still takes a week), parameterization optimization was completely ignored at Wi-Fi, the low-cost end of the wireless communication. Wi-Fi IEEE 802.11 as a technology was not designed for the scale at which it is now deployed, and the limitations on the number of devices that can simultaneously communicate in a given Wi-Fi cell render optimization of the use of Wi-Fi technology crucial.

There exist two areas that address frequency assignment problem: discrete mathematics and telecommunications. Unfortunately the interrelations among them are scarce. This seems to be a long-standing issue, as similar observations were posed already by the pioneering 1980 work on frequency assignment by Hale [4], see also [10, 7, 1]. A possible reason may be that mathematics thrives in well-defined, structured problems, whereas in telecommunications bibliography, the detailed use cases abound. We propose to circumvent this problem by an attempt to claim the best of both worlds: by classifying the landscape of problems, models, and methods for solving these optimization use cases independently of their underlying mathematical model, our contribution allows the telecommunications researchers and practitioners to position their

[1]Faculty of Natural Sciences and Mathematics, University of Maribor, Slovenia
[2]Faculty of Information Studies, Novo mesto, Slovenia
[3]Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[4]Corresponding authors: tanja.gologranc@um.si, drago.bokal@um.si.

specific use case into a mathematically relevant context. The mathematicians would thus be able to study a well-defined problem in a technologically relevant context, and the telecommunications researchers would be able to optimize the settings of their Wi-Fi network using state-of-the-art mathematically based methodology. While the details of this project extend beyond the limited space of this paper, we describe the flavour of the approach and illustrate its potential.

2 TWO DOMAINS, TWO MAIN MODELS AND TASKS

A central parameter of wireless networks is throughput, i.e. the amount of information the network can transfer to and from connected devices. We focus on optimization of this parameter, which we study using two essentially different models of the wireless network.

The first model is *voxel-access point graph* that structures detailed information on signals of access points in relevant volumes of the studied space (voxels, discussed in Section 2.1). The second model is the *interference graph*, which aggregates the signal information into interference weights of edges between access points - graph vertices. It is discussed in Section 2.2.

2.1 Telecommunications: maximizing throughput in voxel-access point graph

A natural model for throughput maximization is an (enriched) bipartite graph. Its vertices are (i) access points in network (the A-set of the partition) and (ii) voxels, i.e. equal volumes of the space covered by the Wi-Fi network (the V-set of the partition). Edges of the graph connect access points with voxels that receive their signal, the weight of the edge being the strength of the signal. This detailed model is usually used in telecommunications bibliography [8, 6], as it allows to include the information about the problem at the greatest level of detail.

In this model, frequencies are assigned to the access points (A-vertices) maximizing the average throughput in over all the voxels (V-vertices). The Wi-Fi technology postulates that a device connects to the unique access point that has the strongest signal at that point, and other access points using the same or nearby frequency channel are contributing to the noise in the communication. The ratio of the (strongest) signal to the sum of all the noise signals determines the throughput through a piecewise constant function, presented in Table 1.

SINR lower bound (dB)	6.4	8.5	9.4	11.2	16.4	18.2	22.7	24.4
Throughput (Mbps)	6	9	12	18	24	36	48	54

Table 1: A sample conversion of SINR to TP for a specific router. Note the (almost) linear dependence of throughput on the lower bound of intervals.

The detailed model of voxel-access point graph allows for optimization of throughput through integer programming and other global optimization techniques, and as such presents a relevant problem in the domain of combinatorial optimization. However, the structure of the model allows for plethora of constraints and augmentations that vary between use cases, and in the absence of their characterization, the mathematicians tend to study frequency assignment from a different, more classical point of view, as proposed in the next section.

2.2 Mathematics: minimizing interference in access point sinr graph

Early models of frequency assignment [4, 3] have adopted a less detailed approach using *interference graph*. Its vertices are access points, two of them connected with an edge if and only if the areas of the corresponding access points overlap. The edges of a graph are weighted and the weight of the edge between two access points is exactly the amount of interference between these two access points.

Assignment of frequencies to access points corresponds to colouring the vertices of the interference graph. A common assumption is that access points at different frequencies do not interfere, implying that a proper colouring of the graph is equivalent to assignment of frequencies with zero interference and therefore maximum throughput. Obviously, the interference graph presents a model of optimization with information organized in a setting that is much more compact than the setting of voxel-access point graph.

There are three issues inhibiting Wi-Fi frequency assignment using the interference graph. First, the general questions that colouring problems solve discuss the number of colours required to properly colour the graph, but Wi-Fi has a fixed number of 20 MHz channels from 11 to 13 in the 2.4 GHz ISM band. This corresponds to constructing a maximum-weight k-cut in the interference graph, as the edges not in the cut are the edges with remaining interference. Thus applicable frequency assignment problems are related more to maximum (weighted) k-cut, i.e. minimum deficiency k-colouring, than to proper (generalized) colouring of interference graph.

Second, any two Wi-Fi channels at integer distance four or less interfere, and there are at most three different channels that do not interfere pairwise (e.g. 1, 6, and 11). The strength of interference depends on the integer distance between the channels. While proper generalized colouring models allow for finding minimum number of channels with zero interference, the direct weighted max-k-cut problems do not directly solve the Wi-Fi frequency assignment, as the weight of an edge (i.e. strength of the interference) depends on the colours assigned to the end-vertices (i.e. frequencies assigned to access points).

Third, while throughput may have been continuously dependent on interference in analogue stages of telecommunications, the relation between throughput and interference in modern routers is piece-wise constant (cf. Table 1). For this reason, the average throughput may increase at the expense of interference of distant APs, as long as the increasing interference stays within the same constant throughput interval.

In the next paragraph we will define a graph that models frequency assignment problem of fixed wireless network in which max k-cut problem will yield the same solution as the problem of maximizing throughput does in the voxel-access point graph.

Sinr graph G_S is a graph in which vertices are access points and two of them are connected with an edge if and only if there exists a voxel, in which the signal from either of the access points is the strongest signal and the signal from the other access point is more than the background noise. The weights of edges in this graph are set in such way that the problem of maximization of throughput in voxel-access point graph is equivalent to max k-cut problem of G_S.

3 Taxonomy

We propose facilitating the knowledge exchange between Wi-Fi network optimization and combinatorial optimization using a taxonomy of use cases. This taxonomy is under development, and we present the current stage in Table 2.

We distinguish eight main dimensions in which the application cases differ. First two describe the use case: (U) in which context was the use case's data acquired (from single AP (U:P) to country(U:L)). There is a modifier I that distinguishes the cases when a new instance is added into the existing context. Second dimension (P) describing the use case specifies which access points are parameterizable (all (P:A), selected access points while keeping others as fixed noise (P:AP), or all access points with others ignored (P:PI)).

Next two dimensions describe the data analysed within the use case. Following above discussion, the key entity (E) can be either voxel-access point graph (E:V) or interference graph with access points as the key entity (E:A). The data (D) could come from a variety of processes, for instance access points locations could be taken from a blueprint and used in a simulated propagation model, (D:BS).

Dimension		Dimension values	
Description	**Designation**	**Description**	**Designation**
Use case area addressed	U	Single access point	P
		Building	B
		Street segment	S
		City quarter	Q
		City	C
		Region	R
		State / legislative unit	L
Parametrizable APs	P	All parametrizable	A
		All of given provider	AP
		Provider's parametrizable, other ignored	PI
Detailed entity	E	Voxels	V
		Access points	A
Data acquisition process	D	Measuring	M*
		Triangulation of measurements	T*
		Simulating propagation	S*
		Acquired from blueprint	B*
		Artificially generated	A*
QoS/QoE parameter addressed	Q	Throughput	T
		Latency/real time guarantees	L
		Jitter	J
		Packet loss	P
		User satisfaction	U
Optimization criterion	C	Throughput [Mbps]	T
		Interference power [dBm]	D
		Interference power [mW]	W
Optimization space	S	Measuring points	P
		2D area pixels	2
		3D space voxels	3
		Graph	G
Optimization method	M	Global optimization	G
		Approximation	A
		Local optimization	L
		Hybrid approach	H

Table 2: Taxonomy of Wi-Fi channel assignment use cases. * symbols can be combined with others from their group.

Fifth dimension (Q) specifies the QoS/QoE parameter that is addressed by the use case. The most fundamental is throughput (Q:T), sometimes there are latency or even real-time requirements that need to be guaranteed (Q:L), and the most vague parameter that may be studied by, for instance, setting minimum voxel TP, is User Satisfaction (Q:U).

While above dimensions describe telecommunications details of the use case, of which mathematicians may be unaware, the final three address mathematical properties of the model used in parameterization. Optimization criterion (C) is either throughput (C:T), interference in dBm (C:D) or in mW (C:W), (cf. cases discussed in Section 4). The second mathematical dimension is the space over which the criterion is optimized (S). Most relevant case is the set of measuring points (S:P), presumably related to the actual locations of devices, but it can also be the complete 2- or 3-dimensional space (S:2), (S:3), respectively, or the aggregated conflict graph (S:G). The dimension most relevant in comparison of optimization approaches is method used in optimization (M). We distinguish families of global optimization (M:G), approximation (M:A) or heuristic local optimization (M:L), as well as hybrid combinations of these (M:H). Later, this dimension will be detailed to allow for comparison of efficiency of specific methods.

4 SPECIFIC INSTANCE: INDIVIDUAL AP (U:P,P:A,E:V,D:M,Q:T,C:*, S:P,M:G)

In this section we present a real life case of Wi-Fi network optimization, in which individual access points are (ceteris paribus) optimized one at a time. This very basic case is therefore classified as (U:P,P:A), but even very simple optimization technique yields significant improvements. The experiment is based on real data, measured (D:M) in the centre of the city of Ljubljana. We optimize using the measurements directly (E:V) and are interested in largest average throughput of the access point (Q:T). We use three different optimization criteria, throughput (C:T), interference in dBm (C:D) and interference in mW (C:W), and we average the criterion over the space of measured locations (S:P). We seek global optima (M:G).

Data sample included 28,398 access points whose signal was acquired in the 2.4 GHz ISM band. Only 14,937 were close enough to the publically accessible measuring spots to have positive throughput, and we computed their settings that would maximize average throughput or minimize interference in the points at which we measured their signal. We used the other stations data for computing interference only. For each of optimal channels, we computed the average throughput of the optimized access point. We averaged these results over all the 14,937 access points, and they are summarized in Table 3.

Parameter averaged over APs	**Unit**	**Before optimization**	**Throughput (C:T)**		**Interference in dBm (C:D)**		**Interference in mW (C:W)**	
			Abs	**%**	**Abs**	**%**	**Abs**	**%**
Throughput Improvement	Mbps		0.96	100%	0.83	86%	0.74	77%
Throughput	Mbps	5.61	6.57	117%	6.44	115%	6.35	113%
Interference	dBm	-94.50	-95.92	102%	-96.07	102%	-95.98	102%
Interference	mW	1.04E-08	5.83E-09	56%	3.70E-09	36%	4.21E-10	4%

Table 3: Summary of the results of optimizing individual access points in the context of a city quarter.

The table shows effect of the three optimizations on the three relevant parameters. Row "Throughput Improvement" shows improvement of average throughput of a voxel in the three cases. As expected, the largest improvement of 0.96 Mbps results from optimizing average throughput. Optimizing interference in dBm results in second best improvement of 0.83 Mbps, which is only 86% of the former. Worst results are obtained by optimizing interference in mW, 0.74 Mbps or 77% of the maximum improvement. All these improvements are statistically pairwise significantly different at statistical significance $\alpha = 0,05$ The second row shows absolute values of average throughput before and after each optimization, showing that setting an optimum channel on average improves the throughput by 17%. Third and fourth row show values of interference before and after optimization, measured in dBm or mW respectively.

We have also computed the number of equal channels for each subset of optimization cases. At 79% of the APs, the optimum channel is the same for all three optimization cases; at 81% APs, the same channel is assigned in the cases (C:T) and (C:I); at 88% APs, the same channel is assigned in the cases (C:T); and (C:D), and, finally, the same channel is assigned to 89% of APs in the cases (C:D) and (C:W). In each case, there are only few APs whose channel is not affected by the optimization: 13% in the case (C:T), and 10% in the other two cases.

5 DISCUSSION AND CONCLUSIONS

In the paper, we have presented three obstacles preventing wider applications of mathematical frequency assignment models to Wi-Fi parametrization, as the Wi-Fi use cases violate the corresponding assumptions of mathematical models. However, analysis of real measurements shows that even in the simplest case of optimizing a single access point by accounting for interference of neighbouring ones, average improvement of 17% increase in throughput can be obtained. This motivates further research into the mathematics of Wi-Fi parametrization use cases, which needs structure that is provided by a proposed taxonomy, allowing for focus on the specifics of the use cases analysed.

6 Acknowledgements

The authors are grateful to Tilen Marc for interesting discussions. The project has been funded by the Research Agency of Republic of Slovenia, grant L7-5459.

References

[1] K. I. Aardal, S. P. M. van Hoesel, A. M. C. A. Koster, C. Mannino, A. Sassano, Models and Solution Techniques for Frequency Assignment Problems, ZIB-Report 1–40, 2001.

[2] P. Cooney, More Than 30 Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020, Internet of Everything (IoE) report, May 2013.

[3] J. R. Griggs, R. K. Yeh, Labeling graphs with a condition at distance two, SIAM J. Discrete Math. 5 (1992) 586–595.

[4] W. K. Hale, Frequency assignment: Theory and application, Proc. IEEE 68 (1980) 1497–1514.

[5] E. Kaplan, iPass Wi-Fi Growth Map Shows 1 Public Hotspot for Every 20 people on Earth by 2018, iPass & Maravedis-Rethink report, November 2014.

[6] B. Kauffmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, C. Diot, Measurement-Based Self Organization of Interfering 802.11 Wireless Access Networks, 26th IEEE International Conference on Computer Communications, INFOCOM 2007.

[7] A. M. C. A. Koster, Frequency assignment Models and Algorithms, Proefschrift Universiteit Maastricht, 1999.

[8] M. Koyuncu, M. Kazim Gercek, T. Ercan, Host Based Dynamic Throughput Maximization Model for IEEE 802.11 WLAN, Communications in Computer and Information Science, 84 (2010) 315–326.

[9] K. S. Lynn, Wi-Fi Customer Premise Equipment, Wi-Fi Market Data report, 2013.

[10] R. A. Murphey, P. M. Pardalos, M. G. C. Resende, Handbook of combinatorial optimization pp. 295–377, Cluwer Academic Publishers, 1999.

TRAFFIC INCIDENT CLASSIFICATION USING BAYESIAN APPROACH AND MARKOV CHAINS

Lukáš Rapant, Kateřina Slaninová, Jan Martinovič
VŠB - Technical University of Ostrava, IT4Innovations, 17. listopadu 15/2172, Ostrava, Czech Republic
{jan.martinovic,katerina.slaninova,lukas.rapant}@vsb.cz

Abstract: One of the most challenging parts of the traffic modeling is how to model traffic behavior during traffic incidents. One of the possible approaches to this problem is to use historical data to identify typical incidents and use this knowledge to classify future time series. This classification can be utilized in traffic incident length prediction and analysis. This procedure requires solution to several problems. These problems are how to cluster historic time series, how to parametrize these clusters and how to classify new series. Main aim of this article is to propose solution to the second and the third problem. These methods are called Markov chains and Bayesian classification. Data utilized in this article comes from the RODOS project from Czech Republic highways.

Keywords: Time series classifications, Traffic incidents, Bayesian classifier, Markov chains

1 INTRODUCTION

Todays volume of traffic is increasing steadily and is producing various challenges for the field of traffic modeling. One of the products of this increase is growing number of traffic incidents. While it is not that difficult to model free flow traffic (thanks to its periodicity and physical behavior), it is vastly different problem to model traffic during the incident. There are many kinds of incidents from the traffic jam caused by road repairs to chain traffic accident. They are also influenced not only by cause but also by place and time. Therefore it is difficult to use standard micro and macro models. Therefore some other approach is required.

These approaches usually utilize historical data about traffic incidents and exploit their statistical properties. The most useful type of data are FCD (floating car data), because the better represent dynamism of the traffic, than stationary sensors. There are some works, that have already been done in this field. Akira Kinoshita et. al. [2], for example, use probabilistic topic model. Ruimin Li et. al. [4] on the other hand use mixture models to compute traffic incident duration. Yangbeibei Ji et. al. [7] present slightly different approach to traffic incident length predictions in form of cell transmission model.

Like most of the existing work, our method uses historical data to learn the types of the incidents. These data come from Czech Republic highways and are available thanks to the RODOS project [1]. Traffic incidents are represented by speed time series from certain part of the road (called TMC segment). Our approach can be divided into two parts. In the first part, traffic incidents are identified in the historical data and these incidents are clustered by hierarchical clustering with dynamic time warping. This clustering is not the main aim of this article and is described here [3]. In the second part we tried to parametrize these clusters of incidents and propose algorithm to classify future incidents. We have found that Markov chains can be used for parameterization of the clustered time series. The later step is realized by Naive Bayes classification.

[1] http://www.rodos-it4i.cz/defaultEN.aspx

2 THEORETICAL BACKGROUND

2.1 Markov Chains

A Markov chain model can be described as a set of states denoted $S = s_1, s_2, ..., s_n$. The process starts in one of these states and at every time step it moves successively from one state to another. If the model is currently in state s_i, then it moves to another state s_j at the next time step with a probability of p_{ij}. This probability does not depend upon previous states of the model before the current state. This is called Markov property. The probabilities p_{ij} are called transition probabilities and are usually stored in $n \times n$ matrix called transition matrix. It is also possible for the process to remain in the state it is in with probability p_{ii}. Starting state is specified by an initial probability distribution. This can be done by specifying a particular state as the starting state.

Markov chains can be described by a sequence of directed graphs. The edges of n^{th} graph are labeled by the probabilities of moving from the state at time n to the other states at time $n + 1$. This can be denoted as $\Pr(X_{n+1} = x \mid X_n = x_n)$. This information can also be represented by the transition matrix from time n to time $n + 1$. However, Markov chains are usually assumed to be time-homogeneous.

The type of the Markov chain that we shall use in this article is called an absorbing Markov chain. A state s_i of a Markov chain is called absorbing if it is impossible to leave it ($p_{ii} = 1$). A Markov chain is absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state (not necessarily in one step). More information on Markov chains can be found here [1].

2.2 Naive Bayes Classifier

Naive Bayes classifiers are a simple probabilistic classifiers based on application of Bayes' theorem with independence assumptions between the features (more thorough description can be found in [6]). Naive Bayes is a simple concept: models are assign to problem instances, represented as vectors of feature values, where the model labels are drawn from some finite set of models. Basic principle connecting all naive Bayes classifiers states that the value of a particular feature is independent of the value of any other feature, given the model. Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers are proven to work quite well in many complex problems.

From strictly statistical perspective, naive Bayes is a conditional probability model. We are given a problem instance to be classified, represented by a vector of n features $\mathbf{x} = (x_1, \ldots, x_n)$. Classifier assigns to this instance probabilities

$$p(C_k|x_1, \ldots, x_n),$$

for each of k possible models.

However, this calculation may become unfeasible if the number of features n is large or if a feature can take on a large number of values. Therefore some reformulation of the model is needed. The conditional probability can be decomposed by Bayes' rule as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

Because the denominator does not depend on C and the values of the x are given, the denominator is effectively scaling constant. By application of assumption of conditional independence between features we can rewrite equation for model probabilities as:

$$p(C_k|x_1, \ldots, x_n) = \frac{1}{Q} p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

where $Q = p(\mathbf{x})$ is a scaling factor. The naive Bayes classifier combines this model with some decision rule. The most common rule is to pick the hypothesis that is most probable. This rule is known as the maximum a posteriori decision rule (MAP). The corresponding Bayes classifier assigns a model label $z = C_k$for some k by following equation:

$$z = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k). \tag{1}$$

3 ALGORITHM DESCRIPTION

Now let us present you how our algorithm works. As it has been mentioned in Introduction, our algorithm is based on analysis of historical traffic data. These data are represented by speed time series for segments of the road. Progress of our general classification algorithm for traffic time series can be summarized into following steps:

1. Find traffic incidents in historical time series and perform a clustering on these incidents.
2. **Find Markov chains for parametrization of these clusters.**
3. Detect traffic incident in current traffic speed time series.
4. **Classify these incidents by Bayesian classification.**

The main aim of this paper is to present a solution to the third and fifth step of the algorithm. Other steps are currently under development or were already presented here [3].

Cluster parameterization is done by Markov chains. Their main advantage is their simplicity and their low computational complexity. As Markov chains are finite state models, we have to transforms speed values from the time series to a number of states. Based on our experiments, we have decided to use one state for a span of ten km per hour (i.e. one state for speeds 0-9 km/h and so on). Starting state is the state that represents the value of speed just after the speed dropped under the critical level for incident detection. Absorbing state (i.e. state that we consider the end of speed time series for traffic incidents) is one state above the critical speed value. Transition probabilities are simply calculated from number of transitions between states in clustered time series.

Next part of our work is Bayesian classification of the new traffic incidents. We are using the classifier in form presented in Equation 1. It is even simplified by the fact, that we have just a single feature of the data (its time series) so we can drop the product from the equation. As a decision rule, we are using standard maximum posteriori probability rule. A priory probability of the model $p(C_k)$ is calculated from the size of appropriate cluster, models of large clusters of historical time series are more probable than those of smaller clusters. More difficult part of the computation is the likelihood $p(x|C_k)$. It represents probability, that the model generates the data and in general is not exactly known. Therefore it must be simulated. We have used the Monte Carlo approach [5] by drawing n time series from the model C_k and comparing them to the data x. From a large number of these comparison, we can compute the likelihood of the data. This comparison, however, creates another problem. These series will usually be of different length, but it does not mean that they are not similar. They can have only slightly different length but their shape is very similar. We can solve this problem problem by application of dynamic time warping (DTW) for measuring time series similarity. Unlike some other similarity measures (like Euclid,...), DTW can compare time series of different length. If the similarity is greater than some threshold, we can consider that the data can be generated by the model. This entire algorithm for computation of a posteriori probability of a model can be described by following pseudo code:

Algorithm 1 Algorithm for computing a posterior probability of model C_k

```
1: procedure APOSTERIORPROBABILITY(C_k, p(C_k), x, n, treshold)
2:     for i = 1 : n do
3:         ts=simulate(C_k);                          ▷ Monte Carlo simulation
4:         sim=dtwsimilarity(ts,x);     ▷ DTW similarity of data and simulated time series
5:         if sim > treshold then
6:             mc++;
7:         end if
8:     end for
9:     likelyhood=mc/n;                                ▷ Likelihood computation
10:    aposterior=likelihood · p(C_k);          ▷ A posterior probability computation
11: end procedure
```

4 EXPERIMENTAL RESULTS

Our algorithm was tested on time series of speed from Czech Republic highways. Historical data used for training come from the April 2015 and testing traffic incidents come from May 2015. We have identified traffic incidents in these time series and extracted them. Standard DTW clustering performed on traffic incidents from April produced 25 clusters, which we numbered from 1 to 25(small clusters with just one or two incidents were not taken into account). These clusters have been transformed into 25 Markov chain models representing these clusters and numbered accordingly to their originating cluster. Then we have taken time series from May and clustered them by DTW to provide a benchmark (these clusters were again numbered for their identification). We took series from the three largest clusters created by this clustering and tried to classify them using our algorithm and Markov models obtained from April. It can be expected, that all time series from each cluster should be classified to come from the same model. Results can be seen in Table 1.

Number of model representing classified series	Number of originating cluster
1	1
3	1
1	1
1	1
24	1
1	1
6	10
10	10
10	10
10	10
5	5
7	5
5	5

Table 1: Comparison of originating cluster and results of classification

As it is evident from the Table 1, our algorithm can successfully classify most of the time series but not all of them. While the success rate suggests that algorithm is functioning, it is also showing some flaws. Let us analyze the first six time series, which should all belong to the same model, to see the reason behind the misclassified series. In Figure 1 you can see all six time series. Correctly classified ones are black, the one from the second row is dashed gray and the one from the fifth row is dashed black.

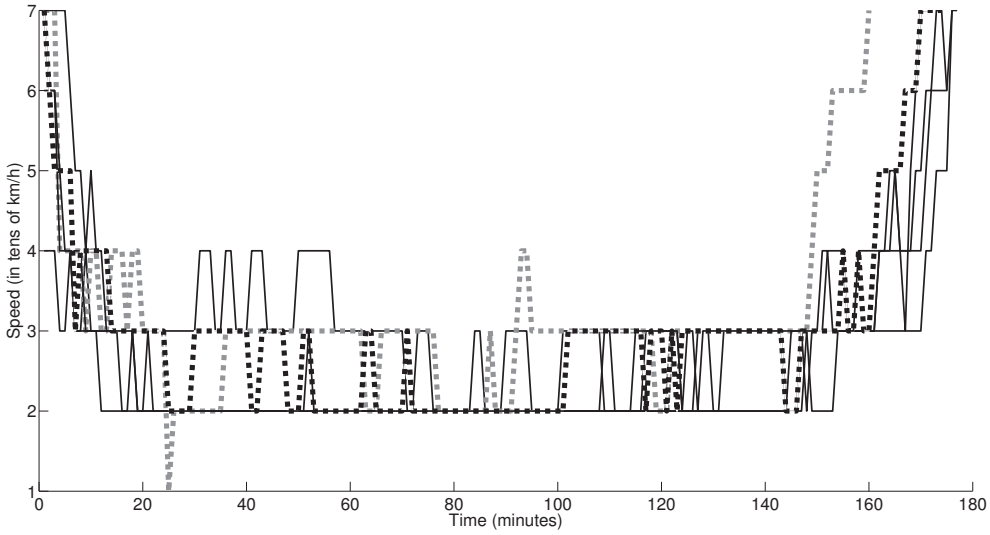


Figure 1: Time series of speed from cluster 1

The dashed gray time series is noticeably shorter than the other ones. It may have been on a border of the cluster and therefore it is possible that some other model describe it better. The dashed black one is, however, very similar to the other ones. Problem here lies with its a posterior probability which is shown in Figure 2 (axis x shows ID number of model and axis y its a posterior probability).

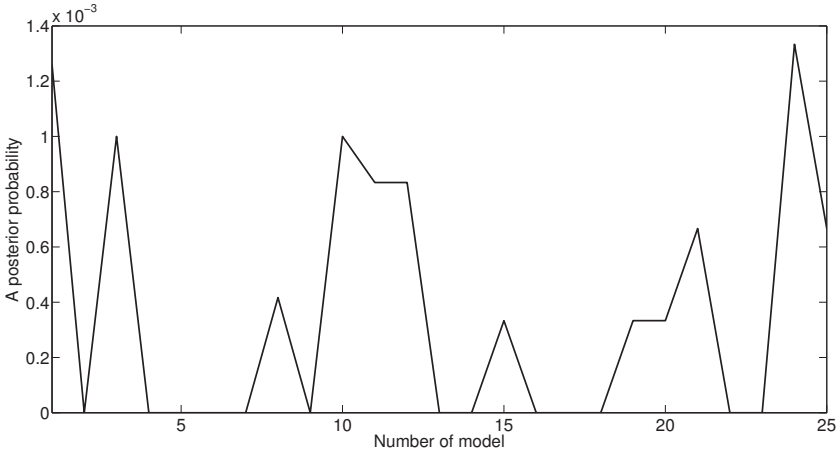


Figure 2: A posterior probability distribution of time series from the fifth row of Table 1

Its MAP peaks at model 24, but it is only marginally higher that a posterior probability at model 1. This time, it is clearly misclassification, as series from cluster 24 are shorter. This problems are probably caused by nature of Monte Carlo simulation used for likelihood computation. Due to Monte Carlo simulation computational complexity, our implementation can run only limited number of simulations (it was parallelized only on two cores). As likelihood values are very small, it is easily possible to misclassify the time series because likelihood tended to be in favor of wrong model, but in long run it should be in favor of the right model. Causes of this problem are discussed in the conclusion.

5 CONCLUSION

In this paper, the algorithm for classification of traffic incidents was presented. This algorithm is based on the clustering of historical traffic incidents, parametrization of these clusters by Markov chains and classification of new traffic incidents by Naive Bayes classifier. While results have proven that this approach is feasible, it is far from flawless. Especially borderline time series are sometimes misclassified. This problem can be solved by greater level of parallelization, and it will be done in a near future. Another possible cause of this problem may by the simplicity of Markov chain. While it was proven to be good model for modeling the time series, it cannot capture all complexities within time series. It is possible that time series of traffic incident are simply too complex to be described by Markov chain, which may be demonstrated by blurry borders of the models. Other, more complex Markov models will be tested in a near future. We are also planning to compare current approach to the clustering of the historical time series with clustering performed by traffic expert. It is possible that this approach also improves the quality of Markov chain models.

6 Acknowledgements

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and the national budget of the Czech Republic via the Research and Development for Innovations Operational Programme, by the project 'New creative teams in priorities of scientific research' (reg. no. CZ.1.07/2.3.00/30.0055), supported by Operational Programme Education for Competitiveness, and co-financed by the European Social Fund and supported by 'Transport Systems Development Centre' co-financed by Technology Agency of the Czech Republic (reg. no. TE01020155) and by the internal grant agency of VŠB Technical University of Ostrava, Czech Republic, under the project no. SP2015/114 'HPC Usage for Analysis of Uncertain Time Series'.

References

[1] Snell J. L. Grinstead C. M. *Introduction to Probability.* American Mathematical Society, 2003.

[2] Adachi J. Kinoshita A., Takasu A. Traffic incident detection using probabilistic topic model. *Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference*, 2014.

[3] Podhorányi M. Kocyan T., Martinovič J. Searching and indexing distorted data collections. *Proceedings of 26th European Modeling and Simulation Symposium, EMSS 2014*, 2014.

[4] Ben-Akiva M. E. Lia R., Pereirab F. C. Competing risks mixture model for traffic incident duration prediction. *Accident Analysis & Prevention Volume 75*, 2015.

[5] Kroese D. P. Rubinstein R. Y. *Simulation and the Monte Carlo Method.* Wiley-Interscience, 2007.

[6] Mitchell T. *Machine Learning.* McGraw-Hill, Inc., 1997.

[7] Xiaoning Zhang Lijun Sun Yangbeibei Ji, Winnie Daamen. Traffic incident recovery time prediction model based on cell transmission model. *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems*, 2009.

PROBABILISTIC TIME-DEPENDENT VEHICLE ROUTING PROBLEM

Tomáš Režnar, Jan Martinovič, Kateřina Slaninová, Ekaterina Grakova
IT4Innovations, VŠB - Technical University of Ostrava,
17. listopadu 15/2172, Ostrava, Czech Republic
{tomas.reznar,jan.martinovic,katerina.slaninova,ekaterina.grakova.st}@vsb.cz

Abstract: Probabilistic Time-Dependent Vehicle Routing Problem is focused on logistic situations dealing with vehicle fleets with limited load capacity and a set of customer requirements served by these fleets with the minimum overall time of delivery for the whole request. Particular demands may be restricted by a defined time window and may have a set of probabilities of the delivery time fulfillment. The delivery time between customers or between a customer and a depot depends on the traveling distance, time of arrival and probability of the arrival time fulfillment for load delivery through each particular vehicle route.

The paper presents a solution to the mentioned Probabilistic Time-Dependent Vehicle Routing Problem with the focus being on the individual setting of particular customer requests. This enables the specification of input parameters like a fixed guaranteed time for the selected requests with higher or lower importance defined by the probability level of its arrival time. This is a new variant of the Vehicle Routing Problem, which is time-dependent with time windows and a probability of arrival in designated time.

Keywords: Vehicle Routing, Time-Dependent Vehicle Problem, Probability of Arrival

1 INTRODUCTION

Vehicle Routing Problem (VRP) was the first defined fifty years ago. The VRP is aimed at minimizing transportation costs, transportation time, pollution emissions, and others. There are many VRP variants and formulations. Recently, the formulation of the VRP has been widened due to various problems arisen. The paper proposes a novel variant of VRP.

Among others, Tasa et al. [10] worked with the VRP with Time-Dependent and Stochastic Travel Times. They achieved minimum arrival times and they counted the total transportation costs and service costs. The authors of [13] proposed a transportation route so that a delivery service time window can be in the middle of the normal probability distribution; it is medium-probable that the vehicle visits the client in this solution. Dynamic Vehicle Routing Problem with fuzzy time windows (DVRPFTW) was presented in [3] so that the overall size of depots, the total arrival time and the waiting time can be minimized. In paper [5], the authors optimized the departure time of every vehicle in order to minimize the duty time of each driver and fulfill driving hours regulations. The authors of [7] minimized a customer waiting time, the total driving distance, covered dynamic information both random demand and dynamic network, and they maximized the number of customers they attend to serve. Time dependent vehicle routing problem with hard time windows was used in [2] to minimize the overall time of transportation and distance from a depot. The authors also maximized the number of customers served. In paper [12], the authors introduced load-dependent VRP which considers simultaneous pick-ups and deliveries. In paper [1], the Vehicle Routing Problems with Time Windows-Probabilistic model has been developed taking an uncertainty of transportation time into account. However, the model was time-independent.

All the literary resources presented above deal with a lot of VRP variants. However, none of them allows to define a probability of arrival in designed time for each one customer request combined with time-dependency of routes. This paper is focused on defining a new variant of the VRP. The new variant of VRP is time-dependent VRP with time windows and probability of arrival in designed time (PTDVRP). Shipping companies may need to ensure probability

of arrival of shipping vehicles at customer locations in designed time. Some deliveries may have a fixed guaranteed time (e.g. a delivery of production process critical inputs which, if not delivered in time, may cause a production process suspension). Some deliveries may be standard, but if they are not delivered in time, it causes negligible damages. This task can be solved as PTDVRP. The problem inputs consist of a list of available vehicles; each vehicle has a given capacity, and a list of customers' requests. Each request has given capacity demand, location, service time window and probability of arrival in designed time. Moreover, another inputs are time-dependent and request probability-dependent driving times between locations. The output then consists of routes of the vehicles minimizing total route times. The routes respect requested time windows, probabilities of arrival in designed time and vehicle capacities.

2 PROBLEM FORMULATION

The following formulation of PTDVRP is based on the formulation of TDVRP published in [6]. A weighted directed multigraph $G(V, E, w)$ is given, where V is set of nodes, E is set of arcs, and w is step function, which assigns weight (travel time) to each arc.

Node 1 represents the depot and the rest of nodes represent locations of customers. Weight of each arc represents travel time from node i to node j in a concrete time interval and with a concrete probability that travel time will be less or equal to the weight. Set of arcs is defined as $E = \{(i,j)_p^m : i,j \in V, p \in P, m \in M\}$, where M is set of distinct time intervals, and P is set of available probabilities. Exactly $|M| \times |P|$ arcs exist from node i to node j. It is not useful if P contains lower probabilities than the minimum probability of given requested probabilities of arrival in designed time of customers' demands, because the arcs with such probabilities could not be used in any feasible solution due to the constraints (11) and (12).

Step function w returns travel time from node i to node j when a vehicle departures from node i within a concrete time interval and with a given probability that the real travel time will not exceed designed travel time. Travel time is vehicle independent.

Each customer is served exactly by one vehicle with a given probability of the vehicle arrival in designed time. K is set of all vehicles. Each vehicle has a given capacity. Each customer request (pickup or delivery) requires a part of vehicle capacity. Vehicle waiting at the customer location is allowed. Each vehicle route starts and ends in the depot.

Graph $G'(V', E', w')$ is created from graph G as follows. Every inbound arc of node 1 (depot) is removed. Node 1 is the first node of each vehicle route. $|K|$ new nodes are inserted. Set of the newly inserted nodes is $\{n+1, \ldots, n+|K|\}$. Each newly inserted node is the last node of the exactly one vehicle route. It means that each vehicle route starts in node 1 and ends in one of the newly inserted nodes. Arcs from all nodes that represent customer locations to the newly inserted nodes are inserted. Weights of the newly inserted arcs are as follows:

$$\forall i \in \{2, \ldots, n\}, \forall j \in \{n+1, \ldots, n+|K|\} : w'((i,j)_p^m) = w((i,1)_p^m). \tag{1}$$

Arcs from node 1 to the all newly inserted nodes are inserted (for each time interval and probability). Weights of those arcs are 0. Vehicle routes with no customer location to visit are allowed. In that case, routes of those vehicles go directly from node 1 to the end node.

Objective function (2) minimizes total route times, where t_{n+k} is arrival time to the last node of the route of each vehicle.

$$\min \sum_{k=1}^{|K|} t_{n+k} \tag{2}$$

subject to

$$\forall j \in \{2,\ldots,n+|K|\} : \sum_{\substack{i=1 \\ i\neq j}}^{n} \sum_{m\in M} \sum_{p\in P} x_{i\,j}^{m\,p} = 1 \tag{3}$$

$$\forall i \in \{2,\ldots,n\} : \sum_{\substack{j=2 \\ j\neq i}}^{n+|K|} \sum_{m\in M} \sum_{p\in P} x_{i\,j}^{m\,p} = 1 \tag{4}$$

$$\sum_{j=2}^{n+|K|} \sum_{m\in M} \sum_{p\in P} x_{1\,j}^{m\,p} = |K| \tag{5}$$

Variable $x_{i\,j}^{m\,p}$ is binary, if arc $(i,j)_p^m$ is traversed by any vehicle, then $x_{i\,j}^{m\,p} = 1$, otherwise $x_{i\,j}^{m\,p} = 0$. Constraints (3) and (4) ensure that each customer is served exactly once (exactly one inbound and one outbound traversed arc exist for each customer location). Constraint (5) allows vehicles to go directly from the start node to the end node. Vehicles which go directly to the end nodes are considered as not used.

$$t_1 = t \tag{6}$$

Variable t_i represents departure time of any vehicle from node i, while t is a given starting time at the depot. Constraint (6) sets departure time from depot for the all vehicles.

$$\forall i \in \{1,\ldots,n\}, \forall j \in \{2,\ldots,n+|K|\}, \forall m \in M, \forall p \in P :$$
$$i \neq j \implies t_j - t_i - B x_{i\,j}^{m\,p} \geq w'((i,j)_p^m) + s_j - B \tag{7}$$

$$\forall i \in \{1,\ldots,n\}, \forall j \in \{2,\ldots,n+|K|\}, \forall m \in M, \forall p \in P :$$
$$i \neq j \implies t_i + B x_{i\,j}^{m\,p} \leq T^m + B \wedge t_i - T^{m-1} x_{i\,j}^{m\,p} \geq 0 \tag{8}$$

Constraint (7) computes the departure time at node j ([4] and [6]). s_j is a given service duration at node j. B is given a large number to ensure that constraint (7) is binding only if nodes are incident with a traversed arc. Constraint (8) ensures that appropriate arc for departure time t_i is chosen, while T^m is upper bound for time interval $m \in M$.

$$\forall i \in \{1,\ldots,n+|K|\} : L_i \leq t_i - s_i \leq U_i \tag{9}$$

Constraint (9) ensures that arrival time $t_i - s_i$ at node i is between the earliest (L_i) and the latest (U_i) given arrival time at node i.

$$y_1 = 1 \tag{10}$$

$$\forall i \in \{2,\ldots,n\} : y_i \geq a_i \wedge \forall i \in \{n+1,\ldots,n+|K|\} : y_i \geq a_1 \tag{11}$$

$$\forall i \in \{1,\ldots,n\}, \forall j \in \{2,\ldots,n+|K|\} :$$
$$i \neq j \implies y_i \cdot \Big(\sum_{m\in M} \sum_{p\in P} x_{j\,i}^{m\,p} \cdot p\Big) - B \sum_{m\in M} \sum_{p\in P} x_{i\,j}^{m\,p} \geq y_j - B \tag{12}$$

y_i is probability of arrival to node i in designed time. Constraint (10) sets an initial probability at the depot. Constraint (11) ensures a probability of arrival in designed time for customer locations and a probability of return in time to the depot. a_i is a given probability of arrival in designed time for the customer i or the depot. Constraint (12) ensures that the

arc with high enough probability of arrival in designed time is chosen based on the probability of arrival at the antecedent node.

$$l_1 = 0 \tag{13}$$

Variable l_i represents vehicle load (in some unit, for example pieces or weight) at node i, so $l_i \geq 0$. Constraint (13) sets vehicle load at the depot.

$$\forall i \in \{1, \ldots, n\}, \forall j \in \{2, \ldots, n + |K|\} : i \neq j \implies l_j - l_i - B \sum_{m \in M} \sum_{p \in P} x_{i\,j}^{m\,p} \geq d_j - B \tag{14}$$

$$\forall k \in \{1, \ldots, |K|\} : l_{n+k} \leq b_k \tag{15}$$

Constraints (14) and (15) ensure that capacity of each vehicle is not exceeded. d_j is a given customer demand at node j, b_k is a given capacity (or maximum load) of vehicle $k \in K$, and l_{n+k} is a load at the last node of the route of vehicle k.

3 EXPERIMENT

The following experiment compares travel times of the PTDVRP solutions to the TDVRP solutions. Data sets described below were solved as the PTDVRP (respecting given probabilities of arrival) and then as the TDVRP (ignoring given probabilities of arrival and using average travel times between customer locations).

Three experimental data sets were created[1]. The data sets differ in the number of customers. The first set has 10, the second set 50 and the third set 100 defined customers. Each customer was described by random location, random capacity demand, random time window for vehicle arrival and random probability of designed time arrival. Each customer has service time 300 seconds. All the sets have the same 25 vehicles available. Travel times between the nodes were generated on the basis of random customer locations, time interval and the probability of arrival in designed time.

Adaptive large neighborhood search algorithm (ALNS) based on [8] and [9] was implemented. Base principle of the algorithm is ruining a part of current feasible solution and then recreation of the solution. This happens in cycles until some condition is met. The solutions of testing sets were searched with the algorithm in 9000 cycles in each run. Table 1 (for the PTDVRP) and 2 (for the TDVRP) shows travel times of the found solutions of the experimental data sets. Each set was run 20 times.

Number of	**Used Vehicles**				**Travel Time [hours]**			
Customers	**min**	**max**	**mean**	ϱ	**min**	**max**	**mean**	ϱ
10	2	4	3.2	0.6	31.8	41.0	36.4	2.8
50	13	18	15.7	1.3	149.8	187.3	167.0	9.9
100	23	25	24.5	0.7	281.6	317.7	297.1	10.1

Table 1: PTDVRP Travel Times of the Found Solutions for Testing Sets

Travel times of the TDVRP solutions are lower. This is the expected result caused by higher travel times used by the PTDVRP. The PTDVRP uses travel times between customer locations based on requested probabilities of arrival in designed time, while the TDVRP uses average travel times between the customer locations. Figure 1 shows an example of travel times between two customer locations for the PTDVRP. In the case of Figure 1, the PTDVRP would count with proper travel time in range from 1865s to 2667s, while the TDVRP would always count with 1925s (which is average of the travel time function on figure 1) in contrast.

[1] All sets are available at http://dataanalysis.vsb.cz/Collections/Ptdvrp

Number of Customers	Used Vehicles min	max	mean	ϱ	Travel Time [hours] min	max	mean	ϱ
10	1	1	1	0	19.2	19.2	19.2	0
50	3	3	3	0	35.6	40.8	37.0	1.7
100	6	6	6	0	69.9	82.3	74.4	2.9

Table 2: TDVRP Travel Times of the Found Solutions for Testing Sets

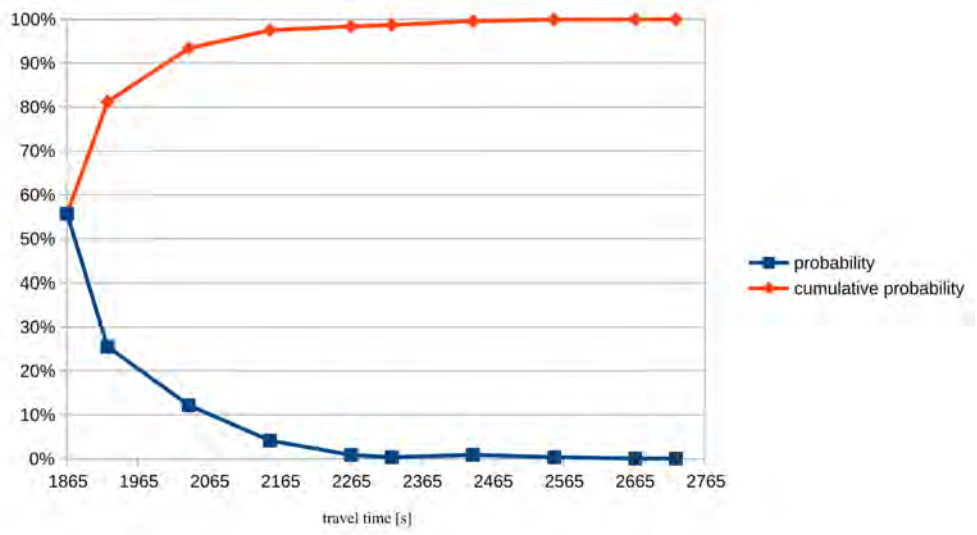


Figure 1: Travel Time Probability and Cumulative Probability between Two Customer Locations for the PTDVRP

4 CONCLUSION

In this paper, the Probabilistic Time-Dependent Vehicle Routing Problem (PTDVRP) has been proposed as a new variant of the VRP. The ALNS algorithm has been used to solve the experimental testing data sets in the experimental part of the paper. The experimental sets were solved as the PTDVRP and the TDVRP to express the difference between the both solutions. Although the PTDVRP feasible solution travel times of routes are higher than travel times of routes of the TDVRP, the PTDVRP allows to ensure the probability of arrival in designed time to the customer locations. Meeting these constraints is much more important then only obtaining the fastest solutions in the real situations. In our future work we will concentrate on the comparisons of the different algorithms for solving the PTDVRP. Data from data sources which provide online traffic flow monitoring will be used as input for computation of the probabilistic distribution of travel time between particular nodes [11]. This probabilistic distribution of travel time on selected path is based on the simulation which utilizes traffic data regarding incidents on roads.

5 Acknowledgment

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and the national budget of the Czech Republic via the Research and Development for Innovations Operational Programme, by the project New creative teams in priorities of scientific research (reg. no. CZ.1.07/2.3.00/30.0055), supported by Operational Programme Education for Competitiveness and by the internal grant agency of VŠB Technical University of Ostrava, Czech Republic, under the project

no. SP2015/157 'HPC Usage for Transport Optimisation based on Dynamic Routing'.

References

[1] Ando, N., and Taniguchi, E. Travel time reliability in vehicle routing and scheduling with time windows. *Netw Spat Econ 6* (2006), 293–311.

[2] Figliozzi, M. A. The time dependent vehicle routing problem with time windows: Benchmark problems, an efficient solution algorithm, and solution characteristics. *Transportation Research Part E: Logistics and Transportation Review 48*, 3 (2012), 616–636.

[3] Ghannadpour, S. F., Noori, S., Tavakkoli-Moghaddam, R., and Ghoseiri, K. A multi-objective dynamic vehicle routing problem with fuzzy time windows: Model, solution and application. *Applied Soft Computing 14, Part C* (2014), 504–527.

[4] Golden, B. L., Magnanti, T. L., and Nguyen, H. Q. Implementing vehicle routing algorithms. *Networks 7*, 2 (1977), 113–148.

[5] Kok, A., Hans, E., Schutten, J., and Zijm, W. A dynamic programming heuristic for vehicle routing with time-dependent travel times and required breaks. *Flexible Services and Manufacturing Journal 22*, 1-2 (2010), 83–108.

[6] Malandraki, C., and Daskin, M. S. Time dependent vehicle routing problems: Formulations, properties and heuristic algorithms. *Transportation Science 26*, 3 (1992), 185–200.

[7] Qiuyun, W., Wenbao, J., and Gang, Z. A novel model and algorithm for solving dynamic vehicle routing problem on goods distribution. *Journal of Applied Sciences 13*, 22 (2013), 5410–5415.

[8] Ropke, S., and Pisinger, D. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. *Transportation science 40*, 4 (2006), 455–472.

[9] Schrimpf, G., Schneider, J., Stamm-Wilbrandt, H., and Dueck, G. Record breaking optimization results using the ruin and recreate principle. *Journal of Computational Physics 159*, 2 (2000), 139–171.

[10] Tasa, D., Dellaert, N., van Woensel, T., and de Kok, T. The time-dependent vehicle routing problem with soft time windows and stochastic travel times. *Transportation Research Part C: Emerging Technologies 48* (2014), 66–83.

[11] Tomis, R., Martinovič, J., Slaninová, K., Rapant, L., and Vondrák, I. Time-dependent route planning for the highways in the czech republic. In *Proceedings of the 14th International Conference on Computer Information Systems and Industrial Management, CISIM 2015* (2015).

[12] Zachariadis, E. E., Tarantilis, C. D., and Kiranoudis, C. T. The load-dependent vehicle routing problem and its pick-up and delivery extension. *Transportation Research Part B: Methodological 71*, C (2015), 158–181.

[13] Zhang, J., Lam, W., and Chen, B. A stochastic vehicle routing problem with travel time uncertainty: Trade-off between cost and customer service. *Networks and Spatial Economics 13* (2013), 471–496.

Segmentation and Detection of Text in Document Images

Darko Zelenika, Janez Povh
Faculty of Information Studies, Laboratory of Data Technologies,
Ulica talcev 3, SI-8000 Novo mesto, Slovenia
{darko.zelenika, janez.povh}@fis.unm.si
Bernard Ženko
Jožef Stefan Institute, Department of Knowledge Technologies,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
bernard.zenko@ijs.si

Abstract: Text detection in document images plays an important role in optical character recognition systems and is a challenging task. The proposed text detection method uses self-adjusting bottom-up segmentation algorithm to segment a document image into a set of connected components. The segmented connected components are then described in terms of 27 features and a machine learning algorithm is used to classify these components as text or non-text. We have collected a dataset (called ASTRoID), which contains 500 images of text blocks and 500 images of non-text blocks in order to test the method. We empirically compare performance of the proposed text detection method with seven different machine learning algorithms; the best performance is obtained with the radial support vector machine.

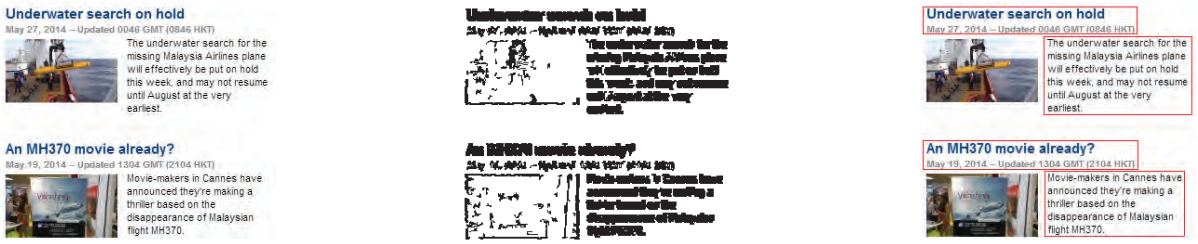
Keywords: text detection, document segmentation, text/non-text classification, machine learning

1 INTRODUCTION

Today a lot of potentially useful textual information is stored in an unstructured form of images of documents such as invoices, contracts, web pages, etc. In order to effectively recognize and extract this text with Optical Character Recognition (OCR) technology, location of the text within the image must be detected first. The first step of text detection in document images is the document segmentation, which is followed by a classification of segments obtained in the first step. Document segmentation is a task which splits a document image into segments or blocks of interest, as shown in Fig. 1b. Here, each group of black pixels (called connected component (CC)) represents one block. Blocks of interest can usually be classified as text or non-text. We are mainly interested in text blocks, so our goal is to identify them and separate them from non-text blocks (see Fig. 1c). It is important to understand that document segmentation and classification (identification) of segmented blocks can hardly be treated as independent tasks and are often merged together in a "(physical) layout analysis" [8].

Document segmentation techniques are traditionally classified into three categories: top-down, bottom-up and hybrid approaches. The top-down approach starts the segmentation from bigger blocks and repetitively segments the document image into smaller blocks until the document image is segmented into the smallest possible blocks [2, 7, 9, 12, 13, 15, 18]. The bottom-up approach is the opposite of the top-down approach. It starts from the smallest segments (characters) and then joins them into bigger and bigger blocks (words, paragraphs, etc.)[3, 4, 5, 16, 20]. An approach that does not fit into a top-down or bottom-up strategy or uses the combination of both is called a hybrid approach [10, 14]. Hybrid approaches often try to combine the speed of top-down approaches and the robustness of bottom-up approaches.

The task of the classification algorithm is to classify the results of the segmentation algorithm. The accuracy of the classification highly depends on the quality of the results of the segmentation algorithm. A standard approach for detecting text and non-text blocks is to extract features from segmented CCs in document images and classify them with some machine learning (ML) algorithm [2, 3, 5, 20]. In this paper we use a similar approach.



(a) (b) (c)

Figure 1: a) Image document, b) Segmented image and c) Detected text

The main contributions of this paper are: (1) a self-adjusting segmentation algorithm for finding text CCs that is independent of the image resolution and font size, (2) a new set of features which describe differences between text and non-text image blocks, (3) a custom benchmark dataset ASTRoID of text and non-text image blocks, and (4) investigation of performance of seven different ML algorithms for separation between text and non-text image blocks.

The rest of the paper is organized as follows. In section two, we introduce the document segmentation algorithm, the ASTRoID dataset and the classification algorithm. The classification results obtained with different ML algorithms are presented and discussed in section three. Finally, section four concludes the paper.

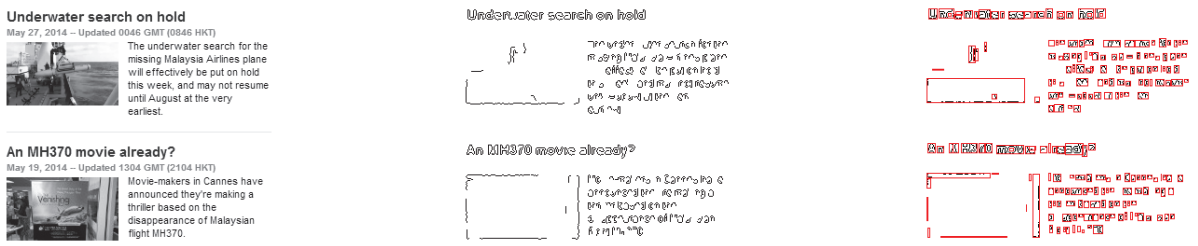
2 TEXT DETECTION

Our text detection method performs two tasks: document image segmentation and classification. The segmentation algorithm extracts image blocks, these are described in terms of the selected features, and then classified with a ML algorithm as either text or non-text blocks.

2.1 Segmentation

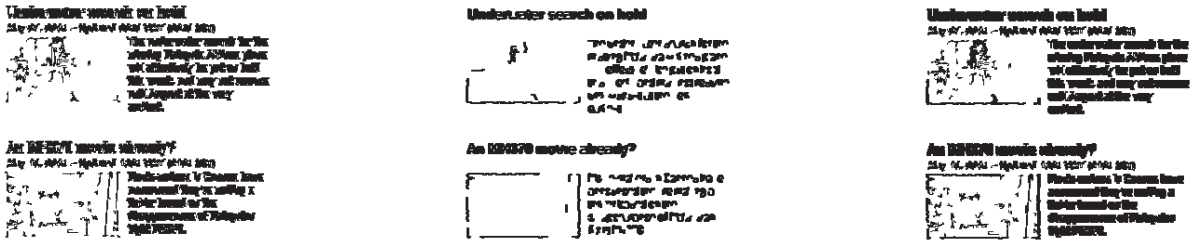
The segmentation algorithm follows the bottom up strategy. It segments the document into CCs, which are constructed with a combination of the Sobel edge detection and dilation methods [8, 17], and is composed of three parts: (1) search for an optimal rectangular kernel, (2) edge detection and (3) extraction of standalone document image blocks. The input is a grayscale image (Fig. 2a) and the output is a binary image.

The first part of the segmentation algorithm finds an optimal rectangular kernel, which is then used by the other two parts of the algorithm. The optimal rectangular kernel highly depends on the height of the dominant text in the document image. In the second part of the segmentation algorithm we apply different Sobel kernels (vertical, horizontal and diagonal) on the grayscale document image. The resulting images of different Sobel kernels are dilated by the optimal rectangular kernel and combined into one image by logical *AND* and *OR* operations, as shown in Fig. 3a. In the third part of the segmentation algorithm we localize all CCs as illustrated in Fig. 2b (red rectangles in Fig. 2c) and keep only those CCs that intersect with two or less than two other CCs (due to characters such as "B" and "8") and we call them standalone CCs. In such a way most of CCs which do not belong to text blocks are removed. This image then is dilated by the optimal rectangular kernel, as shown in Fig. 3b. The final segmented image is obtained by combining the resulting images of the second (Fig. 3a) and third (Fig. 3b) part of the algorithm with logical *OR* operation (see Fig. 3c). In such a way we segment the document image into blocks that can either be text or non-text blocks.



(a) (b) (c)

Figure 2: a) Grayscale image, b) Sobel edges and c) Localized CCs



(a) (b) (c)

Figure 3: Segmentation steps, a) 2nd part, b) 3rd part and c) Final part

2.2 Text vs. non-text classification

In order to evaluate our method we created our custom dataset of text and non-text image blocks, which we called ASTRoID. The dataset includes 500 image blocks which contain plain text of different sizes, lengths, colors, font styles, etc., and 500 equally diverse image blocks which do not contain text. The ASTRoID dataset is available for download at: `http://dk.fis.unm.si/ASTRoID.zip`.

In document images most of the text blocks are uniformly structured and have a regular shape. On the other hand, non-text blocks have a lot of shape variability, i.e., mostly they have an irregular shape. But only shape information is not enough for classification, we also need to take into account the information on the context of these blocks. In this paper we used features similar to the ones proposed in [2, 3, 5, 12]. However, we took a different approach to calculate some of the proposed features and introduced a new feature "color density". In our approach we extract features from multiple images of the same segmented block (Fig. 4), which are obtained by different preprocessing methods, unlike the approaches found in the literature where features are extracted from usually one binary image. Before the actual feature extraction, each segmented block is resized to 100 pixels in height while maintaining the width to height aspect ratio. In our approach we used the following features: number of CCs [2], aspect ratio [2, 3, 12], foreground density [2, 3, 5, 12], color density, standard deviation of the heights and widths of CCs [3, 5, 12], and standard deviation of the lengths of horizontal and vertical runs [12]. Most of the features are extracted from different binary images (Fig. 4b - 4h), which are obtained with the following methods: skeletonization (Fig. 4d) [1], horizontal Sobel kernel (Fig. 4e), vertical Sobel kernel (Fig. 4f), diagonal Sobel kernels (Fig. 4g) and Canny edge detection method (Fig. 4h) [6, 17].

The number of CCs is computed after the binarization of grayscale image with the Otsu's [11] thresholding algorithm (Fig. 4b). The feature foreground density is calculated two times, i.e., once for Fig. 4b and once for Fig. 4h. The color density feature is calculated from the color image (Fig. 4a). The standard deviation of the heights of CCs is calculated eight times, i.e., once for each of the figures: Fig. 4b, 4c, 4d, 4f, 4g and 4h, and for the Sobel vertical lines which are extracted from Fig. 4c and 4h. The standard deviation of the widths of CCs is caluclated

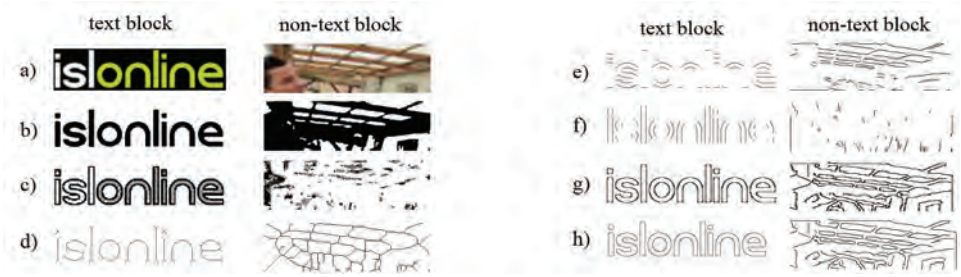


Figure 4: a) Color image, b) Binary image, c) Image of extracted colors, d) Skeleton image, e) Horizontal Sobel, f) Vertical Sobel, g) Diagonal Sobel and h) Canny

two times, i.e., once for Fig. 4e and for the Sobel horizontal lines which are extracted from Fig. 4h. Standard deviation of the vertical runs is calculated nine times, i.e., once for each of the following figures: Fig. 4b, 4c, 4d, 4f, 4g and 4h, and for the Sobel vertical lines which are extracted from Fig. 4c, 4g and 4h. Standard deviation of the horizontal runs is calculated three times, i.e., once for each of the following figures: Fig. 4b and 4e, and for the Sobel horizontal lines which are extracted from Fig. 4h.

All together we extracted 27 features from each image block of the ASTRoID dataset and assigned a class label (text or non-text) to each one. In this way we created a dataset, which we used for classification with seven popular ML algorithms that are frequently used in practical applications and typically give good results, in order to evaluate our choice of features. We used: Naïve Bayes, C4.5 (decision tree), k-Nearest Neighbors (k-NN), Random Forest, Linear Support Vector Machine (SVM), Polynomial SVM and Radial SVM. The accuracy of each of the above mentioned algorithms is estimated by 10-fold cross-validation. We used the implementations of ML algorithms in the WEKA data mining suite [19]. Naïve Bayes and C4.5 (decision tree) algorithms are used with default parameters while parameter k for the k-NN algorithm is set to 1 and parameter number of trees for the Random forest algorithm is set to 100. For SVMs we normalized data by setting normalize parameter. The parameter C for Linear SVM is set to 15. The parameters C, *degree*, *gamma* and *coefficient* for Polynomial SVM are set to 20, 3, 1 and 1, respectively. The parameters C and *gamma* for Radial SVM are set to 20 and 1, respectively.

3 RESULTS & DISCUSSION

Table 1 shows classification results of all ML algorithms. The accuracy of all ML algorithms is higher than 90% which suggests that the choice of our features for text/non-text differentiation is appropriate. The ML algorithm that has the highest accuracy 98.2% is the Radial SVM. The results obtained by the proposed text detection method are promising. Other authors who worked on similar problems also obtained comparable classification results. In [20] authors used SVM and classified text from non-text blocks with the accuracy of 96.62%. In [12] authors obtained 97.5% classification accuracy by using genetic programming to classify the document blocks as text, image, drawing and table. In [3] authors used Multilayer Perceptron to classify text from non-text blocks and obtained 97.25% of accuracy. The only disadvantage of approaches in [12, 3, 20] is that they fail to detect text in documents with complex layout, due to limitations of their segmentation algorithms. The advantage of our segmentation algorithm is that it self-adjusts to the document image regardless the image resolution and font size, it does not need any input parameters and it also works on documents with complex layouts. The only disadvantage is that it fails to detect (segment) some text blocks of very light text color, and also some text blocks with very complex (with a lot of details) background and very decorated text strings. In the future we plan to test our text detection method on other datasets and

compare it with other segmentation methods.

Table 1: Classification results

Classifier	Naive Bayes	k-NN	C4.5 - decision tree	Random forest	Lin. SVM	Poly. SVM	Rad. SVM
Accuracy	90.1%	93.6%	94.3%	97%	97%	97.3%	98.2%
Precision of text blocks	0.860	0.911	0.937	0.976	0.964	0.961	0.978
Precision of non-text blocks	0.953	0.964	0.949	0.964	0.976	0.986	0.986
Recall of text blocks	0.958	0.966	0.950	0.964	0.976	0.986	0.986
Recall of non-text blocks	0.844	0.906	0.936	0.976	0.964	0.960	0.978

4 CONCLUSION

We have presented a text detection method, which consists of document segmentation and feature extraction algorithms. The proposed segmentation algorithm is based on the bottom-up strategy of analysis and segmentation is done by using the Sobel edge detection method. We created ASTRoID dataset of images, which consists of 500 image blocks of text and 500 image blocks of non-text. The proposed feature extraction algorithm extracts features from each image in ASTRoID dataset. We used 27 different features in order to differentiate text from non-text regions. In order to classify text from non-text blocks we used seven ML algorithms. The accuracy of all ML algorithms is higher than 90% which suggests that the choice of our features is appropriate. The classification results show that SVM with radial kernel has the highest accuracy 98.2%.

5 Acknowledgements

The presented work was supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

References

[1] Basilis, G.G. (2014). Imaging Techniques in Document Analysis Processes. In Doearmann, D. and Tombre, K. (Eds.) Handbook of Document Image Processing and Recognition (pp. 73–131). London: Springer.

[2] Barlas, P., Kasar, T., Adams, S., Chatelain, C. and Paquet, T. (2014). A typed and handwritten text block segmentation system for heterogeneous and complex documents. In: 11th IAPR International Workshop on Document Analysis Systems, pp. 46–50, IEEE Press, Tours, France.

[3] Bukhari, S.S., Azawi, M.A., Shafait, F. and Breuel, T.M. (2010). Document image segmentation using discriminative learning over connected components. In: 9th IAPR International Workshop on Document Analysis Systems, pp.183–190. Boston, MA, USA.

[4] Bukhari, S.S., Shafait, F. and Breuel, T.M. (2011). Improved Document Image Segmentation Algorithm using Multiresolution Morphology. In: 18th Document Recognition and Retrieval Conference, pp. 1–10., San Jose, CA, USA.

[5] Bukhari, S.S., Asi, A., Breuel, T.M. and El-Sana, J. (2012). Layout analysis for arabic historical document images using machine learning. In: Int. Conference on Frontiers in Handwriting Recognition, pp. 639–644.

[6] Burger, W. and Burge, M.J. (2009). Principles of Digital Image Processing. Springer, London.

[7] Coppi , D., Grana, C. and Cucchiara , R. (2014). Illustrations Segmentation in Digitized Documents Using Local Correlation Features. In: 10th Italian Research Conference on Digital Libraries, pp. 76–83. Procedia Computer Science, Vol. 38, Padua, Italy.

[8] Kise, K. (2014). Page Segmentation Techniques in Document Analysis. In Doearmann, D. and Tombre, K. (Eds.) Handbook of Document Image Processing and Recognition (pp. 135–175). London: Springer.

[9] Kruatrachue, B., Moongfangklang, N. and Siriboon, K. (2005). Fast Document Segmentation Using Contour and X-Y Cut Technique. In: The Third World Enformatika Conference. pp. 27–29. Turkey.

[10] Kundu, M.K., Dhar, S. and Banerjee, M. (2012). A New Approach for Segmentation of Image and Text in Natural and Commercial Color Document. In: Proc. of International Conference on Communication, Devices and Intelligent Systems, pp. 85–88. IEEE Press, India.

[11] Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, Vol 9: 62–66.

[12] Priyadharshini, N. and Vijaya, M.S. (2013). Genetic Programming for Document Segmentation and Region Classification Using Discipulus. Int. Journal of Advanced Research in A.I., Vol. 2: 15–22.

[13] Priyanka, N., Pal, S. and Mandal, R. (2010). Line and Word Segmentation Approach for Printed Documents. International Journal of Computers and Applications, Vol. 1: 30–36.

[14] Roy, P.P., Pal, U. and Lladós, J. (2010). Touching Text Character Localization in Graphical Documents Using SIFT. In: Proceedings of the 8th International Conference on Graphics Recognition: Achievements, Challenges, and Evolution, pp. 199–211. Springer-Verlag, France.

[15] Shafait, F., Keysers, D. and Breuel, T. (2008). Performance evaluation and benchmarking of six-page segmentation algorithms. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 941–954. IEEE Press.

[16] Sumathi , C.P. and Priya , N. (2013). A Combined Edge-Based Text Region Extraction from Document Images. International Journal of Advanced Research in Computer Science and Software Engineering, 827–835.

[17] Vasuki, S. and Ganesan , L. (2005). Performance Measure for Edge Based Color Image Segmentation in Color Spaces. In: Proceedings of the International Conference on Emerging Technologies in Intelligent System and Control: Exploring, Exposing, and Experiencing the Emerging Technologies, pp. 621–626. Allied Publishers, Coimbatore.

[18] Vikas , J.D. and Vijay , H.M. (2011). Devnagari Document Segmentation Using Histogram Approach. Int. Journal of Computer Science, Engineering and Information Technology, Vol. 1: 46–53.

[19] WEKA (Open source, Data Mining software in Java) (2015). University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka [Accessed 02/06/2015]

[20] Zagoris, K., Chatzichristofis, S.A. and Papamarkos, N. (2011). Text Localization using Standard Deviation Analysis of Structure Elements and Support Vector Machines. EURASIP Journal on Advances in Signal Processing, Vol. 47: 1–2.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 1:

Mathematical Programming and Optimization

ANALYSIS OF USE CASES USING HPC SERVICES IN THE CLOUD

Nejc Bat, Marko Kobal and Tomi Ilijaš
Arctur d.o.o.
Industrijska cesta 5
nejc.bat@arctur.si; tomi.ilijas @arctur.si

Abstract: High performance computing (HPC) systems are traditionally hosted in private facilities, available to particular and specific research groups. The main reason for this phenomenon lies in the scale of such systems. Significant funds for CAPEX and OPEC costs are required to establish and run HPC systems, leaving smaller research groups and small and medium enterprises (SMEs) out of reach for such systems. Cloud computing offers a solution for these users. While general services for general use IT (web services, business applications, mobile applications, etc.) and other smaller scale solutions are easy to migrate into the Cloud, this is not the case in HPC. Specific environment, fast interconnect, custom tailored configurations and complex applications makes HPC solutions hard to port into the cloud. In this paper we present successful use cases of using HPC services in the cloud while pointing out the main pre-requirements that should be met for a successful use of HPC applications in the cloud.

Keywords: high performance computing, cloud computing, use case, evaluation

1 INTRODUCTION

HPC includes a diverse range of scientific and engineering applications with a combination of top-end computing hardware. Supercomputers, the core component of HPC systems, are evolving along with the development of computer science. The first supercomputers have been presented already in the beginning of modern computer era, some 50 years ago. From a computer architecture perspective we can separate HPC systems from early specialized architectures with superscalar processors, vector processors and special systems with shared memory, up to today's well-established and dominant massively parallel and cluster supercomputers [1], [2]. The latter are taking the advantage of low-priced commodity processors and the architecture of distributed memory. Exactly the same basic principle is used in general-use cloud computing with all large scale cloud computing platforms being built on low-priced commodity servers. This intersection between general-use cloud computing and HPC makes interesting initiative to explore migration from traditional HPC systems to cloud based HPC systems.

Besides the hardware computer architecture one must also take into account the software stack, needed to run a supercomputer. This stack is built from the operating system, the middleware and end-user applications. The vast majority of the supercomputers from the current top500.org list [3], [4] are based on Linux operating system. This fact had also an impact on the development of user end software and must be considered when evaluating the process of porting particular application to the HPC Cloud.

2 RELATED WORK

Several benchmarking studies and performance analysis have already been presented in the related work, but are mainly focusing on pure performance or technological aspects rather than real-life use cases experiences. The related research HPC cloud studies can be divided into three major groups, the ones that are applicable to a particular virtualization technology, the ones on general common studies and those related to public cloud offering. General HPC virtualization studies, for example [5] covers general benefits and pitfalls related to High-Performance Cloud Computing, outlining many still open up questions such as scalability

issues and performance variations depending of a particular application. Work, related to public cloud offering, most often discusses Amazon EC2, such as [6], [7].

The most comprehensive related work on actual use-cases has been done in the UberCloud experiment [8]. UberCloud is the online community and marketplace. Engineers and scientists with computational and data-intensive tasks can discover, try and buy on demand computing power and expertise. The UberCloud started in July 2012 with the free voluntary HPC Experiment which today has over 1000 participating organizations and individuals, from 68 countries. To explore the challenges of the end-to-end process for an end-user to access and use remote computing resources, the project members built "Teams of Four", i.e. industry end-user, software provider, resource provider, and HPC expert, to work together on the end-user's application, defining the requirements, getting the licenses and implementing the software on the remote system, running and monitoring it, getting the results back to the end-user, and writing a short case study about their experience, lessons learned, and recommendations, for the benefit the UberCloud community.

3 PLATFORM AND BENCHMARKING CASE

Due to the fact that the most widespread type of supercomputers are based on systems with distributed memory [4] we describe a typical platform, used as a basis for cloud enabled HPC system. Arctur-1 [9] is a typical representative of HPC systems with distributed memory. A part of the system is converted into heterogeneous system, equipped with GP-GPU accelerators [10] in order to support modern applications, for processing as well as for visualization purposes.

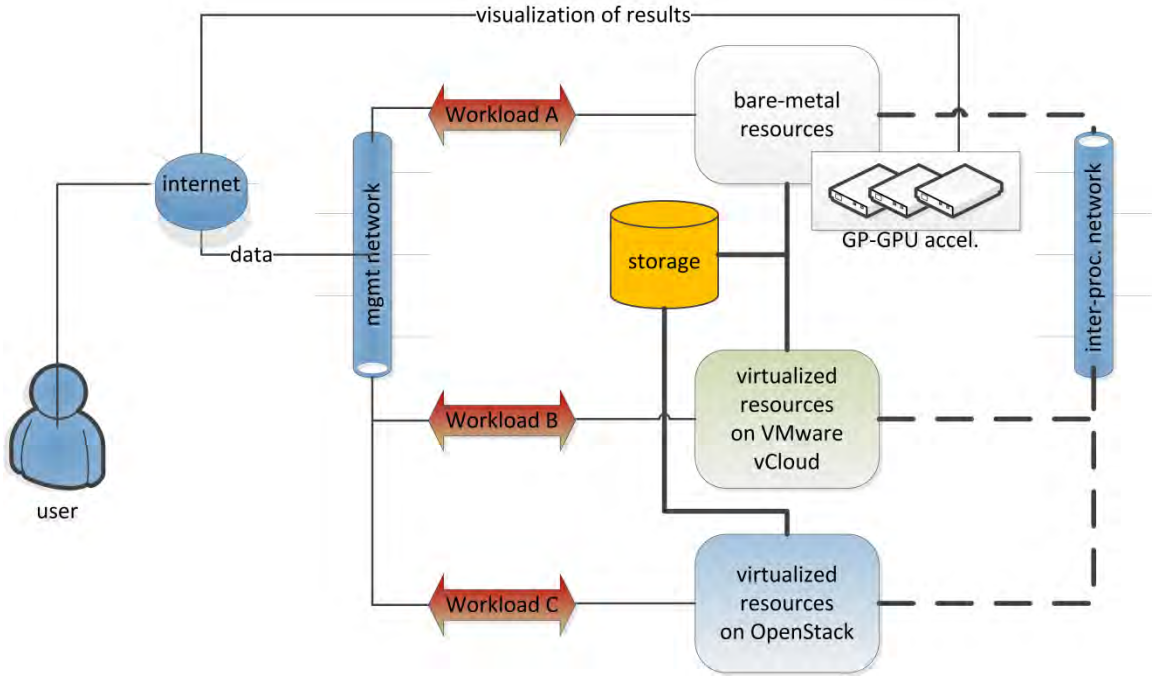


Figure 1: Arcur-1 multi-heterogeneous Cloud HPC system

Typical traditional general-use cloud computing systems are usually based on completely virtualized infrastructure. However, since virtualization brings inevitable performance overhead, this is not always a feasible solution for running HPC applications. Therefore cloud enabled HPC systems are typically composed of virtual as well as physical infrastructure. Compute, specifically CPU or interconnect-intensive workloads are submitted

directly to physical (bare-metal) infrastructure, while some other workloads, which are not so CPU or interconnect-intensive or require higher level of elasticity [11] are typically deployed on fully virtualized HPC platforms. Virtualized infrastructure delivers better flexibility and elasticity by leveraging fast deployment and instantiation from templates. It also provides means of live migration of workloads (virtual machines) between separate physical systems to provide elasticity and some other higher-level added values such as proactive fault-tolerance [12]. In order to support various kinds of workloads, we converted Arctur-1 into such multi-heterogeneous system as show on Figure 1.

Various workloads are submitted by the user to the system through the Cloud. We define "Cloud" as user interfaces, available through web browser graphical user interfaces, SSH, and VNC as well as application programmable interfaces (API), available through various API protocols, based on internet. In dependence of workload type, a particular workload is either executed on bare-metal, bare-metal with graphic processing accelerators (GP-GPUs), commercial virtualized platform (VMware vCloud) or open-source (OpenStack) virtualization platforms. GP-GPUs are used to accelerate workloads, designed and programmed for GP-GPU executable environments (CUDA, OpenCL) and also for results visualization. In this case the actual rendering for visualization is done on GP-GPUs while fast and responsive protocols are used to deliver rendered display to the end user.

The biggest challenge in virtualizing a HPC system, more specifically, it's particular components such as CPU, memory, accelerators, storage interfaces and network, is when dealing with fast interconnect, used for inter-process communication, for example, MPI [13]. The most commonly used interconnect is InfiniBand, which is also the case in Arctur-1 Supercomputer. There are several methods for virtualizing InfiniBand network adapter: direct attaching the device to virtual machine, direct attaching the device to virtual machine by leveraging SR-IOV technology, by using eIPoIB paravirtualization and by using general paravirtualization. We have conducted some basic performance benchmarks by using a sample application in the field of molecular analysis - GROMCAS.

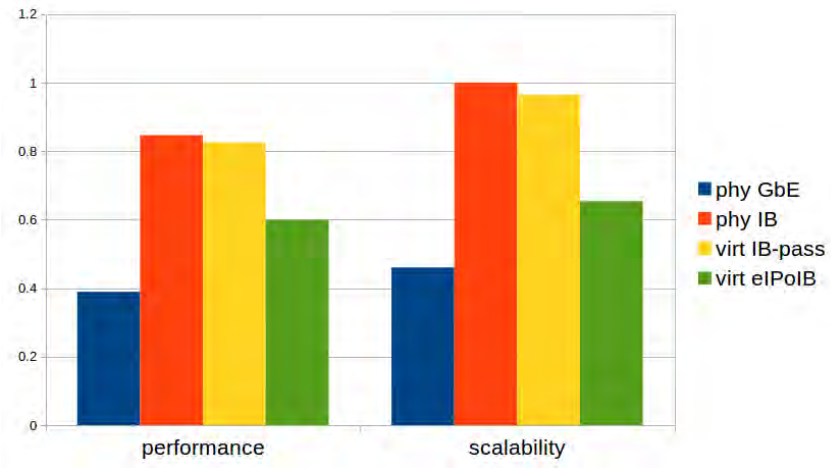


*Figure 2: P*erformance and scalability ratio for 96-core 8-node HPC cluster on GROMACS application benchmark

The results of our tests on GROMCAS application benchmark are shown in ***Figure 2*****.** Our testbed was configured with 8 HPC nodes and total 96 cores with physical (native, bare-metal) configuration and with KVM virtualization for the comparison. The ***Figure 2*** is showing performance ratio of virtualized InfiniBand on 8 nodes versus 1 node and the scalability ratio versus physical (native) InfiniBand. As shown, the direct attached InfiniBand virtualized adapter shows near the same performance as non-virtualized (phy), whereas

paravirtualization via eIPoIB mothod shows significant performance overhead. Nevertheless, the main issue of direct attached method is that InfiniBand adapter can only be attached to a single virtual machine. This effectively disables the ability of live migration which makes its use very limited and does not provide the promise of elasticity for virtual environments. That said we decided not to use InfiniBand for our virtualized parts of Cloud HPC system and only allow paravirtualized Gigabit Ethernet which fully supports live migration and other high-level virtualization features, providing elasticity and dynamic environment. This setup allows only non-MPI applications to be run on virtualized environments (due to high performance penalty while running MPI applications on Gigabit Ethernet or paravirtualized InfiniBand). Therefore workloads are scheduled accordingly when passing through management network and system on Arctur-1 Cloud HPC system.

4 USE CASES

As outlined in the introduction, the main goals of using HPC resources in the cloud is to reduce or eliminate CAPEX and OPEX costs and to provide operational flexibility. Arctur-1 system, as described in ***Chapter 3*** does provide such environment. Several use cases were executed to experiment with Arctuir-1 Cloud HPC system, therefore we are presenting a few use cases from two major projects, funded by EU.

Fortissimo [14] is a collaborative project that will enable European SMEs to be more globally competitive through the use of simulation services running on a cloud-based High-Performance-Computing (HPC) infrastructure. The importance of advanced simulation to the competitiveness of both large and small companies is well established. However, large companies have easier access to advanced simulation than SMEs, which are facing both technological hurdles and financial challenges. This means that SMEs are often not able to take advantage of this technology, even though it would clearly make them more competitive. The goal of Fortissimo is to overcome this impasse through the provision of simulation services and tools running on a cloud infrastructure. A “one-stop-shop” greatly simplifies access to advanced simulation, particularly to SMEs. This makes hardware, expertise, applications, visualisation and tools easily available and affordable on a pay-per-use basis. In doing this, Fortissimo created and demonstrated a viable and sustainable commercial ecosystem. Fortissimo will contribute to the increased competitiveness of European manufacturing industry through the innovative infrastructure that it will develop and test. It will also create commercial opportunities for European Independent Software Vendors and for service and High Performance Computing infrastructure providers, through the creation of a new market for their products and services. Fortissimo places emphasis on the exploitation of opportunities at all levels of the value chain all the way from the SME end-users to the providers of High Performance Computing infrastructure.

Cloudflow [15] aims to empower the different engineering disciplines with on-demand access to scalable computational services, allowing them to start any process when desired and without the need for a complex local infrastructure of cutting-edge high performing computers. The new engineers’ workplaces do not need to be equipped with expensive software (CAD, CAM, CAE, PLM, data archival, etc.) and their required operating systems and versions, or with special hardware (CPU, RAM, GPU) and their dedicated drivers. Therefore, the engineer will not need multiple computers for different tasks; or in the case of limited computers, the engineers will not need to wait for the availability of the resources. Cloudflow will build on existing technologies and standards, integrating and configuring them in such a way that engineers will be able to continue with their standard workflows and with their standard formats, but with a scalable amount of resources on-demand. Cloudflow will go beyond the mere provision of individual computational services on the cloud. It has

the ambition to support chains of services in integrated workflows and to allow the development and simulation of complex products such as mechatronic systems where mechanics, software and electronics work together. The future engineering workplaces will literally be newly conceived. The engineer will become flexible and powerful. They will be able to develop and assess complex systems in an integral manner, combining the different behaviours for integrated workflows within chains of services. The engineers will be relieved from being present at a fixed workplace to perform any of their tasks. In contrast, they will be able to work anywhere and with light resources, able to access all the technology and the needed computational power.

4.1 Evaluation and use cases preliminary results

Both before mentioned projects (Fortissimo and CloudFlow) are End User driven projects. This means that they are building upon actual use cases and on real end user data. Further on the most important results from this end user cases are presented.

Pipistrel was one of th use-cases form the Fortissimo project. It is a company from Slovenia, manufacturing lightweight aircrafts. Computer Aided Design (CAD) and Computational Fluid Dynamic (CFD) simulations are needed to shape an aircraft with perfect aerodynamic characteristics. Normally, the required High Performance Computing (HPC) infrastructure is only affordable to large companies. However, the aircraft manufacturer Pipistrel demonstrated the benefits of a cloud-based HPC infrastructure in a Fortissimo experiment. The medium-sized enterprise was able to speed up their design cycles with increased precision in simulation, while the costs for the computation power were twenty-five times less compared to an in-house solution.

The GridWorker software tool by Fraunhofer EAS (a German research institute as use-case in the CloudFlow project) is being used to parallelize computations as much as possible to reduce the overall time used. Through GridWorker the available HPC resources are deployed in Cloud on a number of virtual machines to exploit the power of many computational cores at the same time. As a result, the HPC resources can enable users to simulate more complex machining tasks more quickly. In fact, the time to compute a best possible toolpath is now only 1/3 of what was necessary before. This provides the opportunity to increase the quality of the machining. Tool paths are now calculated in parallel. The CAM workflow allows the end user to prepare all data sets at once to produce a good machining plan and execute them at once and in parallel in the Cloud instead of having to wait for each individual result in front of his desktop before the next variant can be computed.

4.2 The future – deliver HPC services through self-service web portal

The use-cases, described in the previous section have been realized by semi-automated or manual interaction with the Cloud, using SSH, VNC and other remote access tools and protocols. The next logical step is to provide a completely web-based portal with seamless integration into HPC Cloud thus allowing users to submit workload, monitor progress and visualize results through a single web portal. The user experience with the CloudFlow Portal should resemble the positive aspects of traveling with any of budget airlines. The purchase of the Cloud HPC resources trough CF Portal should be automated, simple, fast and affordable.

We derive from the hypothesis that the end-users main concern is time to solution. Other underlying drivers, such as financial, are also important and should be kept in mind, but our main focus is on promising as short time as possible from the end-user's initial contact with the portal and obtaining the required results. Therefore we proposed an architecture as depicted in the ***Figure 3***. The portal is the main focus point of the CloudFlow workflow

enabling system. It encapsulates the entire technical and management infrastructure to provide the end-user with a complete solution. The portal provides added value by enabling the symbiosis of key competencies and knowledge. This represents the point of contact between the end-user and the gateway to the collaboration through the portal. The portal consist of several entry points for the end-user through the domain experts, infrastructure providers and the ISVs are the next entry point. The latter have a direct channel of contact to the users that have not yet started to use any HPC infrastructure in their work process or are not yet aware of this possibility.

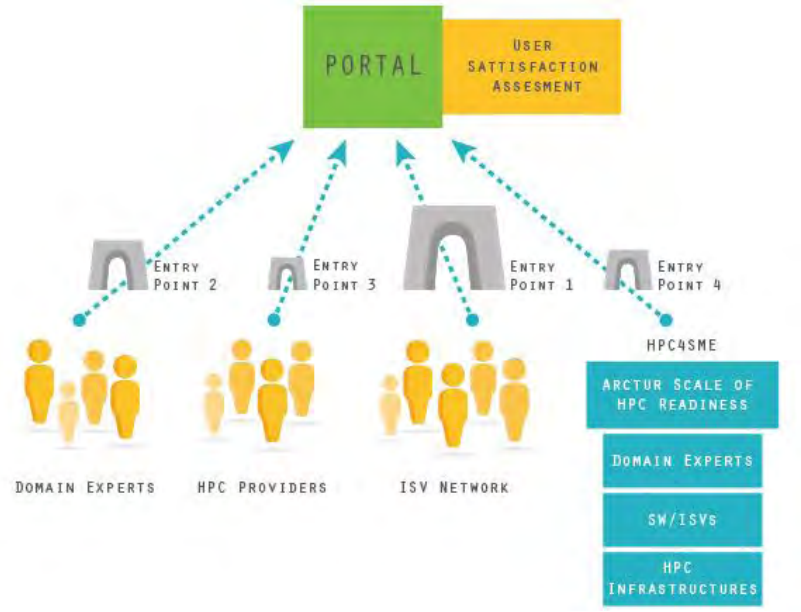


Figure 3: High level overvie of future HPC Cloud Portal

6 CONCLUSION

In this paper we presented the current state of the art of using HPC resources through the cloud. The architecture of Arctur-1 Cloud HPC platform was described and particular technological limitations have been presented, such as inability for efficient and elastic InfiniBand virtualization to provide a fully virtualized and elastic Cloud HPC environment. We also presented real-live use-case experiments with results, showing significant improvement for small research groups as well as small and medium enterprises to accelerate their workflows and provide faster and cheaper ways to get from idea to a solution by using HPC services through the cloud. The future work will mainly be related to unification and simplification of the workflows by providing HPC services through advanced self-service web portal, enabling users to get even easier and faster access to a wide range of HPC services.

References

[1] R. Brightwell, L. A. Fisk, D. S. Greenberg, T. Hudson, M. Levenhagen, A. B. Maccabe, and R. Riesen, “Masively parallel computing using commodity components,” *Parallel Comput.*, vol. 26, no. 2, pp. 243–266, 2000.

[2] G. Bell and J. Gray, “High Performance Computing : Crays , Clusters , and Centers . What Next ?,” *Commun. ACM*, no. August, pp. 1–7, 2001.

[3] J. J. Dongarra, H. W. Meuer, and E. Strohmaier, “TOP500 Supercomputers,” *Rechenzentrum der Univ. Mannheim*, vol. 33, pp. 93–124, 1993.

[4] “Top500 November 2014 list.” [Online]. Available: http://www.top500.org/lists/2014/11/. [Accessed: 05-Aug-2014].

[5] D. Tomić, D. Ogrizović, and Z. Car, “Cloud solutions for high performance computing: oxymoron or realm?,” *Tech. Gaz.*, vol. 20, no. 1, pp. 177–182, 2013.

[6] K. R. Jackson, L. Ramakrishnan, K. Muriki, S. Canon, S. Cholia, J. Shalf, H. J. Wasserman, and N. J. Wright, “Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud,” in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, 2010, pp. 159–168.

[7] Y. Zhai, M. Liu, J. Zhai, X. Ma, and W. Chen, “Cloud versus in-house cluster: Evaluating Amazon cluster compute instances for running MPI applications,” in *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2011, pp. 1–10.

[8] “UberCloud Experiment.” [Online]. Available: https://www.theubercloud.com/hpc-experiment/. [Accessed: 14-Jun-2015].

[9] “Arctur-1 supercomputer.” [Online]. Available: http://en.wikipedia.org/wiki/Arctur-1. [Accessed: 10-Aug-2014].

[10] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn, and T. J. Purcell, “A survey of general-purpose computation on graphics hardware,” *Comput. Graph. Forum*, vol. 26, no. August, pp. 80–113, 2007.

[11] G. Galante and L. C. E. De Bona, “A survey on cloud computing elasticity,” in *Proceedings - 2012 IEEE/ACM 5th International Conference on Utility and Cloud Computing, UCC 2012*, 2012, pp. 263–270.

[12] A. Nagarajan and F. Mueller, “Proactive fault tolerance for HPC with Xen virtualization,” in *Proceedings of the 21st annual international conference on supercomuting*, 2007, pp. 23–32.

[13] D. Walker and J. Dongarra, “MPI: a standard message passing interface,” *Supercomputer*, vol. 12, pp. 56–68, 1996.

[14] “Fortissimo EU Project.” [Online]. Available: http://www.fortissimo-project.eu. [Accessed: 15-Jan-2015].

[15] “CloudFlow EU Project.” [Online]. Available: http://www.eu-cloudflow.eu/. [Accessed: 15-Jan-2015].

INTERVAL CONVEX QUADRATIC PROGRAMMING PROBLEMS IN A GENERAL FORM

Milan Hladík
Charles University, Faculty of Mathematics and Physics,
Department of Applied Mathematics,
Malostranské nám. 25, 11800, Prague, Czech Republic,
e-mail: milan.hladik@matfyz.cz

Abstract: This paper addresses the problem of computing the minimal and the maximal optimal value of a convex quadratic programming (CQP) problem when the coefficients are subject to perturbations in given intervals. Contrary to the previous results concerning on some special forms of CQP only, we present a unified method to deal with interval CQP problems. The problem can be formulated by using equation, inequalities or both, and by using sign-restricted variables or sign-unrestricted variables or both. We propose simple formulas for calculating the minimal and maximal optimal values. Due to NP-hardness of the problem, the formulas are exponential with respect to some characteristics. On the other hand, there are large sub-classes of problems that are polynomially solvable.

Keywords: convex quadratic programming, interval analysis, best case, worst case, uncertainty modeling.

1 INTRODUCTION

Optimization problems with interval data has been an intensively studied subject in recent years. Most of the authors deal with interval-valued linear programming problems; see Hladík [4] for a survey. In this paper, we focus on convex quadratic programming problems with interval data. Despite importance of this topic, there are considerable less results; see, e.g., [3, 6]. Therein, the problem of computing the best case and the worst case optimal values was discussed for convex quadratic programming problems of certain forms. The purpose if this paper is to generalize the above results and to propose algorithms for computing the optimal value range of interval-valued convex quadratic programming problems in a general form.

Notation. An interval matrix is defined as

$$\boldsymbol{A} := \{A \in \mathbb{R}^{m\times n};\ \underline{A} \le A \le \overline{A}\},$$

where $\underline{A}$ and $\overline{A}$, $\underline{A} \le \overline{A}$, are given matrices and the inequality is understood entrywise. The midpoint and radius matrices are defined as

$$A_c := \frac{1}{2}(\underline{A} + \overline{A}), \quad A_\Delta := \frac{1}{2}(\overline{A} - \underline{A}).$$

The set of all interval $m \times n$ matrices is denoted by $\mathbb{IR}^{m\times n}$. The diagonal matrix with entries $s_1, \dots, s_n$ is denoted by $\operatorname{diag}(s)$. In accordance with Fiedler et al. [1], for an interval matrix $\boldsymbol{A} \in \mathbb{IR}^{m\times n}$ and sign vectors $r \in \{\pm 1\}^m$, $s \in \{\pm 1\}^n$, we define the real matrix $A_{rs} := A_c - \operatorname{diag}(r)A_\Delta \operatorname{diag}(s) \in \boldsymbol{A}$. Similarly, for an interval vector $\boldsymbol{b} \in \mathbb{IR}^m$ and a sign vector $r \in \{\pm 1\}^m$, we define the real vector $b_r := b_c + \operatorname{diag}(r)b_\Delta \in \boldsymbol{b}$.

Absolute value applied for matrices and vectors is understood entry-wise. Eventually, $e = (1, \dots, 1)^T$ stands for the vector of ones with suitable dimension.

Problem formulation. Consider the convex quadratic programming (CQP) problem in a general form

$$\min \ \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} P & R \\ R^T & S \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + a^T x + c^T y \ \text{ subject to } \ Ax + By = b, \ Cx + Dy \leq d, \ x \geq 0, \tag{1}$$

where $P \in \mathbb{R}^{n \times n}$, $R \in \mathbb{R}^{n \times n'}$, $S \in \mathbb{R}^{n' \times n'}$, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m \times n'}$, $C \in \mathbb{R}^{m' \times n}$, $D \in \mathbb{R}^{m' \times n'}$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^{n'}$ and $d \in \mathbb{R}^{m'}$. For shortening, we will also use the form

$$f(Q, q, W) := \min \ z^T Q z + q^T z \ \text{ subject to } \ z \in M(W), \tag{2}$$

where

$$z := (x^T \ y^T)^T, \quad Q := \begin{pmatrix} P & R \\ R^T & S \end{pmatrix}, \quad q := \begin{pmatrix} a \\ c \end{pmatrix},$$

$M(W)$ is the feasible set and W gathers A, B, C, D, b and d.

Let respective interval domains $\boldsymbol{Q}$, $\boldsymbol{q}$ and $\boldsymbol{W}$ for the coefficients be given. The problem is to determine the range of the optimal values when the coefficients are subject to perturbations in the given interval domains. Formally, we want to compute

$$\underline{f} := \min \ f(Q, q, W) \ \text{ subject to } \ Q \in \boldsymbol{Q}, \ q \in \boldsymbol{q}, \ W \in \boldsymbol{W},$$
$$\overline{f} := \max \ f(Q, q, W) \ \text{ subject to } \ Q \in \boldsymbol{Q}, \ q \in \boldsymbol{q}, \ W \in \boldsymbol{W}.$$

We will assume that Q is positive definite for all $Q \in \boldsymbol{Q}$. For positive semi-definiteness of interval matrices see, e.g., Rohn [8, 9].

The problem of computing $\underline{f}$ and $\overline{f}$ was considered in Hladík [3] for nonnegative variables only, and in Li et al. [6], it was proposed another method that avoids some assumptions. The general case dealing with variables unrestricted in sign has not been considered yet. We fill this gap by developing methods for the most general model stated above.

2 LOWER BOUND

The lower bound $\underline{f}$ can be expressed as follows

$$\underline{f} := \min \ \left(\min_{Q \in \boldsymbol{Q}, \, q \in \boldsymbol{q}} (z^T Q z + q^T z) \right) \ \text{ subject to } \ z \in \bigcup_{W \in \boldsymbol{W}} M(W).$$

According to Hladík [5], the set $\cup_{W \in \boldsymbol{W}} M(W)$ is described by the nonlinear system

$$\underline{A} x + B_c y \leq B_\Delta |y| + \overline{b}, \tag{3a}$$
$$-\overline{A} x - B_c y \leq B_\Delta |y| - \underline{b}, \tag{3b}$$
$$\underline{C} x + D_c y \leq D_\Delta |y| + \overline{d}, \ x \geq 0. \tag{3c}$$

or, equivalently, as union of linear systems

$$\underline{A} x + B_{es} y \leq \overline{b}, \tag{4a}$$
$$-\overline{A} x - B_{-es} y \leq -\underline{b}, \tag{4b}$$
$$\underline{C} x + D_{es} y \leq \overline{d}, \ x \geq 0 \tag{4c}$$

over $s \in \{\pm 1\}^{n'}$.

The objective function

$$\min_{Q \in \boldsymbol{Q}, \, q \in \boldsymbol{q}} (z^T Q z + q^T z) = \min_{Q \in \boldsymbol{Q}} z^T Q z + \min_{q \in \boldsymbol{q}} q^T z$$

is discussed below.

Proposition 2.1 *We have*

$$\begin{aligned}
\min_{Q\in\boldsymbol{Q}} z^TQz &= z^TQ_cz - |z|^TQ_\Delta|z|,\\
\max_{Q\in\boldsymbol{Q}} z^TQz &= z^TQ_cz + |z|^TQ_\Delta|z|,\\
\min_{q\in\boldsymbol{q}} q^Tz &= q_c^Tz - q_\Delta^T|z|,\\
\max_{q\in\boldsymbol{q}} q^Tz &= q_c^Tz + q_\Delta^T|z|.
\end{aligned}$$

Proof. We estimate

$$z^TQz = z^TQ_cz + z^T(Q-Q_c)z \geq z^TQ_cz - |z|^T\cdot|Q-Q_c|\cdot|z| \geq z^TQ_cz - |z|^TQ_\Delta|z|.$$

On the other hand, denoting $s := \operatorname{sgn}(z)$, we derive

$$z^TQ_cz - |z|^TQ_\Delta|z| = z^TQ_cz - \operatorname{diag}(s)z^TQ_\Delta\operatorname{diag}(s)z = z^TQ_{ss}z.$$

Thus, $z^TQ_cz - |z|^TQ_\Delta|z|$ is attained for $Q := Q_{ss} \in \boldsymbol{Q}$.

The second part of the statement is analogous, and the remaining two parts are well known [1, 7]. □

Based on these observation, we arrive at

$$\underline{f} = \min\ z^TQ_cz - |z|^TQ_\Delta|z| + q_c^Tz - q_\Delta^T|z| \ \text{ subject to } \ (3). \tag{5}$$

By using an orthant-by-orthant decomposition, we obtain a characterization of $\underline{f}$ by means of solving $2^{n'}$ convex quadratic programs:

Theorem 2.2 *We have*

$$\underline{f} = \min_{s\in\{\pm1\}^{n'}} f_s,$$

where

$$f_s := \min\ z^T\begin{pmatrix}\underline{P} & R_{es}\\ (R_{es})^T & S_{ss}\end{pmatrix}z + \begin{pmatrix}\underline{a}^T & c_{-s}^T\end{pmatrix}z \quad \textit{subject to} \quad (4).$$

Proof. By using (5), we can express $\underline{f}$ as

$$\underline{f} = \min_{s\in\{\pm1\}^{n'}}\ \min\ z^TQ_cz - |z|^TQ_\Delta|z| + q_c^Tz - q_\Delta^T|z| \ \text{ subject to } \ (3),\ \operatorname{diag}(s)y \geq 0.$$

Since $|x| = x$ and $|y| = \operatorname{diag}(s)y$, we can write the above inner problem as f_s. □

By the above theorem, $\underline{f}$ can be calculated by solving $2^{n'}$ real CQP problems. Thus our approach is exponential with respect to the number of free variables. This drawback, however, can hardly be overcome since the problem of determining $\underline{f}$ is NP-hard. Moreover, NP-hardness was proved even for the special case of determining the best optimal value of the linear programming problem $\min c^Ty$ subject to $Dy \leq d$, where $c \in \boldsymbol{c}$ and D, d are fixed (see [2, 4]). On the other hand, $\underline{f}$ is effectively computed provided all variables are restricted in sign.

It might be useful to know for which concrete realization of interval coefficients the best case value $\underline{f}$ is attained. The corresponding realization for P, R, S, a, c is given in Theorem 2.2. The remaining values of A, B, C, D, b, d can be determined in the way as in Hladík [5]. If x^*, y^* is an optimal solution to the minimal f_s from Theorem 2.2, then the value of $\underline{f}$ is attained for

$$\begin{aligned}
A &= A_c - \operatorname{diag}(u)A_\Delta, & B &= B_c - \operatorname{diag}(u)B_\Delta\operatorname{diag}(\operatorname{sgn}(y^*)),\\
C &= \underline{C}, & D &= D_c - D_\Delta\operatorname{diag}(\operatorname{sgn}(y^*)),\\
b &= b_c + \operatorname{diag}(u)b_\Delta, & d &= \overline{d},
\end{aligned}$$

where $u \in [-1,1]^m$ is defined as

$$u_i = \begin{cases} \frac{(A_c x^* + B_c y^* - b_c)_i}{(A_\Delta x^* + B_\Delta |y^*| + b_\Delta)_i} & \text{if } (A_\Delta x^* + B_\Delta |y^*| + b_\Delta)_i > 0, \\ 1 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, m.$$

3 UPPER BOUND

Theorem 3.1 *We have*

$$\overline{f} = \max_{s \in \{\pm 1\}^{n'},\, r \in \{\pm 1\}^m} f_{s,r},$$

where $f_{s,r}$ *is the optimal value of* (1) *with input data*

$$\begin{array}{llll} A := A_{re}, & B := B_{rs}, & C := \overline{C}, & D := D_{-es}, \\ a := \overline{a}, & b := b_r, & c := c_s, & d := \underline{d}, \\ P := \overline{P}, & R := R_{-e,s}, & S := S_{-s,s}. & \end{array}$$

Proof. The upper bound $\overline{f}$ can be expressed as

$$\overline{f} = \max_{s \in \{\pm 1\}^{n'}} g_s,$$

where

$$g_s := \max_{Q \in \boldsymbol{Q},\, q \in \boldsymbol{q},\, W \in \boldsymbol{W}} \ \min \ z^T Q z + q^T z \ \text{ subject to } \ z \in M(W),\ \operatorname{diag}(s) y \geq 0.$$

By substitution $y' := \operatorname{diag}(s) y$, this problem takes the form of an interval CQP problem with nonnegative variables. By Li et al. [6], its worst case optimal value g_r can be expressed as

$$g_r = \max_{r \in \{\pm 1\}^m} (1),$$

where the instances of (1) are as stated in the theorem. □

Theorem 3.1 shows that $\overline{f}$ can be determined by solving $2^{m+n'}$ real CQP problems, and so the algorithm is exponential with respect to the number of equations and free variables. This exponential complexity is not surprising since the problem of computing $\overline{f}$ contains several NP-hard problems as sub-classes. For example:

- Testing whether the linear system $By = b$ is unsolvable for some $B \in \boldsymbol{B}$ and $b \in \boldsymbol{b}$ (see Fiedler et al. [1]).
- Computing $\overline{f}$ in linear programming with interval right-hand side

  $$\min a^T x \ \text{ subject to } \ Ax = b,\ x \geq 0,$$

  where a and A is fixed and $b \in \boldsymbol{b}$ (see [2, 4]).
- Maximizing a convex quadratic function on a box, which is equivalent to compute $\overline{f}$ for the subclass

  $$\min y^T S y \ \text{ subject to } \ y = b,$$

  where S is fixed and $b \in \boldsymbol{b}$ (see Vavasis [10]).

On the other hand, still there might be an algorithm being exponential with respect to m and polynomial with respect to n'. A natural approach would be to utilize duality theory. However, contrary to duality in linear programming, the dual problem here has some correlations between coefficients, which makes the problem with interval input very hard. Anyway, this makes a motivating question for further research.

In contrast, there are also completely polynomially solvable subclasses such as

$$\min\ x^T Px + a^T x \ \text{ subject to } \ Cx \leq d,\ x \geq 0,$$

where $P \in \boldsymbol{P}$, $a \in \boldsymbol{a}$, $C \in \boldsymbol{C}$ and $d \in \boldsymbol{d}$. Moreover, an easy to solve problem might possess also equations and free variables provided the corresponding coefficients are fixed and not interval-values. To be specific, the problem

$$\min\ \begin{pmatrix} x^T & y^T \end{pmatrix} \begin{pmatrix} P & R \\ R^T & S \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + a^T x + c^T y \ \text{ subject to } \ Ax + By = b,\ Cx + Dy \leq d,\ x \geq 0,$$

where $P \in \boldsymbol{P}$, $a \in \boldsymbol{a}$, $C \in \boldsymbol{C}$ and $d \in \boldsymbol{d}$, and the other matrices and vectors are fixed, is reduced to solving just one CQP problem (1) with $P := \overline{P}$, $a := \overline{a}$, $C := \overline{C}$ and $d := \underline{d}$.

4 CONCLUSION

We proposed formulas for computing the best case and the worst case optimal value of an interval-valued CQP problem. We considered the general model with arbitrary linear constraints. Thus, we extended the previous results relying on nonnegative variables only.

In the future research, we aim at developing some approximation algorithms for intractable cases as well as considering even more general model of quadratically constrained interval-valued CQP.

5 Acknowledgements

The author was supported by the Czech Science Foundation Grant P402/13-10660S.

References

[1] Fiedler, M., Nedoma, J., Ramík, J., Rohn, J. and Zimmermann, K. (2006). Linear Optimization Problems with Inexact Data. New York: Springer.

[2] Gabrel, V., Murat, C. and Remli, N. (2010). Linear programming with interval right hand sides. Int. Trans. Oper. Res., 17(3):397–408.

[3] Hladík, M. (2011). Optimal value bounds in nonlinear programming with interval data. TOP, 19(1):93–106.

[4] Hladík, M. (2012). Interval linear programming: A survey. In Mann, Z.A. (Ed.). Linear Programming - New Frontiers in Theory and Applications (chapter 2, pp. 85–120). New York: Nova Science Publishers.

[5] Hladík, M. (2013). Weak and strong solvability of interval linear systems of equations and inequalities. Linear Algebra Appl., 438(11):4156–4165.

[6] Li, W., Xia, M. and Li, H. (2015). New method for computing the upper bound of optimal value in interval quadratic program. J. Comput. Appl. Math., 288(0):70–80.

[7] Neumaier, A. (1990). Interval Methods for Systems of Equations. Cambridge: Cambridge University Press.

[8] Rohn, J. (1994). Positive definiteness and stability of interval matrices. SIAM J. Matrix Anal. Appl., 15(1):175–184.

[9] Rohn, J. (2012). A manual of results on interval linear problems. Technical Report 1164, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague.

[10] Vavasis, S.A. (1991). Nonlinear Optimization: Complexity Issues. New York: Oxford University Press.

HYBRID GENETIC ALGORITHMS FOR THE CAPACITATED *p*-MEDIAN PROBLEM

Ľudmila Jánošíková
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 1, 010 26 Žilina, Slovak Republic
Ludmila.Janosikova@fri.uniza.sk

Michal Haviar
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 1, 010 26 Žilina, Slovak Republic
Michal.Haviar@st.fri.uniza.sk

Abstract: The paper suggests several ways how to combine a genetic algorithm with integer programming to improve the quality of the problem solution. The motivation is that today's integer programming solvers are very sophisticated and efficient and they are worth utilizing in combination with metaheuristics to solve hard combinatorial optimization problems. The capacitated *p*-median problem is chosen as an example of such a problem that is intractable for an exact method and needs a heuristic or metaheuristic method, e.g. a genetic algorithm to get a near-optimal solution. A genetic algorithm can be combined with integer programming in such a way that the metaheuristics acts at a higher level and controls the calls to the solver. The solver can be used for: (i) fitness calculation, (ii) improving the best solution, and (iii) generating elite solutions. Several variants of the hybrid genetic algorithm are proposed and tested using benchmark instances.

Keywords: capacitated *p*-median problem, genetic algorithm

1 INTRODUCTION

The capacitated *p*-median problem arises in cluster analysis where a set of entities is to be partitioned into a set of clusters [5, 8]. It is one of the models that can be used to locate centres serving spatially distributed customers. In [2, 3] it was studied in the context of emergency medical service (EMS) stations location in a large territory.

We are given a set I of candidate locations, where emergency stations can be placed. Each station is equipped with one ambulance. The population of the region served by one ambulance cannot exceed Q people. Potential patients live in municipalities spatially spread in the given territory. We denote the set of municipalities by J and the number of inhabitants of municipality $j \in J$ by b_j. Further let t_{ij} be the shortest travel time of an ambulance from node $i \in I$ to node $j \in J$. The goal is to locate p stations in nodes from the set I in order to minimise the total travel time needed to reach all potential patients.

The decision on opening a station must be done for each candidate location $i \in I$. To model this decision we need a binary variable y_i, which takes the value 1 if a station is located in node i, otherwise it takes the value 0. The assignment of municipality j to the station located in node i is modelled by binary variables x_{ij}. Variable x_{ij} takes value 1, if municipality j will be served by an ambulance located in node i, otherwise $x_{ij} = 0$. The model of the capacitated *p*-median problem can be written as:

$$\textit{minimise} \quad \sum_{i \in I} \sum_{j \in J} t_{ij} b_j x_{ij} \tag{1}$$

$$\textit{subject to} \quad \sum_{i \in I} x_{ij} = 1 \qquad \textit{for } j \in J \tag{2}$$

$$x_{ij} \leq y_i \qquad \textit{for } i \in I,\ j \in J \tag{3}$$

$$\sum_{j \in J} b_j x_{ij} \leq \text{Q} \qquad \text{for } i \in I \tag{4}$$

$$\sum_{i \in I} y_i = p \tag{5}$$

$$x_{ij}, y_i \in \{0,1\} \qquad \text{for } i \in I,\ j \in J \tag{6}$$

The objective function (1) minimises the weighted travel time between stations and municipalities. Constraints (2) ensure that every municipality *j* will be assigned to exactly one station *i*. Constraints (3) ensure that if a municipality *j* is assigned to a node *i*, then a station will be open in the node *i*. Constraints (4) limit the total number of inhabitants in the region served by one ambulance. Constraint (5) limits the total number of stations that can be sited. The remaining obligatory constraints (6) specify the definition domains of the variables.

2 A HYBRID GENETIC ALGORITHM

The capacitated *p*-median problem is known to be an *NP*-hard problem. It means that it cannot be solved to optimality in a reasonable time and a heuristic or metaheuristic approach should be used instead. An efficient decomposition heuristic is proposed in [7]. The weakness of a decomposition method is that it can hardly ensure global constraints. For example if the *p*-median model is used to relocate current stations one may require that only a predefined percentage of the current station locations can be changed. Population-based metaheuristics may be more suitable in such situations since they can easily control global constraints by some mutation operator.

In recent years, a lot of attention has been devoted to the integration of metaheuristics with exact methods. The motivation behind such hybridizations of different algorithmic concepts is usually to obtain better performing systems that exploit and unite advantages of the individual pure strategies [6]. Hybrid algorithms often exhibit significantly higher performance with respect to solution quality and time.

We decided to examine abilities of a combined algorithm consisted of a genetic algorithm (GA) and integer programming (IP) based methods.

The genetic algorithm is a well-known population-based metaheuristics. It evolves the population of candidate solutions to a given problem by repeatedly applying operators based on natural selection and genetic recombination to the current population. Our implementation of the GA for the capacitated *p*-median problem follows the implementation details described in [1]. A crossover operator is applied in every generation to create two new individuals from a pair of selected individuals. To ensure a diversity of population, the offspring goes through mutation with a pre-set probability. The population is renewed in a steady-stated method: the offspring is included in the population if it has better fitness value than the worst individual in the old population. The process finishes after a pre-defined number of generations has evolved or a pre-defined amount of computation time has elapsed. The individual in the last population with the lowest fitness value represents the best solution.

We do not design a special integer programming method but we exploit the strength of a general IP solver.

We propose three combinations of the GA and IP:

A. **Integer programming is used to calculate the fitness value of an individual**. A candidate solution is represented in the chromosome incompletely as a list of indices of those nodes, where medians are located. To evaluate the fitness value, customers must be assigned to medians. In the original GA, the assignment is calculated by a heuristic. But the assignment problem can also be formulated as an IP problem and solved exactly. From

mathematical programming point of view, the model (1)-(6) is to be solved where variables $\boldsymbol{y}$ are fixed according to the individual's chromosome and variables $\boldsymbol{x}$ need to be calculated. However the exact calculation takes by an order more time than the heuristic one and the time consumption is not compensated by the quality of the final solution. The time needed to more precise fitness evaluation can be devoted to explore more solutions. Therefore we use the solver only once after the GA termination to calculate the assignment of customers to the medians represented by the best individual.

B. **Integer programming is used as a postprocessing technique. A**fter the GA has finished, the *p*-median problem (1)-(6) is solved starting from the location of medians given by the best individual.

C. **Integer programming is embedded in the GA to generate elite individuals.** The idea is that in the neighbourhood of some good quality individuals there may exist even better individuals. Such an operation that searches for elite individuals in the neighbourhood of selected individuals is called hypermutation in [1]. It is applied with a very low probability to each iteration of the steady-state method. In our implementation, the hypermutation starts with randomly selecting an individual out of the best 10% individuals and then explores its neighbourhood in an exact fashion. It means the *p*-median problem (1)-(6) is solved starting with the initial solution given by the selected individual. The neighbourhood of the individual is defined such that only β closest locations to each current median are regarded as candidates. When the GA terminates, the local improvement of the best solution is again performed by the solver.

3 COMPUTATIONAL EXPERIMENTS

We tested the proposed three versions of the hybrid GA on two sets of benchmark instances Both sets were proposed by Lorena and Senne and are available at site http://www.lac.inpe.br/~lorena/instancias.html. The first set comprises six real instances named *sjc*1 to *sjc*4*b* of size ranging from 100 to 402 nodes and 10 to 40 medians. These instances can be solved exactly by a solver in several minutes. The second set contains five instances named *p*3038_600 to *p*3038_1000. They consists of 3038 nodes and 600 to 1000 medians are to be placed. Optimal solutions of these instances have not been published so far, so our results are compared to lower bounds obtained using a column generation method [4] and published in [7]. The sets of candidate locations and customers are identical in all instances, i.e. every customer can be a median. The number of customers is denoted by *n* in the presented results.

The computational experiments were performed on a personal computer equipped with the Intel Core i7 processor with 1.60 GHz and 8 GB RAM. The main application was implemented in Java language and the solver Gurobi Optimizer 6.0.0 was used.

The goal of the first experiment was to find out whether the final solution of the GA can be improved using the LP solver. It was performed on the first set of *sjc* instances. The parameters of the GA were set as follows: population size – 100 individuals, mutation rate – 20%, stopping criterion – 1000 generations, $\beta = 10$.

Tables 1 and 2 compare three methods: the original GA, a hybrid GA in version A, where the solver is used to calculate the optimal assignment of customers to medians present in the best individual, and a hybrid GA in version B, where the best solution found by the GA is improved by the solver. Table 1 compares the methods in terms of the objective function. The best solutions out of 10 runs and standard deviation of the best solutions are presented for all three implementations of the GA. The best results are emphasized in bold. Table 2 contains computation time that Gurobi Optimizer takes to generate optimal solution and average computation time of one run of the GA.

Table 1: Comparison of the original and hybrid GAs in terms of the objective function

			Optimal	*Original GA*		*Hybrid GA v. A*		*Hybrid GA v. B*	
Inst.	***n***	***p***	***solution***	***Best sol.***	***St. dev.***	***Best sol.***	***St. dev.***	***Best sol.***	***St. dev.***
sjc1	100	10	**17288.99**	17564.10	236.34	17454.38	168.74	**17288.99**	**13.06**
sjc2	200	15	**33270.94**	33861.24	403.11	33649.16	383.65	**33293.40**	**38.54**
sjc3a	300	35	**45335.16**	47320.11	406.61	47201.54	363.78	**45338.01**	**105.55**
sjc3b	300	30	**40635.90**	41550.29	596.17	41910.20	463.47	**40635.90**	**101.51**
sjc4a	402	30	**61925.51**	65752.96	448.19	64858.11	921.04	**62069.75**	**151.78**
sjc4b	402	40	**52458.02**	55122.93	359.51	54599.66	572.92	**52487.22**	**103.06**

Table 2: Comparison of the original and hybrid GAs in terms of the computation time (in seconds)

Inst.	***n***	***p***	***Solver***	***Original GA***	***Hybrid GA v. A***	***Hybrid GA v. B***
sjc1	100	10	14.10	0.67	0.73	9.47
sjc2	200	15	20.75	1.16	1.23	11.07
sjc3a	300	35	76.48	2.01	2.28	55.25
sjc3b	300	30	21.09	2.05	2.28	22.83
sjc4a	402	30	713.70	2.85	3.20	236.59
sjc4b	402	40	61.09	3.08	3.40	16.56

As can be seen, the best implementation of the GA is version B, where the best solution after 1000 generations is improved by the solver. Average gap between the best solution and the optimal one is 3.54% for the original GA, 3.03% for the hybrid GA version A and only 0.06% for the hybrid GA version B. In two benchmark instances (*sjc*1 and *sjc*3*b*) the hybrid GA version B found the optimal solution at least once out of 10 replications. At the same time, version B produces the most balanced results (standard deviation is the least). On the other hand, the computation time of version B increases compared to the original GA or version A by one or even two orders. However, it is still faster compared to the exact method.

Based on these results, we proposed the second experiment on larger instances. Here version B is compared with version C, where the solver supplies elite solutions by the hypermutation operation.

If the instance has thousands of candidates and customers, the number of assignment variables $\boldsymbol{x}$ is huge. In order to decrease computational complexity, we propose a reduction of the solution space. The mathematical programming model is reduced by heuristic elimination of those variables $\boldsymbol{x}$ which are less likely to belong to a good or optimal solution. The elimination is based on the assumption that customers will not be served by those centres that are too far away. That is why only those variables $\boldsymbol{x}$ are included in the model for which coefficient t_{ij} is less than a predefined threshold. The threshold is defined by the value $\alpha \cdot t^{\max} / \sqrt{p}$ where $t^{\max} = \max\{t_{ij} : i \in I, j \in J\}$ and α is a parameter (α = 1.5 in our experiments).

As in the former tests, population has 100 individuals and mutation rate is 20%. Hypermutation rate is 0.001%. In order compare the implementations mutually, the stopping criterion is the total running time. Due to large size of instances, the computation time of the solver is limited. These time limits are set so that the total computation time is about 1 hour. In version B, the GA runs 45 minutes and then the solver runs 15 minutes. In this limited time the GA performs approximately 128000 iterations for the 600 median problem and 54000 iterations for the 1000 median problem. In version C the solver runs 150 seconds. To keep the running-time of 1 hour, the GA runs 3450 seconds and then the solver starts from the best solution and runs 150 seconds. This way the GA performs about 20000 iterations and 20 hypermutations. Since the computation time for the solver is rather long and the

hypermutation is called randomly, the substantial differences in the number of generations for various instances are not observed. The results are summarised in Table 3.

Table 3: Comparison of versions B and C of the hybrid GA for larger instances

Instance	*n*	*p*	*Lower bound*	*Hybrid GA v. B*		*Hybrid GA v. C*	
				Best sol.	*St. dev.*	*Best sol.*	*St. dev.*
***p*3038_600**	3038	600	122020.66	125345.34	597.11	**123980.88**	**116.92**
***p*3038_700**	3038	700	108685.59	112000.00	888.00	**111021.47**	**300.21**
***p*3038_800**	3038	800	98530.99	**102216.16**	981.56	102854.88	**487.53**
***p*3038_900**	3038	900	90239.65	**96137.00**	**2804.22**	96194.96	3674.77
***p*3038_1000**	3038	1000	83231.58	**91085.83**	2922.47	93108.42	**2404.42**

Comparison of two versions of the GA for large-sized instances does not reveal the prevalence of one version. Average gap between the best solution and the lower bound is 5.10% for version B and 5.32% for version C. However the quality of the solutions for version C is more balanced (standard deviation is in 4 cases lower than for version B). That is why we consider the combination of the GA with the solver producing elite solutions to be the most appropriate.

4 CONCLUSIONS

Our research suggests that combination of the genetic algorithm with integer programming is a promising method for solving a large-sized capacitated *p*-median problem. It reveals that the most appropriate combination is an integrative algorithm where the exact mathematical programming based method is embedded into the genetic algorithm to generate high-quality solutions. The solver is called with a small probability at each generation of the genetic algorithm. It solves the *p*-medial problem starting with the initial solution given by a randomly chosen good individual. The strength of this method becomes apparent in a parallel environment where elite solutions may be generated concurrently on one or more computation nodes and distributed to the other nodes that perform the genetic algorithm. Further research on this topic will concentrate on parallelization of the proposed hybrid algorithm and on development of a more efficient crossover operator that would exploit the knowledge about the problem being solved.

Acknowledgement
This research was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences under project VEGA 1/0518/15 “Resilient rescue systems with uncertain accessibility of service” and by the Slovak Research and Development Agency under project APVV-0760-11 “Designing of Fair Service Systems on Transportation Networks”.

References

[1] Correa, E.A., Steiner, M.T.A., Freitas, A.A. and Carnieri, C. (2004). A genetic algorithm for solving a capacitated *p*-median problem. Numerical Algorithms, 35: 373–388.

[2] Gábrišová, L. and Janáček, J. (2015). Design of capacitated emergency service system. Communications : Scientific letters of the University of Žilina, 17(2): 42–48.

[3] Jánošíková, Ľ. and Žarnay, M. (2014). Location of emergency stations as the capacitated *p*-median problem. Proceedings of the International Scientific Conference Quantitative Methods in Economics – Multiple Criteria Decision Making XVII, Virt, Slovak Republic, 28 – 30 May 2014. Bratislava: Ekonóm. pp. 116–122.

[4] Lorena, L.A.N. and Senne, E.L.F. (2004). A column generation approach to capacitated *p*-median problems. Computers & Operations Research, 31: 863–876.

[5] Osman, I.H. and Christofides, N. (1994). Capacitated clustering problems by hybrid simulated annealing and tabu search. International Transactions in Operational Research, 1(3): 317–336.

[6] Raidl, G.R. (2006). A unified view on hybrid metaheuristics. In Almeida, F., Blesa Aguilera, M.J., Blum, C., Moreno Vega, J.M., Pérez, M.P., Roli, A. and Sampels, M. (Eds.). Hybrid Metaheuristics, volume 4030 of Lecture Notes in Computer Science (pp. 1–12). Berlin Heidelberg: Springer.

[7] Stefanello, F., de Araújo, O.C.B. and Müller, F.M. (2012). Matheuristics for the capacitated *p*-median problem. Presentation at the 4th International Workshop on Model Based Metaheuristics. 17 – 20 September, 2012, Rio de Janeiro, Brazil.

[8] Taillard, É.D. (2003). Heuristic methods for large centroid clustering problems. Journal of Heuristics, 9: 51–73.

On robustness in deterministic linear programming

Latif Pourkarimi
Department of Mathematics, Razi University, Kermanshah, Iran;
E-mail: lp_karimi@yahoo.com.com

Majid Soleimani-damaneh
School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran & School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P.O. Box: 19395-5746, Tehran, Iran; E-mail: soleimani@khayam.ut.ac.ir

Abstract:
This paper deals with the robustness issue in deterministic linear programming from two new standpoints. Corresponding to an optimal solution, a new quantity, robustness order, is defined with respect to the degree of interiority of the cost vector in the binding cone. A linear programming problem is provided to calculate the robustness order of a given optimal solution. The second part of the paper is devoted to investigating the robustness with respect to the eligible angle deviation of the cost vector in the binding cone. Some theoretical results are given to obtain the maximum eligible angle deviation, and finally the connection between two above-mentioned robustness standpoints is established.

Keywords: Linear programming, Robustness, Angle deviation, Relative interior

1 INTRODUCTION

In this paper, the following linear programming (LP) problem is considered

$$LP(c): \quad \max \quad c^t x \quad \text{s.t.} \quad Ax \leq b, \tag{1}$$

where $c \in R^n$, $A \in R^{m\times n}$, and $b \in R^m$. The feasible set of this problem is denoted by $X = \{x \in R^n : \ Ax \leq b\}$. By LP($c$) we denote an LP with objective coefficient vector c. The superscript "t" stands for transpose. In the whole paper, the inequalities between two vectors are componentwise. For matrix A, the i-th row is denoted by a^i and the j-th column is denoted by a_j. Let $x^* \in X$ be a feasible solution of LP (1). $I(x^*)$ denotes the index set of binding constraints of LP (1) at x^*, i.e. $I(x^*) = \{i : \ (a^i)^t x^* = b_i\}$. Also, $A_{I(x^*)}$ is the sub-matrix of A whose rows are a^i's with $i \in I(x^*)$. The convex cone generated by the rows of $A_{I(x^*)}$, denoted by $\mathcal{A}^*$, is called the *binding cone*: $\mathcal{A}^* = cone\left(\left\{(a^i)^t : \ i \in I(x^*)\right\}\right)$. This cone plays a vital role in our work.

Studying robust solutions, the optimal solutions of LP(c) which are insensitive to some changes of the problem data (changes of the cost vector in the present paper), is an important issue from an applied point of view; see e.g. Deb and Gupta (2006), Ben-Tal et al. (2009), and Zamani et al. (in press). Since a necessary and sufficient condition for optimality of a feasible solution x^* is $c \in \mathcal{A}^*$, there is a strong connection between the robustness of an optimal solution and preserving c in $\mathcal{A}^*$ by some alterations in c. According to this fact, we define the robustness with respect to the value of possible alterations in the cost vector c under remaining this vector in the binding cone. The value of this alteration is dealt with from two sights: the eligible angle deviation which preserves c in the binding cone and the degree of interiority of c in the binding cone. Due to this, corresponding to an optimal solution, a new quantity, robustness order, is defined using the degree of interiority. An LP problem is given to calculate the robustness order of a given optimal solution. Then the robustness is defined with respect to the eligible angle deviation of the cost vector in the binding cone. Robust solutions in two above-mentioned senses are characterized and finally the connection between

two robustness standpoints is established. Moreover, we show that a robustness notion existing in the literature is equivalent to the uniqueness of the optimal solution.

2 PRELIMINARIES

We say that $x^0 \in S \subseteq R^n$ is a relative interior point of order k for S if there exist scalar $0 < \epsilon \in R$ and an affine subspace H with $\dim(H) = k$ such that $B(x^0;\epsilon) \cap H \subseteq S$, where $B(x^0;\epsilon) = \{x \in R^n : \ \|x - x^0\| < \epsilon\}$ is the ball centered at x^0 with radius ϵ. For the sake of convenience, in this paper we use the Euclidean norm for vectors. The largest k with this property is called the degree of interiority of x^0 in S. If $k = \dim(S)$, then x^0 is called a relative interior point of S. It is clear that, if $k = n$, then x^0 is an interior point of S. We denote the set of relative interior points of S by $ri(S)$. Also, the set of relative interior points of order k for S is denoted by $ri(S;k)$. The notation $|S|$ is used to denote the cardinality of S. The convex cone generated by $\{a_1, a_2 \ldots, a_m\}$ is denoted by $cone(\{a_1, a_2 \ldots, a_m\})$.

Let $\alpha \geq 0$ and $a_1, a_2, \ldots, a_m$ be some given vectors in R^n. For a representation $\sum_{j=1}^m \alpha_j a_j$, we use the following notations hereafter: A_α denotes the matrix whose columns are a_j's with $\alpha_j > 0$; and

$$J(\alpha) = \{j : \alpha_j > 0\}. \tag{2}$$

Also, for row vectors $a^1, a^2, \ldots, a^m$ and a representation $\sum_{i=1}^m \alpha_i a^i$, the matrix A^α denotes the matrix whose rows are a^i's with $\alpha_i > 0$.

Lemma 2.1 *Let $\mathcal{F} = cone(\{a_1, a_2, \ldots, a_m\})$ and $x \in \mathcal{F}$. Then, $x \in ri(\mathcal{F};k)$ if and only if there exists nonnegative vector $\alpha \in R^m$ such that $x = \sum_{j=1}^m \alpha_j a_j$ and $rank(A_\alpha) = k$.*

3 ROBUSTNESS W.R.T. THE INTERIORITY OF c

The feasible solution $x^* \in X$ is an optimal solution of LP (1) if and only if $c \in \mathcal{A}^*$. According to this fact, for an optimal solution x^*, increasing the dimension of $\mathcal{A}^*$ increases the potentiality of robustness with respect to varying the objective vector.

Definition 3.1 *$x^* \in X$ is called a potentially robust solution of order k ($k \leq n$) if $dim(\mathcal{A}^*) = k$. The scalar k is called the potential robustness order of x^* and it is denoted by $PRO(x^*)$. Moreover, x^* is called a potentially robust solution if $dim(\mathcal{A}^*) = n$.*

Definition 3.2 *Let $x^* \in X$ be an optimal solution of LP(c). Then x^* is called a robust optimal solution (robust solution briefly) of order k ($k \leq n$) if $c \in ri(\mathcal{A}^*;k)$. Moreover, the largest k with this property is called the robustness order of x^* and it is denoted by $RO(x^*)$.*

Definition 3.3 *$x^* \in X$ is called a robust solution of LP(c) if it is robust of order n.*

Example 3.4 *Consider the following LP problem:*

$$\begin{array}{ll} \max & x_1 + x_2 \\ s.t. & x_1 + x_2 \leq 1, \ -x_1, -x_2 \leq 0. \end{array}$$

It can be seen that $x^1 = (0,1)^t$, $x^2 = (\frac{1}{2}, \frac{1}{2})^t$ and $x^3 = (1,0)^t$ are optimal solutions to this problem. Also, $A_{I(x^1)} = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$, $A_{I(x^2)} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and $A_{I(x^3)} = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$. Thus, $RO(x^1) = RO(x^2) = RO(x^3) = 1$ while $PRO(x^1) = PRO(x^3) = 2$ and $PRO(x^2) = 1$.

The following definition, which measures the alterations in the cost vector by a norm, has been investigated in the literature; See Definition 3.1 in Georgiev et al. (2013).

Definition 3.5 *(Georgiev et al. 2013) Let x^* be an optimal solution of LP(c). It is said to be robust in the norm sense if there exists scalar $\epsilon > 0$ such that x^* is an optimal solution to LP($c + d$) for each $d \in R^n$ with $\|d\| \leq \epsilon$.*

The following theorem shows that the robustness defined in Definition 3.3 is equivalent to that given in Definition 3.5 (in the norm sense) and its is equivalent to the uniqueness of the given optimal solution.

Theorem 3.6 *Let $x^* \in X$ be a feasible solution of LP (1). Then the following three statements are equivalent:*
(i) $c \in int(\mathcal{A}^*)$ (i.e. robustness in the sense of Definition 3.3),
(ii) There exists scalar $\epsilon > 0$ such that x^* is an optimal solution to LP($c + d$) for each $d \in R^n$ with $\|d\| \leq \epsilon$ (i.e. robustness in the sense of Definition 3.5),
(iii) x^* is a unique optimal solution of LP(c).

Due to the above theorem, the robustness concept defined in Definition 3.5 does not seem convenient for expressing the robustness in linear programming. In fact, the problem in this definition is that the optimality in linear programming strongly depends on the direction of c and not on the norm of c. Since the direction of c is important in changing/preserving the optimality and the direction changes when there is a positive angle deviation between the cost vector and the altered one, in Section 5 we define the robustness with respect to the eligible angle deviation of the cost vector such that this vector remains in the binding cone. On the other hand, if $c \in ri(\mathcal{A}^*; k)$, then c belongs to a face of the (polyhedral) binding cone with dimension k. Hence, there are k linear independent vectors in the binding cone which have positive eligible angle deviation with c. Due to this, we defined the robust solution of order k and the robustness order in Definition 3.2 using the interiority order.

4 CALCULATING THE ROBUSTNESS ORDER

Computing the potentially robustness order (PRO) is equivalent to calculating the rank of a matrix. Therefore, there are various approaches in Linear Algebra literature to calculate it. The following theorem provides an LP problem to compute the robustness order (RO). In the LP given in this theorem, $(w, y, \alpha) \in R^{|I(x^*)|} \times R^{|I(x^*)|} \times R$ is the variable vector.

Theorem 4.1 *Let $x^* \in X$ be an optimal solution to LP (1). Consider the following LP problem*

$$\begin{array}{ll} \max & \sum_{i \in I(x^*)} y_i \\ s.t. & \sum_{i \in I(x^*)} (y_i + w_i) a^i = \alpha c^t \\ & w_i \geq 0, \quad 0 \leq y_i \leq 1, \ i \in I(x^*), \\ & \alpha \geq 1. \end{array} \tag{3}$$

Let $(w^*, y^*, \alpha^*) \in R^{|I(x^*)|} \times R^{|I(x^*)|} \times R$ be an optimal solution of Problem (3), and E^* be the submatrix of A whose rows are a^i's with $y_i^* > 0$. Then $RO(x^*) = rank(E^*)$.

Theorem 4.2 *Let $x^* \in X$ be an optimal solution of LP(c) and $RO(x^*) = k$. Then x^* is an optimal solution of LP($c + d$) with $RO(x^*) \geq k$ for each $d \in \mathcal{A}^*$.*

5 ROBUSTNESS W.R.T THE ANGLE DEVIATION

We start this section with eligible angle deviation notion. For two vectors $x, y \in R^n$, the notation $\angle(x, y)$ stands for the angle between x and y.

Definition 5.1 *Let $0 \neq c, d \in R^n$ and $\theta \in (0, \pi]$. Also, let $x^* \in X$ be an optimal solution of LP(c). Then, θ is said to be an eligible angle deviation of c at x^* in direction d if there exists scalar $\alpha > 0$ such that*
i. $c + \alpha d \neq 0$,
ii. $\angle(c, c + \alpha d) \geq \theta$, *and*
iii. x^* *is an optimal solution of LP*$(c + \alpha d)$.

Definition 5.2 *Let $x^* \in X$ be an optimal solution of LP(c). Then, $\theta \in (0, \pi]$ is said to be an eligible angle deviation of c at x^* if it is an eligible angle deviation of c at x^* in some direction $d \in R^n$.*

The following proposition provides a system to characterize the eligible angle deviation. In the provided system, $w \in R^{|I(x^*)|}$ is the variable vector.

Proposition 5.3 *θ is an eligible angle deviation of c at x^* if and only if there exists some $w \in R^{|I(x^*)|}$ satisfying the following system:*

$$\begin{cases} w^t A_{I(x^*)} c \leq \cos\theta \|w^t A_{I(x^*)}\| \|c\|, \\ \|w^t A_{I(x^*)}\| \geq 1, \\ w \geq 0. \end{cases} \tag{4}$$

Remark 5.4 *Let $x^* \in X$ be an optimal solution of LP(c). Then, $PRO(x^*) > 1$ if and only if there exists some $\theta \in (0, \pi)$ such that θ is an eligible angle deviation of c at x^*.*

If $RO(x^*) > 1$, then by Remark 5.4 and due to $RO(x^*) \leq PRO(x^*)$, there exists some $\theta \in (0, \pi)$ such that θ is an eligible angle deviation of c at x^*. The converse of this assertion does not hold necessarily.

Proposition 5.5 gives an optimization problem to calculate the largest eligible angle deviation of c at x^*. In this problem, $(w, z) \in R^{|I(x^*)|} \times R$ is the variable vector.

Proposition 5.5 *Let $x^* \in X$ be an optimal solution of LP(c), and z^* be the optimal value of the following optimization problem:*

$$\begin{array}{rll} z^* = \min & z & \\ s.t. & w^t A_{I(x^*)} c \leq z \|w^t A_{I(x^*)}\| \|c\|, & \\ & \|w^t A_{I(x^*)}\| \geq 1, & \\ & w \geq 0. & \end{array} \tag{5}$$

Let θ^* denote the largest eligible angle deviation of c at x^*. Then
(i) $z^* \in [-1, 1]$.
(ii) If $z^* = 1$, *then there is no eligible angle deviation of c at x^*.*
(iii) If $z^* < 1$, *then* $\theta^* = \arccos(z^*)$. *Moreover,* $d = A^t_{I(x^*)} w^* - c$ *is a direction with the largest eligible angle deviation, where w^* is a part of an optimal solution of problem (5).*

Example 5.6 *Model (5) corresponding to optimal solution x^1 in example 3.4 is as follows:*

$$\begin{array}{rl} \min & z \\ s.t. & 2w_1 - w_2 \leq \sqrt{2} z \sqrt{(w_1 - w_2)^2 + w_1^2}, \\ & (w_1 - w_2)^2 + w_1^2 \geq 1, \\ & w_1, w_2 \geq 0. \end{array}$$

An optimal solution of this problem is $(w_1^*, w_2^*, z^*) = (0.0000, 385.4493, -0.7071)$. *Hence, the largest eligible angle deviation of c at x^* is* $\theta = \arccos(-0.7071) = 2.3562$ *radiant or equivalently* $\theta = 135.0000^\circ$.

Solving the optimization problem (5) can be difficult computationally, due to the second constraint of this problem. Fortunately, in some special cases the largest angel deviation can be obtained by some simple calculations. The following theorem deals with this issue. In this result, e_i stands for the i-th unit vector.

Theorem 5.7 *Let $x^* \in X$ be an optimal solution of LP(c). Then*
(a) For any $i \in I(x^*)$ with $\angle(c, a^i) \neq 0$, the scalar $\angle(c, a^i)$ is an eligible angle deviation of c at x^* in direction $a^i - c$.
(b) Assume that $\|a^i\| = 1$ and $a^i c \geq 0$ for all $i \in I(x^*)$. Set $a^{i_0}c := \min_{i \in I(x^*)} a^i c$. Then $\left(w^* = e_{i_0}, z^* = \cos\left(\angle(c, a^{i_0})\right)\right)$ is an optimal solution of problem (5) and $\angle(c, a^{i_0})$ is the largest eligible angle deviation of c at x^*.

Example 5.8 *Consider the LP problem*

$$\begin{array}{ll} \max & c^t x \\ s.t. & \begin{pmatrix} A \\ -I \end{pmatrix} x \leq \begin{pmatrix} b \\ 0 \end{pmatrix}, \end{array} \tag{6}$$

with $c = (-8, -2, 12, 0, -7, -2, 7, 0)^t$,

$$A = \begin{pmatrix} -4 & -4 & 7 & -5 & -1 & -3 & 5 & 1 \\ 0 & 3 & -2 & -2 & -3 & 0 & 2 & -1 \\ 1 & 7 & 0 & 0 & -6 & 5 & -1 & -2 \\ -4 & 1 & 1 & 1 & -6 & -3 & -2 & -1 \\ 1 & 1 & -1 & 0 & 1 & -4 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \text{ and } b = \begin{pmatrix} 1 \\ 8 \\ 4 \\ 2 \\ 8 \\ 40 \end{pmatrix}.$$

It can be seen that $x^* = (0, 0, 6.1944, 21.1389, 0, 0, 12.6667, 0)$ is an optimal solution of Problem (6) (with the above A, b, and c) and $I(x^*) = \{1, 4, 6, 7, 8, 11, 12, 14\}$. It is seen that $A_{I(x^*)}c = (172, 76, 0, 8, 2, 7, 2, 0) \geq 0$. Based on the notation used in Theorem 5.7,

$$\begin{aligned} &a^{i_0} = (1, 1, 1, 1, 1, 1, 1, 1), \\ &a^{i_0}c = \min\{172, 76, 0, 8, 2, 7, 2, 0\} = 0. \end{aligned}$$

Hence, $(w^*, z^*) = \left((0, 0, 1, 0, 0, 0, 0, 0)^t, 0\right)$ is an optimal solution of Problem (5) and $\angle(c, a^{i_0}) = \arccos(0) = \pi/2$ is the largest eligible angle deviation of c at x^*.

The following theorem provides an important connection between the robust order and eligible angle deviation.

Theorem 5.9 *Let x^* be an optimal solution of LP(c) and k be a natural number. If there exist k linear independent vectors $d^1, d^2, \ldots, d^k \in R^n$ such that there is a positive eligible angle deviation of c at x^* in direction $\pm d^i$ for each $i \in \{1, 2, \ldots, k\}$, then $RO(x^*) \geq k$. If k is the biggest natural number with this property, then $RO(x^*) = k$.*
Conversely, if $RO(x^*) = k \geq 2$, then there exist k linear independent vectors $d^1, d^2, \ldots, d^k \in R^n$ such that there is a positive eligible angle deviation of c at x^* in direction $\pm d^i$ for each $i \in \{1, 2, \ldots, k\}$.

6 Conclusion

In linear programming, the optimality of a feasible solution is equivalent to that the cost vector belongs to the binding cone. Due to this, in the present work the degree of interiority of the cost vector in the binding cone has been used for defining the robust solutions. A robustness order

has been defined and an LP-based method has been presented for calculating it. In another part of the work, robustness has been defined with respect to the eligible angle deviation of the cost vector in the binding cone. In addition to the theoretical results concerning the maximum eligible angle deviation, the connection between two above-mentioned robustness standpoints has been addressed as well. Although we have provided an efficient procedure for calculating the largest eligible angle deviation in special cases, calculating this quantity is difficult computationally in general case. It can be worth studying in future.

7 Acknowledgements

This research was in part supported by a grant from IPM (No. 93260124).

References

[1] Ben-Tal, A., Ghaoui, L.-E., and Nemirovski, A. (2009). Robust optimization. Princeton and Oxford: Princeton University Press.

[2] Deb, K., and Gupta, H. (2006). Introducing robustness in multi-objective optimization. Evolutionary Computation, 14: 463-494

[3] Georgiev, P.G., Luc, D.T., and Pardalos, P.M. (2013). Robust aspects of solutions in deterministic multiple objective linear programming. European Journal of Operational Research, 229(1): 29–36.

[4] Zamani, M., Soleimani-damaneh, M., and Kabgani, A. (in press) Robustness in nonsmooth nonlinear multi-objective programming. European Journal of Operational Research, In Press.

A BASIS FOR TAXONOMY OF FUZZY LINEAR PROGRAMMING METHODS

Maja Repnik
University of Maribor, Faculty of Energy Technology, Hočevarjev trg 1, 8270 Krško, Slovenia,
maja.repnik@um.si
Drago Bokal
University of Maribor, Faculty of Natural Sciences and Mathematics, Koroška cesta 160,
2000 Maribor, Slovenia, drago.bokal@um.si

Abstract: Linear programming is one of the most frequently used techniques in operations research, but often in real world problems, some parameters of LP cannot be precisely determined. To overcome this problem, fuzzy parameters were introduced into LP and many researchers started to investigate fuzzy linear programming. In past decades, various types of fuzzy linear programming problems have been defined and several approaches to solving these problems were proposed. Despite this variety, a thorough recent survey allowing for detailed overview of the landscape of fuzzy linear programming techniques and applications of the methods seems not to be available, and we intend to overcome this with our research, aimed to propose a taxonomy allowing researchers and practitioners to navigate the knowledge domain of Fuzzy Linear Programs techniques, technologies, and applications. In this contribution, we focus on the fuzzy linear programming techniques.

Keywords: fuzzy linear programming, fuzzy constraints, fuzzy numbers.

1 INTRODUCTION

It has been over 40 years since in 1970, Bellman and Zadeh introduced the concept of fuzzy decision [2]. Tanaka et. al. extended this concept into fuzzy mathematical programming in 1973 and Zimmermann proposed the formulation of Fuzzy Linear Programming (FLP) problems [20] in 1978. Since then, researches developed a relatively large number of different methods to solve FLP problems.

A general model of linear program (LP) problem is written as

$$\begin{array}{ll} \text{maximize} & cx \\ \text{subject to} & Ax \leq b \\ & x \geq 0 \end{array}$$

Frequently in real world problems, some of the coefficients (c, A, b) can not be precisely defined. One can avoid this problem by treating them as fuzzy numbers (as opposed to the precise "crisp" numbers). FLP is LP where some of the coefficients involved in the objective or constraint functions are fuzzy numbers (notation for fuzzy number a is $\widetilde{a}$). Based on the place where fuzzy numbers appear, we can divide FLP problems into 4 main categories:

- FLP with fuzzy resources $(\widetilde{b_i})$: model of FLP where instead of crisp numbers, fuzzy numbers appear on right hand side.
- FLP with fuzzy coefficients in objective function $(\widetilde{c_i})$
- FLP with fuzzy technological coefficients $(\widetilde{a_{ij}})$
- FLP with fuzzy variables $(\widetilde{x_i})$.

By combining these four categories, we get many different types of FLP problems and each of them is solvable by different method. Authors use different indices to compare fuzzy numbers in their methods, some of them even two different indices. Considering all this issues, it is easy

to understand that there is a plethora of different methods to solve FLP problems, and a practitioner or novice researcher has difficulty in finding direction in this vast landscape. With our research, which is at this stage limited to the most significant references, we first address the taxonomy of fuzzy linear programming approaches, which we intend to extend to include taxonomy of technologies (i.e. software implementations) and real world applications. The goal is to design a decision support system for practitioners, who will be able to efficiently find the most suitable FLP technique and technology for their problem, building upon experience from past scientific research in the area.

2 TIP OF THE BIBLIOGRAPHY ICEBERG

In 1985, Slowinski analysed 19 of most significant methods in the first decade of FLP [16]. In 1992, Lai and Hwang classified [11] FLP problems and existing approaches based on an extensive and thorough survey literature until 1992. They distinguished FLP problems from possibilistic LP problems. FLP problems associate fuzzy input data, which are modelled by subjective preference-based membership functions, while in possibilistic LP imprecise data are modelled by possibility distribution. They also made discussions of several practical problems to fill the gap between researchers and real decision makers, since only few methods have been tested with real-world problems by real decision makers.

An extensive research was performed by Rommelfanger [13] in 1996. He collected and studied a large number of methods that appeared in literature until 1996 and presented a good survey on procedures for solving FLP problems. He also listed applications of fuzzy linear programs published in the literature. In 2000, Inuiguchi and Ramık reviewed [6] some FLP methods and techniques from a practical point of view, but they discussed only possibilistic approaches. In 2007 Rommelfanger made a survey of optimization of fuzzy objective functions in FLP, but in [14] he concentrated only on the handling of fuzzy objectives. A more recent survey on FLP has been in 2012 prepared by Shams, et al in [15]. Unfortunately, it proposes only the most basic classification of FLP problems: FLPs with fuzzy constraints, FLPs with fuzzy constraints and fuzzy objective function, FLPs with fuzzy constraints and fuzzy technological coefficients and does not propose a navigation through the research abounding in the recent years. They present some methods and approaches proposed to solve them, unfortunately missing many more. For instance, parameters in objective function can be fuzzy with [4] and [19] being just two of the methods developed to solve these FLP problems. Methods for solving problems where all parameters (c, A and b) are fuzzy are in [7], [8], [12] etc. There are also many researchers dealing with fully FLP, where all coefficients are fuzzy including variable ([3], [5], [18], [9], [10], to name a few.)

3 A PROPOSED TAXONOMY

There are many surveys of FLP problems, but to the best of our knowledge, no extensive survey similar to [11] or [13] has been published for the last 19 years. In this period, however, FLP research has abounded, and a large number of new methods were developed to solve FLP problems especially fully FLP problems. Consequently, there is a need for taxonomy that includes also new methods developed to solve various FLP problems. This is subject of our research.

In table 1, we present five main dimensions in which models of FLP differ. Type of FLP (as our first dimension, “FLP”) answers the question ”What is fuzzy in the LP problem?”. As discussed, this can be resources (b), coefficients in objective function (c), technological coefficients (a), variables (x) or a combination of those.

The second dimension illustrates the method for the ordering of fuzzy quantities (COMP). There is no generally accepted criteria for comparison of fuzzy numbers in FLP. A decade ago,

Table 1: Taxonomy of FLP models

Dimension		Dimension values		Frequency
Description	**Designation**	**Description**	**Designation**	
Type of FLP	FLP	Fuzzy resources	b	39
		Fuzzy coefficients in objective function	c	33
		Fuzzy technological coefficients	a	30
		Fuzzy variables	x	17
Type of comparison	COMP	Ranking function	RF	15
		Pairwise comparison	PC	17
		Reference set	RS	1
Shape of membership function	MF	Linear	LIN	26
		L-R type	LR	8
		Exponential	EXP	2
		Hyperbolic	HYP	1
		Logistic curve	LC	2
Main computational procedure	MCP	Converting to LP	LP	17
		Converting to parametric LP	PLP	4
		Converting to MOLP	MOLP	8
		Convering to NLP	NLP	5
		Converting to semi-infinite LP	SILP	2
		Simplex method	SM	5
Solving also fuzzy MOLP	FMOLP	YES	YES	10
		NO	NO	37

there were already more than 35 different ordering indices presented in the literature [17]. It is possible to classify them into tree categories. Ranking function (RF) is a mapping from the set of fuzzy quantities into set of real numbers. Fuzzy quantities can then be compared according to the corresponding real numbers. Pairwise comparison (PC) is a fuzzy relation, comparing fuzzy quantities involved. In the third class, reference sets are set up and all the fuzzy quantities to be ranked are compared with the reference sets [17].

The third dimension reveals the shape of membership functions (MF) presenting fuzzy numbers used in FLP problem. In first FLP problems, membership functions usually were triangular and trapezoidal, because calculations with triangular and trapezoidal membership functions are easy. We will denote them linear membership functions (LIN), which is only the special case of L-R shape of membership function (LR), which is also the most general and widely used. Some researches are dealing also with exponential (EXP), hyperbolic (HYP) and logistic (LC) shapes of membership functions.

The fourth dimension explains what is the main computational procedure (MCP). Most of researches defuzzify FLP problems, by converting them to crisp optimization problems, which are relatively easy to solve. They prove that crisp model is equivalent to the FLP problem and they use the optimal solution of the crisp model as the optimal solution of the FLP problem. They convert FLP to LP (LP), parametric LP (PLP), multi-objective linear programming problem (MOLP), non-linear programming problem (NLP) or semi-infinite LP. Some researchers claim that we should solve FLPs without converting them to crisp LP problems. They proposed a simplex method (SM) which finds the solution of FLP problem directly.

The last dimension declares if the model is suitable also for solving fuzzy MOLP problems (FMOLP).

So far we analysed around 50 methods, found on-line and in accessible bibliography. Based on the findings, we completed the column Frequency with information about the how many of the analysed methods fit into which value of each dimension. This is an indication of how well-understood a particular approach is, but the details of that will be made available after completion of the detailed analysis of the FLP methods, technologies and problem landscape.

4 CONCLUSION

Since there are relatively a large number of different methods in literature that deal with FLP problems, it is difficult for researches and practitioners who desire to solve real-life FLP problems to choose the most appropriate method. The proposed taxonomy, which is still under development, will be a tool to help them in doing that. We hope that this first iteration will help us gather feedback on the approach and through the feedback improve the taxonomy's use cases as well as content for both practitioners and researchers. In the first iteration, we have presented values of the taxonomy dimensions related to FLP methods, as well numbers of occurrence in our currently investigated papers. We will continue our research with exploring dimensions related to technology (fuzzy LP implementations), real world applications and comparison criteria. This will lead us towards a goal of allowing the researchers with a given real world problem to suggest the most suitable technique and implementation under the criterion of their choice.

References

[1] Note that due to space limitations, some of references have been suppressed.

[2] Bellman, R. E. and Zadeh, L. A. (1970). Decision-making in a fuzzy environment. Management science, 17(4), B-141.

[3] Cheng, H., Huang, W. and Cai, J. (2013). Solving a fully fuzzy linear programming problem through compromise programming. Journal of Applied Mathematics, 2013.

[4] Delgado, M., Verdegay, J. L. and Vila, M. A. (1987). Imprecise costs in mathematical programming problems. Control and Cybernetics, 16(2), 113-121.

[5] Ezzati, R., Khorram, E. and Enayati, R. (2013). A new algorithm to solve fully fuzzy linear programming problems using the MOLP problem. Applied Mathematical Modelling. doi.org/10.1016/j.apm.2013.03.014

[6] Inuiguchi, M. and Ramık, J. (2000). Possibilistic linear programming: a brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem. Fuzzy sets and systems, 111(1), 3-28.

[7] Jamison, K. D. and Lodwick, W. A. (2001). Fuzzy linear programming using a penalty method. Fuzzy Sets and Systems, 119(1), 97-110.

[8] Jiménez, M., Arenas, M., Bilbao, A. and Rodrı, M. V. (2007). Linear programming with fuzzy parameters: an interactive method resolution. European Journal of Operational Research, 177(3), 1599-1609.

[9] Khan, I. U., Ahmad, T. and Maan, N. (2013). A simplified novel technique for solving fully fuzzy linear programming problems. Journal of Optimization Theory and Applications, 159(2), 536-546.

[10] Kumar, A., Kaur, J. and Singh, P. (2011). A new method for solving fully fuzzy linear programming problems. Applied Mathematical Modelling, 35(2), 817-823.

[11] Lai Y. L. and Hwang C. L. (1992). Fuzzy mathematical programming-methods and applications. Berlin/Heidelberg: Springer.

[12] Maleki, H. R., Tata, M. and Mashinchi, M. (2000). Linear programming with fuzzy variables. Fuzzy sets and systems, 109(1), 21-33.

[13] Rommelfanger, H. (1996). Fuzzy linear programming and applications. European journal of operational research, 92(3), 512-527.

[14] Rommelfanger, H. J. (2007). Optimization of Fuzzy Objective Functions in Fuzzy (Multicriteria) Linear Programs-A Critical Survey. Theoretical Advances and Applications of Fuzzy Logic and Soft Computing. Springer Berlin Heidelberg. 324-333.

[15] Shams, H., Mogouee, M. D., Jamali, F. and Haji, A. (2012). A Survey on Fuzzy Linear Programming. American Journal of Scientific Research, (75), 117-133.

[16] Słowiński, R. (1986). A multicriteria fuzzy linear programming method for water supply system development planning. Fuzzy sets and systems, 19(3), 217-237.

[17] Wang, X., and Kerre, E. E. (2001). Reasonable properties for the ordering of fuzzy quantities (I). Fuzzy Sets and Systems, 118(3), 375-385.

[18] Yang, X. P., Zhou, X. G., Cao, B. Y. and Nasseri, S. H. (2014). A MOLP Method for Solving Fully Fuzzy Linear Programming with LR Fuzzy Parameters. Mathematical Problems in Engineering, 2014. Article ID 782376, 10 pages, 2014. doi:10.1155/2014/782376

[19] Zhang, G., Wu, Y. H., Remias, M. and Lu, J. (2003). Formulation of fuzzy linear programming problems as four-objective constrained optimization problems. Applied Mathematics and Computation, 139(2), 383-399.

[20] Zimmermann, H. J. (1978). Fuzzy programming and linear programming with several objective functions. Fuzzy sets and systems, 1(1), 45-55.

DESIGN OF WEB-BASED INFORMATION SYSTEM FOR OPTIMIZATION OF PORTFOLIO

Galina Stoyanova Panayotova
University of Library Studies and Information Technologies
bul.Tsarigradsko shose 119, Sofia 1784, Bulgaria
panayotovag@gmail.com

Georgi Petrov Dimitrov
University of Library Studies and Information Technologies
bul.Tsarigradsko shose 119, Sofia 1784, Bulgaria
geo.p.dimitrov@gmail.com

Abstract: This paper briefly introduces an integrated framework for portfolio selection and implementation of the framework for Web-based information system. The structure of Web-based information system for optimization of portfolio is investigated. Portfolio optimization in MATLAB environment is given. The optimization programme created from the existing software packages uses single-period Mean-Variance Optimization model.

Keywords: Portfolio theory, portfolio optimization, Web-based information system, MATLAB.

1 INTRODUCTION

The Portfolio theory concerns an optimal allocation of financial resources buying and/or selling assets (securities) from the financial markets [1]. The allocation of the financial resources is called “investment”, which investors regard as a prospect for future return. The better combination of assets (securities) in the portfolio provides better future rate of return to the investor. The portfolio theory develops models that allow the best combinations to be found. This theory extends the classical economic model of uncertainty.Modern portfolio theory stems from Markowitz’s [2] great insights of the Mean–Variance model (MVO), which states that the key information of a portfolio can be derived from three measurements: expected returns (taken as the arithmetic mean), standard deviations and correlations among those returns.

1.1 Modern Portfolio Theory

The revolution in financial investing known as modern portfolio theory was initiated in the 1950's by Nobel Prize winner Harry Markowitz. Markowitz showed that investors can obtain significantly greater return at lower risk if, instead of choosing stocks and other financial assets based on their individual potentials, they make choices based on calculating the impact on the risk and return generated by the portfolio as a whole. Certain combinations of investments (portfolios) are efficient (they lie on an "efficient frontier") in that they create the greatest possible value for the least risk. Inefficient portfolios should be avoided. Which of the efficient portfolios is best depends on the investor's risk tolerance or willingness to accept risk.

1.2 Portfolio optimization.

As the main stage of the framework, portfolio optimization is to ensure optimal risk-adjusted returns by analyzing the portfolio and managing the assets. The output of this stage is an efficient frontier, and the investors can have the optimal result on the efficient frontier according to their own risk preferences. There are many comparative optimization models to

be considered in this stage, such as Mean–Variance Optimization (MVO) model [2] that takes transaction cost (not fixed) into consideration

1.3 Optimization Model

In the process of selecting an efficient portfolio, the optimization model plays an important role [1].

The task of portfolio optimization is to determine these values $x_1, x_2, ..., x_n$, which minimize the objective function

$$\min_x \{-E_p + \lambda V_p\}, \quad (1)$$

where E_p is the expected portfolio return, V_p is the portfolio risk and $\sum_{i=1}^{N} x_i = 1, \quad x_i > 0$.

That means that the investment x_i can have positive and negative values. For x_i >0, the investor has to buy security *i* with the relative value x_i of the total amount of the investment. For the case x_i < 0, the investor has to sell security x_i. This "short sale" means that the investor has to borrow security *i*, to sell it and later has to restore it.

The portfolio return $E_p(x_i)$ and risk $V_p(x_i)$ give one point of the "efficient frontier". Different values of the "efficient frontier" are calculated giving different values to the parameter λ in problem (1) and respectively solving (1). (1) is linear-quadratic optimization problems of the mathematical programming.

In the process of the project portfolio selection, portfolio optimization model is a key component, and the modelling process is knowledge- intensive and time consuming.
In fixing the value of the parameter, the task of portfolio optimization is analytically determined and can be solved. The difficulties in solving (1) originate from the fact that the "efficient frontier" can be found only via numerical recursive computations by changing λ. The requirements for solving (1) in real time to implement portfolio optimization as an informational service in Internet stress the necessity to design fast computational algorithms.

2 STRUCTURE OF WEB-BASED INFORMATION SYSTEM FOR OPTIMIZACION

The information services have been intensively developed in the last 15 years. The launch of Internet as universal and global communication environment gave ground to the information services in raising their functionality [4].

The information service of the portfolio optimization is implemented as a hierarchical information system with 4 tiers. In Fig. 1, we give a structure of this system. The server side is developed under Window operation system, MySQL database and Apache server.

The optimization methods developed are implemented by PHP programming technology. The server application suite consists of PHP based programs that apply business logic of the investment process, MySQL data storage, and GD graphical library. The algorithmic server is developed on C++ technology. It performs the investment calculations and the optimization algorithms.

The forth level performs the specific and complex data processing, performs complex mathematical evaluations, support on-line control functionality. The forth tier in the research

is called “Algorithmic Server” [3]. It performs tasks that cannot be supported by the server-side programming tools or performs fast algorithmic information processing.

A model database is used to support variety of modeling techniques for portfolio selection optimization. Optimization programs written by us or from existing software packages can be plugged into the database for this system to use.

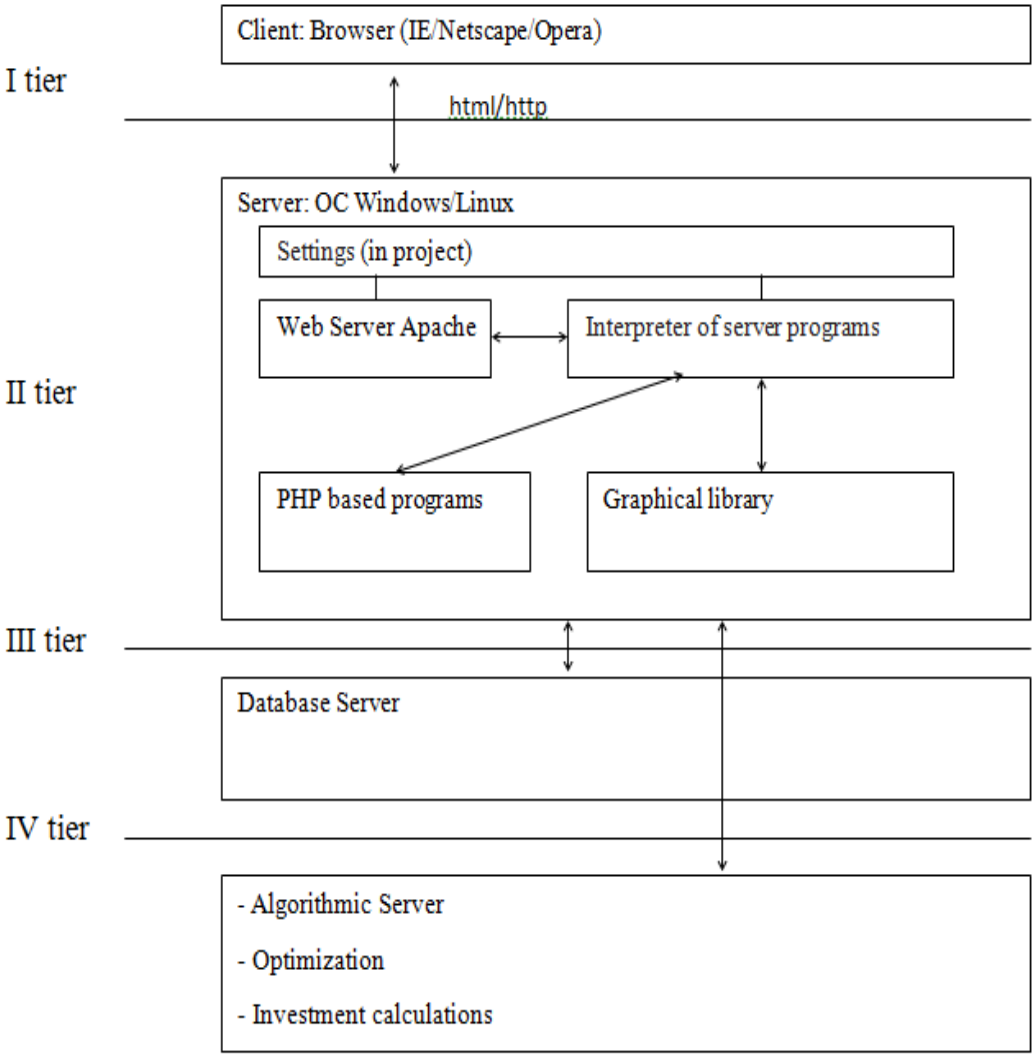


Figure 1: Fourth-tier client server model of WEB based information system

Fig. 2 shows results of the portfolio optimization using single-period MVO model from an existing software package. It is generated dynamically by a Java applet from information retrieved from our system database that stores historical data. The output is an efficient frontier, and the users can have the optimal decision on the efficient frontier according to their own risk preferences.

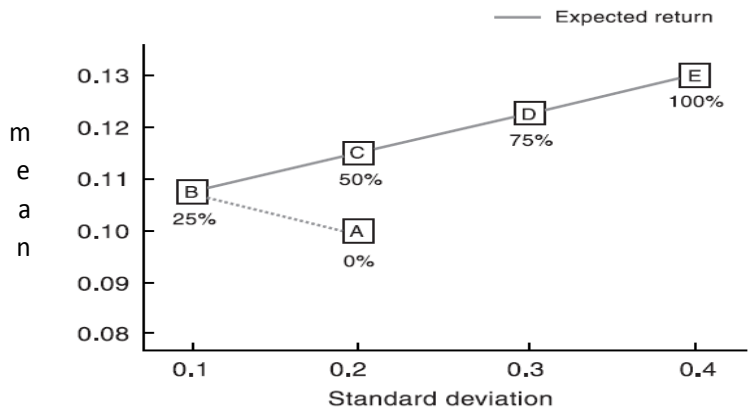


Figure 2: Portfolio optimization results

3 PORTFOLIO OPTIMIZATION IN MATLAB ENVIROMENT

The performances of the algorithms, which solve the portfolio optimization problem, are given in MATLAB environment [6].

Common steps in optimizing portfolios include [5]:

- Estimating asset return and total return moments from price or return data;
- Computing portfolio-level statistics;
- Performing constrained mean-variance, conditional value –at-risk, and mean-absolute-deviation optimization;
- Examining the time evolution of efficient portfolio allocations;
- Accounting for turnover and transaction costs.

For example: Asset allocation

The portfolio optimization problem, start with basic definitions of known quantities associated with the structure of this problem. Each asset class is assumed to have a tradeable asset with a real-time price. The initial portfolio with holdings in each asset that has a total of $7.5 million along with an additional cash position of $60,000. These basic quantities and the costs to trade are set up in the following variables with asset names in the cell array Asset, current prices in the vector Price, current portfolio holdings in the vector holding, and transaction costs in the vector Unit Cost (Tab. 1).

Table 1: Selected characteristics of the assets

	Price	*InitHolding*	*InitPort*	*UnitCost*
Bonds	52,4	42938	0,3	0,001
Large-Cap Equities	122,7	24449	0,4	0,001
Small-Cap Equities	35,2	42612	0,2	0,001
Emerging Equities	46,9	15991	0,1	0,004

Simulating Asset Prices

Since this is a hypothetical example, to simulate asset prices from a given mean and covariance of annual asset total returns for the asset classes, portsim is used to create asset returns with the desired mean and covariance. Specifically, portsim is used to simulate five years of monthly total returns. The mean and covariance of annual asset total returns are maintained in the variables AssetMean and AssetCovar. The simulated asset total return

prices (which are compounded total returns) are maintained in the variable Y. All initial asset total return prices are normalized to 1 in this example.

```
AssetMean = [ 0.05; 0.1; 0.12; 0.18 ];

AssetCovar = [ 0.0064 0.00408 0.00192 0;

    0.00408 0.0289 0.0204 0.0119;

    0.00192 0.0204 0.0576 0.0336;

    0 0.0119 0.0336 0.1225 ];

X = portsim(AssetMean'/12, AssetCovar/12, 60); % monthly total returns for 5
years (60 months)

[Y, T] = ret2tick(X, [], 1/12);                % form total return prices
```

This plot shows the log of the simulated total return prices (Fig. 3)

```
plot(T, log(Y));

title('\bfSimulated Asset Class Total Return Prices');

xlabel('Year');

ylabel('Log Total Return Price');

legend(Asset,'Location','best');
```

Figure 3: Simulated total return prices

An important step in portfolio optimization is to validate that the portfolio problem is feasible and the main test is to ensure that the set of portfolios is nonempty and bounded. Use the estimate Bounds function to determine the bounds for the portfolio set. See Fig.4

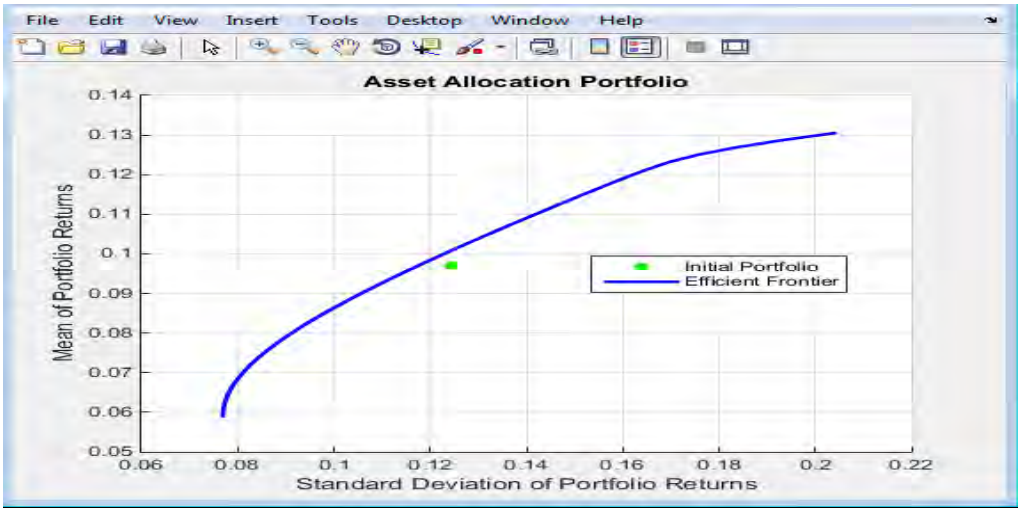


Figure 4: Asset Allocation Portfolio

4 CONCLUSION

The paper addresses the development of the information services in the global network Internet. The new tendencies in the implementation of the services as multi-tier software client-server system insist decomposition and specialization of the server side programming suit on several tiers. This concept is applied for the implementation of the portfolio design and optimization as an e-service in Internet. Due to the lack of full available data for the investment optimization, soft computing approach is applied in solving the bi-level optimization problem. Linear-quadratic approximation and fuzzy logic is applied for solving the portfolio problem. Thus the optimization problem for the investment allocation is solved promptly, respecting the requirements for real time response to requests, arising in the Internet environment.

References

[1] Chen, A. Chianglin, A. and Chung, H. (2001) "Establishing an index arbitrage model by applyin g neural networks method-A case study of Nikkei 225 index," Journ. Internat.Journ. of Neural Systems, No11(5), pp. 489-496.

[2] Markowitz, H. (1952) Portfolio selection, Journal of Finance, Vol 7: pp 77–91.

[3] Curbera F. et al, (2002) Unravelling the web services. Web and introduction to SOAP, WSDL, UDDI," IEEE Internet computing,vol.6, ISSUE 2, pp. 86-93

[4] Cash, J. Eccles, R. Nohria, N. and Nolan, R. (1993) Building the information age organization, IRWIN, Chicago, p. 498

[5] Panayotova, G. (2014) Mathematical modelling, Sofia, Bulgaria, ISBN 976-619-185-037-2

[6] Garvanov, I. (2014) MATLAB. Sofia, Bulgaria, ISBN 976-619-185-056-3

DUALITY IN MONOPOLY

Ilko Vrankić and Mira Krpan
University of Zagreb, Faculty of Economics and Business
Trg J. F. Kennedy 6, 10000 Zagreb, Croatia
{ivrankic, mkrpan}@efzg.hr

Abstract: This paper follows our theoretical research on duality in microeconomic theory and applies the duality principles, which rest on the price taking behaviour of economic subjects, on the monopolistic behaviour. The standard approach of deriving the profit function for the monopolist from the production function and defined pseudo production function is accompanied by the alternative approach in which the starting point is the pseudo cost function. Finally, starting from the derived profit function the pseudo production function is recovered and a version of Hotelling's lemma is given. All results are illustrated with a numerical example.

Keywords: duality, pseudo production function, pseudo cost function, profit function, Hotelling's lemma.

1 INTRODUCTION

Rationality of economic agents in microeconomic theory is described by the optimization problems in which convex sets play an important role. The possibility of characterizing them in two ways naturally gives rise to duality in microeconomic theory [3,4]. Duality in microeconomic theory includes derivation and recovering of the alternative representations of the consumer preferences and the production technology. Since duality in this case rests on the price taking behaviour, the question is how can the principles of duality be applied in the case of a monopolistic firm where the single producer has influence on the price which he charges for the product [1,2,5,7]? In this paper we show how the profit function for the monopolist for the given demand function can be obtained by introducing the pseudo cost function [6] and solving the profit maximization problem. The standard approach in which the production function and defined pseudo production function is the starting point is also derived. Finally, starting from the derived profit function we show how the pseudo production function can be recovered. Although Hotelling's lemma, a famous result from the microeconomic theory, defined below, cannot be directly applied in obtaining the monopolist's supply function for the given demand function, a version of Hotelling's lemma is illustrated.

2 FROM THE PSEUDO PRODUCTION FUNCTION TO THE PROFIT FUNCTION

The emphasis is on a monopolist whose objective is to maximize profit. The monopolist faces the following negatively sloped inverse demand function [5]

$$p_0 = wD\left[x_0\right], \tag{1}$$

where p_0 is the price of the monopolist's product and $w > 0$ represents the influence on demand of "other variables", for example income, and D is a function of x_0 for which $D'(x_0) < 0$. The monopolist's technology is described by the production function $x_0 = F(\mathbf{x})$, where $\mathbf{x}$ is the vector of inputs used in the production of the monopolist's product x_0. Since the total revenue function is described by $p_0 x_0 = wD\left[F(\mathbf{x})\right]F(\mathbf{x}) = wF(\mathbf{x})D\left[F(\mathbf{x})\right]$, the profit maximization model for the monopolist in which the quantities of inputs are the choice variables reduces to

$$\max_{\mathbf{x}} \; wF(\mathbf{x})D\left[F(\mathbf{x})\right] - \mathbf{wx} \;, \tag{2}$$

where **w** is the vector of input prices. If we form the pseudo production function [5]

$$F^*(\mathbf{x}) = F(\mathbf{x})D\left[F(\mathbf{x})\right] = \frac{p_o x_0}{w} \tag{3}$$

as deflated revenue function and interpret the parameter *w* in (2) as the price of the pseudo product, then the known results from duality theory in microeconomics could be applied. It is assumed that sufficient regularity conditions are satisfied so that the maximum exists [1,5,7]. The first order necessary conditions for the profit maximization problem reduce to

$$w\frac{\partial F^*(\mathbf{x})}{\partial x_i} = w_i, \;\; \forall i \tag{4}$$

Therefore, the monopolist will hire the levels of inputs for which the marginal revenue of the corresponding input, $w\dfrac{\partial F^*(\mathbf{x})}{\partial x_i}$, is equal to its marginal cost, w_i. An interior solution is assumed. It is assumed also that the input market is perfectly competitive so that the monopolist takes input prices as given.

Solving the first order necessary conditions leads us to the input demand functions. By substituting the derived input demand functions in the production function, for the given demand function, we obtain the monopolist's supply function. The profit function is obtained by substituting the derived functions in the goal function in (2). To illustrate this procedure, the production function

$$x_0 = f(x_1, x_2) = x_1^{\frac{1}{4}} x_2^{\frac{1}{4}} \tag{5}$$

is chosen [6]. Let us start from the linear inverse market demand function,

$$p_0 = w(a - bx_0), \tag{6}$$

for which the pseudo production function is defined as follows

$$F^*(\mathbf{x}) = (a - bx_1^{\frac{1}{4}} x_2^{\frac{1}{4}}) x_1^{\frac{1}{4}} x_2^{\frac{1}{4}} = ax_1^{\frac{1}{4}} x_2^{\frac{1}{4}} - bx_1^{\frac{1}{2}} x_2^{\frac{1}{2}}. \tag{7}$$

The profit maximization model reduces to

$$\max_{x_1, x_2} \; w(ax_1^{\frac{1}{4}} x_2^{\frac{1}{4}} - bx_1^{\frac{1}{2}} x_2^{\frac{1}{2}}) - w_1 x_1 - w_2 x_2 \,. \tag{8}$$

First order necessary conditions are expressed by the following system of equations

$$\begin{aligned} w\frac{\partial F^*(\mathbf{x})}{\partial x_1} &= w\left(\frac{a}{4} x_1^{-\frac{3}{4}} x_2^{\frac{1}{4}} - \frac{b}{2} x_1^{-\frac{1}{2}} x_2^{\frac{1}{2}}\right) = \frac{w}{4} x_1^{-\frac{3}{4}} x_2^{\frac{1}{4}} (a - 2bx_1^{\frac{1}{4}} x_2^{\frac{1}{4}}) = w_1 \\ w\frac{\partial F^*(\mathbf{x})}{\partial x_2} &= w\left(\frac{a}{4} x_1^{\frac{1}{4}} x_2^{-\frac{3}{4}} - \frac{b}{2} x_1^{\frac{1}{2}} x_2^{-\frac{1}{2}}\right) = \frac{w}{4} x_1^{\frac{1}{4}} x_2^{-\frac{3}{4}} (a - 2bx_1^{\frac{1}{4}} x_2^{\frac{1}{4}}) = w_2 \end{aligned}, \tag{9}$$

Dividing both expressions in (9) gets us to $\frac{x_2}{x_1} = \frac{w_1}{w_2}$. We can express x_2 as a function of x_1, $x_2 = \frac{w_1}{w_2} x_1$, whose substitution in one of (9) leads us to $\frac{w}{4} x_1^{-\frac{3}{4}} \frac{w_1^{\frac{1}{4}}}{w_2^{\frac{1}{4}}} x_1^{\frac{1}{4}} (a - 2b x_1^{\frac{1}{2}} \frac{w_1^{\frac{1}{4}}}{w_2^{\frac{1}{4}}}) = w_1$. By solving it we get the monopolist's demand function for the first input, $x_1^* = \frac{a^2 w^2 w_2^{\frac{1}{4}}}{4 w_1^{\frac{1}{2}} (2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2}$. By inserting it in $x_2 = \frac{w_1}{w_2} x_1$, we get the monopolist's demand function for the second input, $x_2^* = \frac{a^2 w^2 w_1^{\frac{1}{4}}}{4 w_2^{\frac{1}{2}} (2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2}$. The supply function for the monopolist, for the given demand function, is then $x_0^* = \frac{aw}{2(2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)}$. By inserting the derived supply and demand functions in the goal function of a producer, the profit function is obtained,

$$\pi^* = \frac{a^2 w^2}{4(2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)}. \tag{10}$$

3 FROM THE PSEUDO COST FUNCTION TO THE PROFIT FUNCTION

Another method for obtaining the monopolist's profit function for the given demand function is to start from the monopolist's cost function, $c(x_0, \mathbf{w})$, which is an alternative way of describing technology [9]. The decision variable in this model of the monopolist's profit maximization problem is the quantity of production,

$$\max_{x_0} p_0 x_0 - c(x_0, \mathbf{w}) = \max_{x_0} wD(x_0)x_0 - c(x_0, \mathbf{w}). \tag{11}$$

Taking into account the definition of the pseudo production function, $F^*(\mathbf{x}) = D(x_0)x_0 = y$, the quantity of production can be expressed as the function of pseudo production, $x_0 = g(y)$. In this case the profit maximization problem reduces to

$$\begin{aligned} &\max_y wy - c[\mathbf{w}, g(y)] \\ &= \max_y wy - c^*(y, \mathbf{w}), \end{aligned} \tag{12}$$

where $c^*(y, \mathbf{w})$ can be called the pseudo cost function. The first order necessary condition for the profit maximization problem gives the following equation which needs to be solved to get us to the pseudo production function,

$$w = \frac{\partial c^*(y, \mathbf{w})}{\partial y}. \tag{13}$$

Therefore, profit maximization for the given demand function in this model is characterized by the equality between the price of the pseudo product and the marginal pseudo cost. To illustrate how the profit function for the monopolist can be derived by starting from the pseudo cost function, we start from the chosen production function in (5). The cost function is derived from the model of cost minimization for the given level of production[1],

$$\begin{aligned} c(x_0, \mathbf{w}) &= \min_{\mathbf{x}} \ \mathbf{wx} \\ subject\ to\ & x_1^{\frac{1}{4}} x_2^{\frac{1}{4}} = x_0. \end{aligned} \tag{14}$$

From the theory of production it is known that the economic efficiency is characterized by the equality between the marginal rate of technical substitution $MRTS = \frac{\frac{\partial f(x_1, x_2)}{\partial x_1}}{\frac{f(x_1, x_2)}{\partial x_2}}$ and the input price ratio $\frac{w_1}{w_2}$ [8]. In our case it is $MRTS = \frac{x_2}{x_1} = \frac{w_1}{w_2}$, from which the long-run expansion path is derived, which describes the optimal combinations of inputs at each output level as output expands, $x_2 = \frac{w_1 x_1}{w_2}$. Substituting it in the constraint, we get the conditional input demand functions, which give the cost minimizing input levels for the given output level, $x_1(w_1, w_2, x_0) = w_1^{-\frac{1}{2}} w_2^{\frac{1}{2}} x_0^2$ and $x_2(w_1, w_2, x_0) = w_1^{\frac{1}{2}} w_2^{-\frac{1}{2}} x_0^2$, whose substitution in the goal function of (14) gives us the cost function

$$c(w_1, w_2, x_0) = 2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} (x_0)^2. \tag{15}$$

Our goal is to express it in terms of the pseudo production function and to obtain the pseudo cost function. For the given inverse demand function in (6) the pseudo production function is defined as $y = x_0(a - bx_0) = ax_0 - bx_0^2$. Therefore, the quantity of production is related to the pseudo production function by the following quadratic equation $b(x_0)^2 - ax_0 + y = 0$ from which it follows $x_0 = \frac{a \pm \sqrt{a^2 - 4by}}{2b}$ and the pseudocost function as a function of pseudoproduction function collapses to

$$c^*(y, w_1, w_2) = 2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} \left(\frac{a \pm \sqrt{a^2 - 4by}}{2b}\right)^2. \tag{16}$$

Therefore, the following optimization model needs to be solved

$$\max_y \ wy - 2 w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} \left(\frac{a \pm \sqrt{a^2 - 4by}}{2b}\right)^2. \tag{17}$$

The first order necessary condition is

[1] More on derivation and the properties of the cost function can be found in [8].

$$\frac{\partial c}{\partial y} = \pm \frac{2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} a}{b\sqrt{a^2 - 4by}} - \frac{2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}}}{b} = w. \tag{18}$$

Solving the equation for *y* gives the pseudoproduction function,

$$y = \frac{a^2}{4} \frac{4ww_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw^2}{\left(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw\right)^2}, \tag{19}$$

whose substitution into the goal function leads to the same profit function as before,

$$\pi^* = w\frac{a^2}{4} \frac{4ww_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw^2}{\left(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw\right)^2} - 2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} \frac{a^2 w^2}{4(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2} = \frac{a^2 w^2}{4(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)}. \tag{20}$$

4 HOTELLING'S LEMMA

According to Hotelling's lemma [1,2,5,7,8], the derivative of the profit function for a perfectly competitive, price-taking firm with respect to the product price of a firm is equal to the firm's quantity supplied. Since the monopolist is not the price taker, the question is how can be Hotelling's lemma applied in this case. Starting from the monopolist's profit function and looking at the parameter *w* as the price of the pseudo product, we can apply Hotelling's lemma and get the pseudo production function [6],

$$y = \frac{\partial \pi^*}{\partial w} = \frac{a^2}{4} \frac{2w(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw) - bw^2}{(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2} = \frac{a^2}{4} \frac{4ww_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw^2}{(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2}. \tag{21}$$

The supply function of a monopolist can be obtained from the pseudo production function for the given demand function, $\frac{\partial \pi^*}{\partial w} = \frac{p_0^* x_0^*}{w} = \frac{w(a - bx_0^*)x_0^*}{w} = ax_0^* - b(x_0^*)^2$ or $b(x_0^*)^2 - ax_0^* + \frac{\partial \pi^*}{\partial w} = 0$. So we got the quadratic equation which needs to be solved to get the supply function, x_0^*, for the given demand function

$$x_0^* = \frac{a \pm \sqrt{a^2 - 4b\frac{\partial \pi^*}{\partial w}}}{2b} = \frac{a \pm \sqrt{a^2 - a^2 b \frac{4ww_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw^2}{(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)^2}}}{2b} = \frac{aw}{2(2w_1^{\frac{1}{2}} w_2^{\frac{1}{2}} + bw)}. \tag{22}$$

5 FROM THE PROFIT FUNCTION TO THE PSEUDO PRODUCTION FUNCTION

Starting from the profit function for the monopolist, how can we go back and recover the pseudo production function? Since the profit function gives the maximum profit for every combination of input prices, w_1 and w_2, and the parameter *w*, its definition brings us to the following inequality

$$\pi(w, w_1, w_2) \geq wF(\mathbf{x})D\big[F(\mathbf{x})\big] - w_1x_1 - w_2x_2 \qquad \forall w_1, w_2, w. \tag{23}$$

Definition of the pseudo production function enables us to rewrite the previous inequality as

$$\pi(w, w_1, w_2) \geq wF^*(\mathbf{x}) - w_1x_1 - w_2x_2 \qquad \forall w_1, w_2, w, \tag{24}$$

from which it follows that the pseudo production function is the result of the following optimization problem

$$F^*(\mathbf{x}) = \max\{y : wy \leq w_1x_1 + w_2x_2 + \pi(w, w_1, w_2)\ \}. \tag{25}$$

By normalizing the price of the pseudo product and dividing all the input prices by w, $W_1 = \frac{w_1}{w}, W_2 = \frac{w_2}{w}$, the previous optimization problem reduces to

$$F^*(\mathbf{x}) = \max\{y : y \leq W_1x_1 + W_2x_2 + \pi(1, W_1, W_2)\ \},\ \forall W_1, W_2 \tag{26}$$

and

$$F^*(\mathbf{x}) = \min_{W_1, W_2} W_1x_1 + W_2x_2 + \pi(1, W_1, W_2)\ . \tag{27}$$

Its solution is the monopolist's pseudo production function. Below we illustrate how to recover the pseudo production function from the profit function. Starting from our derived profit function in (10), the normalized profit function[2] is $\pi(1, W_1, W_2) = \dfrac{a^2}{4(2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b)}$ and the pseudo production function can be obtained as the solution to the following optimization problem

$$F^*(\mathbf{x}) = \min_{W_1, W_2} W_1x_1 + W_2x_2 + \frac{a^2}{4(2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b)}\ . \tag{28}$$

The system of equations that expresses the first order necessary conditions follows,

$$\begin{aligned} \frac{\partial f}{\partial W_1} &= x_1 - \frac{a^2}{4}(2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b)^{-2}W_1^{-\frac{1}{2}}W_2^{\frac{1}{2}} = 0 \\ \frac{\partial f}{\partial w_2} &= x_2 - \frac{a^2}{4}(2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b)^{-2}W_1^{\frac{1}{2}}W_2^{-\frac{1}{2}} = 0 \end{aligned} \tag{29}$$

[2] The profit function $\pi(1, \frac{w_1}{w}, \frac{w_2}{w})$ is actually the conjugate function to the pseudoproduction function $F^*(\mathbf{x})$ if $F^*(\mathbf{x})$ is concave [5]. More on the conjugacy approach to duality theory can be found in [7].

Multiplying quantities of inputs gives $x_1 x_2 = \frac{a^2}{16}(2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b)^{-4}$ which implies $2W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} + b = \frac{a}{2}x_1^{-\frac{1}{4}}x_1^{-\frac{1}{4}}$ and $W_1^{\frac{1}{2}}W_2^{\frac{1}{2}} = \frac{a}{4}x_1^{-\frac{1}{4}}x_1^{-\frac{1}{4}} - \frac{b}{2}$. The product of the corresponding input with its price reduces to $W_1 x_1 = \frac{a}{4}x_1^{\frac{1}{4}}x_2^{\frac{1}{4}} - \frac{b}{2}x_1^{\frac{1}{2}}x_2^{\frac{1}{2}}$ and $W_2 x_2 = \frac{a}{4}x_1^{\frac{1}{4}}x_2^{\frac{1}{4}} - \frac{b}{2}x_1^{\frac{1}{2}}x_2^{\frac{1}{2}}$. Inserting the given results in the goal function gives

$$F^*(\mathbf{x}) = 2\frac{a}{4}x_1^{\frac{1}{4}}x_2^{\frac{1}{4}} - 2\frac{b}{2}x_1^{\frac{1}{2}}x_2^{\frac{1}{2}} + \frac{a}{2}x_1^{\frac{1}{4}}x_2^{\frac{1}{4}} = ax_1^{\frac{1}{4}}x_2^{\frac{1}{4}} - bx_1^{\frac{1}{2}}x_2^{\frac{1}{2}}, \tag{30}$$

which is the pseudo production function we started from.

6 CONCLUSION

Since duality in microeconomics rests on price taking behaviour, the main idea was to apply the known duality principles to the monopolistic case. In deriving the profit function for the monopolist the standard approach is shown which includes starting from the production function and the defined pseudo production function. We give another approach by starting from the pseudo cost function. Finally, starting from the derived profit function the pseudo production function is recovered. A version of Hotelling's lemma, important from the empirical standpoint, is given. Application of the duality theory in monopoly to the real data is left for the future research.

References:

[1] Appelbaum, E. (1975). Essays in the Theory and Applications of Duality in Economics. PhD Thesis. Vancouver:University of British Columbia.

[2] Appelbaum, E. (1979). Testing Price Taking Behaviour. Journal of Econometrics, 9: 283–294.

[3] Blume, L. E. (2008). Convex Programming. In Durlauf, S. N., Blume, L. E. (Eds.). The New Palgrave Dictionary of Economics Online.
<http://www.dictionaryofeconomics.com/article?id=pde2008_C000348>
doi:10.1057/9780230226203.0314.

[4] Blume, L. E. (2008). Duality. In Durlauf, S. N., Blume, L. E. (Eds.). The New Palgrave Dictionary of Economics Online.
<http://www.dictionaryofeconomics.com/article?id=pde2008_D000196>
doi:10.1057/9780230226203.0411.

[5] Diewert, W.E. (1982). Duality Approaches to Microeconomic Theory. In Intriligator, M.D., Arrow, K..J. (Eds.). Handbook of Mathematical Economics (pp. 535–599). Amsterdam: North Holland.

[6] Krpan, M. (2011). Dualnost u mikroekonomskoj teoriji. PhD Thesis. Zagreb: Sveučilište u Zagrebu, Ekonomski fakultet-Zagreb.

[7] Lau, L. J. (1978). Applications of Profit Function. In Fuss, M., McFadden, D. (Eds.). Production Economics: A Dual Approach to Theory and Applications (pp. 133–216). Vol II, Part IV. Amsterdam: North-Holland.

[8] Mas – Colell, A., Whinston, M. D., Green, J. (1995). Microeconomic Theory. New York: Oxford University Press.

[9] Shepard, R.W. (1970). Theory of Cost and Production Functions. Princeton: Princeton University Press.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 2:

Graphs and Their Applications

LOGISTICS CYCLE INDEXES: SUPPLY NETWORK ANALYSIS TAKES A STEP FURTHER

Anna Azzi, Daria Battini, Alessandro Persona and Fabio Sgarbossa
Department of Management and Engineering, University of Padua, Stradella San Nicola 3, 36100, Italy.
Email: annaazzi.uni@gmail.com, daria.battini@unipd.it, alessandro.persona@unipd.it, fabio.sgarbossa@unipd.it

Abstract: Supply Network Analysis (SNA) is a methodology inspired by Ecological Network Analysis and based on the concept of *information entropy*. By measuring flows of goods and interaction costs between different sectors of activity, within the supply chain, a network of flows can be empirically built and successively investigated by SNA, a powerful tool for supply chain analysis and mapping. To reach a complete comprehension of the complexity dimension, as well as to extend the analysis to reverse logistics, SNA is here extended to the study of supply chain cycling, with the introduction of ten different Logistics Cycle Indexes (LCIs), meant to identify complexity drivers and quantify how much a supply chain contains closing loops. The theoretical model, properly developed to investigate forward and reverse logistics flows, provides a diagnosis of supply chains structure in presence of closed cycles.

Keywords: Supply Network Analysis, Closed Loop Supply Chain, Reverse Logistics, Material Cycle, Logistic cycles.

1 INTRODUCTION ON LOGISTIC CYCLING CATEGORIES

Reverse logistics is a hot topic, which has had a significant economic impact on industry as well as society [6], and its function in supply chain management has received great attention in recent years due to increased awareness and implementation of legal requirements [3, 9]. Although manufacturers show a growing interest in extracting value from product returns, the need to make the appropriate reverse supply chain design choices, has not inspired much research [3]. Starting from this perspective, the opportunity to develop a methodology to analyze and compare these kinds of network, make the topic even more challenging. As previously mentioned, the purpose is not just to extend SNA to the analysis of reverse logistics flows, but in a broader sense, to include every cause of cycling inside a supply network, which cause an increase of logistics complexity. From a literature review, ten different categories of logistics flows have been identified, and summarized in Figure 1. It represents a draft generalization of possible logistics flow that involves cycling. Of course it is not to be intended as a hard configuration, but rather a theoretical and conceptual framework of real flows; in other word, supply networks are neither necessarily characterized by every possible cycling flow nor portrayed by flows having exactly the schematized paths. The aim is rather to be comprehensive of all possible cycling categories previously detailed. As illustrated in figure 1, customers may return products to the renter or reseller for several reasons: in such a case products are returned to the forward distribution channel. In some cases, diagnostic tests are performed to determine what action would recovers the most value from the returned product: this is done also in case of product returns due to defects or failures. Returns from costumers may also be related to maintenance tasks, for repairing and refurbishing. Cannibalizing and remanufacturing are activities that may be conducted in common plants or structures that pool all remanufacturing activities in a separate plant. Remanufactured products may be sold in the same market or in a secondary markets, often to a marketing segment unwilling or unable to purchase a new product. At last, returns may also be used to recover spare parts for warranty claims, to reduce the cost of providing these services for customers. Products which are not either partially reused or remanufactured are designated for scrap or recycling, usually after physically destroying the product. Reverse

logistics flows headed for recycling may also come directly from a municipal waste collection or a third party recycler. Other possible logistics cycling, feasible at every supply chain tier, are those related to returnable goods, such as returnable transportation items, returnable packaging materials and certain categories of reusable products. At the same time, at every tier, supply chain members can be part of materials and products cycling due to returns for non-conformity (both related to quantity and quality). Last, but not least, logistics cycling may be related to part processing subcontracting.

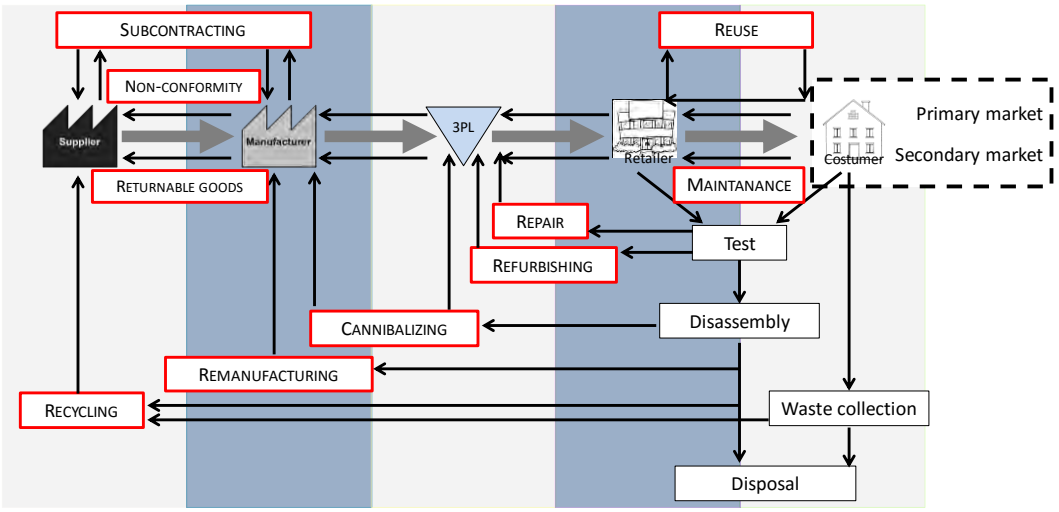


Figure 1. Logistics cycling categories framework.

2 FINN CYCLING INDEX

A cycle is defined as a path, through which energy and matter flow through the chain to return to the starting point, and is often examined in ecosystem ecology, especially as it relates to the behavior of autonomous systems, i.e. characterized by reduced dependence on external energy absorption. Although the presence of trophic cycles was discovered and reported early in ecological studies [4], the first model aiming to quantify the amount of cycling occurring was not proposed until the end of the nineteen-seventies, by Finn in 1976 [2] in the context of Ecological Network Analysis (ENA).What became known as Finn's cycling index (FCI) accounts for the percentage of all fluxes that is generated by cycling, and has been applied in a wide range of ecological studies [1, 5]). The chief advantage of FCI has been its simplicity, as its computation requires but a single matrix inversion, and its dimensionless, a feature that allows ecologists to directly compare diverse ecosystems. Before giving a deeper dissertation of Finn's model and its development and computation, we should take a back step, since Finn's methodology starts by employing the so called "Input-Output" technique to quantify the amount of recycling in ecosystems. In economics, an Input-Output model is a quantitative economic technique that represents the interdependencies between diverse branches of the national economy or between branches of different, and even competing economies. Wassily Leontief in 1967 [7] was credited with Nobel Memorial Prize in Economic Sciences for the development of this model. The method consists in the construction of a matrix reflecting the economic structure of inter-branches flows in an economic system, aiming to estimate the amount of raw materials and services required to produce a certain quantity of goods. Input-Output analysis of ecosystems [2, 8, 10] is an ecological adaptation of the original Input-Output analysis proposed by Leontief [7] and can be considered the starting point for Finn's methodology. Given a matrix of

exchanges T, one can normalize its columns by dividing each coefficient $T_{i,j}$ by its corresponding inflow $S_{i,}$.

$$S_i = T_{.i} + X_{i.} \tag{1}$$

or, in other words, defining a fractional inflow matrix, [G],where [G] are obtained from the elements of the flow matrix, [T], and the input vector, (X), by normalizing the inter-compartmental exchanges using the total input to the receiving node, j,

$$g_{i,j} = \frac{t_{i,j}}{\sum_k t_{k,j} + X_i} \tag{2}$$

Element $g_{i,j}$ represents the fraction of j's inflows that is comprised by i. Reading column j of [G], information about the percentages for each logistics flow coming from i and entering a node j, which constitutes of the full intake by j.

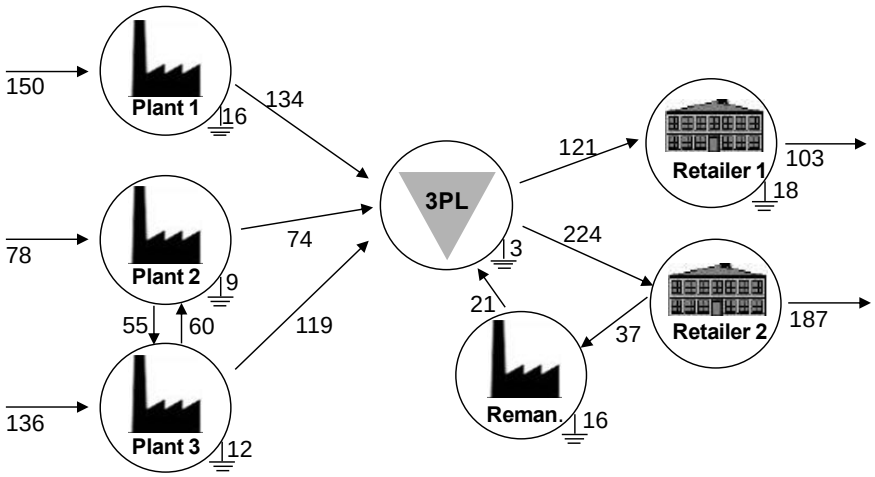


Figure 2. Example of an industrial supply chain.

For example, in the very simple network in figure 2, the nine non-zero values of $T_{i,j}$, generate corresponding nine elements in the matrix [*G*].

Multiplying the matrix $[G]$($[G]\times[G]=[G]^2$), the reader's attention is drawn to the fact that each of the non-zero elements of $[G]^2$ corresponds to the collection of pathways of length 2 that connect i with j. For example, the 1–5 element of $[G]^2$ reveals how much gets to 5 from 1 over the two step pathway 1→4→5, i.e. plant1→3PL→retailer 1. Multiplying $[G]^2$ by $[G]$ once more yields the matrix $[G]^3$. Again, non-zero elements of $[G]^3$ correspond to the three step pathways in the graph. For exemple, 1-7 element of matrix $[G]^3$ match with the path 1→4→6→7, i.e plant 1→3PL→retail 2→recovery product plant (remanufacturing). Thus, the m^{th} power of [*G*] contains contributions from each and every pathway of exactly length *m* in the graph. The sequence of powers of $[G]$ truncates with $[G]^k = [0]$, whenever there are no pathways $> k$ in the network.

$$G = \begin{vmatrix} 0 & 0 & 0 & g_{14} & 0 & 0 & 0 \\ 0 & 0 & g_{23} & g_{24} & 0 & 0 & 0 \\ 0 & g_{34} & 0 & g_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & g_{45} & g_{46} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & g_{67} \\ 0 & 0 & 0 & g_{74} & 0 & 0 & 0 \end{vmatrix}$$

$$G^2 = \begin{vmatrix} 0 & 0 & 0 & 0 & g_{14} \cdot g_{45} & g_{14} \cdot g_{46} & 0 \\ 0 & g_{23} \cdot g_{34} & 0 & g_{23} \cdot g_{34} & g_{24} \cdot g_{45} & g_{24} \cdot g_{46} & 0 \\ 0 & 0 & g_{34} \cdot g_{23} & g_{34} \cdot g_{24} & g_{34} \cdot g_{45} & g_{34} \cdot g_{46} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & g_{46} \cdot g_{67} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & g_{67} \cdot g_{74} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & g_{74} \cdot g_{45} & g_{74} \cdot g_{46} & 0 \end{vmatrix}$$

When logistics cycling flows are in the network, the sequence of power of $[G]$ does not vanish, though growing progressively smaller. Recalling that the geometric series:

$$\sum_{n=0}^{\infty} q^n = 1 + q + q^2 + q^3 + q^4 + \ldots = \frac{1}{1-q} \tag{3}$$

Whenever $-1 < q < 1$, it is possible to demonstrate (Higashi et al., 1991) that, whenever $0 \le G_{ij} \le 1$, then

$$\sum_n G^n = I + G + G^2 + \ldots G^n \rightarrow [I\text{-}G]^{-1} \tag{4}$$

Where $[I]$ is the identity matrix (i.e. it consists of ones along its diagonal and zeroes elsewhere.) This limit, $[L] = [I - G]^{-1}$,is called the Leontief structure matrix. The i–jth component of [L] provides the fraction of the total input to j from i over all pathways of all lengths per unit of final demand, which plays a key role in economic theory. The discovery of the [S] matrix enabled economists to estimate the necessary production in various economic sectors in order to satisfy any vector of final demands. The Leontief matrix can be interpreted as follows: the number of times a quantum entering i^{th} will visit i^{th} compartment (the diagonal elements) is at least 1, where any coefficient greater than unity indicates that the compartment participates in the cycles.

The Finn cycling index (Finn, 1976) utilizes the Leontief matrix to assess the amount of material cycling within the supply chain. The formula, derived from the inverse matrix L is straightforward and simple:

$$FCI = \sum_{i=1}^{N} \frac{S_i}{TST} \frac{l_{ii} - 1}{l_{ii}} \tag{5}$$

where l_{ii} is the i^{th} coefficient along the diagonal of the Leontief matrix, TST is the Total System Throughput $TST = T_{..} + X_{.} + E_{.} + D_{.}$and S_i is the total inflow to the i^{th} supply chain

member, where $S_i = T_{.i} + X_{i.}$ The FCI related to the simple distribution network under study turns out to be: $FCI = 0.0495$ meaning that the 4.95% of logistics flow are due to cycling.

3 LOGISTICS CYCLE INDEXES AND FINAL CONCLUSION

Up to here, it is clear how SNA is successful in highlighting the complexity of a logistics network, however, what is not obvious is how to discriminate this complexity and how to relate it to the proper kind of logistics flows, since cycles certainly increase network complexity and FCI is capable of measuring the amount of material and product cycling within a supply network. Based on these considerations, therefore, a methodology to identify which fraction of FCI can be ascribed to a category of cycling rather than another is proposed. These "fraction" are called Logistics Cycle Indexes (LCIs) and are named after the typology of cycling flows involved, i.e. Reuse LCI, Maintenance LCI, Subcontracting LCI, Non-conformity LCI, Returnable goods LCI, Repair LCI, Refurbishing LCI, Remanufacturing LCI, Cannibalization LCI, Recycling LCI. The flow chart describing the computation procedure is illustrated in figure 3, together with a very simple example.

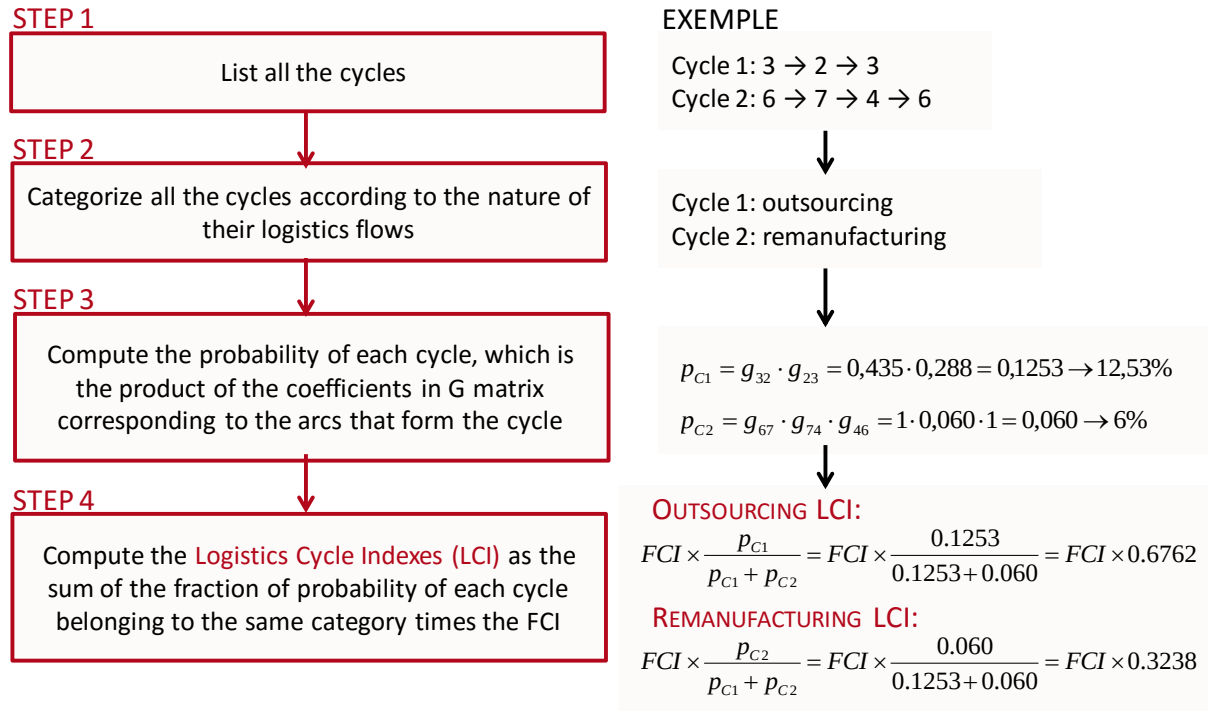


Figure 3. LCIs computation: procedure flow-chart.

The procedure includes the 4 steps reported on the left in figure 3. In the sample network reported in figure 2, two kinds of logistics cycles can be identified:

- Plant 3 → Plant 2 → Plant 3, in case of subcontracting;
- Retailer 2 → Recovery product plant → 3PL → Retailer 2, in case of remanufacturing.

Then, the probability of each cycle, which is the product of the coefficients in G matrix corresponding to the arcs formed by the cycle, can be computed. In this case the first cycle's probability is $p_{C1} = g_{32} \cdot g_{23} = 0.435 \cdot 0.288 = 0.1253 \rightarrow 12.53\%$. While the second cycle probability is: $p_{C2} = g_{67} \cdot g_{74} \cdot g_{46} = 1 \cdot 0.060 \cdot 1 = 0.060 \rightarrow 6\%$

The fraction of probability of each cycle is:

– $\frac{p_{C1}}{p_{C1}+p_{C2}} = \frac{0.1253}{0.1253+0.060} = 0.6762 \rightarrow 67.62\%$

– $\frac{p_{C2}}{p_{C1}+p_{C2}} = \frac{0.060}{0.1253+0.060} = 0.3238 \rightarrow 32.38\%$

Therefore, LCIs in this simple numerical example can be computed for the Outsourcing and Remanufacturing cycles as reported in figure 3:

$$Outsourcing LCI = 0.6762 \cdot FCI = 0.03347$$

$$\mathrm{Re}\, manufacturing LCI = 0.3238 \cdot FCI = 0.01602$$

This simple numerical application aims to demonstrate the real feasibility to mathematically compute the impact of closed loops in a supply chain by a network analysis point of view. Further investigations are under development to extend the methodology to real industrial cases.

References

[1] Allesina, S., Ulanowicz, R. (2004). Cycling in ecological networks: Finns index revisited. Computational Biology and Chemistry 28, 227-233.

[2] Finn, J. T. (1976). Measures of ecosystem structure and function derived from analysis of flows. Journal of Theoretical Biology, 56 (2), 363-380.

[3] Guide Jr., V.D.R., Souza, G., Van Wassenhove, L.N., Blackburn, J.D. (2006). Time value of commercial product returns. Management Science, 52 (8), 1200–1214.

[4] Hutchinson, G. E. (1948). Circular causal systems in ecology. Annals of the New York Academy of Sciences, 50 (4), 221 -246.

[5] Kay, J.J., Graham, L.A., Ulanowicz, R.E. (1989). A detailed guide to network analysis. In: Wulff, F., Field, J.G., Mann, K.H. (Eds.), Network Analysis in Marine Ecology: Methods and Applications. Springer-Verlag, Heidelburg, 15–61.

[6] Krumwiede, D.W. and Sheu, C. (2002). A model for reverse logistics entry by third-party providers, Omega, 30 (5), 322-33.

[7] Leontief, W. (1963). The structure of American Economy 1919-1939. Oxford University Press, New York.

[8] Shannon C.E and Weaver W (1948). The mathematical Theory of Communication, The University of Illinois Press, Urbana, IL.

[9] Simpson, D. (2010). Use of supply relationships to recycle secondary materials. International Journal Of Production Research, 48 (1), 227-249.

[10] Szyrmer, J. M. (1984). Total flow in input-output models. Ph.D. thesis, University of Pennsylvania.

IMPROVING THE EXACT ALGORITHM FOR SOLVING THE PUBLIC SERVICE SYSTEM DESIGN IN THE BRANCHING

Ján Bendík
University of Žilina, Faculty Management Science and Informatics,
Department of Mathematical Method and Operations Research
Univerzitná 8215/1, 01026 Žilina, Slovakia
Jan.Bendik@fri.uniza.sk

Abstract: The paper deals with the problem of designing the optimal structure of a public service system. The public service system design is often related to the p-median location problem. We designed the exact algorithm based on the branch and bound method with the Erlenkotter approach. We try to decrease a gap between the lower and upper bound with a suitable choice of the candidate to the fixation. We designed and verified the various variants of the function for obtaining the candidate to the branching on the Slovak road network. We chose the most feasible function for our exact algorithm.

Keywords: public service system design, p-median location problem, Erlenkotter approach, candidate for the facility location

1 INTRODUCTION

A public administration system, fire-brigade deployment, police service system and medical emergency system [4] represent the typical public service system. The public service system structure is formed by deployment of a limited number of the service centres and the associated objective is to minimize the total costs. The public service system design are often related to the uncapacitated facility location problem or the p-median location problem. The p-median location problem is the NP-hard problem [6], [7], which became one of the most well-known problems in the field of the facility location. This problem is formulated as a task of determination of at most p network nodes as facility locations. The number of the possible service center locations seriously impacts a computational time. Balinski [2] provided an early integer programming formulation of the plant location problem which has historically been adapted to the p-median problem. Finding the optimal solution of the p-median location problem consists of a formulation of the mathematical model and applying some mathematical programming method. Avella, Sassano and Vasil´ev [1] presented a branch-and-price-and-cut algorithm to solving the large-scale instances of the p-median problem. Reese [11] summarized the exact solution methods for the p-median problem. Mladenovic [10] summarized the heuristic methods for solving the p-median problem. This problem is very similar with the uncapacitated facility location problem. Erlenkotter [5] used knowledge from the theory of duality and proposed one of the most effective algorithms DualLoc for solving the uncapacitated facility location problem. The algorithm DualLoc realizes the branch and bound method. Inspired by this approach Korkel [9] improved the Erlenkotter approach and designed the algorithm PDLoc. Janacek and Buzna [8] improved the Erlenkotter and Korkel approach and designed the algorithm BBDual for solving the uncapacitated facility location problem. These algorithms exploit the relation between the primary and dual formulation of the strong linear programming relaxation of the original problem. Erlenkotter [5] tried to minimize a gap between values of the primary and dual solution with the fixation of locations in the branch and bound algorithm and to obtain the better lower bound.

We generalized the Erlenkotter dual approach to the lower bounding to be able to solve p-median location problem in [3]. We formulated the dual model of the p-median location problem and verified the procedure for obtaining the lower bound. We designed the exact algorithm pMedBBDual based on the branch and bound method. The designed algorithm processed big number of the processed nodes which increase the computational time. The

better lower bound in the branch and bound method is addicted to a choice of the candidate to the fixation. Obtaining the good candidate can markedly decrease number of the processed nodes. We design the various variants of the function for obtaining the candidate to the branching. We verify the individual variants of the function in the computational time and the frequency of execution on the benchmarks from Slovak road network. We choose the most effective function for our exact algorithm pMedBBDual.

2 PROBLEM FORMULATION AND OUR SOLUTION

2.1 Mathematical model of p-median location problem

The p-median location problem can be modeled using the following notation. Let the decision of the service center location at the place $i \in I$ be modeled by a zero-one variable $y_i \in \{0, 1\}$ which takes the value of 1, if a center is located at i, otherwise it takes the value of 0. In addition, the variables $z_{ij} \in \{0, 1\}$ for each $i \in I$ and $j \in J$ represent to assign a customer j to a possible location i by the value of 1. The maximal number of the facility locations represents the constant p. The mathematical model for the p-median location problem is formulated as follows (1-6):

$$Minimize \qquad F_P = \sum_{i \in I} f_i \, y_i + \sum_{i \in I} \sum_{j \in J} c_{ij} \, z_{ij} \tag{1}$$

$$Subject\ To: \qquad \sum_{i \in I} z_{ij} = 1 \qquad for\ j \in J \tag{2}$$

$$z_{ij} \leq y_i \qquad for\ i \in I, \quad j \in J \tag{3}$$

$$\sum_{i \in I} y_i \leq p \tag{4}$$

$$y_i \in \{0,1\} \qquad for\ i \in I \tag{5}$$

$$z_{ij} \in \{0,1\} \qquad for\ i \in I, \quad j \in J \tag{6}$$

The objective function (1) represents the minimization of the total costs of the p-median location problem which consists of the fixed charges f_i and the costs c_{ij}. The constraints (2) ensure that each customer is assigned to the exactly one possible service center location. Binding constraints (3) enable to assign a customer to a possible location i, only if the service center is located at this location. The constraint (4) bounds the maximal number of the located service centers. The obligatory conditions in the mathematical model are (5) and (6).

2.2 Dual model of p-median location problem

We applied the theory of duality and formulated the associated dual model of the p-median location problem than Erlenkotter for solving the uncapacitated facility location problem. The dual model of the p-median location problem has the following notation (7-11):

$$Maximize \qquad F_D = \sum_{j \in J} v_j + px \tag{7}$$

$$Subject\ To: \qquad \sum_{j \in J} max\{0, v_j - c_{ij}\} + x + u_i = f_i \quad for\ i \in I \tag{8}$$

$$v_j \geq 0 \qquad for\ j \in J \tag{9}$$

$$u_i \geq 0 \qquad for\ i \in I \tag{10}$$

$$x \leq 0 \tag{11}$$

The dual variables v_j correspond to the constraints (2). The dual variable x corresponds to the constraint (4). A lower bound for the optimal solution of the problem (1-6) constitutes the objective function value of the arbitrary feasible solution of the dual problem (7 – 11).

2.3 Complementary conditions

A dual solution and an induced primal feasible solution can be obtained by applying the complementary constraints (12 - 15):

$$(y_i - z_{ij})\ max\{0, v_j - c_{ij}\} = 0 \qquad for\ i \in I, \quad j \in J \tag{12}$$

$$u_i y_i = 0 \qquad for\ i\ \in I \tag{13}$$

$$max\{0, c_{ij} - v_j\}\ z_{ij} = 0 \qquad for\ i \in I, \quad j \in J \tag{14}$$

$$(p - \sum_{i \epsilon I} y_i)\ x = 0 \tag{15}$$

According to the weak duality theorem, the objective function value F_D (7) of any feasible solution is smaller or equal to any objective function value F_P (1) of any feasible solution of the linear relaxation of the problem (1-6). Our approach for construction of the associated primal solution to the dual solution minimizes the difference between the primal and dual value of the objective function as follows (16):

$$F_p - F_D = \sum_{i \in I} \sum_{j \in J} (y_i - z_{ij})\ max\{0, v_j - c_{ij}\} + (\sum_{i \in I} y_i - p)\ x + \sum_{i \in I} u_i y_i + \sum_{i \in I} \sum_{j \in J} z_{ij}\ max\{0, c_{ij} - v_j\} \tag{16}$$

F_p represents a value of the objective function of the p-median location problem (1-6). F_D represents a value of the objective function of the associated dual problem (7-11).

2.4 Choice the candidates

Our algorithm creates the nodes of the branching tree in the branch and bound methods with the 0-1 fixation of the candidate for the facility location. It means, one fixation of the selected candidate distributes one problem to two subproblems. We solve subproblems and try to decrease a gap (16) between the lower and upper bound which is addicted to a choice of the candidate to the fixation and a minimal set of locations. We designed the function for obtaining the minimal set of locations. This function ensures that condition (15) will be respected and tries to obtain the minimal set of locations with the minimal gap (16). It is not possible to ensure always the satisfaction of conditions (12-14) when we construct of a minimal set. Otherwise, the created gap is possible to decrease with the good choice of the candidate for the location. Obtaining a good candidate also decreases a number of the processed nodes and the computational time. The function for obtaining the candidate is based on the evaluation of the complementary conditions (12-14) and finding the first location from the set of locations which does not satisfy the conditions (12-14). This location gives a candidate to the fixation. We designed 6 variants of the function for obtaining the candidate which differ in the order of the evaluation of the complementary conditions (see Table 1).

Table 1: Order of evaluation of complementary conditions for created variants of function

	Evaluation of conditions		
	1.	*2.*	*3.*
Variant V1	12	13	14
Variant V2	12	14	13
Variant V3	14	13	12
Variant V4	13	14	12
Variant V5	13	12	14
Variant V6	14	12	13

We search the best variant of the function for obtaining of the candidate to the fixation in our exact algorithm pMedBBDual. If we do not obtain a candidate with the evaluation of conditions (12-14), then we choose the candidate $i \in I$ with a minimal value x_{min} (17) of the variable x.

$$x_{min} = \min_{i \in I} \left\{ f_i - \sum_{j \in J} max\{0, v_j - c_{ij}\} \right\} \tag{17}$$

3 EXPERIMENTS AND EVALUATION

We tested and verified our algorithm pMedBBdual with designed variants of function for obtaining the candidate on the benchmarks from the Slovak road network. Tested benchmarks were BB100x100 - 100 candidates for the facility location and 100 costumers from district of Banska Bystrica (see Table 3), ZA315x315 - 315 candidates for the facility location and 315 costumers (Table 2 and Table 3) and others (NR350x350 - district of Nitra and TN276x276 – district of Trenčin). We verified the variants in the computational time and the number of the obtained candidates (frequency of function execution).

The experiments from district of Banská Bystrica showed us the small differences of the variants in the computational times and the number of the candidates. So we tested the larger benchmarks from the Slovak road network (Nitra, Trenčín, Žilina). We selected for the demonstration of the obtained results some cases from district of Žilina, where the differences in the computational time between the variants were the largest (see Table 2). The experiments in the Table 2 showed us that obtaining the optimal solution in some cases is time-consuming (symbol *). We obtained with the variants V1, V2, V4 and V5 the shortest time in the tested cases. Otherwise the variants V3 and V6 did not provide us the best computational times.

Table 2: Comparing variants of function for obtaining candidates on the benchmark ZA315x315

p	***Computational times in the seconds (s)***					
	V1	*V2*	*V3*	*V4*	*V5*	*V6*
24	511	644	928	105	81	914
27	151	94	501	451	403	490
30	78	125	206	546	379	178
51	843	1 222	918	273	424	916
54	194	160	2 035	743	861	2 027
55	317	312	3 600*	1 046	1 059	3 600*
58	3 600*	2 118	3 600*	1 468	1 729	3 600*
59	3 600*	3 058	2 721	1 102	899	2 676
60	3 313	3 600*	1 954	614	495	1 962
62	710	1 205	3 600*	1 040	1 075	3 600*
63	591	641	2 242	484	462	2 217
65	524	891	3 600*	713	597	3 600*

We want to choose the most feasible variant of function, so we use the total statistics (absolute and relative number of the candidates, absolute and relative computational time) for the better evalution of the results and the feasibility (see Table 3). Absolute number of candidates gives the frequency of the function execution for all values of the constant *p*. Relative number of candidates gives a number of the candidates of the actual variant to the minimal number of the candidates obtained by some variant for all values of the constant *p*. Absolute computational time gives sum of the times of actual variant for all values of the constant *p*. Relative computational time gives times of the algorithm with the actual variant to the times of the algorithm for the variant with the shortest time for all values of the constant *p*.

Table 3: Comparing variants of function for obtaining candidates in total statistics on the benchmarks BB100x100 and ZA315x315

Variant	**BB100x100**				**ZA315x315**			
	Number of candidates		***Computational time***		***Number of candidates***		***Computational time***	
	absolute	*relative*	*absolute*	*relative*	*absolute*	*relative*	*absolute*	*relative*
V1	16 204	191,71	39,38	123,70	1 165 219	992	44 648	1 244
V2	43 026	462,51	58,24	161,02	1 436 251	1 025	56 145	1 304
V3	17 100	211,54	49,63	152,58	3 436 204	1 553	98 432	2 054
V4	16 557	194,96	46,30	132,91	584 149	480	23 735	510
V5	16 633	195,67	51,15	146,21	605 787	490	22 716	469
V6	16 898	207,06	53,97	163,89	3 399 542	1 431	90 738	1 996

The total statistics in the Table 3 for the benchmark BB100x100 showed us that the best variant of function gives V1. Otherwise, variant V1 for the larger-scale benchmark ZA315x315 is doubly time-consuming than variants V4 and V5. The variants of function evaluate condition (13) on the first place. The differences between V4 and V5 are small in the computational time and the frequency of the candidates. Based on all executed experiments from the Slovak road network and the total statistics the variant V5 provides us the better computational time in the most of cases than the variant V4. The variant V5 is the most feasible for our designed algorithm pMedBBDual.

4 CONCLUSION

The public service system design is the NP-hard problem. This problem is often related to the p-median location problem. We designed the exact algorithm pMedBBDual for solving the p-median location problem based on the Erlenkotter approach. Erlenkotter designed one of the most effective algorithm for solving the uncapacitated facility location problem. We generalized the Erlenkotter dual approach to the lower bounding to be able to solve the associated location problem with restricted number of the located service centers. The obtaining the good lower bound is addicted on the effective procedures for obtaining the dual solution and the good choice of the candidate to the fixation in the branch and bound method. The obtaining the good candidate also decreased the number of the processed nodes and the computational time. We designed six various variants of the function for obtaining the candidate to the branching. We tested and verified our algorithm pMedBBdual with the designed variants of the function for obtaining the candidate on the benchmarks from the Slovak road network in the computational time and the number of the obtained candidates. Based on all executed experiments from the Slovak road network and the total statistics the variant V5 provided us the better computational time in the most of cases. The variant V5 is for us the most feasible for our designed algorithm pMedBBDual.

In the future we would like to design new functions for the construction of the minimal set of the facility locations and compare with the actual used function.

Acknowledgement
This work was supported by the research grants VEGA 1/0518/15 "Resilient rescue systems with uncertain accessibility of service" and VEGA 1/0339/13 „Advanced microscopic modelling and complex data sources for designing spatially large public service systems" and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks". We would also like to thank to „Centre of excellence for systems and services of intelligent transport"(ITMS 26220120050) for built up the infrastructure, which was used.

References

[1] Avella, P., Sassano A., Vasilev I. (2007). Computational Study of Large-Scale p-Median Problems. Mathematical Programming, 109(1): 89-114.

[2] Balinski, M. (1965). Integer programming, methods, uses and computation. Management Science 12(3): 254–313.

[3] Bendík, J. (2014). Generalization of the Erlenkotter approach for solving of the public service system design, Proceedings of the 2nd international conference on Traffic and Transport Engineering (ICTTE), Belgrade, Serbia, pp. 12-17.

[4] Drezner, Z., Hamacher, H.W. (2002). Facility location: Applications and theory. Berlin, Germany: Springer.

[5] Erlenkotter, D. (1978). A Dual-Based procedure for uncapacitated facility location. Operations Research, 26(6): 992-1009.

[6] Garey, M. R., Johnson, D. S. (1979). Computers and intractability: A guide to the theory of NP-completeness, A Series of Books in the Mathematical Sciences. San Francisco, USA: W.H. Freeman and Company.

[7] Hakimi, S. L., Kariv, C. (1979). An algorithmic approach to network location problems. SIAM Journal on Applied Mathematics, 37(3): 539–560.

[8] Janacek, J., Buzna, L. (2008). An acceleration of Erlenkotter-Korkel´s algorithms for the uncapacitated facility location problem. Annals of Operations Research, 164(1): 97-109.

[9] Körkel, M. (1898). On the exact solution of large – scale simple plant location problem. European Journal of Operational Research, 39(2): 157-173.

[10] Mladenovic, N., Brimberg, J., Hansen, P., Moreno-Perez, J.A. (2007). The p-median problem: a survey of metaheuristic approaches. European Journal of Operational Research, 179(3): 927–939.

[11] Reese, J. (2006). Solution methods for the p-median problem. Networks, 48(3): 125–142.

VARIOUS MODIFICATIONS OF GREEDY ALGORITHM FOR THE ZONE PARTITIONING PROBLEM IN ZONE TARIFF SYSTEMS

Michal Koháni
University of Zilina, Faculty of Management Science and Informatics
Univerzitna 8215/1, 01026 Zilina, Slovakia
Michal.Kohani@fri.uniza.sk

Abstract: Zone partitioning problem is one of many problems solved when designing zone tariff. It can be described by the mathematical model similar to the p-median problem, but main drawback of this model is a nonlinear objective function. This paper deals with various modifications of greedy heuristic approach to the zone partitioning problem which, on the basis of specified criteria, gradually merge small areas into larger zones. We make a computational study on test data from selected region in Slovak Republic to compare effectivenes of proposed heuristic methods with exact solution.

Keywords: tariff planning, tariff zones design, zone partitioning, heuristic method, greedy approach.

1 INTRODUCTION

Planning of the regional public transportation is connected with a large set of optimization problems. One of problems that transport planners deal with is the problem of the tariff and the ticket prices. Various types of tariffs were mentioned in [4] and [10]. In this paper we are dealing with the counting zones tariff system, where the region is divided into smaller sub-regions - tariff zones and the price for travelling depends on the origin and destination zone and on the number of travelled zones on the trip.

Figure 1: Example of the counting zones tariff system in Bratislava region [12]

When we want to design zone tariff system, there are several decisions that need to be made. It is necessary to design zones and to fix new fares. Several approaches for designing zone tariff system can be found in the literature. Hamacher and Schöbel in [4], Schöbel in [10] and Babel and Kellerer in [1] proposed exact solution approaches for the counting zones tariff system where the goal was to design zones such that new and old price for most of the trips are as close as possible. A note on fair fare tariff on the bus line was mentioned also by Palúch in [9]. Exact algorithm to solve the problem was mentioned also in [5] and [6].

This paper deals with various modifications of greedy based heuristic approach to the zone partitioning in the counting zone tariff system design and will be organizes as follows. In section 2 we present mathematical model of the zone partitioning for counting zones tariff system and we briefly describe solution approaches. In section 3 we introduce greedy based heuristic method with various criteria of merging small areas into larger zones. In section 4 we present a computational study on the test networks to compare the results obtained by greedy heuristic method with the solution of exact method in terms of computation time and preciseness of the solution.

2 MATHEMATICAL MODEL OF THE ZONE PARTITIONING PROBLEM WITH GIVEN PRICES AND NUMBER OF ZONES

Let all stations in the public transport network constitute the set of nodes *I*. Stations *i* and *j* from the set *I* are connected by the edge $(i,j) \in E$, if there is a direct connection by public transport line between these two stations. Symbol *E* denotes the set of edges. Distance between stations *i* and *j* is denoted as d_{ij}. For each pair of stations *i* and *j* is c_{ij} the current or fair price of travelling between these two stations. We assume that c_{ij} is equal to c_{ji} for each pair of stations *i* and *j*. Number of passengers between stations *i* and *j* is b_{ij} (OD matrix). To describe passenger flows we introduce parameter a_{ij}^{rs}, where used paths will be observed. Value of the parameter a_{ij}^{rs} is equal to 1 if the edge (r,s) is used for travelling from station *i* to station *j* and 0 otherwise.

Zone partitioning problem is similar to the general graph partitioning problem [3], where we want to divide the vertex set of a graph into a number of nonempty subsets so that the total weight of edges connecting distinct subsets is minimized. In the zone partitioning problem we use different objective function compared to general graph partitioning problem.

Construction of the zone partitioning model was inspired by the model of the p-median problem [5]. Based on this model, we introduce binary variables y_i, which represent a "fictional" centre of the zone and is equal to 1 if there is a centre of the zone in node *i* and 0 otherwise. For each pair of stations *i* and *j* we introduce variable z_{ij} which is equal to 1 if a station *j* is assigned to the zone with centre in the node *i* and 0 otherwise. We expect to create at most p_{max} tariff zones.

When we want to set a new price for travelling in proposed system, there are more possibilities how to do it. Hamacher and Schöbel in [4] and Schöbel in [10] proposed solution of a fare problem with fixed zones to obtain new fares for trips with various number of travelled zones. In [5] and [7] a unit price for travelling per one zone was set. In this paper we use two different unit prices, as was mentioned in [5] – price f_1 for travelling in the first zone and unit price f_2 for travelling in each additional zone. If we want to calculate new price of the trip between nodes *i* and *j* in the system, we need to calculate number of zones crossed on this trip. The calculation of the number of crossed zones can be easily replaced by the calculation of crossed zone borders accordingly to [4] and [10]. We assume that station can be assigned only to one zone and the border between zones is on the edge. We introduce binary variable w_{rs} for each existing edge $(r, s) \in E$, which is equal to 1 if stations *r* and *s* are in different zones and is equal to 0 otherwise. For the calculation of crossed borders number we need to determine the used path for travelling between stations *i* and *j*. New price n_{ij} determined by the number of crossed zones is calculated as follows (1):

$$n_{ij} = f_1 + \sum_{(r,s)\in E} f_2 a_{ij}^{rs} w_{rs} \tag{1}$$

When we want to suggest the objective function of the model, there are many possible ways. In presented model we will use the average deviation between current and new price for all passengers, according to the advices of experts in [10]. In objective function (2) value of n_{ij} depends on variables *z* and *w* as explained above. Mathematical model of the zone partitioning with fixed prices and number of zones can be written in the following form:

$$Minimize\ dev_{avg} = \frac{\sum_{i\in I}\sum_{j\in I}\left|c_{ij} - n_{ij}\right| b_{ij}}{\sum_{i\in I}\sum_{j\in I} b_{ij}} \tag{2}$$

$$subject\ to \sum_{i\in I} z_{ij} = 1,\ for\ j \in I \tag{3}$$

$$z_{ij} \le y_i,\ for\ i, j \in I \tag{4}$$

$$\sum_{i \in I} y_i \leq p_{\max} \tag{5}$$

$$z_{ij} - z_{ik} \leq w_{jk}, \; for\, i \in I, (j,k) \in E \tag{6}$$

$$z_{ij} \in \{0,1\}, \; for\, i, j \in I \tag{7}$$

$$y_i \in \{0,1\}, \; for\, i \in I \tag{8}$$

$$w_{ij} \in \{0,1\}, \; for\, (i, j) \in E \tag{9}$$

Conditions (3) ensure that each station will be assigned exactly to one zone. Conditions (4) ensure that the station *j* will be assigned only to the existing centre of the zone. Condition (5) ensures that we will create at most *p* tariff zones. Conditions (6) are coupling between variables for allocation of the station to the zone and variables for determining the zone border on the edge *(j,k).*

Objective function (2) in this model is not a linear function. In [5] and [6] the linearization of the model was proposed and the model was subsequently solved using IP solver Xpress [11]. To determine the optimal values of parameters in the model, a two-phase procedure was used. In the first phase the optimal number of zones was determined and the model with different settings of parameters f_1 and f_2 and with given number of zones *p* was solved in the second phase. Optimal parameter setting was determined from parameters of the solution with the best value of objective function. A major drawback of this process was the time complexity of the problem where the computation time for given parameter *p* grows rapidly with the increase of the problem size, as was written in [6].

Another approach to solve this problem was proposed in [1], [4] and [10]. In the first stage optimal price of travelling is calculated and subsequently, three algorithms were used to calculate the zone partitioning. First algorithm is based on the clustering theory, second algorithm is a greedy algorithm and the last algorithm is based on the spanning tree approach.

3 GREEDY HEURISTIC APPROACH

Based on ideas of previously mentioned methods we proposed a greedy based heuristic method to solve the zone partitioning problem. The algorithm starts with the situation, where each station forms separate zone and the number of zones is equal to the number of stations. In the algorithm we subsequently select two smallest neighbouring zones which are merged into the one new zone. The goal is to create zones that are approximately the same size. Algorithm terminates when given maximum number of zones is reached. Greedy heuristic algorithm with approximately same zone size (*GH_size*) can be described as follows:

STEP 1: Start with a partition *P* consisting of $|I|$ zones, each zone contains a single station. For each zone Z_i from *P* calculate the parameter e_i to express size of the zone.

STEP 2: Determine two neighboring zones Z_i and Z_j, where the sum $e_i + e_j$ is minimal.

STEP 3: Merge zones Z_i and Z_j to the new zone Z_k and get a new partition *P*. For new zone Z_k calculate parameter e_k.

STEP 4: If maximum number p_{max} of zones is reached, then terminate, else go to STEP 2.

For the calculation of zone size parameter e_i we can use various approaches. In [7] and [8] there were proposed various formulas and we select two formulas which give best solutions. Formula (10) calculates parameter e_i as the average distance of all stops in the zone Z_i to the neighbouring zones. Formula (11) use number of inhabitants of all stations in the zone Z_i as the zone size parameter e'_i. In all formulas we use following notation: S_i is the set of stops, which are connected with at least one stop in the zone Z_i, $|Z_i|$ represents the number of nodes in the zone Z_i and parameter b_k represents the number of inhabitants in the node *k.*

$$e_i = \frac{\sum_{k \in Z_i} \sum_{j \in S_i} d_{kj}}{|Z_i| . |S_i|} \tag{10}$$

$$e'_i = \sum_{j \in Z_i} b_j \tag{11}$$

To obtain the best settings of parameters f_1 and f_2 in this approach we apply the *GH_size* method to obtain the zone partitioning with given p_{max} parameter and then we subsequently calculate average difference between new and old prices for all possible settings of parameters f_1 and f_2. As the best solution we choose the one that has the smallest value of the average difference.

Formulas (10) and (11) do not include the price for traveling, what can be the drawback in the preciseness of solution. Therefore we propose the modification of the greedy algorithm, which takes this issue into the consideration.

For the price difference calculation we use two formulas. Formula (12) calculates price difference g_{ij} according to the difference between old and new price only and formula (13) uses the number of passengers between stations i and j the zone as a weight to the price difference g_{ij}.

$$g_{ij} = |c_{ij} - n_{ij}| \tag{12}$$

$$g_{ij} = |c_{ij} - n_{ij}| b_{ij} \tag{13}$$

Greedy heuristic algorithm with price criterion (*GH_price*) can be then described as follows:

STEP 1: Start with given parameters f_1 and f_2, a partition P consisting of $|I|$ zones, each zone contains a single station. For each zone Z_i from P calculate the parameter e_i using formula (10).

STEP 2: For all pairs $i, j \in I$ calculate the price difference g_{ij}. Find i' and j' where $g_{i'j'}$ is maximal.

STEP 3: For all $(r,s) \in E$, where $a_{i'j'}^{rs} = 1$, determine two neighboring zones Z_r and Z_s, where the sum $e_r + e_s$ is minimal.

STEP 4: Merge zones Z_r and Z_s to the new zone Z_k and get a new partition P. For new zone Z_k calculate parameter e_k using formula (10).

STEP 5: If maximum number p_{max} of zones is reached, then terminate, else go to STEP 2.

4 COMPUTATIONAL STUDY

The goal of numerical experiments is to compare proposed greedy heuristic methods with previously mentioned exact approach [5] [6]. We compare the effectiveness of heuristic and exact approach for selected values of parameter p_{max}.

We make the computational study on the data sets created from the real public transportation network in the Zvolen Region in Slovak Republic. Stations in networks are represented by municipalities or parts of municipalities. We use two networks with 25 or 51 stations, which are shown in the Figure 2. Current prices were calculated according to real prices depending on the distance for travelling by regional buses. The OD matrix was estimated using the gravity model as in [2]. Numerical experiments were performed on the personal computer equipped with Intel Core 2 Duo E6850 with parameters 3 GHz and 3.5 GB RAM. Experiments were performed in Xpress Optimizer solver [11].

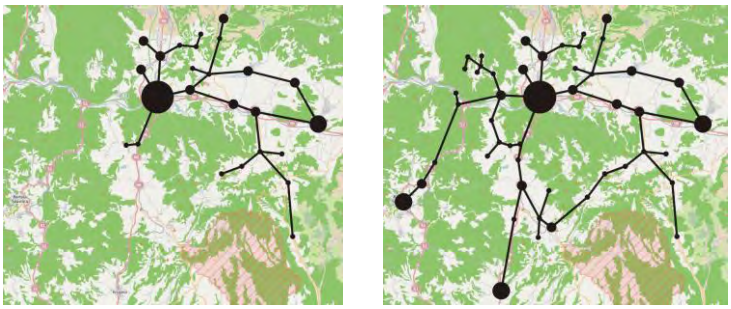


Figure 2: Test networks with 25 and 51 stations, Map source: openstreetmap.org

In the computational study we want to calculate best values of fare prices for selected values of parameter p_{max}. According to the current fare prices, we set the values of parameter f_1 from 0.3 to 0.9 with step by 0.1 and values of parameter f_2 from 0.1 to 0.6 with step by 0.1 for all the experiments.

Table 1 represents results of numerical experiments and Table 2 represents computation time of numerical experiments. In all tables we denote columns as follows. Columns denoted as *Exact* represent results of the model (2) – (9) using exact approach described in [5]. Columns denoted as *GH_sz1* represent solution for *GH_size* with criterion (10), columns denoted as *GH_sz2* for the *GH_size* with criterion (11). Columns denoted as *GH_pr1* represents results *GH_price* with formula (12) and columns denoted as *GH_pr2* for the *GH_price* using formula (13). For the exact solution we denoted as *F** best values of objective function for average deviation calculated for given parameter p_{max} in previous research. Solutions of heuristic approaches are represented by the gap in % between optimal solution obtained by exact method and the best solution obtained by given heuristic approach.

Table 1: Best values of average deviation – networks with 25 (a) and 51 (b) stations

a)

p_{max}	Exact	GH_sz1	GH_sz2	GH_pr1	GH_pr2
	F*	Gap [%]	Gap [%]	Gap [%]	Gap [%]
4	4276	2.4	1.9	3.3	1.7
6	4124	2.7	5.6	10.1	2.6
8	4095	2.6	5.6	5.4	2.1
10	4054	2.4	6.3	4.7	1.2
13	3990	2.2	9.5	8.3	1.7
16	3999	2.2	9.2	7.1	1.4
20	4031	3.1	8.1	8.1	1.6

b)

p_{max}	Exact	GH_sz1	GH_sz2	GH_pr1	GH_pr2
	F*	Gap [%]	Gap [%]	Gap [%]	Gap [%]
6	9656	6.2	10.1	10.5	1
10	9230	3	3.2	10.3	2.3
13	8455	5.3	3.6	15.3	2.6
16	8266	4.5	7.8	15	2.2
20	7860	4.8	5.2	13.9	1.3
25	7991	7.1	10.7	9.1	3.1
30	8383	4.5	2.2	7.1	0.6

Table 2: Computation time in seconds – networks with 25 (a) and 51 (b) stations

a)

p_{max}	Exact	GH_sz1	GH_sz2	GH_pr1	GH_pr2
4	19.2	1.6	1.7	21.2	20.9
6	17.9	1.6	1.7	18	17.9
8	19.2	1.6	1.6	16.3	16.9
10	21.9	1.6	1.7	14.5	15
13	18.9	1.7	1.7	12	12.7
16	18.7	1.7	1.7	10.9	10.4
20	17.6	1.6	1.7	8.5	8.6

b)

p_{max}	Exact	GH_sz1	GH_sz2	GH_pr1	GH_pr2
6	145.8	27.4	24.1	145.8	146
10	136.7	27.2	24	135	134.9
13	137.8	27.2	24.1	130.7	130
16	150.3	27.7	24	120.5	119.9
20	140.7	27	24.4	110.1	110.9
25	210	26.9	24.3	95.6	96.5
30	164	27.1	24.3	82.7	81.9

5 CONCLUSION

In the paper we proposed two variants of greedy based heuristic method to solve the zone partitioning problem in zone tariff system and we performed numerical experiments on the data sets created from the public transport network in Zvolen Region in Slovakia.

From the results of the numerical experiments in the Table 1 and Table 2 we can see, that the best solution of proposed heuristic approaches we can obtain using greedy heuristic algorithm with price criterion *GH_price* and formula (13). Comparing computation time we find that computational time in the case of *GH_price* approach increases with the decreasing of p_{max} value. This is caused by the fact, that zone partition calculation in this case depends on the prices and we cannot calculate zone partition in the first step as in *GH_size*, but we need to calculate zone partitioning for all the parameters settings separately.

In the future we would like to verify this approach on the different networks with real OD matrix. Numerical experiments will be also extended to the larger test network to study behavior of designed heuristics on larger problems, because from the numerical experiments it is obvious, that computation time of *GH_price* with best obtained heuristic solutions approaches computation time of exact method for instances with smaller values of parameter p_{max}.

Acknowledgement

This work was supported by the research grants VEGA 1/0339/13 "Advanced microscopic modeling and complex data sources for designing spatially large public service systems" and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks". We would like to thank to "Centre of excellence for systems and services of intelligent transport" (ITMS 26220120050) for built-up infrastructure, which was used.

References

[1] Babel, L. and Kellerer, H. (2001). Design of tariff zones in public transportation systems: Theoretical and practical results. Technical report. University of Graz, Austria.

[2] Cenek, P., Janáček, J. and Jánošíková, Ľ. (2002). Location of Transportation Districts at Modeling of Tranportation Processes in a Region. Communications - Scientific Letters of the University of Zilina., Vol. 4: 1-2.

[3] Fan, N., Pardalos, P. M. (2010): Linear and quadratic programming approaches for the general graph partitioning problem, Journal of global optimization. Vol. 48(1): 57-71.

[4] Hamacher, H. W. and Schöbel, A. (2004). Design of Zone Tariff Systems in Public Transportation. Operations Research. Vol. 2 (6): 897-908.

[5] Kohání, M. (2013a). Designing of zone tariff in integrated transport systems. Communications: scientific letters of the University of Žilina. Vol. 15 (1): 29-33.

[6] Kohání, M. (2013b). Zone partitioning problem with given prices and number of zones in counting zones tariff system. SOR ´13: proceedings of the 12th International Symposium on Operational Research: Dolenjske Toplice, Slovenia., Vol. 13: 75-80.

[7] Kohání, M. (2014a). Heuristic Approach to the Zone Partitioning Problem in Counting Zones Tariff System. MME: proceedings of the 32th international conference. 2014, Vol. 32: 449-454.

[8] Kohání, M. (2014b). Greedy metóda na riešenie úlohy rozdelenia oblasti do tarifných zón (in Slovak). Úlohy diskrétní optimalizace v dopravní praxi 2014: SW podpora rozhodování v inteligentních dopravních systémech. Univerzita Pardubice: 51-62.

[9] Palúch, S. (2013). On a fair fare rating on a bus line. Communications: scientific letters of the University of Žilina. Vol. 15 (1): 25-28.

[10] Schöbel, A. (2006). Optimization in Public Transportation: Stop Location, Delay Management and Tariff Zone Design in a Public Transportation Network, Springer, 2006.

[11] XPRESS-Mosel User guide. Fair Isaac Corporation, Birmingham, 2012.

[12] BID (2015). Bratislavská integrovaná doprava (Bratislava Integrated Transport Office). http://www.bid.sk [Accessed 5/5/2015]

MEASURING ENVIRONMENTAL EFFICIENCY OF EUROPEAN REGIONS: A META-FRONTIER APPROACH

Lukáš Melecký
VŠB-Technical University of Ostrava, Faculty of Economics, Department of European Integration
Sokolská třda 33, 701 21 Ostrava, Czech Republic
lukas.melecky@vsb.cz.cz

Jana Hančlová
VŠB-Technical University of Ostrava, Faculty of Economics, Department of Science Engineering
Sokolská třda 33, 701 21 Ostrava, Czech Republic
jana.hanclova@vsb.cz

Abstract: This paper presents an analysis framework to evaluate the effects of technology heterogeneities and undesirable outputs on environmental efficiency measurement. The proposed framework combines directional distance function and a meta-frontier analysis. Firstly, production units are divided into four groups using the factors of regional competitiveness based on Regional Competitiveness Index approach. In the second step the meta-frontier is set down. The obtained results indicate a significant improvement of meta-technology ratio holds within the European regions and support the evidence that the traditional differences of technological frontiers formation are more significant in comparison with group frontiers constitution.

Keywords: DEA, Directional Distance Function, Efficiency, Meta-frontier, NUTS 2 region, MTR.

1 INTRODUCTION

Environmental issues have become one of the most important problems related with social and economic sustainable development. Modern comprehensive evaluation of environmental efficiency begins with the problem of the limitation of energy and the carbon dioxide emissions caused in the process of energy production in 1980s. Evaluating environmental efficiency in different regions and sectors can have strong practical implications and therefore a comprehensive environmental efficiency measurement must be appropriately developed and computed. Regional-scale environmental assessments require integrating data sets from a variety of sources collected for diverse purposes and having inconsistent spatial or temporal scales. Moreover, the environmental processes and the relationships among variables in the assessment tend to be poorly understood [9]. Regional assessments often use multivariate statistics to describe the relationships between these variables, but multivariate analyses frequently reduce data dimensionality; see e.g. [16] or [15]. However, the current evaluation methods for environmental efficiency are mostly based on the determination of the inputs and outputs evaluation index. Some evaluation methods (such as the Data Envelopment Analysis or Stochastic Frontier Analysis) are proposed based mainly on the macro data or micro data.

The paper is focused on efficiency evaluation of the EU28 NUTS 2 regions by selected *regional data* included in one composite indicator - Regional Competitiveness Index 2013 (RCI 2013). This synthetic indicator has been finally performed by [1]. The roots of the RCI 2013 lay down in the most known competitiveness indicator, the Global Competitiveness Index reported by the World Economic Forum; see [17]. RCI 2013 is based on a set of 80 candidate indicators of which 73 have been included in the index. In the paper, we understand measuring environmental efficiency as measuring of production environment efficiency just using selected socio-economic indicators based on competitiveness approach and included undesirable outputs in form of negative impacts on health and long-term unemployment. Efficiency of the production units might be evaluated by the parametric and non-parametric approach. The classic non-parametric approach used in the analysis of data

set is defined by the certain combination of outputs and inputs. Unlike the second one – parametric approach evaluates the efficiency through the estimated parameters defined production function in advance. Our paper is focused on non-parametric approach to the efficiency evaluation by the production multiple desirable and undesirable economic, social, environmental and infrastructural outputs based on the RCI 2013 approach. The production function makes no assumptions about functional relationship among variables but by quantifying the extent to which desirable outputs can extend desirable outputs and contracted inputs and undesirable outputs. Furthermore, there is distinguished from the production range – the production function with the constant return to scale (CRS) or variable return to scale (VRS) revenue from the range of production function. The paper follows CRS production function. For measurement of the decision making unit (DMU) distance from the efficiency frontier the directional output distance function that takes into account the presence heterogeneity is used.

2 METHODOLOGY

Directional distance function (DDF) in empirical analysis can be calculated in several forms and approaches: [6] specify the DDF as a quadratic distance function and employ linear programming (LP), [3] employ the non-parametric approach using Data envelopment analysis (DEA) - type of LP, [11] provides modified DDF to define and decompose the meta-frontier Malmquist-Luenberger productivity index, [18] used DDF for measuring the environmental efficiency and constructing the environmental efficiency index and [7] explains DDF in primal and dual spaces and appendixes the parametric and non-parametric form of DDF. There are two main differences between the above mentioned approaches. The first approach can easily calculate shadow prices however it requires the assumption on the functional form of the directional distance function and imposes a lot of restrictions on the parameters. The latter approach does not require any functional form of directional distance function nor does it place any restrictions on the parameters. We employ the latter approach in our paper. Assuming for each production unit a productive process using an input vector $\boldsymbol{x} \in \boldsymbol{R}_{+}^{H}$, to obtain a set of desirable outputs denoted by the vector $\boldsymbol{y} \in \boldsymbol{R}_{+}^{N}$ and a vector of undesirable outputs $\boldsymbol{b} \in \boldsymbol{R}_{+}^{F}$ through a production technology given by following equation:

$$P = \{(\mathbf{x},\mathbf{y},\mathbf{b}) : \textit{can produce}(\mathbf{y},\mathbf{b})\}. \quad (1)$$

The described technology (1) indicates all technological feasible relationships between inputs and outputs. A number of assumptions are required in the form of axioms on the output side; see e.g. [14], [3] or [11]. We model the idea that reduction of bads may be costly to be diverted to reduce undesirable outputs (weak disposability). This condition allows for the reduction of the undesirable outputs only when accompanied by the simultaneous reduction of the desirable outputs. In words, this states that a reduction in bads is feasible only if goods are simultaneously reduced, given a fixed level of inputs. In addition, we assume that the good or desirable outputs are freely disposable. In words, says that the good outputs are "null-joint" with the bad outputs if the only way to produce no bads is by producing zero good outputs. Alternatively, this means that if a good output is produced in a positive amount some bad output must also be produced. These conditions will be incorporated into our computational model discussed further. We also consider group heterogeneity in production activities and we assume that technology of one group is different from those of other groups and there are G different groups ($g=1,\ldots,G$) in the whole sample. [3] provide the basis to represent the joint production of good and bad outputs by extending the Shepherd's output distance function [13] to the directional output distance function. This distance is in our case defined for *o*-unit in the *g* group as follows:

$$\vec{\boldsymbol{D}}_o^g\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{b};\boldsymbol{g}\right)=\max\left\{\boldsymbol{\beta}^g:\left(\boldsymbol{x},\boldsymbol{y}+\boldsymbol{\beta}^g\boldsymbol{g}_y,\boldsymbol{b}-\boldsymbol{\beta}^g\boldsymbol{g}_b\right)\in P\right\}, \tag{2}$$

where $\boldsymbol{g}=(\boldsymbol{g}_y,\boldsymbol{g}_b)=(\boldsymbol{y},\boldsymbol{b})$ is direction vector. This distance function searches for maximum attainable expansion of desirable outputs in the $\boldsymbol{g}_y$ direction and the largest feasible contraction of undesirable outputs in the $-\boldsymbol{g}_b$ direction, which is *negative* and therefore as consequences of the reduction of undesirable production output. We also replace $\vec{\boldsymbol{D}}_o^g\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{b};\boldsymbol{g}\right)=\vec{\boldsymbol{D}}_o^g\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{b}\right)$ in the remainder of this paper. The directional distance function of the *o*-th DMU in the *g*-th group is represented by β_o^g reported in [12].

We assume that we have sample of $k=1,2,...,K$ production units, a vector of $h=1, 2, ..., H$ inputs to obtain a vector of $m=1, 2, ..., M$ desirable outputs and $f=1, 2, ..., F$ undesirables outputs. Assuming that the production process satisfies a constant return to scale (CRS) and strong disposability of all inputs and outputs, the output set can be under the above mentioned assumptions and using the directional output distance function is set by optimization model:

$$\vec{\boldsymbol{D}}^g\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{b}\right)=\max_{\beta^g\geq 0}\beta^g \tag{3}$$

$$\text{s. t.}\sum_{k=1}^{K}x_{hk}\lambda_k\leq x_{ho}\quad h=1,2,...,H \tag{4}$$

$$\sum_{k=1}^{K}y_{mk}\lambda_k\geq(1+\beta^g)y_{mo}\quad m=1,2,...,M \tag{5}$$

$$\sum_{k=1}^{K}b_{fk}\lambda_k=(1-\beta^g)b_{fo}\quad f=1,2,...,F \tag{6}$$

$$\lambda_k\geq 0\quad k=1,2,...,K, \tag{7}$$

where $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,...,\lambda_K)$ is a vector of intensity variables. The first constrain (4) ensures that resultant input is no more than what is actually being used. The second constrains (5) ensure that the resultant desirable output is no lower than what is actually being produced. The third constrain (6) expresses that the resultant undesirable output could be higher than what is actually being produced. Weak disposability and null-jointness hypotheses are imposed in equation (6) through the strict equality of undesirable outputs constrain. Symbol β^g represents the *technical inefficiency* denoted by distance from the observation point to the production frontier in *g*-th group. Obviously, β^g takes value of zero for a technically efficient production unit on the frontier, whereas a positive value implies an inefficient production unit below the frontier. The higher the value, the more inefficient the production unit is set.

Within using the DDF for the whole group of regions, there is supposed, that all regional producers possess the same level of production technology, which is unreal, and confirmed in the study [12] or [18]. Study [10] proposed and applied the *meta-frontier concept* in estimation of DEA efficiency. They estimated a meta-frontier through the use of overall samples and also they divided the DMUs into the groups and estimated group frontiers of group samples. These groups should capture better the heterogeneity of the production processes with regard to the technologies or variable production environment of the input resources or desirable and also undesirable outputs. All EU28 regions are divided into *1, 2, ..., G* groups and for each group is defined $K_1,K_2, ..., K_g, ..., K_G$ regions and there is valid, that their sum is equalled to the number of units in meta group, i.e. *K*. Further there will be distinguish for each *o-th* DMU in one time period of meta-efficiency $MEE_o=1-\beta_o^m$ and group-efficiency $GEE_o=1-\beta_o^g$, where parameters β_o^m or β_o^g are obtained through optimization model in equations (3) to (7), which is analysed for the whole sample or particular group *g*. Because the meta-frontier is enveloped in the *G* group-specific frontiers, the efficiency

measured on the basis of the meta-frontier is less than that of the group frontiers i.e. $MEE_o \le GEE_o$. We can also measure the ratio between the efficiency of the meta-frontier (MEE) and group frontier (GEE), (see e.g. [2]) using Meta-technology ratio (MTR) given by (8):

$$0 < MTR_o = \frac{MME_o}{GEE_o} \le 1. \tag{8}$$

The closer MTR_o is to *1*, the smaller is technology heterogeneity, which means that the efficiency of that group's frontiers is closer to the meta-frontier. The MTR_o evaluated based on the meta-frontier and group-frontier indicates that DMU_o in group-specific frontiers are only distinguished by technology heterogeneity between the two frontiers, so the source of the inefficiency of the meta-frontier inefficiency cannot be identified.

3 DATA AND EMPIRICAL RESULTS

The data set for empirical analysis was collected from regional statistics database of Eurostat [5] and from annexes of RCI 2013 draft report [4]. All values of indicators have been normalized by z-scores; for more details see [1]. The whole tested sample includes 258 NUTS 2 regions instead of 276 in the whole EU28 countries under current NUTS 2013 classification because of the unavailability of data for 18 NUTS 2 regions placed in Germany, Spain, France, Greece, Portugal, Romania and Croatia. All selected RCI 2013 indicators were collected between years 2006 and 2011 and values have been normalized by z-scores. The distribution of RCI 2013 scores between 1.5 (highest level) and (-1.5) (lowest level). Zero value of RCI 2013 shows the average value of regional competitiveness. For the purpose of evaluation of meta-efficiency and group-efficiency of EU28 NUTS 2 regions based on concept of RCI 2013 – 10 indicators of inputs were selected (x_1, …, x_{10}), 4 desirable indicators of outputs (y_1, …,y_4) and 3 undesirable indicators of outputs (b_1, b_2 and b_3) has been chosen. Selected dataset of RCI 2013 indicators include 17 indicators placed in 11 pillars of RCI 2013 (see Tab. 1).

Table 1: Selected Inputs (*x*), Desirable (*y*) and Undesirable (*b*) Outputs of selected EU28 NUTS 2 Regions

x/y/b	***RCI 2013 Pillar***	***RCI 2013 Indicator***	***Unit***	***Mean***	***Standard Deviation***
x_1	Institutions	Corruption	Index	5.255	0.997
x_2		Rule of Law	Index	5.228	0.937
x_3	Infrastructure	Motorway potential accessibility	Index	4.897	0.990
x_4		Railway potential accessibility	Index	4.846	0.992
x_5		Number of passenger flights	Number	4.834	0.956
x_6	Higher education	Accessibility to universities	%	4.886	1.124
x_7	Labour market efficiency	Labour productivity	GDP/person employed	4.974	1.002
x_8	Technological readiness	Households access to broadband	%	5.071	1.006
x_9	Innovation	Knowledge workers	%	4.940	0.967
x_{10}		Total intramural R&D expenditure	%	4.890	1.051
y_1	Market size	Potential GDP in PPS	GDP (PPS)	4.815	0.932
y_2	Business sophistication	GVA (K-N sectors)	%	4.780	0.964
y_3	Innovation	Total patent applications	Number	4.944	0.954
y_4	Health	Healthy life expectancy	Number	5.046	0.985
b_1	Health	Road fatalities	Number	4.860	1.105
b_2		Cancer disease death rate	%	4.994	1.053
b_3	Labour market efficiency	Long-term unemployment	%	5.169	0.987

All selected regions were firstly ranked in ascending order and divided into *4 groups* based on the scores of RCI 2013. Zero value of RCI 2013 shows the average value of selected 258

NUTS 2 regions. The new formed groups of selected NUTS 2 regions can be described as follows:

1. Group 1 includes the regions with the minus RCI 2013 value under the first quartile (N_1=69 regions). Structure of NUTS 2 regions within first group is represented mostly by regions from 'new' EU Member States (EU13) plus regions of Portugal and Greece.
2. Group 2 contains the regions with below-average regional competitiveness compared to the RCI 2013 index scores between the first and second quartile (median) (N_2= 35 regions). This group is presented only by regions from 'old' EU Member States (EU15) and most of Italian and Spanish regions are included here.
3. Group 3 introduces the regions with above-average RCI 2013 scores compared to the EU28 and RCI 2013 scores between the second and the third quartile (N_3=72 regions). The most variable structure of NUTS 2 regions from EU15 and EU13 Member States is presented in this group, especially in the regions of Austria, Cyprus, Germany, Finland, France, Ireland, Spain, Sweden and United Kingdom.
4. The majority of the competitive regions from EU15 Member States is placed in the Group 4 and the level of RCI 2013 is above the third quartile of RCI 2013 scores (N_4 = 82). There are NUTS 2 regions from Belgium, France, Germany, Netherland and United Kingdom.

The efficiency of the meta-frontier and group-frontiers are estimated using optimization model for 258 NUTS 2 regions and further for each group *g=(1,…,4)* using free version of *R software* (version 3.2.0) with package 'direc.dea {nonparaeff}'solving the DDF with undesirable outputs; see [3]. Table 2 summarizes the amounts of efficient NUTS 2 regions by the meta-environmental efficiency (MEE) and group environmental efficiency (GEE). For the whole sample the share of the efficient regions reached almost 70%. In the second group there are all NUTS 2 regions defined as efficient. In the other groups, the share of efficiency units is in the range of 91.3 to 98.6%, which is in compliance with other comparative empirical studies focused on EU Member States or regions, e.g. [11], [2] or [8].

Table 2: Meta and Group Frontiers Efficiencies of selected EU28 NUTS 2 Regions

	Number of DMUs	***Technical Efficiency***	
		Number	***%***
Meta	258	179	69.4
Group 1	69	63	91.3
Group 2	35	35	100.0
Group 3	72	71	98.6
Group 4	82	77	93.9

Table 3 completes further descriptive statistics of the efficiency indices for analysed group samples and confirms the following relation: meta-environmental efficiency is lesser or equal to the group-environmental efficiency: $MEE^{g}_{average} \leq GEE^{g}_{average}$ .The lower average group-efficiency was indicated in the first group of the regions and there were identified the highest standard deviations. On the contrary, in the second group the efficiency was almost unitary with the lowest variability inside the group. A share of the efficient units in the fourth group of the regions was in the amount of 0.998% and variability of the efficiency in the group was lower compared to the second group. The differences between the efficiencies of the meta-frontier for various groups were examined using non-parametric statistic. We applied the *Kruskal-Wallis test* to examine the technology frontier differences between all groups of regions using IBM SPSS Statistics 22. The results shows that the Kruskal-Wallis value is 6.133 and we reject the null hypothesis that the distribution of MEE is the same across group

samples at 5% level of significance (sig.= 0.105). This indicates that the four group-samples are formed different populations and that technology heterogeneity exists between them.

Table 3: Descriptive Statistics of the Meta-frontier and Group-frontiers Efficiencies

Descriptive statistics	***Group***	$MEE^{g}_{average}$	$GEE^{g}_{average}$	$MTR^{g}_{average}$
Mean	1	0.9940	0.9980	0.9959
	2	0.9969	1.0000	0.9969
	3	0.9913	0.9998	0.9915
	4	0.9888	0.9992	0.9896
Standard deviation	1	0.0150	0.0078	0.0100
	2	0.0074	0.0000	0.0074
	3	0.0162	0.0018	0.0158
	4	0.0195	0.0035	0.0186

Conclusions of the Kruskal-Wallis test also confirm the *MTR* indicator that is the lowest for the fourth group with the highest standard deviation. The lowest average level of the MTR is for the first group of analysed NUTS 2 regions, but variability is in the group higher compared to the second group. We can also briefly mention individual results of environmental efficiency of the meta-frontier MEE^{g}_{0} in different group-samples. Among the worst efficient NUTS 2 regions belong Groningen (NL11), Bremen (DE50), South Western Scotland (UKM3), Zachodniopomorskie (PL42) and Dolnośląskie (PL51), where the reduction of undesirable outputs and increasing desirable outputs should be at the minimal level of 6%. The following NUTS 2 regions Prov. Oost-Vlaanderen (BE23), Pomorskie (PL63), Prov. Namur (BE35), East Wales (UKL2), Lorraine (FR41) and Lietuva (LT00) should increase desirable outputs and decrease undesirable outputs minimally by 5%. Significant share between the environmental efficiency in relation to the meta-frontier (*MEE)* and group-frontier (*GEE*) for the above mentioned NUTS 2 regions demonstrates that there exist a higher technology heterogeneity of the production process with comparison to the group and meta-frontier.

4 CONCLUSION

Most of empirical studies that analysed the environmental efficiency of the production process of territories are devoted to the relation between CO_2 emissions and economic activity. Our paper deals with the evaluation of the efficiency of the production process of 258 NUTS 2 regional units in the EU28 countries. Production process comes out of the concept of the input and output indicators designed by the RCI 2013 approach focused on evaluation of regional competitiveness. As benefit of the paper we find an inclusion of selected undesirable outputs in a form of negative impacts on health (e.g. car accidents linked with the increasing traffic load, lifestyle diseases as a cancer following increasing stress situation in work) and long-term unemployment. Preliminary analysis of the relation between meta-frontier and group-frontier could distinguish a variable heterogeneity of the production process in the groups of NUTS 2 regions. Within the analysis of meta-frontier it was identified, that 69.4% of the regional production units was efficient. Average level of Meta-environmental efficiency was higher (0.997) in the second group of EU15 regions with the below-average level of RCI 2013 compared to the EU28 average and also variability of the efficiency was the lowest compared to the other groups of regions. On the other side, the lowest average meta-efficiency (0.989) was indicated by the fourth group of regions in the EU15 countries with the above-average of RCI 2013 scores in the EU28, but with the highest efficiency variability in the group. Average level of the Meta-technology ratio in the groups

ranged from 0.9896–0.9969 and therefore it seems heterogeneity between groups is low, but the Kruskal-Wallis has indicated that four group-samples are formed different populations and that technology heterogeneity exists between them at 5% level of significance. The preliminary results of the paper include a wide range of assumptions that should be analysed *further*. For instance: a creation of wider database of indicators, classification of regions into the groups, conditions of the constant revenues from the range of production process, the same intensity of the decline in undesirable and increase in desirable outputs, free desirable outputs and possibility of reduction the undesirable outputs.

Acknowledgement
The paper is supported by SGS project (SP2015/106) of Faculty of Economics, VŠB-TUO and Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296.

References

[1] Annoni, P. and Dijkstra, L. (2013). EU Regional Competitiveness Index 2013. Luxembourg: Publication Office of the European Union.
[2] Chiu, Ch. et al. (2012). Decomposition of the environmental inefficiency of the meta-frontier with undesirable output. Energy Economics, 34, pp. 1392-1399.
[3] Chung, Y. et al. (1997). Productivity and undesirable outputs: A directional distance function approach. Journal of Environmental Management 51(3):229–240.
[4] European Commission. (2013). Regional Policy Inforegio. Draft EU Regional Competitiveness Index: RCI 2013. Ec.europa.eu/regional_policy [19/06/2015]. http://ec.europa.eu/regional_policy /cs/information/publications/studies/2013/draft-eu-regional-competitiveness-index-rci-2013.
[5] Eurostat. (2015). Data. Regional Statistics by NUTS classification. Epp.eurostat.ec.europa.eu [12/04/2015]. http://ec.europa.eu/eurostat/data/database.
[6] Färe, R. et al. (2006). Productivity growth and convergence in the European Union. Journal of Productivity Analysis, 25: 111–141.
[7] Färe, R. et al. (2015). Distance Functions in Primal and Dual Spaces. In Zhu, J. et al. (Eds.). Data Envelopment Analysis. A Handbook of Models and Methods (pp. 12-32). New York: Springer.
[8] Foddi, M. and Usai, S. (2013). Technological Catching Up among European Regions. Lessons from Data Envelopment Analysis. WP4/02 Search Working Paper.
[9] Macpherson, A. et al. (2010). A Directional distance function approach to regional environmental-economic assessments. Ecological Economics, 69, pp. 1918–1925.
[10] O'Donell, C.J. et al. (2008). Metafrontier Frameworks for Study of Firm-level Efficiencies and Technology ratios. Empirical Economics, 34, pp. 231–255.
[11] Oh, D. (2010). A Metafrontier Approach for Measuring an Environmentally Sensitive Productivity Growth Index. Energy Economics, 32, pp. 146–157.
[12] Picazo-Tadeo, et al. (2005). Directional Distance Functions and Environmental Regulation. Energy Economics, 27, pp. 131–142.
[13] Shephard, R.W. (1970). Theory of Cost and Production Functions, Princeton University Press, Princeton.
[14] Shephard, R.W. and Färe, R. (1974). The Law of Diminishing Returns. Zeitschrift für Nationalökonomie, 34, pp. 69-90.
[15] Staníčková, M. (2014). Competitiveness Factors in Regional Growth: the Case of Central and Eastern European Countries. In Honová, I. et al. (Eds.). Proceedings of the 2nd International Conference on European Integration 2014 (pp. 634– 643). Ostrava: VŠB-TUO.
[16] Tran, L.T. et al. (2006). A generalized distance measure for integrating multiple environmental assessment indicators. Landscape Ecology, 21(4): 469–476.
[17] World Economic Forum. (2013).The Global Competitiveness Report 2013–2014. Lausanne: WEF.
[18] Yuan, P. et al. (2013). Measuring the environmental efficiency of the Chinese industrial sector: A directional distance function approach. Mathematical and Computer Modelling, 58: 936–947.

STAR COLORING OF FIBONACCI CUBES

Aleksander Vesel
Faculty of Natural Sciences and Mathematics,
University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia

Abstract: The $L(1,1)$-labeling of a graph is an abstraction of assigning integer frequencies to radio transceivers such that transceivers that are one or two units of distance apart receive frequencies that differ by at least one. The least span of frequencies in such a labeling is referred to as the $\lambda_{1,1}$-number of the graph. Similar labeling problem, known as star-coloring problem, has been also studied in the context of the problem of data distribution in parallel memory systems. The hypercube of order d, denoted by Q_d, is the graph $G = (V, E)$ where the vertex set $V(G)$ is the set of all binary strings $u = u_1u_2 \dots u_d$, $u_i \in \{0, 1\}$, and two vertices $x, y \in V(G)$ are adjacent in Q_d if and only if x and y differ in precisely one place. A Fibonacci string of length d is a binary string $u = u_1u_2 \dots u_d$, $u_i \in \{0, 1\}$, with $u_i \cdot u_{i+1} = 0$ for $i \in \{0, 1, ..., , d-1\}$. The Fibonacci cube Γ_d is the subgraph of Q_d induced by the Fibonacci strings of length d. In this paper, an upper bound for the $\lambda_{1,1}$-number of Fibonacci cubes is established.

Keywords: Fibonacci cube, $L(1,1)$-labeling, star labeling

1 INTRODUCTION

Hypercube is a popular interconnection scheme for multicomputers. Routing in a hypercube is simple by using the labels of vertices in the hypercube, i.e. a shortest path from a node x to a node y, where I is the set of bit positions in which x and y differ, can be found easily by flipping bits of x with positions I in some arbitrary order. The Fibonacci cube is a communication network that possesses many suitable properties which are important in network design and application. Its major advantage is that it uses fewer links than the comparable hypercube, while its size does not increase as fast as the hypercube's. In other words, they allow more alternatives to build networks of various sizes. Note also that the Fibonacci cube can emulate many hypercube algorithms. Moreover, they emulate other topologies, such as trees, rings and meshes very efficiently and can therefore find applications in fault-tolerant computing.

Efficient assignment of frequencies for wireless communication in different network scenarios has become an extremely important topic of recent research. The main reason is the enormous growth of wireless networks which causes the issue of efficiently using the radio spectrum - an expensive and scarce resource - very important. The wide class of frequency assignment problems (FAPs) asks for assigning frequencies (channels) from the available radio spectrum to the transceivers of the network, so that unconstrained simultaneous transmissions cannot cause interference. The objective is to minimize the used radio spectrum. While problems of these type arose some decades ago in radio and TV broadcasting, they today play a growingly significant role due to the wide diffusion of telephone and satellite communication. Since the problems are of vital importance, they have been largely explored in the literature [3, 7, 11].

In a broadcasting network, each transceiver is assigned a frequency channel for its transmissions. Two transmissions can interfere if their channels are too close. That means that even if two transceivers use different channels, there still may be interference if the two transceivers are located close to each other.

For the interference graph of the network such kind of problem can be seen as a variation of a vertex coloring problem [1, 4]. More formally, for a given graph $G = (V, E)$ modeling the wireless network, where vertices represent transceivers and edges between two transceivers represent possible interference, we consider a function from the set of vertices of G to the set of nonnegative integers such that the given separation constraints are satisfied.

Among the variants of the frequency assignment problem, the $L(p,q)$-labeling problem where adjacent vertices must be assigned colors of distance at least p apart and vertices of distance two must be assigned colors at least q apart has attracted considerable interest. In this paper, we are interested in $L(1,1)$-labeling problem, i.e. vertices of distance at most two must be assigned different colors.

Similar labeling problems have been also studied in the context of the problem of data distribution in parallel memory systems such that a high degree of data-parallelism is achieved. The goal is to store data items in such a manner that they can be accessed with zero conflicts assuming that the multiprocessor architecture is able to request a number of data items from the memory subsystem.

We define the *data distribution problem* as follows. Let G be a host graph representing the data structure corresponding to an application and H be a subgraph of G. Copies of H in G, called the *templates*, describe memory modules which are to be accessed together. The data distribution problem asks for an efficient algorithm for mapping the data items onto memory modules such that the number of modules for accessing the specified template is minimized when conflict-freeness must be guaranteed.

It is straightforward to see that minimizing the conflicts for arbitrary templates is computationally hard, since this problem corresponds to a variant of a graph coloring where a color represents a memory module while nodes in a template instance must have different colors to guarantee conflict-freeness.

A natural choice for data distribution schemes of conflict-free access is the *star* template which consists of a vertex of a host graph G together with all of its neighbors. The goal is to color the vertices of the host data structure in such a way that vertices in any arbitrary star are assigned different colors. The data distribution problem for the star template in a host graph G is called the *star-coloring problem*. It is not difficult to see that the star-coloring problem for a graph G corresponds to a $L(1,1)$-labeling of G.

Star templates appear in many classical algorithms. For instance, graph algorithms based on breadth-first search explore in each iteration a vertex's immediate neighbors yet to be visited.

These types of problems have been studied for various types of host graphs: circular lists, trees, tori and hypercubes (see [2] for the details).

It is also well known, that a $L(1,1)$-labeling of a graph G is equivalent to the vertex coloring of the square of G ([6, 8]).

2 PRELIMINARIES

The *hypercube* of order d, denoted by Q_d, is the graph $G = (V, E)$, where the vertex set $V(G)$ is the set of all binary strings $u = u_1u_2 \dots u_d$, $u_i \in \{0,1\}$, and two vertices $x, y \in V(G)$ are adjacent in Q_d if and only if x and y differ in precisely one place.

We will use $[n]$ for the set $\{1, 2, \dots, n\}$ in this paper.

The *Fibonacci numbers* form a sequence of non-negative integers F_n, where $F_0 = 0$, $F_1 = 1$ and for $n \geq 0$ satisfy the recurrence $F_{n+2} = F_{n+1} + F_n$.

A *Fibonacci string* of length d is a binary string $u = u_1u_2 \dots u_d$, $u_i \in \{0,1\}$, with $u_i \cdot u_{i+1} = 0$ for $i \in [d-1]$. In other words, a Fibonacci string is a binary string without two consecutive ones. The *Fibonacci cube* Γ_d is the subgraph of Q_d induced by the Fibonacci strings of length d. The name is obtained from the appealing Zeckendorf's theorem which asserts that any positive integer can be uniquely written as the sum of nonconsecutive Fibonacci numbers. Hence Fibonacci strings are representations of integers in this number system. It is well known that $|V(\Gamma_d)| = F_{d+2}$. The Fibonacci cubes Γ_1, Γ_2, and Γ_3 are shown in Fig. 1.

For two vertices u and v in a graph G, we denote by $d(u,v)$ the distance between u and v. An $L(1,1)$-labeling of a graph G is a function f from the vertex set $V(G)$ to a set C of non-negative integers (called *labels*) such that

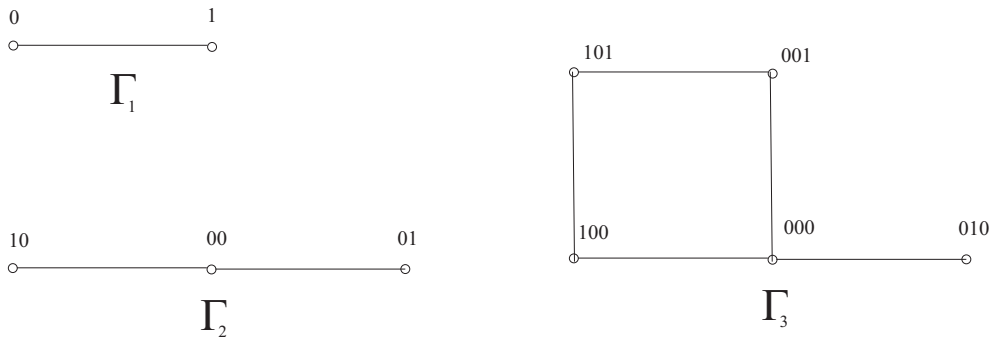


Figure 1: The Fibonacci cube Γ_1, Γ_2, and Γ_3

$f(u) \neq f(v)$ if $d(u, v) \leq 2$.

A k-$L(1,1)$-*labeling* is a $L(1,1)$-labeling of G such that $C = \{0, \ldots, k\}$. An optimal $L(1,1)$-labeling of G is a k-$L(1,1)$-labeling with smallest possible k. The largest label used by an optimal $L(1,1)$-labeling is called the $\lambda_{1,1}$-*number* of G and denoted by $\lambda_{1,1}(G)$.

Let $G = (V, E)$ be a graph. The square G^2 of a graph G is given by $V(G^2) := V(G)$ and $uv \in E(G^2)$ if and only if $1 \leq d_G(u, v) \leq 2$. It is straightforward to see that a $L(1,1)$-labeling of a graph G is equivalent to the vertex coloring of G^2.

3 BOUNDS ON THE STAR-COLORING

A lower bound for the star-coloring of a Fibonacci cube is given by the following lemma.

Lemma 3.1 *[9] If $d \geq 1$, then $\lambda_{1,1}(\Gamma_d) \geq d$.*

The following lemma is also easy to obtain.

Lemma 3.2 *If H is a subgraph of G, then $\lambda_{1,1}(H) \leq \lambda_{1,1}(G)$.*

It is proved in [2], that $\lambda_{1,1}(Q_d) \leq 2d$. From Lemma 3.2 then it follows

Lemma 3.3 *[9] If $d \geq 1$, then $\Gamma_{1,1}(\Gamma_d) \leq 2d$.*

Note that it is confirmed in [9] by an exhaustive search that for $d \leq 10$ we have $\Gamma_{1,1}(\Gamma_d) = d$. It is therefore a natural question, whether a better upper bound on the star-coloring of Fibonacci cubes can be provided.

4 NEW PRESENTATION OF FIBONACCI CUBES

In order to provide a better upper bound on the star-coloring of Fibonacci cubes , we introduce the following presentation of this class of graphs.

Let $v = v_1 v_2 \ldots v_k$, where $v_i \in \{0, 1, 2\}$, $i \in [k]$, denote a *ternary string* of length k. As usual, $v_i \in \{0, 1, 2\}$ is called a *trit*. Let also $\mathcal{T}_k$ denote the set of all ternary strings of length k. Since two adjacent bits of a vertex of Γ_d cannot be both equal to one, the vertices of Γ_d can be represented as ternary strings of length $\lceil \frac{d}{2} \rceil$ in a natural way. More formally, we define a maping $t : V(\Gamma_d) \to \mathcal{T}_{\lceil \frac{d}{2} \rceil}$ such that for $u \in V(\Gamma_d)$ we have $t(u) := v_1 \ldots v_{\lceil \frac{d}{2} \rceil}$, where $v_i := 2u_{2i} + u_{2i-1}$ (note that $2u_{2i} + u_{2i-1} \in \{0, 1, 2\}$.) The mapping t is called the *ternary representation* of Γ_d. The ternary representation of Γ_5 is depicted in Fig. 2.

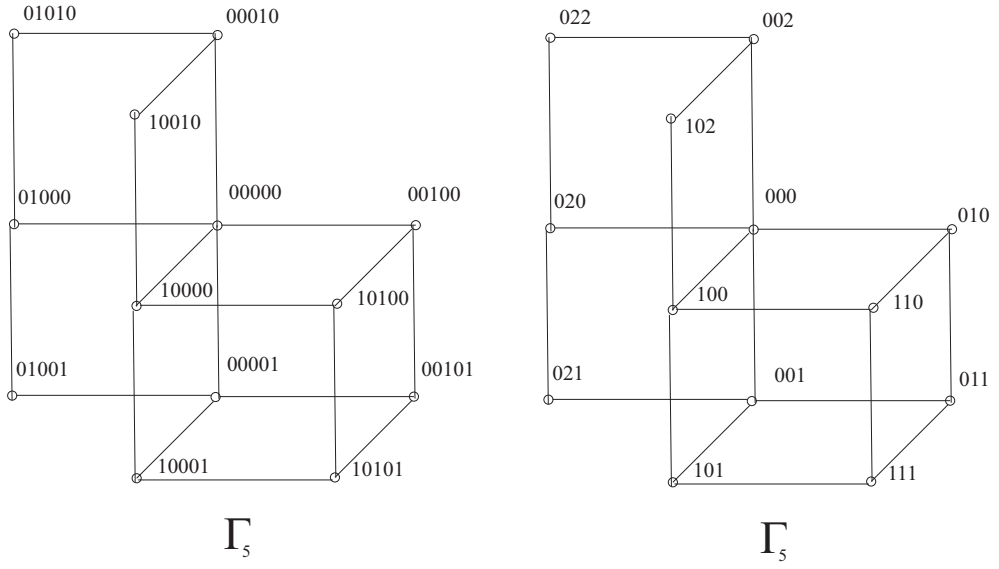


Figure 2: The Fibonacci cube Γ_5 and its ternary representation

5 ALGORITHM

Let $\oplus^3$ define a tritwise sum modulo 3 operator, i.e. $\oplus^3$ takes two ternary strings of equal length and performs the sum modulo 3 on each pair of corresponding trits. For instance, $210 \oplus^3 221 = 101$. Let also 0^h stands for the vertex with zero at all coordinates and $0^{h-i-1}10^i$ (resp. $0^{h-i-1}20^i$) for the vertex with one (resp. two) exactly at the i-th coordinate for $i \in [h]$.

Let $s_3(n)$ denote a ternary representation (i.e. a ternary string) of a nonnegative integer n. By a slight abuse of notation, a nonnegative integer n will also sometimes stand for $s_3(n)$ and vice-versa. In particular, for $i \in [n]$, the integer 3^{i-1} (resp. $2 \cdot 3^{i-1}$) will stand for $0^{h-i}10^{i-1}$ (resp. $0^{h-i}20^{i-1}$) and vice-versa in algorithm Fibonacci cube star-coloring. The algorithm for the Fibonacci cube Γ_d computes a star labeling of Γ_d. The algorithm is based on the ternary representation of Γ_d.

Algorithm 1: Fibonacci cube star-coloring(d)

```
Data: The dimension d of Γ_d.
Result: A star labeling c of Γ_d.
begin
    k := ⌈d/2⌉;
    t := the ternary representation of Γ_d;
    c(0^d) :=0;
    for i := 0 to k − 1 do
        for j := 1 to p do
            for all u ∈ V(Γ_d) with j3^i ≤ t(u) < (j + 1)3^i do
                c(t(u)) := (2i + j) ⊕^3 c(t(u) − j3^i)
            end
        end
    end
end
```

As an example, consider the running of the algorithm for Γ_5. We get:
$c(t(00000)) = c(000) = 0$,
$c(t(00001)) = c(001) = 1$,

$c(t(00010)) = c(002) = 2,$
$c(t(00100)) = c(010) = 3,$
$c(t(00101)) = c(011) = 4,$
$c(t(01000)) = c(020) = 4,$
$c(t(01001)) = c(021) = 5,$
$c(t(01010)) = c(022) = 3,$
$c(t(10000)) = c(100) = 5,$
$c(t(01001)) = c(101) = 3,$
$c(t(10100)) = c(110) = 4,$
$c(t(10101)) = c(111) = 2,$
$c(t(10010)) = c(102) = 6.$

We can prove [9] the following

Theorem 5.1 *Algorithm Fibonacci cube star-coloring computes in linear time a $L(1,1)$-labeling of Fibonacci cube Γ_d using $3^{\lfloor log_3 d \rfloor + 1}$ colors.*

References

[1] Barany, M., Tuza Z. (2014) Circular coloring of graphs via linear programming and tabu search. Cent. Eur. J. Oper. Res. To appear. DOI 10.1007/s10100-014-0345-8.

[2] Das, S. K., Finocchi, I., Petreschi R. (2006). Conflict-free star-access in parallel memory systems. J. Parallel Distrib. Comput. 66: 431 – 1441.

[3] Dupont, A., Linhares, A.C., Artigues, C., Feillet, D., Michelon, P., Vasquez, M. (2009). The dynamic frequency assignment problem. European J. Oper. Res. 195: 75–88.

[4] Govorčin, J., Gvozdenović, N., Povh, J. (2013). New heuristics for the vertex coloring problem based on semidefinite programming. Cent. Eur. J. Oper. Res. 21: 13–25.

[5] Hsu,W.-J. (1993). Fibonacci cubes - a new interconnection technology. IEEE Trans. Parallel Distrib. Syst. 4: 3–12.

[6] Shao, Z. and Vesel, A. (2013). A note on the chromatic number of the square of the Cartesian product of two cycles, Discrete Mathematics 313: 999-1001.

[7] Smith, D.H., Hurley, S., Thiel, S.U. (1998). Improving heuristics for the frequency assignment problem. European J. Oper. Res. 107: 76–86.

[8] Śopena, E., Wu, J. (2010), Coloring the square of the Cartesian product of two cycles. Discrete Math. 310: 2327–2333.

[9] Vesel, A. (2015). On the L(1,1)-labeling of Fibonacci cubes. Manuscript.

[10] Vesel, A. (2015). Linear recognition and embedding of Fibonacci cubes. Algorithmica 49: 81–93.

[11] Žerovnik, J. (1998). On the convergence of a randomized algorithm frequency assignment problem. Cent. Eur. J. Oper. Res. Econ. 6: 135–151.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 3:

Multiple Criteria Decision Making

GENERALISATION OF METHODS FOR CRITERIA WEIGHT DERIVATION BASED ON DISCORDANCE INFORMATION

Andrej Bregar
Informatika d.d., Vetrinjska ulica 2, 2000 Maribor
andrej.bregar@informatika.si

Abstract: In decision analysis, the importance of criteria may be determined either directly by compensatory criteria weights, or indirectly based on the selective characteristics of criteria. The paper discusses the correlation between the noncompensatory influence of veto and criteria importance weights. It builds on the methods for automatic derivation of weights from the veto related preferential information that is modelled in the special case of dichotomic sorting analysis on the basis of the outranking relation. It formally adapts and extends these methods to the general problematics of sorting and ranking, and to different types of preference models.

Keywords: Multiple criteria decision analysis, Criteria importance modelling, Outranking relation, Utility function, Ranking, Sorting

1 INTRODUCTION

One of key steps in the process of decision-making is the specification of criteria importance weights. Several studies have discovered that it is a difficult and time-consuming task [7, 9, 12, 14, 15], and have therefore introduced various structured techniques to facilitate the elicitation of weights. Most of these techniques aim primarily at reducing the cognitive load of the decision-maker, and at improving his ability of analysis, expression and structuring of information. Few, however, derive weights automatically according to other complementary preferential parameters that model the characteristics of a problem situation [11, 17].

Within the scope of our past research work, two methods have been defined that derive criteria weights from the discordance related preferential information, i.e. with regard to veto thresholds [4]. As these thresholds have a noncompensatory influence on the evaluation of alternatives, a single criterion veto can prevent the selection of any alternative independently of the assessment of other preferential parameters. The more alternatives that a criterion disqualifies as inappropriate choices by opposing a veto, the higher selective strength it has. In this way, it contributes to a greater extent to the final decision, and hence exhibits a higher relative importance.

The two proposed methods derive criteria weights in the forms of selective strengths and dominance indices, respectively. Their limitation is that they are designed for a specific decision-making problematic and a specific preference model, i.e. the dichotomic sorting of alternatives based on the concepts of pseudo-criterion and outranking relation [6]. Although this approach has several advantages [1, 2, 3], other decision-making problematics are often addressed, such as rank ordering of alternatives or sorting of alternatives into an arbitrary number of categories [18], and many different preference models and methods for multiple criteria decision analysis are used, especially from the domain of utility theory [8]. It is therefore the goal of this research paper to adapt the inference of criteria weights form the veto related preferential information (1.) to the general problematics of sorting and ranking, and (2.) to different types of preference models. An additional purpose of the presented study is to experimentally evaluate the efficiency of the derived criteria weights that are based on selective strengths and dominance indices. Due to limitations in the scope, empirical results of the evaluation study will be presented in a follow up paper.

The rest of the paper is organized as follows. Section 2 briefly describes fundamental concepts by discussing the correlation between veto thresholds and criteria weights, and by presenting the weight derivation mechanisms for the case of localized alternative sorting. In

Sections 3 to 5, the weight derivation mechanisms are generalised to different decision-making problematics and preference models. Finally, Section 6 concludes the paper with a resume and some directions for further work.

2 WEIGHT DERIVATION FOR THE CASE OF DICHOTOMIC SORTING

One of fundamental approaches to decision analysis is outranking [13]. It derives a partial, weak or complete rank order of alternatives by constructing outranking relations with regard to thresholds of pseudo-criteria. The preference and indifference thresholds determine to what compensatory extent an alternative outperforms another alternative. The veto threshold describes the noncompensatory conditions under which an alternative may not outperform a compared one regardless of any other threshold. If alternatives are sorted into preordered categories, criterion-wise thresholds are modeled in the neighbourhood of a reference profile that delimits categories [10]. In the case of dichotomic sorting, the set of alternatives is partitioned into two exclusive categories – the positive category of all acceptable alternatives and the negative category of unacceptabe alternatives [1, 4, 6]. The localization principle calls for the prevention of the incomparability relation. Otherwise, it cannot be clearly stated whether an alternative should be assigned to the positive or to the negative category. Only one discordance index $d_j(a_i)$ is hence defined for each criterion x_j.

Because the veto threshold characterizes the conditions for a single discordant criterion to prevent an outranking relation, and because it has a noncompensatory effect, it can, on its own, eliminate any alternative from the positive category. The more selective it is, the more the corresponding criterion contributes to the final decision. The importance of a criterion is therefore determined by both its weighting coefficient and its veto capacity.

As a result of past research work, two localized weight derivation mechanisms have been introduced: the selective strength based approach and the binary relation based approach [4]. Both construct a fuzzy veto relation by organizing partial discordance indices for the set of alternatives A and the set of criteria X:

$$V = \left(d_j(a_i)\right)_{i=1..m,\, j=1..n},\ a_i \in A, x_j \in X.$$

2.1 Selective strengths based approach

The approach consists of three required and three optional steps (steps 1 to 3, and steps 4 to 6, respectively):

1. All possible α-cuts of the fuzzy relation V are taken.
2. Partial selective strengths are calculated for each α-cut.
3. Partial selective strengths are joint with an algorithm or an equivalent operator.
4. The obtained complete selective strengths are transformed so that ratios of pairs of weights are reflected through a comparison matrix.
5. The decision-maker modifies the proposed ratios.
6. Weights are computed from the adjusted pairwise comparison matrix.

The partial selective strength indicates the degree to which the *j*-th criterion outperforms the weakest criterion. It is obtained according to the *i*-th alternative and the α_k cut-level:

$$\varphi_{ji}^{k} = \begin{cases} \mathrm{card}\left(x_l \in X \backslash \{x_j\} : d_l(a_i) < \alpha_k\right), & d_j(a_i) \geq \alpha_k \\ 0, & d_j(a_i) < \alpha_k \end{cases}.$$

The algorithm that aggregates partial selective strengths into complete selective strengths is based on several presumptions:

1. The contribution to total strengths is proportional to the α_k cut-level.

2. The criterion x_j gains the highest achievable selective strength according to a_i at the first cut for which the discordance degree $d_j(a_i)$ exceeds the α_k threshold.
3. The logical maximum concept is obeyed. Only the cut with the highest level α_{k_1} is considered when $\varphi_{ji}^{k_1} = \cdots = \varphi_{ji}^{k_h}$ for adjacent $\alpha_{k_1} > \cdots > \alpha_{k_h}$.

The algorithm for the derivation of complete selective strengths may be substituted with a straightforward operator:

$$\Phi_j = \sum_{k=1..l} \sum_{i=1..m} \alpha_k \cdot \left(\varphi_{ji}^{k} - \varphi_{ji}^{k-1}\right), \varphi_{ji}^{0} = 0.$$

Selective strengths do not have to be directly interpreted as criteria weights. Computed strengths can be modified in order to properly reflect an individual's personal beliefs. Ratios of criteria weights are transformed with a linear or exponential function in a consistent AHP pairwise comparison matrix. Each ratio increases according to the difference $\Delta_{ij} = \Phi_i - \Phi_j$.

2.2 Dominance indices based approach

A fuzzy binary relation on the criteria set is first constructed with a triangle superproduct composition so that Lukasiewitcz's implication compares two criteria with regard to their restrictive veto effects on a single alternative. Then, the transitive closure of the fuzzy binary relation is found. Every α-cut of the transitive closure is analysed to obtain a unique partial order of criteria. Partial orders are combined into one weak order. Finally, dominance indices are computed, which measure how influential different criteria are. A criterion is the more influential the more are relations in which it is with other criteria distant from the antiideal considering all cut-levels and standard distance metrics. More details on the method may be found in the literature [4].

3 SORTING OF ALTERNATIVES INTO ARBITRARY MANY CATEGORIES

In the case of sorting alternatives into $p + 1$ ordered categories with regard to p reference profiles the criterion-wise influence of veto is locally restricted to two adjacent categories. This means that different values of the veto threshold can change the assignment of an alternative for at most one category. Therefore, sorting depends globally on the evaluations of referential profiles, and locally on the evaluations of veto thresholds.

Theorem
The influence of veto depends on the veto threshold v_j only in the neighbourhood of one referential profile, while it depends on the absolute evaluation of the profile $g_j(b_h)$ for all other referential profiles.

Proof
For $p \geq 2$ and a maximized criterion, referential profiles are ordered in the ascending order, so that $g_j(b_1) < g_j(b_2) < \cdots < g_j(b_p)$ and $g_j(b_h) < g_j(b_{h+1}) - v_j < g_j(b_{h+1})$. The evaluation of the alternative a_i with regard to the criterion x_j is generally in one of the intervals $[g_j(b_{h-1}), g_j(b_h)]$, or in the intervals $[D_j^-, g_j(b_1)]$ or $[g_j(b_p), D_j^+]$, where D_j^- and D_j^+ are the lower and upper bounds of the j-th criterion domain. Then, h is the index of the only profile at which the veto is dependent on the threshold v_j. It can be observed that an alternative may not be subjected to veto of the lower $h - 1$ successive profiles regardless of v_j, because:

$$g_j(a) > g_j(b_i), \forall i = 1, \dots, h-1.$$

Analogously, the veto always has a full effect according to the upper $p-h$ profiles for each possible value of v_j:

$$g_j(a) < g_j(b_h) \Rightarrow g_j(a) < g_j(b_i) - v_j, \forall i > h, \forall v_j < g_j(b_i) - g_j(b_{i-1}).$$

Therefore, the selection of the veto threshold v_j influences only the assignment into the adjacent categories C_h and C_{h+1}, respectively. This means that for any allowed modification of v_j the assignment of each alternative can improve or deteriorate for at most one category. If the veto threshold v_j is set close to the preference threshold p_j and consequently has a moderate strength, then alternatives that are slightly outperformed by the profile $g_j(b_h)$ according to the criterion x_j are sorted into the lower category C_h. And at the opposite, if v_j is more distant from p_j, alternatives that are considerably worse than $g_j(b_h)$ may be sorted into the higher adjacent category C_{h+1}. When many categories are applied, the veto threshold consequently has a globally small contribution to the overall decision.

Based on the theorem and its proof, it is possible to derive criteria weights in the case of arbitrary many categories in such a way that the profile with the limited local veto effect is identified. If it exists, it is processed analogously as in the case of dichotomic sorting.

4 RANKING OF ALTERNATIVES

It is a slightly more difficult task to extend the mechanisms for criteria weight derivation for the decision-making problematic of ranking alternatives based on the outranking relations. In this case, a single alternative may impose a veto on zero, one or many other alternatives with regard to the i-th criterion. Since there is a set of m alternatives, m discordance degrees are required for the combination of the i-th criterion and j-th alternative. The discordance index $d_i(a_j), i = 1, \dots, n, j = 1, \dots m$ is hence substituted with indices $d_i(a_j, a_k)$ where $i = 1, \dots, n$ and $j, k = 1, \dots, m$. These indices express the fuzzy opposition to a set of assertions: »The alternative a_j is at least as good as the alternative a_k according to the criterion x_i.« They imply an additional third dimension compared to the case of dichotomic sorting. Three dimensions may be dealt with in two ways: (1.) directly by processing the fuzzy veto relation in the three-dimensional space, or (2.) by the reduction to a two-dimensional problem with the use of an appropriate fuzzy aggregation operator. It is more straightforward to implement the latter approach by defining the transformation:

$$d_i(a_j, a_k) \rightarrow d_i(a_j),$$

which leads to the basic two-dimensional veto relation with the interpretation: »The criterion x_j imposes a veto on the alternative a_j.« The rationale for this interpretation is that weighting applies to measuring the influence of criteria instead of the influence of alternatives. For this reason, the following questions are relevant to determine the importance of a criterion:

1. how many alternatives in total are subjected to the i-th criterion's veto;
2. how many alternatives are disqualified because of the i-th criterion's veto according to the single j-th alternative, and with what intensity;
3. according to how many different alternatives the single j-th alternative is disqualified because of the i-th criterion's veto, and with what intensity;
4. how many other criteria impose a veto on the same j-th alternative.

As a consequence of the above four rules, a criterion is selectively strong if it dismisses many alternatives, if other criteria impose no veto or, at most, a weak veto on the same set of

alternatives, and if many alternatives oppose a veto on the same j-th alternative with regard to this criterion. The second and third rule are correlated because of the symmetry:

$$a_j V_i a_k \Rightarrow \neg\left(a_k V_i a_j\right).$$

The first rule is dependent on the second and third rule, and the last rule is implemented in the dichotomic sorting algorithms for the derivation of selective strengths and dominance indices. It follows that the reduction of dimensionality may be simplified based on the fact that the criterion x_i outranks the alternative a_j by means of other alternatives a_k, for some $k = 1, \dots, m$ and $j \neq k$, which implies that the selective strength of x_i over a_j is proportional to the number of alternatives a_k. Then it is possible to aggregate discordance/veto matrices with the fuzzy »or« operator:

$$d_i\left(a_j\right) = \gamma \cdot \max_{k=1..m} d_i\left(a_j, a_k\right) + (1-\gamma) \cdot \frac{\sum_{k=1..m} d_i\left(a_j, a_k\right)}{m},$$

where $\gamma = 0.5$. The compensation between the maximum and average values overcomes two contradictory drawbacks:

- A criterion with a strong veto according to a single alternative should not necessarily outperform another criterion with weaker vetos according to many alternatives.
- The average of several weak vetos should not necessarily substitute for a strong veto, since the higher the degree of discordance is, the more relevant it is.

To summarize, $d_i\left(a_j\right)$ expresses a weak veto of the criterion x_i on the acceptability of the alternative a_j. It is reflected through fuzzy discordance degrees of $m-1$ alternatives that contradict the preference of a_j over these $m-1$ alternatives. As a result, a two-dimensional veto matrix is obtained that is identical to the one which is used in the case of dichotomic sorting. Two approaches from Section 2 can hence be applied after the transformation of veto matrices in order to derive criteria weights.

5. APPLICATION OF VETO IN RELATION WITH THE UTILITY FUNCTION

Sections 2 to 4 refer to decision-making approaches of the so called European school, which is founded on the outranking relation [13]. However, the utility theory [17] is often applied in practice. In the past, some ideas have been expressed to introduce three key concepts of outranking – constructivism, incomparability and veto based (partial) incompensation – into the utility theory [16]. One of the first theoretically sound and useful approaches to apply the noncompensatory veto effect in the multi-attribute utility function has been designed within the scope of our previous research work [5]. It is based on the veto criterion.

The veto criterion is modelled in accordance with the underlaying concepts of the utility theory. The veto function is specified by obeying the formal axiomatized concept of certain equivalence, so that criterion-wise values of alternatives are monotonously projected on the [0, 1] interval. There are three key distinctions between the »ordinary« criterion and the veto criterion. The first has a relative compensatory effect, exhibits positive characteristics that should be maximised, and is modelled locally on various hierarchical levels of the criteria structure. The latter has an absolute noncompensatory effect, shows negative characteristics that should be minimised, and is modelled globally on the highest hierarchical level.

The operator that aggregates compensatory utilities with noncompensatory discordance information multiplies the total utility with the product of inverse veto degrees:

$$\sigma(a_i) = u(a_i) \cdot \prod\nolimits_{j=1..n} \left(1 - d_j(a_i)\right).$$

It can be observed that the veto degrees $d_j(a_i)$ form the veto matrix V, as is defined in Section 2. The selective strengths and dominance indices based weight derivation methods can therefore be directly applied. Further information may be found in the literature [5].

6 CONCLUSION

The importance of criteria does not need to be specified directly in the form of compensatory criteria weights. It can also be indirectly and (semi)automatically inferred from the values of other preferential parameters, especially the noncompensatory veto thresholds or functions. Such derivation depends on the set of available alternatives. It provides a good insight into the modelled decision-making situation, reduces the decision-maker's cognitive load, and is applicable for autonomous aggregation-disaggregation algorithms.

The paper introduced the methodological foundations of weight derivation in the context of various decision-making problematics and preference models. Because of limitations in the scope, it omitted some aspects that will be presented separately. Hence, the efficiency of the proposed weight derivation methods will be discussed in a follow up paper, and some practical examples will be additionally provided.

References

[1] Bregar, A. (2009). Efficiency of problem localization in group decision-making. Proceedings of 10th International Symposium on Operational Research in Slovenia (pp. 139–149). Nova Gorica.

[2] Bregar, A. (2011). Investigation of the aggregation-disaggregation approach to multi-criteria negotiations: Consolidation of simulation and case studies. Proceedings of 11th International Symposium on Operational Research in Slovenia (pp. 117–124). Dolenjske Toplice.

[3] Bregar, A. (2011). Outranking methods and their application in group decision-making: A case study. Uporabna Informatika, 19(2): 75–90.

[4] Bregar, A., Györkös, J. (2003). Semiautomatic determination of criteria weights according to veto thresholds in the case of the localized alternative sorting analysis. Proceedings of 7th International Symposium on Operational Research in Slovenia (pp. 267–274). Podčetrtek.

[5] Bregar, A., Györkös, J., Jurič, M. B. (2007). Vpeljava nekompenzacijskega učinka veta v večatributno funkcijo koristnosti. Proceedings of 14th Symposium Dnevi Slovenske Informatike (10 pp.). Portorož.

[6] Bregar, A., Györkös, J., Jurič, M. B. (2008). Interactive aggregation/disaggregation dichotomic sorting procedure for group decision analysis based on the threshold model. Informatica, 19(2): 161–190.

[7] Figueira, J., Roy, B. (2002). Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. European Journal of Operational Research, 139(2): 317–326.

[8] Keeney, R. L., Raiffa, H. (1976). Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: John Willey & Sons.

[9] Miettinen, K., Salminen, P. (1999). Decision-aid for discrete multiple criteria decision making problems with imprecise data. European Journal of Operational Research, 119(1): 50–60.

[10] Mousseau, V., Slowinski, R., Zielniewicz, P. (2000). A user-oriented implementation of ELECTRE TRI integrating preference elicitation support. Computers & Operations Research, 27(7–8): 757–777.

[11] Ribeiro, R. A., Marques Pereira, R. A. (2003). Generalized mixture operators using weighting functions: A comparative study with WA and OWA. European Journal of Operational Research, 145(2): 329–342.

[12] Rogers, M., Bruen, M. (1998). A new system for weighting environmental criteria for use within ELECTRE III. European Journal of Operational Research, 107(3): 552–563.

[13] Roy, B. (1996). Multicriteria Methodology for Decision Aiding. Dordrecht: Kluwer Academic.

[14] Saaty, T. L. (2006). Rank from comparisons and from ratings in the analytic hierarchy/network processes. European Journal of Operational Research, 168(2): 557–570.

[15] Shepetukha, Y., Olson, D. L. (2001). Comparative analysis of multiattribute techniques based on cardinal and ordinal inputs. Mathematical and Computer Modelling, 34(1–2): 229–241.

[16] Stewart, T. J., Losa, F. B. (2003). Towards reconciling outranking and value measurement practice. European Journal of Operational Research, 145(3): 645–659.

[17] Zeleny, M. (1982). Multiple Criteria decision Making. New York: McGraw-Hill.

[18] Zopounidis, C., Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. European Journal of Operational Research, 138(2): 229–246.

IGMDEA AND TFNGMDEA METHODS FOR DERIVING WEIGHTS FROM INTERVAL-VALUED COMPARISON MATRICES IN AHP

Petra Grošelj and Lidija Zadnik Stirn
University of Ljubljana, Biotechnical Faculty
Jamnikarjeva 101, 1000 Ljubljana, Slovenia
{petra.groselj, lidija.zadnik}@bf.uni-lj.si

Abstract: Analytic hierarchy process is a suitable method for group multi-criteria decision making. Individual judgments can be aggregated into interval-valued group judgments as intervals could better reflect diversity of decision-makers. Interval-valued judgments can be interpreted as interval judgments or as triangular fuzzy numbers. In the paper we propose two new methods for deriving weights from interval-valued comparison matrices and apply them to real world case study (Pohorje, Slovenia).

Keywords: analytic hierarchy process, group decision making, fuzzy numbers, interval numbers

1 INTRODUCTION

Group decision-making is important in dealing with complex real world problems as diverse knowledge, experiences and perspectives could be included with different stakeholders. The analytic hierarchy process (AHP) [14] is a widely used method in group multi-criteria decision making. AHP is hierarchical approach with hierarchy of goal, criteria and alternatives based on pairwise comparisons of decision-makers valued from 1 to 9 on the AHP scale. The group judgments are rather presented as interval-valued judgments than as scalar-valued as they can better reflect diversity in opinions of individuals. There are two possible interpretations of the group interval-valued judgments, expressed as interval judgments or as triangular fuzzy numbers (TFNs). In the paper we discuss both possibilities, and suggest new approaches for deriving a priority vector from the interval-valued comparison matrices. The presented approaches are applied to a decision making problem of development of Pohorje, a mountain range in northeastern Slovenia.

2 ANALYTIC HIERARCHY PROCESS

Let $A=\left(a_{ij}^{(k)}\right)_{n\times n}$, $i,j=1,...,n,\ k=1,...,r$ be comparison matrices of pairwise comparisons in AHP of r decision-makers comparing n objects and $a_{ji}^{(k)}=1/a_{ij}^{(k)}$. The priority vectors $w^{(k)}=(w_1^{(k)},...,w_n^{(k)})$ can be derived from the comparison matrices by many known methods [5, 14, 16]. The consistency of comparisons is measured by the consistency ratio:

$$CR=\frac{CI}{RI},\ CI=\frac{\lambda_{\max}-n}{n-1}, \tag{1}$$

$\lambda_{\max}$ denoting the principal eigenvalue of comparison matrix A and RI denoting the random index. According to Saaty [15], $CR<0.1$ is considered acceptable.

According to the literature the most common aggregating approach is connected with minimal and maximal individual judgments. Let $A^{group}=\left(a_{ij}^{group}\right),\ i,j=1,...,n$ be a comparison matrix of the group interval-valued judgments.

Group judgments a_{ij}^{group} can be interpreted as interval judgments $a_{ij}^{group}=\left\lfloor l_{ij}^{group},u_{ij}^{group}\right\rfloor$ [1, 19]

$$\text{where } l_{ij}^{group} = \min_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\} \tag{2}$$

$$\text{and } u_{ij}^{group} = \max_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\}. \tag{3}$$

or as TFN judgments $\tilde{a}_{ij}^{group} = \left(l_{ij}^{group}, m_{ij}^{group}, u_{ij}^{group} \right)$ [2, 3, 11]

$$\text{where } l_{ij}^{group} = \min_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\} \tag{4}$$

$$m_{ij}^{group} = \left(\prod_{k=1}^{r} a_{ij}^{(k)} \right)^{1/r} \tag{5}$$

$$\text{and } u_{ij}^{group} = \max_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\} . \tag{6}$$

Interval comparison matrix $A^{group} = \left(\left[l_{ij}^{group}, u_{ij}^{group} \right] \right)_{nxn}$ is reciprocal because

$$1/u_{ji}^{group} = 1/\left(\max_{k=1,2,...,r} \left\{ a_{ji}^{(k)} \right\} \right) = 1/\left(\max_{k=1,2,...,r} \left\{ 1/a_{ij}^{(k)} \right\} \right) = 1/\left(1/\min_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\} \right) = \min_{k=1,2,...,r} \left\{ a_{ij}^{(k)} \right\} = l_{ij}^{group}. \tag{7}$$

Similarly $1/l_{ji}^{group} = u_{ij}^{group}$ for $i, j = 1,...,n$. TFN comparison matrix is reciprocal because

$$1/\tilde{a}_{ji}^{group} = 1/\left(l_{ji}^{group}, m_{ji}^{group}, u_{ji}^{group} \right) = \left(1/u_{ji}^{group}, 1/m_{ji}^{group}, 1/l_{ji}^{group} \right) = \left(l_{ij}^{group}, m_{ij}^{group}, u_{ij}^{group} \right) = \tilde{a}_{ij}^{group}, \tag{8}$$

where usual fuzzy arithmetic operations have been applied.

According to Wang [20] TFN comparison matrix $\tilde{A}^{group} = \left(\left(l_{ij}^{group}, \mathrm{m}_{ij}^{group}, u_{ij}^{group} \right) \right)_{n \times n}$ is acceptably consistent if scalar-valued matrices $A^{m} = \left(\mathrm{m}_{ij}^{group} \right)_{n \times n}$ and $A^{glu} = \left(\left(l_{ij}^{group} u_{ij}^{group} \right)^{1/2} \right)_{n \times n}$ are acceptably consistent. Similarly we define interval comparison matrix $A^{group} = \left(\left[l_{ij}^{group}, u_{ij}^{group} \right] \right)_{nxn}$ to be acceptably consistent if scalar-valued matrix $A^{glu} = \left(\left(l_{ij}^{group} u_{ij}^{group} \right)^{1/2} \right)_{n \times n}$ is acceptably consistent.

One of the important questions in such cases and the main concern of our paper is how to derive a priority vector from the interval-valued comparison matrix. There are several known methods of deriving an interval priority vector from the interval comparison matrix [9, 12, 18, 19, 21, 22], but ranking of interval weights is not always easy if the weights overlap.

The extent analysis method [4] with its improved version [8, 17, 23] is the most popular method for deriving weights from a triangular fuzzy comparison matrix. Its main drawback is that it could produce illogical zero weights [17].

To overcome this drawback we propose a new aggregation method that results in scalar-valued weights. We choose Wang&Chin DEA method [16] as the base for our approach:

$$\begin{aligned} &\max w_0 = \sum_{j=1}^{n} a_{0j} x_j \\ &\textit{subject to } \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij} \right) x_j = 1, \\ &\qquad \sum_{j=1}^{n} a_{ij} x_j \geq n x_i, \;\; i = 1,...,n, \\ &\qquad x_j \geq 0, \;\; j = 1,...,n. \end{aligned} \tag{9}$$

Wang&Chin DEA method has already been expanded to the WGMDEA method [7], the method for deriving group weights from individual scalar-valued judgments and has already been successfully employed in several applications [6, 10, 13].

3 PROPOSED APPROACH FOR DERIVING WEIGHTS FROM INTERVAL-VALUED COMPARISON MATRIX

First, we propose a new method, called IGMDEA, for deriving weights from interval comparison matrix $A^{group} = \left(\left[l_{ij}^{group}, u_{ij}^{group}\right]\right)_{nxn}$:

$$\begin{aligned} &\max w_0 = \sum_{j=1}^{n} \left(l_{0j}^{group} \cdot u_{0j}^{group}\right)^{1/2} x_j \\ &subject\ to\ \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \left(l_{ij}^{group} \cdot u_{ij}^{group}\right)^{1/2} \right) x_j = 1, \\ &\qquad \sum_{j=1}^{n} \left(l_{ij}^{group} \cdot u_{ij}^{group}\right)^{1/2} x_j \geq n x_i, \ \ i = 1, ..., n, \\ &\qquad x_j \geq 0, \ \ j = 1, ..., n. \end{aligned} \tag{10}$$

Second, we propose a new method, called TFNGMDEA, for deriving weights from TFN comparison matrix $\tilde{A}^{group} = \left(\left(l_{ij}^{group}, \mathrm{m}_{ij}^{group}, u_{ij}^{group}\right)\right)_{nxn}$:

$$\begin{aligned} &\max w_0 = \sum_{j=1}^{n} \left(l_{0j}^{group} \cdot m_{0j}^{group} \cdot u_{0j}^{group}\right)^{1/3} x_j \\ &subject\ to\ \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \left(l_{ij}^{group} \cdot m_{ij}^{group} \cdot u_{ij}^{group}\right)^{1/3} \right) x_j = 1, \\ &\qquad \sum_{j=1}^{n} \left(l_{ij}^{group} \cdot m_{ij}^{group} \cdot u_{ij}^{group}\right)^{1/3} x_j \geq n x_i, \ \ i = 1, ..., n, \\ &\qquad x_j \geq 0, \ \ j = 1, ..., n. \end{aligned} \tag{11}$$

In TFNGMDEA method not only lower and upper bounds of TFNs are taken into account but also the middle values, which may reflect that TFNs are not necessary symmetrical.

4 EXAMPLE

We apply the presented methods on an example where stakeholders resolve which sector is most important for the development of Slovenian mountain range Pohorje. This research was already presented in the doctoral thesis [6]. Twelve stakeholders familiar with Pohorje (three experts from each sector) compared four sectors (forestry, agriculture, tourism, nature protection) and estimated which is from their point of view most important for the development of Pohorje. Their judgments are aggregated into interval comparison matrix A^{group} applying (2) and (3)

$$A^{group} = \begin{bmatrix} 1 & [0.5, 8] & [0.33, 3] & [0.25, 3] \\ [0.125, 2] & 1 & [0.2, 2] & [0.2, 2] \\ [0.33, 3] & [0.5, 5] & 1 & [0.25, 4] \\ [0.33, 4] & [0.5, 5] & [0.25, 4] & 1 \end{bmatrix} \tag{12}$$

and into TFN comparison matrix $\tilde{A}^{group}$ applying (4), (5) and (6)

$$\tilde{A}^{group} = \begin{bmatrix} (1,1,1) & (0.5,1,8) & (0.33,0.94,3) & (0.25,0.94,3) \\ (0.125,1,2) & (1,1,1) & (0.2,0.98,2) & (0.2,0.71,2) \\ (0.33,1.06,3) & (0.5,1.02,5) & (1,1,1) & (0.25,0.86,4) \\ (0.33,1.06,4) & (0.5,1.41,5) & (0.25,1.16,4) & (1,1,1) \end{bmatrix}. \quad (13)$$

Both group comparison matrices are acceptably consistent because of $CR\left(A^{glu}\right) = 0.005$ and $CR\left(A^{m}\right) = 0.003$. Priority weights are derived from A^{group} using IGMDEA (10) and from $\tilde{A}^{group}$ using TFNGMDEA (11). They are presented together with results of WGMDEA method [7] and improved extent analysis method for TFN [17] in Table 1.

Table 1: Weights of four sectors, calculated with different methods

	IGMDEA	TFNGMDEA	WGMDEA	Improved extent analysis
Tourism	28.0%	26.8%	24.2%	24.9%
Forestry	16.3%	18.3%	22.8%	23.7%
Agriculture	27.3%	26.4%	24.4%	24.9%
Nature protection	28.4%	28.5%	28.6%	26.6%

The individual results show that the majority of stakeholders slightly favored their sector, but the aggregation negated this phenomenon. The group results show that nature protection received the highest weight, indicating that stakeholders believe that opinion of experts from the sector of nature protection is important, since they strive for maintaining the nature and biodiversity of Pohorje. However, weights do not differ much. The second and the third place were gained by tourism and agriculture. The smallest weight belongs to the forestry.

The differences between methods are not high. Forestry received smaller weight with methods IGMDEA and TFNDEA. This indicates that the differences between stakeholders have been considered to a higher degree. While geometric mean of individual judgments used in WGMDEA obliterate the differences between individual judgments, IGMDEA and TFNDEA take into account the whole range of individual judgments.

5 CONCLUSIONS

In the paper we presented aggregation of individual judgments in AHP into group interval judgments or/and into group TFNs. We proposed two new methods, IGMDEA and TFNDEA, for deriving weights from group interval comparison matrix and group TFN comparison matrix, respectively. The results obtained in the application of the methods on the problem of the development of Pohorje, Slovenia show that methods could be suitable for the application. They better take into account the wide range of individual judgments as some other methods. The future work should be devoted to the further evaluation of the new methods.

Acknowledgement
Petra Grošelj acknowledges financial support from Slovenian Research Agency within the framework of Program P4-0059.

References

[1] Chandran, B., Golden, B. and Wasil, E., 2005. Linear programming models for estimating weights in the analytic hierarchy process. Computers & Operations Research, 32(9), pp. 2235–2254.

[2] Chang, C.-W., Horng, D.-J. and Lin, H.-L., 2011. A measurement model for experts knowledge-based systems algorithm using fuzzy analytic network process. Expert Systems with Applications, 38(10), pp. 12009–12017.

[3] Chang, C.-W., Wu, C.-R. and Lin, H.-L., 2009. Applying fuzzy hierarchy multiple attributes to construct an expert decision making process. Expert Systems with Applications, 36(4), pp. 7363–7368.

[4] Chang, D.-Y., 1996. Applications of the extent analysis method on fuzzy AHP. European Journal of Operational Research, 95(3), pp. 649–655.

[5] Crawford, G. and Williams, C., 1985. A note on the analysis of subjective judgment matrices. Journal of Mathematical Psychology, 29(4), pp. 387–405.

[6] Grošelj, P. (2013) Group methods in analytic hierarchy process and their application to the management of protected area Pohorje. Biotechnical faculty. University of Ljubljana, Ljubljana.

[7] Grošelj, P., Pezdevšek Malovrh, Š. and Zadnik Stirn, L., 2011. Methods based on data envelopment analysis for deriving group priorities in analytic hierarchy process. Central European Journal of Operations Research, 19(3), pp. 267–284.

[8] Heo, E., Kim, J. and Cho, S., 2012. Selecting hydrogen production methods using fuzzy analytic hierarchy process with opportunities, costs, and risks. International Journal of Hydrogen Energy, 37(23), pp. 17655–17662.

[9] Jibin, L., Xinmiao, Y., Mingming, H. and Xiaojing, W. (2011) Obtaining interval weights from interval reciprocal comparison matrices. Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on.

[10] Kuzman, M. K., Grošelj, P., Ayrilmis, N. and Zbašnik-Senegačnik, M., 2013. Comparison of passive house construction types using analytic hierarchy process. Energy and Buildings, 64(0), pp. 258-263.

[11] Larimian, T., Zarabadi, Z. S. S. and Sadeghi, A., 2013. Developing a fuzzy AHP model to evaluate environmental sustainability from the perspective of Secured by Design scheme—A case study. Sustainable Cities and Society, 7(0), pp. 25–36.

[12] Meng, F., Chen, X., Zhu, M. and Lin, J., 2015. Two new methods for deriving the priority vector from interval multiplicative preference relations. Information Fusion, 26(0), pp. 122–135.

[13] Oblak, L. and Glavonjić, B., 2015. A Model for the Evaluation of Radio Advertisements for the Sale of Timber Products. Drvna industrija, 65(4), pp. 303–308.

[14] Saaty, T. L., 1980 The Analytic Hierarchy Process McGraw-Hill, New York.

[15] Saaty, T. L., 2005 Theory and Applications of the Analytic Network Process: Decision Making with Benefits, Opportunities, Costs, and Risks RWS Publications, Pittsburgh.

[16] Wang, Y.-M. and Chin, K.-S., 2009. A new data envelopment analysis method for priority determination and group decision making in the analytic hierarchy process. European Journal of Operational Research, 195(1), pp. 239–250.

[17] Wang, Y.-M., Luo, Y. and Hua, Z., 2008. On the extent analysis method for fuzzy AHP and its applications. European Journal of Operational Research, 186(2), pp. 735–747.

[18] Wang, Y.-M., Yang, J.-B. and Xu, D.-L., 2005. Interval weight generation approaches based on consistency test and interval comparison matrices. Applied Mathematics and Computation, 167(1), pp. 252–273.

[19] Wang, Y.-M., Yang, J.-B. and Xu, D.-L., 2005. A two-stage logarithmic goal programming method for generating weights from interval comparison matrices. Fuzzy Sets and Systems, 152(3), pp. 475–498.

[20] Wang, Z.-J., 2015. Consistency analysis and priority derivation of triangular fuzzy preference relations based on modal value and geometric mean. Information Sciences, 314(0), pp. 169-183.

[21] Xia, M. and Xu, Z., 2014. Interval weight generation approaches for reciprocal relations. Applied Mathematical Modelling, 38(3), pp. 828–838.

[22] Zeshui, X. and Yager, R. R., 2010. Power-Geometric Operators and Their Use in Group Decision Making. Fuzzy Systems, IEEE Transactions on, 18(1), pp. 94–105.

[23] Zhu, K.-J., Jing, Y. and Chang, D.-Y., 1999. A discussion on Extent Analysis Method and applications of fuzzy AHP. European Journal of Operational Research, 116(2), pp. 450–456.

APPROXIMATING RANDOM INDEXES IN THE ANALYTICAL HIERARCHY PROCESS

Slawomir Jarek
University of Economics in Katowice, Department of Operations Research
ul. 1 maja 50, 40-287 Katowice, Poland
slawomir.jarek@ue.katowice.pl

Abstract: Verification of the consistency of the pair wise comparisons matrix plays a key role in the *Analytical Hierarchy Process* (AHP). This test consists of calculating the *consistency ratio* CR, which depends on the *random index* RI. A method for determining these values has been described by Saaty. Since then, a number of studies have been published which give different values of these coefficients. Unfortunately, these values refer to several selected ranges of values and do not take into account the other scales, described by Saaty in [7]. Therefore, the need arises to calculate new RI values which take into account the needs of AHP users. This paper presents an R script, which allows to solve this problem. However, for a large number of decision criteria or variants, this may be a very time-consuming process. To solve this problem an approximated formula was suggested that facilitates the determination of RI.

Keywords: decision making, MCDM, AHP, consistency index, random index

1 INTRODUCTION

In the AHP method described in [8] the creation of a ranking of discrete decision variants starts with pairwise comparisons. The consistency of this process is verified by CR, which can be described by the relation (1).

$$CR = \frac{CI}{RI} \tag{1}$$

Here CI is the consistency index, which has been calculated using the formula (2) where m is the number of decision criteria or variants compared, and λ_{max} is the largest eigenvalue of the pairwise comparisons matrix (PCM) of $\boldsymbol{A}$.

$$CI = \frac{\lambda_{\max} - m}{m-1} \tag{2}$$

If the matrix $\boldsymbol{A}$ is consistent, its rank is equal to exactly 1 and it has only one non-zero eigenvalue. This value is equal to the sum of diagonal elements of $\boldsymbol{A}$ and it equals exactly m. If the matrix $\boldsymbol{A}$ is consistent then both CI and CR are 0. Otherwise, there are additional eigenvalues that are complex numbers, and the value of λ_{max} is greater than m. Of course, the sum of all the eigenvalues of the matrix $\boldsymbol{A}$ is equal to m, thus CI can be interpreted as expressing the degree of non-consistency of pairwise comparisons. This value is dependent on the dimension of the matrix $\boldsymbol{A}$. To obtain a measure that is independent of the size of the matrix $\boldsymbol{A}$, formula (1) includes RI, which reflects the average size of CI for a given m. In [7] Saaty describes a method for determining the value of RI. It involves the generation of n random matrices $\boldsymbol{A}$, which all have the same dimensions as the PCM analyzed. Then, for each of these matrices, CI is calculated. The value of RI is the arithmetic mean of the calculated CI's. AHP users do not have to compute RI in this way, because these values are given in the literature. Unfortunately, since the publication of the work [8] many authors have proposed many significantly different variants of the values of RI. An overview of these studies is included in [2].

Alonso and Lamata in [2] analyzed the values of RI determined by various authors (among them [1], [4], [5] and [6]) and compared nine different versions of RI. Alonso and Lamata studied RI's calculated from samples of 100, 500, 1000, 2500 and 100 000 random matrices. They noted that the achieved results do not differ significantly in samples which consist of 100 000 and 500 000 experiments, respectively. In addition, the paper [2] proposed a polynomial expression, which enables the user to determine RI without conducting computer simulations. However, for large *m*, the values of the polynomial may differ significantly from the values obtained in computer simulation. It is worth noting that the results obtained in [5] by Noble and Sanchez for samples of 500 random matrices are consistent with the values

For determining the RI, Saaty in [8] determined a sample size of only 100 for *m* ranging from 2 to 11, and the sample size was increased to 500 for *m* ranging from 12 to 15. In [3] Donegan and Dodd noticed that the size of these samples seems too small for a proper estimation of confidence intervals for RI. On this basis, they carried out calculations for samples of 1000 elements, resulting in a slightly smaller value of RI. Saaty in his review paper [7] suggests to increase the number of simulation experiments to 50 000, which seems sufficiently large.

These results have been adopted in this study and it was assumed that the sample size is 50 000 elements.

2 EVALUATION OF RI

In chapter 4 of [7] Saaty discusses various scales used to build a PCM. Table 1 on page 257 defines Fundamental Scale of Absolute Numbers [7, p. 257]. The first of these scales is commonly used and admits values from 1 to 9 and their reciprocals. The second scale is used to compare very similar activities. For such problems, the results of pairwise comparisons admit values 1 and 2 only. To further illustrate the diversity of the objects compared, Saaty proposes to introduce another scale, which ranges from 1.1 to 1.9 in steps of 0.1, whose interpretation coincides with that of the first scale. For the second scale, the corresponding values of RI have not been proposed.

The procedure for determining the RI described in chapter 6 of [7] is as follows. We construct *n* random matrices *A*. The values of pairwise comparisons on the main diagonal are 1 while the values above it are selected at random from among all the possible values of the scale. In the case of the standard scale there are seventeen such values, forming the set {1/9, 1/8, 1/7, ..., ½, 1, 2, ..., 9}. If the objects being compared are very similar, the values are drawn from the following set of 19 elements: {1/1.9, 1/1.8, 1/1.7, … , 1/1.2, 1/1.1, 1, 1.1, 1.2, … ,1.9}. Next, the elements below the main diagonal are calculated as the inverses of the corresponding elements above the main diagonal. For such a randomly selected matrix the values of CI are calculated from formula (2). The value of RI is calculated as the mean of the previously calculated values of CI. It is also possible to determine the standard deviation and the limit value RI at a given level of confidence, according to equation (3).

$$RI[m] = mean(ri_m) - 2.33\frac{std(ri_m)}{\sqrt{n}} \tag{3}$$

2.1 A suggested script in R

It is convenient to perform calculations in the statistical package ***R*** which was described in [9]. This package allows for efficient matrix calculations and is free of charge. The proposed script is shown in *Listing 1*. First, the size of the sample (n = 50 000) and the dimension of the matrix $\boldsymbol{A}$ (m = 15) are specified, and the number k of elements located above the main

diagonal is determined. The vector ***scores*** contains all the values of the scale used to compare the objects. Then the individual samples are drawn; they will serve to create further samples of ***A***. The function ***runif*** generates random variables with uniform distributions. In the loops the consecutive values of CI are generated and stored in the vector ci.sampl. At the end, the mean and the standard deviation of ci.sample are determined. Based on this statistics the value of RI can be determined. Listing 1 contains a procedure which allows to determine RI for a standard 1-to-9 scale.

```
n         <- 50000
m         <- 15
k         <- m*(m-1)/2
ci.sampl <- matrix(nrow=n)
scores    <- c(1/9:1, 2:9)
samples   <- matrix(scores[round(runif(n*k,max=16))+1],n,k)

for (p in 1:n) {
  a <- diag(m)
  l <- 1
  for(i in 1:(m-1)) for(j in (i+1):m){
    a[i,j] <- samples[p,l]
    a[j,i] <- 1/samples[p,l]
    l      <- l+1
  }

  b            <- a %*% diag(1/apply(a,2,sum))
  w            <- apply(b,1,mean)
  lambda       <- mean(a %*% w /w)
  ci.sampl[p] <- (lambda-m)/(m-1)
}

mean(ci.sampl)
sd(ci.sampl)
mean(ci.sampl)-2.33*sd(ci.sampl)/sqrt(n)
```

Listing 1: The determination of statistics for the standard 1-to-9 scale

In the case of another scale, with values from 1.1 to 1.9, the vector *scores* and the method of generating the matrix *samples* should be modified. The necessary modifications of Listing 1 are shown in Listing 2.

```
scores  <- c(1/seq(1.9, 1, by=-0.1), seq(1.1, 1.9, by=0.1))
samples <- matrix(scores[round(runif(n*k,max=18))+1],n,k)
```

Listing 2: The modification which allows the determination of statistics for the 1.1-to-1.9 scale

2.2 An approximate formula for RI

Analyzing the results of the calculations contained in Table 1 and the graphs of RI published in [7] and [2] it can be seen that the relevant points are arranged along a curve resembling a logarithmic function. It turns out that a good approximation of the values of RI published in [7] can be obtained from equation (4). In this expression k is the number of elements of the

matrix $\boldsymbol{A}$ below the main diagonal and m is the size of the array. The values obtained from expression (4) are shown in Table 1, in the row marked as Approx.

$$RI[m] = \frac{\log(k)}{\log(m+3)} \tag{4}$$

3 THE MAIN RESULTS

Using the script described in the previous section we can determine RI for m in the range from 3 to 15 and for m equal to exactly 30. The calculations were made for the standard 1-9 scale and for the 1.1-1.9 scale. In both cases, the sample size is 50 000. The calculation results are provided in Tables 1 and Table 2 and illustrated in Figure 1. The sample size is large enough so that the mean value of CI and the expression (3) do not differ. In the course of the experiments, consistent pairwise comparison matrices occur only for small values of m. For larger values of m the values of CI indicate a lack of consistency of the matrix drawn.

3.1 Results for the standard scale

Table 1 contains the following rows:

- m – size of PCM,
- Saaty – values of RI given in [7],
- Approx –values of RI determined from the relation (4),
- mean(ri_m) – average values in ci.sampl,
- sd(ri_m) – standard deviation components of the vector ci.sampl,
- RI[m] – limits determined from formula (3) and
- DD – values published by Donegan and Dodd in [3].

Table 1: The values of RI obtained for the standard 1-to-9 scale

m	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*30*
Saaty	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49	1.52	1.54	1.56	1.58	1.59	-
Approx	0.61	0.92	1.11	1.23	1.32	1.39	1.44	1.48	1.52	1.55	1.57	1.59	1.61	1.74
mean(ri_m)	0.51	0.86	1.07	1.20	1.28	1.34	1.38	1.41	1.43	1.45	1.47	1.48	1.49	1.57
std(ri_m)	0.67	0.61	0.50	0.40	0.33	0.27	0.23	0.20	0.18	0.16	0.14	0.13	0.12	0.06
RI[m]	0.51	0.85	1.06	1.20	1.28	1.34	1.38	1.41	1.43	1.45	1.47	1.48	1.49	1.57
DD	0.49	0.80	1.06	1.18	1.25	1.32	1.37	1.41	1.42	1.45	1.46	1.48	1.50	1.58

The results of the procedure described in Listing 1 (see Table 1 and Fig. 1) are very similar to those given by Donegan and Dodd in [3] and Noble and Sanchez in [5]. It is worth noting that in [3] the sample size is smaller and is equal to only 100 for m < 12.

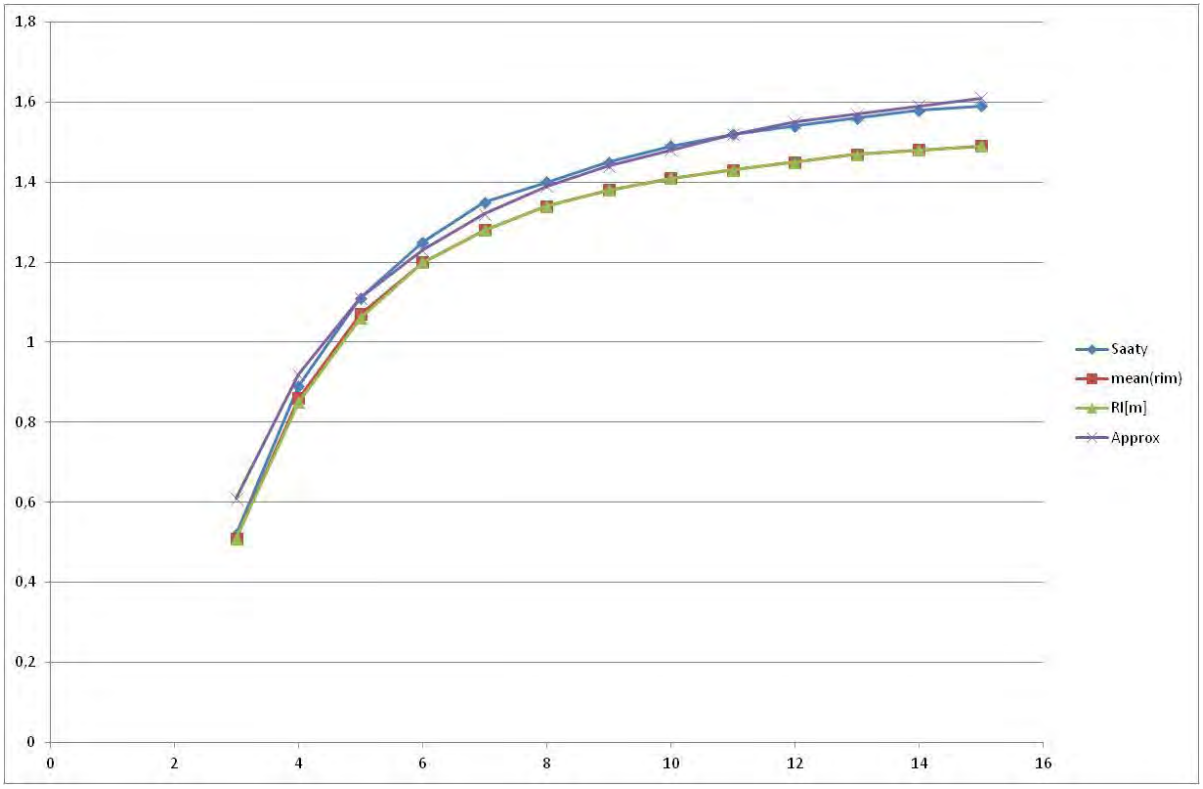


Figure 1: The values of RI obtained for the standard 1-to-9 scale

3.2 Results for the scale of 1.1 to 1.9

Table 2 contains the following rows of values, whose meaning is discussed in section 2.

Table 2: The values of RI obtained for the 1.1-to-1.9 scale

M	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*330*
mean(rim)	0.027	0.041	0.050	0.055	0.059	0.062	0.064	0.066	0.067	0.069	0.070	0.071	0.071	0.077
std(rim)	0.034	0.028	0.022	0.018	0.016	0.014	0.012	0.011	0.010	0.009	0.008	0.007	0.007	0.003
RI[m]	0.027	0.041	0.049	0.055	0.059	0.062	0.064	0.066	0.067	0.069	0.070	0.071	0.071	0.077

It is worth noting that for $m > 11$ the increment of RI is already very slow and reaches its saturation value of around 0.07.

4 CONCLUSIONS

This paper contains R scripts which can be used for two different scales. For the standard scale the results achieved are similar to those obtained in [3] and [5]. Furthermore, the authors of publications on AHP most often included only a dozen or so values of RI. This may constitute an obstacle to more complex problems. The proposed R scripts allow to exceed this limit and to determine the necessary value of RI. Moreover, the construction of the scripts allows to determine RI for those problems in which non-standard scales occur. The results of calculations for a sample assessment scale ranging from 1.1 to 1.9 are given in section 3.2.

The proposed form of the approximated formula (4) used to determine RI for the standard 1-to-9 scale coincides with the value of RI published in the paper [7]. The proposed formula makes it possible to estimate RI for larger problems where it is necessary to compare more

than a dozen elements. In contrast to other proposals based on polynomial functions, the use of a logarithmic function in (4) results in a much slower increase of the estimates of RI for large values of *m*.

References

[1] Aguaron, J. and Moreno-Jimenez, J. M. (2003). The geometric consistency index: Approximated thresholds. European Journal of Operational Research, 147(1): 137-145.

[2] Alonso, J.A. and Lamata, m. T. (2006). Consistency in the Analytic Hierarchy Process: a new approach. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 14(04): 445-459.

[3] Donegan, H. and Dodd, F. (1991). A note on Saaty's Random Indexes. Mathematical and Computer Modelling, 15(10): 135 -137.

[4] Lane, E.F. and Verdini, W.A. (1989). A consistency test for AHP decision makers. Decision Sciences, 20(3): 575-590.

[5] Noble E.E. and Sanchez P.P. (1990). A note on the information content of a consistent pairwise comparison judgement matrix of an AHP decision maker. Theory and Decision, 34: 99-108.

[6] Ozdemir, M.S. (2005). Validity and inconsistency in the Analytic Hierarchy Process. Appl. Math. Comput., 161(3): 707-720.

[7] Saaty, T.L. (2008). Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the Analytic Hierarchy/Network Process. RACSAM - Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas, 102(2): 251-318.

[8] Saaty, T. L. (1980). The Analytic Hierarchy Process. New York: McGraw-Hill.

[9] The R Core Team. R: A Language and Environment for Statistical Computing. (2015). Vienna, Austria: R Foundation for Statistical Computing. http://cran.r-project.org/doc/manuals/fullrefman.pdf [18/06/2015].

A NEW METHOD AIMING AT MORE REALISTIC MODELLING AND DEPENDABLE RESULTS IN MULTICRITERIA DECISION-MAKING

Nolberto Munier
Valencia Polytechnic University, INGENIO -Avenue des Tarongers S/N, Valencia, Spain
Avda. Peris i Valero 157- Valencia, 46006, Spain
nolmunier@yahoo.com

Abstract: This paper considers that most actual models for Multicriteria Decision Making (MCDM) for projects selection present two common drawbacks: Use of subjective considerations that make unlikely for a problem to get a matching result when different models are used to solve it.In addition, it is believed that many relaxing assumptions are made in trying to replicate actual scenarios, not only by simplifying them but also by not taking into account existing conditions. The model presented here addresses these two issues.

Keywords: Linear Programming, SIMUS, Subjectivity, ERM matrix, PDM matrix, Ranking.

1 INTRODUCTION

Decision-Making practice supported by mathematical help has been around for some time now, and different mathematical models grounded on diverse postulates have been developed addressing this issue. They have been devised by authors such as (Dantzig, 1948) [SIMPLEX algorithm][2]based on Linear Programming (LP); (Roy,1968) [ELECTRE method][8][9]based on outranking; (Brans and Vincke (1985) [PROMETHEE method][1] based on outranking; (Saaty,1980) [AHP method][9] based on personal preferences; (Saaty,1996) [ANP method][10] based on personal preferences; (Hwang and Yoon, 1981) [TOPSIS method][4] based on distances. In most of them different degrees of subjectivity are present which produces dissimilar results when solving a problem. In addition, poor modelling is also present.

1.1 Subjectivity.

Given a problem, all of the above mentioned methods reach compromise solutions, although with different results, regarding project selection and rankings. Since all of them use sound mathematical procedures, why different models produce different solutions? Because subjectivity, related to assigning weights, establishing preferences, thresholds acceptance for outranking, unsupported assumptions, distances selection, etc.

Naturally, decision-making is per se a subjective activity depending on the analysis, experience and knowledge of the Decision-Maker (DM), and because the amount and complexity of interrelations, he must be assisted by MCDM tools. Obviously, nobody can blindly rely on what a mathematical model shows, or take it as the true and irrefutable best solution; these models are only tools to process data and to help and support the DM. However, results are based on data, own ideas and experience, supplied by the DM, who of course, is not infallible when using his judgement. Nevertheless, he applies these models to get results obtained by using his own objective and subjective inputs, which may be biased, or influenced by external factors, or simply incorrect because uncertainty. Data reliability, results examination, analysis and decisions pertain to the DM,while processing is left to a model. Consequently, results in some extent depend on the person doing the analysis and this explains the different results. It appears that the only way to get homogenous results, it does not matter the method used, is avoiding subjectivity, and this is the aim of this paper.

1.2 Modelling

In addition, all models are crude approximations to reality, as they are very basic in their formulation by ignoring important aspects of each scenario; therefore, the mathematical model usually does not accurately represent reality, and then results are debatable.The model proposed here tries to improve modelling by using useful characteristics from LP.

2 THE MODEL FOUNDATION AND HOW IT WORKS

'SIMUS' (Munier, 2011a,b),[6][7] is the name of the model introduced in this paper; it is an acronym for *'Sequential Interactive Model for Urban Systems'*, addressing the two issues above mentioned, that is subjectivity and poor modelling.

SIMUS is a hybrid MCDM method designed to select and rank projects or alternatives based on both Linear Programming (LP) and Outranking. It is grounded on the fact that objective functions (which are our goals), and criteria (namely a set of conditions to evaluate projects), have the same mathematical structure, and thus they are interchangeable.

Modelling starts by building a standard '*Decision matrix'* in Excel (Table 1), that is, a table with a set of projects to be evaluated in columns, a set of criteria or targets responsible for this evaluation in rows,and cardinal values at their intersections or cells, denoting the contribution of each project with respect to each criterion. The DM must decide which criteria will be used as objective functions and identify them as *'Targets'* (Table 2). SIMUS works by selecting the first target, extracting and deleting it from the decision matrix, and applying said target as the objective function of a linear programming problem. Then, data is processed by '*Solver'*(Frontline Systems), an Excel add-in, which operates according to the Simplex algorithm. The process produces (it exists) an optimal solution that turns up as a score for each project. This set of scores is saved in the first row of a new matrix called *'Efficient Results Matrix'* (ERM) (Table 2), and then the target previously chosen is returned to the decision matrix.

*Table 1:*Typical Decision Matrix

	Projects		
Criteria	A	B	C
Criterion 1	42.4	39.8	42.0
Criterion 2	0.23	0.15	0.18
Criterion 3	801	726	582

Then, the following target is extracted and follows the same procedure, saving the result as a second row of the ERM. This sequence is repeated until all targets have been tested as objective functions. The resulting ERM is a *'Pareto Efficient Matrix,*composed by a set values or scores that cannot be improved, because they are optimal values.

*Table 2:*Construction of the Efficient Results Matrix (ERM)

	Projects		
Criteria*	A	B	C
Target 1	0.23	-----	0.56
Target 2	0.21	1.23	0.15
Target 3	---	1	---

* *'Criteria' and 'Targets' are the same. 'Criteria' is* a generic name and all of them used to evaluate a set of projects. However, if the DM selects some criteria (or them all), to be also used as objectives, they are called *'Targets'*.

The ERM is then normalized (Table 3), and used as a source of reliable data. From it, the model follows two different procedures or modes that will produce two different results. It is equivalent of treating the decision matrix of a problem by two different methods.

According to the first procedure (ERM mode), the model examines column-wise the normalized ERM.

a) It sums-up the values or scores within each project column (SC) along all criteria,which constitutes a partial measure of importance for each project.
b) Computes the number of times each project in its column has scores or contributions along all criteria, and finds the ratio of this contribution regarding the total number of criteria considered. This is the *'Performance Factor'*(PF), and it is required because it could be that a project participates with a very high score in one criterion and with '0' in others, and then being selected because it has a high (SC), while there could be other projects that participate in many criteria although with smaller values, and thus yielding a smaller (SC). Once computed the (PF) is normalized (NPF).
c) The model multiplies (SC) by the normalized (NPF) and delivers the final score for each project.
d) These final figures (when > 0) not only identify selected projects from the original set, but also rank them according to their decreasing values.

Table 3: Normalized ERM and results according to the ERM mode

	Projects		
Criteria	A	B	C
Target 1	0.29	-----	0.71
Target 2	0.13	0.77	0.09
Target 3	---	1	---
SC	0.42	1.77	0.80
PF	2/3	2/3	2/3
Scores (SC) x(NPF)	**0.28**	**1.17**	**0.53**

Thus, the ranking, using the ERM mode is: **Project B – Project C – Project A**

According to the second procedure (PDM mode), the model also starts from the normalized ERM (Table 2) and analyzes it row-wise. It works as follows:

a) The model builds a new square matrix called *'Project Dominance Matrix'* (PDM) (Table 4) composed by projects in columns and rows.
b) Examining the first row of the normalized ERM (Table 3), finds the highest score on that row (Target 1) and identifies the project it belongs to (in this example, C, with 0.71).
c) Find the differences between this score and each score on the same row corresponding to the other projects; that is, C with A and B, and after that, A and B.
   Project C – Project A = 0.71–0.29 = 0.42
   Project C – Project B = 0.71–0 = 0.71
   On this same row Project A outranks Project B, then
   Project A – Project B = 0.29-0 = 0.29
d) In PDM enters the two first difference values on the row corresponding to project 'C', and at the intersections with the columns pertaining to projects 'A' and 'B' respectively as shown in Table 4.
   The third difference value is placed on PDM on row A and at the intersection with column B.

e) The model then analyzes row two of the normalized ERM and follows the same procedure.
f) The model then analyzes row three of the normalized ERM and follows the same procedure. Notice that '1s' are added on row B to the cells for projects A and C, because the differences are unitary when analyzing target 3.

Table 4: Project Dominance Matrix (PDM) and result according to the PDM mode

	Subordinate Projects					
Dominant Projects	A	B	C	Sum of rows	Net difference	
A		0.29	0.04	0.33	**0.33-2.06 = -1.73**	Third
B	0.64+1		0.68+1	3.32	**3.32-1 = 2.32**	First
C	0.42	0.71		1.13	**1.13-1.72 = -0.59**	Second
Sum of columns	2.06	1	1.72			

Next step is the summation of values on each row. The result indicates the *dominance* of the project corresponding to each row. Thus, project A has a dominance value of 0.33.

Next step is the summation of values on each column. The result indicates the *subordination* of the project corresponding to each column. Thus project A, has a subordinate value of 2.06.The net difference between dominant and subordinate values for a same project gives its importance. In this case, Project B is a clear winner, followed by Project C and Project A.

Thus, the ranking using the PDM mode: **Project B - Project C - Project A**
Observe that the two procedures (ERM mode and PDM mode), deliver the same ranking.

In a more elaborate scenario with say 17 projects, probably there will be as a result a subset of selected projects with values >0, and a subset of unselected projects with values = 0. In this case both ERM and PDM modes will identify the same selected and unselected projects and in addition the same rankings for those selected.

3 SIMUS FEATURES TO IMPROVE ACCURACY AND MODELLING

It is detailed here a list of the model features - that this author believes are not present in other methods - to generate results that are subjectivity free, and that in some extent reflect more precisely real life situations and circumstances, and in so doing helping the DM.

3.1 Regarding subjectivity

It has been stressed at the beginning of this paper that subjectivity is the culprit for which different methods, even when based on mathematical procedures do not coincide in their results when analyzing a problem. Subjectivity is a product of uncertainties and materializes in different ways according to the methods, such as establishing weights for criteria, using personal preferences, establishing thresholds to determine when a project outranks another, selecting a certain distance scheme to reach a solution, etc. LP does not need any of these subjective estimates, except in the selection of the number of criteria and their definitions (that is the decision matrix), which is a common ground to all methods. Consequently,it is believed that LP possesses a definitive advantage, as well as some drawbacks in other aspects, for instance, in working only with one objective function, which is not realistic in many cases. SIMUS, which is heavily based in LP, takes advantage of LP properties regarding absence of subjectivity and tries to overcome the drawbacks.

3.2 Regarding reliability

Probably the most important feature in this model is that for each problem it simultaneously delivers two solutions. This is a unique and very important attribute amongst MCDM models, because considering that if two different procedures deliver the same set of selected projects and the same ranking, it means a large probability that the 'right' result has been achieved. It is equivalent to treat a problem with two different methods and reaching the same outcome. Naturally, this does not guaranty that the method reaches the optimal solution, even if it exists, but obviously offers an additional dimension for acceptance when compared with other methods.

3.3 Approximation to reality

The model acknowledges the fact that everything in our world is limited or restricted in some way or the other, be it funds, man-hours, level of contamination, public acceptance, annual budgets, etc., which in most systems are ignored. In this model these limitations are taken into account by establishing limits or thresholds for each criterion, but which are not arbitrary, because they depend on economical, technical, social and environmental conditions that will be affected by a project. These limits can be physical and known quantities such as available funds, storage capacity, allowable contamination levels, maximum amount of energy to be generated, etc., and corresponding to *quantitative* criteria. However, normally there are also *qualitative* criteria such as for instance *'Maximize Disposable Income'*, or *'Minimize Poverty'*, which do not posses limits, as they are just 'wishes' or goals; however, the method generates limits ased on the extreme values of the respective cardinals. Therefore, the model is capable of representing reality more faithfully than other methods.

3.4 DM participation

The method allows - and in fact demands - strong DM participation, but not related with subjective figures or personal preferences, conditions, or assumptions, unless data come from experts. His duties are associated to the construction of the decision matrix and keeping an eye when examining every partial result and its significance related to selected projects. As a matter of fact, this characteristic allows for solving problems in Group Decision-Making, where the DM participates jointly with his/her staff, proposing amendments and modifications, and most important examining and quantifying these outcomes, which allow for accepting or rejecting changes proposed by any member of the team.

3.5 Projects discrimination

In a portfolio of projects, it usually occurs that two or more projects show the same or very close scores. If this happens, a model is not helping in project selection because it does not discriminate, and then putting the DM in a difficult situation leaving him at square one. SIMUS has various procedures to break these ties. This author is not aware that this discriminatory power exists in other methods.

3.6 Projects dependency

In many situations in which there are several projects, a condition or restriction may exist establishing that if project B is selected, then project D, must also be selected because they are complementary. The same happens when for whatever reasons a project H can only be started when project J is finished, that is, there is a functional or construction dependency between projects that must be honoured. These aspects are easily managed by SIMUS. Again, this author is not aware that this consideration of dependencies exists in other methods.

Considering these comments there is no doubts that SIMUS has the ability of modelling more faithfully actual scenarios than most popular MCDM methods.

4 TESTING

The model has been tested in about 75 different large and small projects involving many different activities such as large Hydroelectric projects (Argentina and Nepal), Non-conventional sources of energy (Canada), Manufacturing and Risk (Spain), Urban highways (Argentina), Urban planning (various cities in the Americas and Europe), City rehabilitation (Spain, Mexico and Palestine), City Planning (China), Environmental indicators (Canada), Social housing development (Ghana, World Bank), Doctoral Thesis (Spain and Cuba), Location analysis (Italy), Airport expansion (The Netherlands), Land use (Taiwan), Recycling policies (Canada), SWOT analysis and Business Administration (Valencia Polytechnic University, Spain), as well as teaching examples on People selection, Crops planning, Broadband options, etc. SIMUS software in β version (Lliso, 2014)[5]is in the Web and freely available.

5 CONCLUSION

This paper proposes and explains a new method called SIMUS for MCDM, which addresses two main observable aspects, namely *subjectivity*, which provokes different results according to the method and the person doing the analysis, and *lack of a proper modelling* of an actual scenario to represent it more faithfully. For the first aspect the proposed model is explained stressing the fact that it does not need any subjective consideration. It also demonstrates step by step how SIMUS generates more accurate results, based in the concordance of two different procedures. The paper comments the different advantages of the proposed method to represent reality more accurately, by supporting scenarios where facts such as project discrimination is needed, project dependency considered, limited resources included, and group participation admitted. Due to page limit it was not possible to illustrate each one of these advantages, which will be the matter of another paper. Section 4 provides information about the different types of projects the different fields where the model has been applied.

References

[1] Brans, J., Vincke, Ph. (1985). A preference ranking organization method. The PROMETHEE method, Management Science 31, 647-656

[2] Dantzig, G. (1948). Linear Programming and extensions. United States Air Force

[3] Frontline Systems. http://www.solver.com/

[4] Hwang, C., Yoon, K. (1981). Multiple attribute decision making: Methods and applications, A State- of - the - Art Survey. Springer-Verlag, Berlin

[5] Lliso, P. (2014) Multicriteria Decision-Making by Simus http://decisionmaking.esy.es/

[6] Munier, N. (2011a). A Strategy for Using Multicriteria Analysis in Decision-Making: A Guide for Simple and complex Environmental Projects. Springer. USA.

[7] Munier, N. (2011b) Tesis Doctoral – Procedimiento fundamentado en la Programación Lineal para la selección de alternativas en proyectos de naturaleza compleja y con objetivos múltiples- Universidad Politécnica de Valencia – España

[8] Roy B. y Bouyssou D. (1993): Aide multicritere a la decision: methodes et cas, Editorial Economica, Paris, Francia.

[9] Roy, B. (1968) Classement et choix en presence de points de vue multiples (la method ELECTRE). Rev. Fr. Inf. Rech. 2 (8) p. 57-75

[10] Saaty, T. (1980): Multicriteria decision making - The analytic hierarchy process, McGraw-Hill, New York

[11] Saaty, T. (1996). Decision Making with Dependence and Feedback: The Analytic Network Process. Pittsburgh, Pennsylvania: RWS Publications. ISBN 0-9620317-9-8.

A NEW LINEARIZATION APPROACH FOR SOLVING MULTI OBJECTIVE LINEAR FRACTIONAL PROGRAMMING PROBLEMS

Tunjo Perić
University of Zagreb, Faculty of Economics & Business
Kennedyjev trg 6, 10000 Zagreb, Croatia
tperic@efzg.hr

Zoran Babić
University of Split, Faculty of Economics
Cvite Fiskovića 5, 21000 Split, Croatia
babic@efst.hr

Abstract: In this paper we propose a new linearization approach for solving multi objective linear fractional programming problems by using goal programming methods. The applicability of the proposed methodology has been tested on the example of financial structure optimization of a company. The obtained results show the advantages of the proposed methodology compared to the existing methods. The proposed methodology is simple for both the analyst and the decision maker. Determining objective function weights by the decision maker, the obtained solutions reflect the decision maker's preferences.

Keywords: Multi objective linear fractional programming; goal programming; financial structure optimization.

1 INTRODUCTION

If the economic problem goals are expressed as the ratio of two economic magnitudes, and if the parameters and variables in the model are linear, then the optimization of economic problems requires multi objective linear fractional programming (MOLFP).

The problem of linear fractional programming with one objective function was extensively researched and efficient methods were developed for solving such problems ([3], [5]).

However, in MOLFP problems determining efficient (Pareto optimal) solutions is technically demanding. Solving multi objective linear fractional programming models is limited to a small number of multi objective programming methods which are not efficient enough either from the viewpoint of the analyst or the decision maker ([1], [2], [6], [7]).

A special problem arises in the application of goal programming methods as by adding the deviation variables d^- and d^+ to the fractional linear functions to form the constraints of the goal programming model, nonlinear constraints are obtained which cause problems in model solving. Several methods have been developed that use goal programming to solve multi objective fractional linear programming model ([4], [8], [9]), but there are few studies in which these methods are applied and tested in solving real economic problems ([7], [9]).

In this paper a new goal programming approach based on the linearization of linear fractional objective functions for solving multi objective fractional programming problems has been proposed. In the proposed method the decision maker is only asked to provide the information on the relative importance and the aspired value of the objective functions. The proposed method has been tested on the example of financial structure optimization of a company.

2 GOAL PROGRAMMING METHODOLOGY FOR SOLVING MOLFP PROBLEMS

2.1. MOLFP model

A MOFLP model can be written as ([8])

$$\max_{x \in S} [f_1, f_2, ..., f_K], \quad S = \{Ax \leq = \geq b\}, \tag{1}$$

where $f_k = \left(\sum_{j=1}^{n} c_{jk} x_j + c_0^k \right) / \left(\sum_{j=1}^{n} d_{jk} x_j + d_0^k \right)$, $k = 1, 2, ..., K$, are linear fractional functions, A is a m x n matrix, x is a n – dimensional nonnegative variable vector, b is a m – dimensional constraint vector.

To solve the model (1) by the goal programming methods the linear fractional objective functions have to be linearized. Many different methods of the linearization of objective functions have been proposed ([9], [10]).

2.2. The proposed goal programming model

To solve the model (1) by the goal programming method we need to linearize the objective functions. Let

$$m_k = \min_{x \in S} f_k(x), \quad M_k = \max_{x \in S} f_k(x), \quad k = 1, 2, ..., K.$$

We have

$$m_k \leq f_k(x) \leq M_k \implies 0 \leq f_k(x) - m_k \leq M_k - m_k.$$

If we divide these inequalities by $M_k - m_k > 0$, we obtain

$$0 \leq \frac{f_k(x) - m_k}{M_k - m_k} \leq 1.$$

Thus, if we define

$$\tilde{f}_k(x) = \frac{f_k(x) - m_k}{M_k - m_k} = \frac{\frac{c_k(x)}{d_k(x)} - m_k}{M_k - m_k} = \frac{c_k(x) - m_k d_k(x)}{(M_k - m_k) d_k(x)}, \quad k = 1, 2, ..., K, \tag{2}$$

then we have

$$0 \leq \tilde{f}_k(x) \leq 1, \quad \min_{x \in S} \tilde{f}_k(x) = 0, \quad \max_{x \in S} \tilde{f}_k(x) = 1, \tag{3}$$

where the optimal points for min and max for $\tilde{f}_k(x)$ are the same as those for $f_k(x)$. In this way we have normalized the objective functions in the sense that the functions $f_k : S \to [m_k, M_k]$ with different sets of values are transformed into the functions $\tilde{f}_k : S \to [0,1]$ with the same set of values. Generally speaking we have done an affine transformation of the given objective functions,

$$f_k(x) \to \tilde{f}_k(x) = \alpha f_k(x) + \beta = \frac{1}{M_k - m_k} f_k(x) - \frac{m_k}{M_k - m_k}, \quad k = 1, 2, ..., K.$$

The same is true for inverse transformation,

$$\tilde{f}_k(x) \to f_k(x) = \tilde{\alpha} \tilde{f}_k(x) + \tilde{\beta} = (M_k - m_k) \tilde{f}_k(x) + m_k, \quad k = 1, 2, ..., K.$$

Now we can transform the initial problem into the problem

$$\max_{x \in S} \left(\tilde{f}_1(x),\ \tilde{f}_2(x),\ ...,\ \tilde{f}_K(x) \right). \tag{4}$$

Let

$$\tilde{f}_k(x) = \frac{\tilde{c}_k(x)}{\tilde{d}_k(x)},\ k = 1, 2, ..., K.$$

By this definition, $\tilde{c}_k(x)$ and $\tilde{d}_k(x)$ are not uniquely defined. Using (2) we have

$$\tilde{c}_k(x) = \tau \left[c_k(x) - m_k d_k(x) \right], \quad \tilde{d}_k(x) = \tau (M_k - m_k) d_k(x), \quad \tau \neq 0.$$

Let

$$\delta_k = \operatorname*{sgn}_{x \in S} \left(d_k(x) \right), \quad k = 1, 2, ..., K.$$

Note that δ_k is well defined because we have assumed that $d_k(x)$ has a constant sign on *S*. In the above expressions we take such τ for which $\operatorname{sgn}(\tau) = \delta_k$, or equivalently $\tau = \delta_k |\tau|$, holds. This yields $\tilde{d}_k(x) > 0$ and, using (2), $\tilde{d}_k(x) \geq \tilde{c}_k(x) \geq 0$. Let

$$\tilde{z}_k(x) = \tilde{d}_k(x) - \tilde{c}_k(x) = \tau \left[M_k d_k(x) - c_k(x) \right] = \delta_k |\tau| \left[M_k d_k(x) - c_k(x) \right], \quad k = 1, 2, ..., K.$$

Obviously $\tilde{z}_k(x)$ is linear function and we have $\tilde{z}_k(x) \geq 0,\ x \in S$ and also $\min_{x \in S} \tilde{z}_k(x) = 0$. Now, we normalize $\tilde{z}_k(x)$ by dividing it by its maximal value on the set S. Thus, we define

$$z_k(x) = \frac{\tilde{z}_k(x)}{\max_{y \in S} \tilde{z}_k(y)}.$$

Since

$$\max_{y \in S} \tilde{z}_k(y) = \max_{y \in S} \tau \left[M_k d_k(y) - c_k(y) \right] = |\tau| \max_{y \in S} \delta_k \left[M_k d_k(y) - c_k(y) \right],$$

we have

$$z_k(x) = \frac{\delta_k \left[M_k d_k(x) - c_k(x) \right]}{N_k}, \quad N_k = \max_{y \in S} \delta_k \left[M_k d_k(y) - c_k(y) \right], \quad k = 1, 2, ..., K, \tag{5}$$

which is now uniquely defined (it does not depend on τ any more). Note also that $\delta_k \left[M_k d_k(x) - c_k(x) \right] = \left| M_k d_k(x) - c_k(x) \right|$. By this construction we obtain

$$0 \leq z_k(x) \leq 1, \quad \min_{x \in S} z_k(x) = 0, \quad \max_{x \in S} z_k(x) = 1, \quad k = 1, 2, ..., K, \tag{6}$$

where minimum is attained in the same point(s) where $\tilde{f}_k(x)$ and $f_k(x)$ achieve maximum. Now, we state the MOLFPP (4) in the following linear form,

$$\min_{x \in S} \left(z_1(x),\ z_2(x),\ ...,\ z_K(x) \right). \tag{7}$$

The problems (4) and (7) are equivalent in the sense that for each $k = 1, 2, ..., K$ we have

$$\max_{x \in S} f_k(x) = f_k(x^{(i)}) \ \Leftrightarrow \ \max_{x \in S} \tilde{f}_k(x) = \tilde{f}_k(x^{(i)}) = 1 \ \Leftrightarrow \ \min_{x \in S} z_k(x) = z_k(x^{(i)}) = 0.$$

Now we can form the following goal programming model:

$$\min_{x \in S'} \ g_k(w_k d_k^-,\ w_k d_k^+),\ S' = \left\{ z_k + d_k^- - d_k^+ = \bar{z}_k;\ Ax \leq = \geq b \right\},\ k = 1, \ldots, K \tag{8}$$

where $\bar{z}_k$ is decision maker's aspired value of *k*th objective function. In the model (8) both the objective function and the constraints are linear. This model can be solved by using the following linear programming model ([11]):

$$\min_{(x,\lambda)\in S''} \lambda,\ S'' = \left\{ \lambda \geq w_k \left(d_k^- + d_k^+ \right); z_k + d_k^- - d_k^+ = \bar{z}_k; Ax \leqq\geqq b \right\},\ k = 1, ..., K \qquad (9)$$

3. PRACTICAL EXAMPLE: FINANCIAL STRUCTURE OPTIMIZATION

3.1. MOLFP model

The data for the problem are taken from [10]. Based on the data the following model is formed:

$$\max_{x\in S} \left(f_1,\ f_2,\ f_3, f_4 \right),$$

$$\text{where } f_1(x_1, x_3) = \frac{-x_1}{x_3},\ f_2(x_3, x_4, x_5, x_6) = \frac{-x_3 - x_4}{x_5 + x_6},\ f_3(x_1, x_2) = \frac{60}{x_1 + x_2},\ f_4(x_6) = \frac{x_6}{60}, \qquad (10)$$

$$S = \left\{ \begin{array}{l} \left(x_1, x_2, x_3, x_4, x_5, x_6 \right) \in R^6 : x_1, x_2, x_3, x_4, x_5, x_6 \geq 0, 150 \leq x_1 \leq 250, x_2 \leq 300, \\ x_1 + x_2 \geq 350,\ 75 \leq x_3 \leq 300, 100 \leq x_4 \leq 300,\ x_3 + x_4 \geq 250, \\ 75 \leq x_5 \leq 125,\ 100 \leq x_6 \leq 140,\ x_1 + x_2 = x_3 + x_4 + x_5 + x_6 . \end{array} \right\}$$

Note that minimization of current and debt ratio is turned into maximization by changing the sign in f_1 and f_2.

3.2. MOLFP model solving

The extreme values m_i and M_i of the model (10) are given in Table 1.

Table 1. Marginal solutions

Extremes m_i and M_i	$(x_1, x_2, x_3, x_4, x_5, x_6, f_i)$
m_1	(250, 300, 75, 210, 125, 140, −10/3)
M_1	(150, 300, 175, 100, 75, 100, −6/7)
m_2	(250, 300, 275, 100, 75, 100, −15/7)
M_2	(250, 265, 150, 100, 125, 140, −50/53)
m_3	(250, 300, 75, 210, 125, 140, 6/55)
M_3	(250, 175, 75, 175, 75, 100, 12/85)
m_4	(250, 300, 75, 300, 75, 100, 5/3)
M_4	(250, 300, 75, 260, 75, 140, 7/3)

The normalization is done by using (5) and the data from Table 1. Now we have

$$\delta_1 \left[M_1 d_1(x) - c_1(x) \right] = (+1) \left| -\frac{6}{7} x_3 - (-x_1) \right| = \frac{7x_1 - 6x_3}{7},$$

$$\delta_2 \left[M_2 d_2(x) - c_2(x) \right] = (+1) \left[-\frac{50}{53} (x_5 + x_6) - (-x_3 - x_4) \right] = \frac{53x_3 + 53x_4 - 50x_5 - 50x_6}{53},$$

$$\delta_3 \left[M_3 d_3(x) - c_3(x) \right] = (+1) \left[\frac{12}{85} (x_1 + x_2) - 60 \right] = \frac{12x_1 + 12x_2 - 5100}{85},$$

$$\delta_4 \left[M_4 d_4(x) - c_4(x) \right] = (+1) \left[\frac{7}{3} \cdot 60 - x_6 \right] = 140 - x_6,$$

and $N_1 = 1300/7,\ N_2 = 11125/53,\ N_3 = 1500/85,\ N_4 = 40,$

$$z_1 = \frac{7x_1 - 6x_3}{1300},\ z_2 = \frac{53x_3 + 53x_4 - 50x_5 - 50x_6}{11125},\ z_3 = \frac{x_1 + x_2 - 425}{125},\ z_4 = \frac{140 - x_6}{40}.$$

The problem (8) is given in the form:

$$\min_{(x,d_k^-,d_k^+)\in S'} \left(w_1(d_1^- + d_1^+); w_2(d_2^- + d_2^+); w_3(d_3^- + d_3^+); w_4(d_1^- + d_1^+) \right),\ k = 1,2,3,4$$

$$S' = \left\{ \begin{array}{l} x = (x_1, x_2, x_3, x_4, x_5, x_6) \in S,\ \dfrac{7x_1 - 6x_3}{1300} + d_1^- - d_1^+ = 1.05, \\ \dfrac{53x_3 + 53x_4 - 50x_5 - 50x_6}{11125} + d_2^- - d_2^+ = 1.1,\ \dfrac{x_1 + x_2 - 425}{125} + d_3^- - d_3^+ = \\ 1.08,\ \dfrac{140 - x_6}{40} + d_4^- - d_4^+ = 1.1\ \ d_k^-, d_k^+ \geq 0,\ k = 1,2,3,4. \end{array} \right\}. \quad (11)$$

The model (11) is transformed to the following linear programming model:

$$\min_{(x,d_k^-,d_k^+,\lambda)\in S''} \lambda$$

$$S'' = \left\{ (x, d_k^-, d_k^+) \in S',\ \lambda \geq 0,\ \lambda \geq w_k(d_k^- + d_k^+), k = 1,2,3,4 \right\}. \quad (12)$$

The aspired value of the objective functions (1.05; 1.1; 1.08; 1.1) is obtained from the decision maker. For different values for w_k, k = 1, 2, 3, 4, the following solutions have been obtained:

Table 3. Goal programming solutions

GP solution	*Variable values*		f_1	f_2	f_3	f_4
Model (12) w_1 = w_2 = w_3 = w_4 = 0.25	x_1 = 153.19	x_2 = 271.82	1.0213	1.4286	0.1412	1.6667
	x_3 = 150.00	x_4 = 100.00				
	x_5 = 75.00	x_6 = 100.00				
Model (12) w_1 = 0.4, w_2 = 0.3, w_3 = 0.2, w_4 = 0.1	x_1 = 150.00	x_2 = 300.00	1.0000	1.2500	0.1333	2.0833
	x_3 = 150.00	x_4 = 100.00				
	x_5= 75.00	x_6 = 125				
Model (12) w_1 = 0.1, w_2 = 0.2, w_3 = 0.3, w_4 = 0.4	x_1 = 219.17	x_2 = 205.83	1.4611	1.4286	0.1412	1.6667
	x_3 = 150.00	x_4 = 100.00				
	x_5 = 75.00	x_6 = 100.00				
Model (12) w_1 = 0.1, w_2 = 0.3, w_3 = 0.4, w_4 = 0.2	x_1 = 166.85	x_2 = 300.00	0.8742	1.6614	0.1285	1.6667
	x_3 = 190.85	x_4 = 100.00				
	x_5 = 75.00	x_6 = 100.00				

The application of the proposed model to solve financial structure optimization problem gives different solutions for different values of the objective function weights. In doing so greater weight of the objective function gives greater value to the objective function, and vice versa (see the Table 3). Therefore the obtained solutions express the decision maker's preferences. The decision maker can give information of the objective function weights as well as information on the preferred value of the objective functions.

Conclusion

This paper proposes the methodology of linearization of linear fractional objective functions in order to solve MOLFP problem by goal programming methods. The proposed methodology has been tested on the problem of optimization of a company's financial structure using the weighted sum deviations approach of the goal programming method.

The obtained results reveal the possibility of an efficient application of the proposed methodology in solving the given problem.

There are many advantages of the proposed methodology compared to the existing methods:

- The methodology is simple for both the analyst and the decision maker.
- The decision maker can determine objective function weights, and thereby the obtained solutions reflect the decision maker's preferences.
- This method allows the analyst to form a set of efficient solutions by varying the objective function weights, from which the decision maker can choose the preferred one.

For the next research we propose using the proposed linearization technique in solving practical multi objective linear fractional programming problems by different multi objective programming methods to test their efficiency.

Acknowledgement

This paper has been **fully** supported by Croatian Science Foundation under the project STRENGTHS Number 9402.

References

[1] Caballero, R. and Hernandez, M. (2006). Restoration of efficiency in a goal programming problem with linear fractional criteria. European Journal of Operational Research, Vol. 172: 31–39.

[2] Chakraborty, M. and Gupta, S. (2002). Fuzzy mathematical programming for multi objective linear fractional programming problem. Fuzzy Sets and Systems, Vol. 125: 335–342.

[3] A. Charnes, A. and W.W. Cooper, W.W. (1961). Management Models of Industrial Applications of Linear Programme (Appendix B) Vol. 1, Wiley, New York.

[4] Kao, C. and Liu, S.T. (2001). Fractional programming approach to fuzzy weighted average. Fuzzy Sets and Systems, Vol. 120: 435–444.

[5] Mehrjerdi,Y.Z. (2011). Solving fractional programming problem through fuzzy goal setting and approximation. Applied Soft Computing, Vol. 11: 1735-1742.

[6] Metev, B. and Gueorguieva, D. (2000). A simple method for obtaining weakly efficient points in multiobjective linear fractional programming problems. European Journal of Operational Research, Vol. 126: 386-390.

[7] Mishra, B. and Singh, S.R. (2013). Linear Fractional Programming Procedure for Multi Objective Linear Programming Problem in Agricultural System. International Journal of Computer Applications, Vol. 60: 0975-8887.

[8] Pal, B.B., Moitra, B.N. and Maulik, U. (2003). A goal programming procedure for fuzzy multiobjective linear fractional programming problem. Fuzzy Sets and Systems, Vol. 139: 395-405.

[9] Perić, T. and Babić, Z. (2014) A new goal programming approach for multi objective fractional linear programming, Proceedings of the 25th DAAAM International Symposium on Intelligent Manufacturing and Automation, DAAAM 2014, (Editor: Branko Katalinić), Vienna, pp. 1-10.

[10] Perić, T. and Babić, Z. (2012). Financial structure optimization by using a goal programming approach. Croatian Operational Research Review, Vol. 3:150-162.

[11] Zimmermann, H.-J. (1978). Fuzzy programming and linear programming with several objective functions. Fuzzy Sets and Systems, Vol. 1: 45-55.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 4:

Econometric Models and Statistics

ALGORITHMS FOR MAXIMUM LIKELIHOOD ESTIMATION OF GARCH MODELS

Josip Arnerić
University of Zagreb, Faculty of Economics and Business Zagreb
Trg J.F.Kennedyja 6, Zagreb, Croatia
E-mail: jarneric@efzg.hr

Ivana Lolić
University of Zagreb, Faculty of Economics and Business Zagreb
Trg J.F.Kennedyja 6, Hrvatska
E-mail: ilolic@efzg.hr

Tea Poklepović
University of Split, Faculty of Economics
Cvite Fiskovića 5, Split, Croatia
E-mail: tpoklepo@efst.hr

Abstract: Analytical solution for the log-likelihood function maximization using first and second derivatives is too complex within GARCH type models. This paper examines NR, BHHH, BFGS and DFP as commonly used numerical algorithms. As solutions of different algorithms are sensitive to the initial values and convergence criteria, the contribution of this paper is to determine which algorithm gives the most stable estimates of the GARCH(1,1) parameters with application to time series of daily returns from Zagreb Stock Exchange. This paper reveals advantages and disadvantages of different iteration procedures according to the approximation of the Hessian matrix.

Keywords: GARCH(1,1), maximum likelihood estimation, Newton-Raphson, BHHH, BFGS, DFP.

1 INTRODUCTION

Maximum likelihood estimation (MLE) is one of the fundamental estimation approach in statistical estimation theory. Researchers have a preference for MLE since it has desirable statistical properties as efficiency and consistency. In most cases explicit formula for the parameter estimates is not available, so various algorithms are employed in order to yield the optimal solution. The fastest algorithm applicable to MLE problem is Newton-Raphson. It uses the first and the second derivative of the likelihood function resulting with rather rare usage in practical applications. The algorithms that are in researcher's main focus use secant updates as DFP (Davidon, Fletcher, and Powel) and BFGS (Broyden, Fletcher, Goldfarb and Shanno) algorithms, or statistical approximation as BHHH algorithm (Berdnt, Hall, Hall and Hausman). Trough years various updates are proposed as switching algorithms (e.g. using one algorithm in 5 iterations, than second etc.), line search technique or trust region approach. Various researchers have dealt with algorithms designed to solve specific MLE problem. Authors in [1] and [14] have investigated algorithms in context of volatility models, particularly about GARCH type models. They have concluded superiority of BHHH algorithm in comparison with DFP and BFGS method.
Probabilistic choice models are considered in [5]. The author compares performance of quasi-Newton algorithms (DFP, BFGS, and BHHH) by including model switching and trust regions approach. Author concludes that choice of specific secant method does not appear to be critical, while setting the initial value of Hessian seems to increase speed significantly.
Authors Mai, Bastin and Toulouse in [10] have investigated maximum likelihood estimation in the framework of discrete choice models (mixed logit and logit based route choice models). They reviewed standard trust region and line search algorithms with emphasis on

different Hessian approximations. They conclude that predictive approach outperforms switching approach and classical optimization algorithms (Newton, DFP, BFGS and BHHH). Carling and Söderberg discussed in [6] maximum likelihood estimation in econometric duration analysis with examining time and number of iterations upon convergence.
This paper attempts to fill a niche which algorithm has shown to be the best for estimating parameters in GARCH(1,1) framework using STATA software, i.e. to determine which algorithm gives the most stable parameters estimates that are robust on changing the initial values and convergence criteria.

2 MAXIMUM LIKELIHOOD ESTIMATION OF GARCH(1,1) MODEL

Most financial time series experience stylized facts, i.e. time varying variance, volatility clustering, ARCH effects (heteroskedasticity), etc. Models used for describing daily volatility are GARCH(p,q) type models. Due to simplicity and sufficiency of GARCH(1,1) to capture all statistical properties of returns standard GARCH(1,1) model is analyzed:

$$\begin{aligned} y_t &= \mu + \varepsilon_t \quad ; \ \varepsilon_t = u_t\sqrt{\sigma_t^{\,2}} \ ; u_t \sim i.i.d \\ \sigma_t^{\,2} &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^{\,2} + \beta_1 \sigma_{t-1}^{\,2} \end{aligned} \tag{1}$$

Log-likelihood function for GARCH(1,1) model assuming Gaussian distribution is:

$$\ln L(\beta) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log\sigma_t^{\,2} - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon_t^{\,2}}{\sigma_t^{\,2}}. \tag{2}$$

In comparison with linear regression model, the system of equations (by setting all partial derivatives equal to zero) does not yield analytical formula for estimator of unknown population parameters $\beta = [\mu \quad \alpha_0 \quad \alpha_1 \quad \beta_1]'$. Therefore, numerical approach is necessary to for maximizing the log-likelihood. Various numerical algorithms can be employed to solve this problem, but the choice of specific algorithm depends on the objective function (model complexity), prior knowledge of initial values and convergence criteria.

3 OPTIMIZATION ALGORITHMS

Since the log-likelihood is infinitely differentiable function, Newton algorithm can be employed. Let β_i be an estimate of β in i-th iteration, than $i+1$-th iteration is computed with second order Taylor's approximation of $\ln L(\beta)$ arround β_i (evaluated in β_{i+1}):

$$\ln L(\beta_{i+1}) = \ln L(\beta_i) + (\beta_{i+1} - \beta_i)' g_i + \frac{1}{2}(\beta_{i+1} - \beta_i)' H_i (\beta_{i+1} - \beta_i) \tag{3}$$

After calculating the first order condition and rearranging, the iterative algorithm is:

$$\beta_{i+1} = \beta_i + (-H_i)^{-1} g_i \tag{4}$$

where $g_i = \left.\frac{\partial \ln L(\beta)}{\partial \beta}\right|_{\beta=\beta_i}$ is the gradient vector and $H_i = \left.\frac{\partial^2 \ln L(\beta)}{\partial \beta^2}\right|_{\beta=\beta_i}$ is Hessian matrix. The procedure is known as Newton-Raphson's algorithm. The algorithm is quadratically convergent, resulting with very fast descent toward the optimal value. The main disadvantages of this method is the usage of the exact Hessian matrix and highly dependent convergence on initial values (local convergence). The analytical formulas for Hessian matrix are not easily computed so various upgrades of the algorithm were introduced. To assure that in each iteration log-likelihood is increased, scalar λ_i is added into the equation (4):

$$\beta_{i+1} = \beta_i + \lambda_i (-H_i)^{-1} g_i . \tag{5}$$

The vector $d_i = (-H_i)^{-1} g_i$ is called direction vector, λ_i is step length and algorithm in (5) is usually referenced as modified Newton's algorithm. Scalar λ_i can be calculated using simpler optimization procedure as golden ratio algorithm [7].

Davidon, Fletcher, and Powel have discovered the following relation:

$$H_{i+1}^{-1} = H_i^{-1} + \frac{p_i p_i'}{p_i' q_i} - \frac{H_i^{-1} q_i q_i' H_i^{-1}'}{q_i' H_i^{-1} q_i} \quad (6)$$

where $q_i = g_{i+1} - g_i$ and $p_i = \beta_{i+1} - \beta_i$. When Hessian matrix in relation (5) is calculated with (6), the algorithm is usually called DFP method.

Broyden, Fletcher, Goldfarb and Shanno have developed the following iteration procedure:

$$H_{i+1}^{-1} = H_i^{-1} + \left(1 + \frac{q_i' H_i^{-1} q_i'}{p_i' q_i}\right) \frac{p_i p_i'}{p_i' q_i} - \frac{p_i q_i' H_i^{-1} + H_i^{-1} q_i p_i'}{q_i' p_i}. \quad (7)$$

This method is usually called BFGS.

Berdnt, Hall, Hall and Hausman have proposed slightly different approach in [4], founded on the information matrix equality. If the model is correctly specified and assuming that β_0 is the vector of true model parameters, the following equality is valid:

$$I(\beta_0) = -E[H(\beta_0)] \quad (8)$$

where $I(\beta_0)$ is the Fisher information matrix, defined as the covariance-variance matrix of the score at β_0. $H(\beta_0)$ denotes Hessian matrix of likelihood function, while the gradient vector is defined as:

$$g(y|\beta) = \frac{\partial}{\partial \beta} \ln f(y|\beta) \quad (9)$$

The score function can be differently written in expanded formulation as:

$$g(y_1, ..., y_T|\beta) = \sum_{t=1}^{T} \frac{\partial}{\partial \beta} \ln f(y_t|\beta) \quad (10)$$

where f denotes individual density. The information matrix can be differently expressed as:

$$I(\beta_0) = Var(g(y|\beta_0)) = E\left[\frac{\partial}{\partial \beta} \ln f(y|\beta_0) \frac{\partial}{\partial \beta} \ln f(y|\beta_0)'\right] \quad (11)$$

For a finite sample, the information matrix $I(\beta)$ can be consistently estimated as

$$I(\hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \beta} \ln f(y_t|\hat{\beta}) \frac{\partial}{\partial \beta} \ln f(y_t|\hat{\beta})' \quad (12)$$

According to these findings, Hessian matrix can be approximated with formula:

$$H_{i+1} = -\sum_{t=1}^{T} h_t h_t' \quad (13)$$

The equation (13) is usually called outer product of gradients (OPG).

All discussed algorithms converge to the optimal value if starting value is "near" the global maximum. If this is not the case, algorithm can converge to a local maximum. The best possible approach is to have a good feeling about the value of parameters. Most algorithms have predefined maximum number of iterations to prevent endless loop. If algorithm stops due to this reason, maximum value is not probably found. Other numerical convergence criterion considers increase in (absolute) parameter value $|\beta_{i+1} - \beta_i|$ or increase in (absolute) function value $|\ln L(\beta_{i+1}) - \ln L(\beta_i)|$.

4 COMPARISON RESULTS

Four algorithms Newton-Raphson, BHHH, BFGS and DFP are compared within parameters estimation of GARCH(1,1) model using daily returns of Croatian stock market index CROBEX from January 05, 2010 to December 20, 2014 (total of 1246 observations).

Table 1: Comparison of Newton-Raphson, BHHH, BFGS and DFP algorithms within parameters estimation of GARCH(1,1) model using different initial values and different convergence criteria

Panel A)	*Convergence criteria* $\lvert\beta_{i+1}-\beta_i\rvert < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.1,\ \beta_1=0.8$							
	Recursion started with expected unconditional variance				*Recursion started with estimated variance from OLS residuals*			
	NR	*BHHH*	*BFGS*	*DFP*	*NR*	*BHHH*	*BFGS*	*DFP*
μ	-0.000039	-0.000039	-0.000039	-0.000039	-0.000045	-0.000045	-0.000045	-0.000045
α_0	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021
α_1	0.063166	0.063158	0.063139	0.063141	0.063149	0.063170	0.063145	0.063151
β_1	1.328317	1.328400	1.328618	1.328601	1.328732	1.328550	1.328728	1.328725
$\ln L$	4386.927	4386.927	4386.927	4386.927	4386.922	4386.922	4386.922	4386.922
No. iter.	6	41	28	27	7	44	26	28
Panel B)	*Converg. criteria* $\lvert\ln L(\beta_{i+1})-\ln L(\beta_i)\rvert < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.1,\ \beta_1=0.8$							
μ	-0.000039	-0.000039	-0.000039	-0.000039	-0.000045	-0.000045	-0.000045	-0.000045
α_0	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021
α_1	0.063166	0.063158	0.063139	0.063141	0.063149	0.063170	0.063145	0.063151
β_1	1.328317	1.328400	1.328618	1.328601	1.328732	1.328550	1.328728	1.328725
$\ln L$	4386.927	4386.927	4386.927	4386.927	4386.922	4386.922	4386.922	4386.922
No. iter.	6	41	28	27	7	44	26	28
Panel C)	*Convergence criteria* $g_i^T(-H_i)^{-1}g_i < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.1,\ \beta_1=0.8$							
μ	-0.000039	-0.000039	-0.000039	Flat log likelihood (cannot find uphill direction)	-0.000045	-0.000045	Flat log likelihood (cannot find uphill direction)	
α_0	-0.000021	-0.000021	-0.000021		-0.000021	-0.000021		
α_1	0.063142	0.063149	0.063138		0.063149	0.063160		
β_1	1.328581	1.328502	1.328627		1.328732	1.328612		
$\ln L$	4386.927	4386.927	4386.927		4386.922	4386.922		
No. iter.	39	81	38		39	82		
Panel D)	*Convergence criteria* $\lvert\beta_{i+1}-\beta_i\rvert < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.2,\ \beta_1=0.7$							
μ	-0.000039	-0.000039	-0.000039	-0.000040	-0.000045	-0.000045	-0.000045	-0.000045
α_0	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021
α_1	0.063147	0.063136	0.063112	0.063151	0.063159	0.063146	0.063157	0.063135
β_1	1.328527	1.328669	1.328904	1.328507	1.328619	1.328788	1.328634	1.328890
$\ln L$	4386.927	4386.927	4386.927	4386.927	4386.922	4386.922	4386.922	4386.922
No. iter.	7	87	40	38	7	89	36	43
Panel E)	*Converg. criteria* $\lvert\ln L(\beta_{i+1})-\ln L(\beta_i)\rvert < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.2,\ \beta_1=0.7$							
μ	-0.000039	-0.000039	-0.000039	-0.000040	-0.000045	-0.000045	-0.000045	-0.000045
α_0	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021	-0.000021
α_1	0.063090	0.063136	0.063112	0.063151	0.063159	0.063146	0.063157	0.063135
β_1	1.329169	1.328669	1.328904	1.328507	1.328619	1.328788	1.328634	1.328890
$\ln L$	4386.927	4386.927	4386.927	4386.927	4386.922	4386.922	4386.922	4386.922
No. iter.	6	87	40	38	7	89	36	43

Panel F)	*Convergence criteria* $g_i^T\left(-H_i\right)^{-1}g_i < 0.0001$ *with initial values* $\mu=\alpha_0=0,\ \alpha_1=0.2,\ \beta_1=0.7$					
μ	-0.000039	-0.000039	Flat log likelihood (cannot find uphill direction)	-0.000045	-0.000045	Flat log likelihood (cannot find uphill direction)
α_0	-0.000021	-0.000021		-0.000021	-0.000021	
α_1	0.063139	0.063136		0.063149	0.063146	
β_1	1.328614	1.328669		1.328738	1.328788	
$\ln L$	4386.927	4386.927		4386.922	4386.922	
No. iter.	40	87		40	89	

Stability of the above-mentioned algorithms is examined due to different convergence criteria (relative change of parameters vector, relative change of the log-likelihood value between two successive iterations and gradient vector), different initial values of unknown parameters vector $\beta=\left[\mu \quad \alpha_0 \quad \alpha_1 \quad \beta_1\right]'$ and different setting of pre-sample variance σ_0^2 for starting recursion (either expected unconditional variance or estimated variance from OLS residuals). In theory the maximum of log-likelihood occurs when the gradient vector is zero. Namely, in practice the calculated gradient vector is never exactly zero, but can be very close. Therefore, $g_i^T\left(-H_i\right)^{-1}g_i$ is often used to evaluate convergence. If inequality $g_i^T\left(-H_i\right)^{-1}g_i < 0{,}0001$ is satisfied, the iterative process stops and the parameters at current iteration are considered as estimates. However, small changes in parameter values, with small increases in log-likelihood function could be evidence that convergence has been achieved. Even so, small changes in accompanied by a gradient vector that is not close to zero indicate that we are not effective in finding the maximum. To investigate if local maximum is the global optimum we should use different starting values and observe whether convergence occurs at the same parameter values. Empirical research has showed that initial vector of parameters as null-vector is not appropriate because it would result in portion of not concave log-likelihood function at null iteration.

For constant terms in mean equation and variance equation initial values are set to zero, i.e. $\mu=\alpha_0=0$, while for ARCH term α_1 and GARCH term β_1, different initial values are taken into account. Model GARCH(1,1) usually indicate high volatility persistence, i.e. the sum of parameters $\alpha_1+\beta_1$ is close to one, describing almost integrated behaviour of the conditional variance. In other words, it is expected that long time is needed for shocks in volatility to die out and that volatility reacts at low intensity on past market movements. Therefore, two combinations of initial values are introduced a) $\mu=\alpha_0=0,\ \alpha_1=0.1,\ \beta_1=0.8$ and b) $\mu=\alpha_0=0,\ \alpha_1=0.2,\ \beta_1=0.7$. Procedure starts with calculating gradient vector with respect to initial parameter values. At each iteration the step size is reduced (stepping backward) or increased (stepping forward) in purpose to calculate "new β". Step size is reduced when the initial step is "bad", and it is increased when the initial step is "good". Procedure stops at last iteration when a convergence criterion is satisfied and the last step is "ignored". Table 1 summarizes comparison results of parameters estimation according to NR, BHHH, BFGS and DFP algorithms.

5 CONCLUSION

This paper attempts to fill a niche which algorithm has shown to be the best for estimating parameters in standard GARCH(1,1) model, i.e. to determine which algorithm gives the most stable parameters estimates that are robust on changing the initial values and convergence criteria. All four algorithms (NR, BHHH, BFGS and DFP), have advantages and disadvantages. Newton-Raphson's algorithm uses the Hessian matrix which needs to be

computed in every iteration. If the second derivatives of log-likelihood function are (numerically) complex, this method should be avoided. DFP and BFGS algorithm use similar approach with approximating Hessian matrix and generating sequence of matrix, where the current one is computed from previous version. The procedure is highly dependent upon initial value. BHHH algorithm also uses Hessian approximation by computing outer product of gradients. It can be noticed that iterative optimization procedure does not have much influence on constant terms in both equations, while parameters α_1 and β_1 differs between algorithms. However, these differences are negligible when a convergence criterion based on the relative change of parameters vector or the log-likelihood value between two successive iterations. When a convergence criterion is based on gradient approximately close to zero more iterations are needed in general, while BFGS and DFP algorithms do not converge due to flat log-likelihood function (cannot find uphill direction). Newton-Rapson and BHHH algorithm have proved to be stable according to different convergence criteria. Further research will be based on simulation techniques to investigate which algorithm gives parameters estimates that are most likely to be the true parameters values.

Acknowledgement
This work has been fully supported by Croatian Science Foundation under the project "Volatility measurement, modeling and forecasting" (5199).

References

[1] Aptech System (2001). Maximum Likelihood Estimation for Gauss, Version 5.0. Maple Valley: Aptech Systems, Inc.
[2] Arnerić, J. and Rozga, A. (2009). Numerical Optimization within Vector of Parameters Estimation in Volatility Models. World Academy of Science, Engineering and Technology, Vol3: 523-527.
[3] Bazarra, M. S., Sherali H. D. and Shetty, C. M. (1993), *NonlinearProgramming - Theory and Algorithms* (second edition), John Wiley & Sons Inc., New York.
[4] Berndt, E., Hall, B., Hall, R. and Hausman, J. (1974), "Estimation and Inference in Nonlinear Structural Models", Annals of Social Measurement, Vol. 3., p. 653-665.
[5] Bunch, D.S. (1988). A comparison of algorithms for maximum likelihood estimation of choice models. Journal of Econometrics, Vol 38 (1–2): 145–67.
[6] Carling, K., and Söderberg, H. (1998) An Experimental Comparison of Gradient Methods in Econometric Duration Analysis. *Computational Statistics & Data Analysis*, 27, 83-97.
[7] Drmač, Z., Hari, V., Rogina, M., Singer, S., Singer, S. and Slapničar, I. (2003). Numerička analiza, predavanja i vježbe. Zagreb: PMF-MO.
[8] Eliason, S. R. (1993). Maximum Likelihood Estimation: Logic and Practice. Iowa City: Sage Publications.
[9] Gould, W., Pitblado, J. and Sribney, W. (2006), *Maximum Likelihood Estimation with Stata* (third edition), College Station, StataCorp LP.
[10] Mai, A. T., Bastin, F., and Toulouse, M. (2014). On Optimization Algorithms for Maximum Likelihood Estimation. CIRRELT, CIRRELT-2014-64.
[11] Neralić, L. (2003), *Uvod u Matematičko programiranje 1,* Element, Zagreb.
[12] Schoenberg R. (2001), "Optimization with the Quasi-Newton Method", Aptech Systems, Inc., Valley WA, working paper, p. 1-9.
[13] Shanno, D. F. (1970), "Conditioning of quasi Newton methods for function minimization", Mathematics of Computation, No. 24., p. 145-160.
[14] Škrabić, B., Arnerić, J. and Babić, Z. (2007). Maximization of the Likelihood Function in Financial Time Series Models. Proceedings of the International Scientific Conference on Contemporary Challenges of Economic Theory and Practice - Quantitative Economics and Finance, Belgrade, Serbia, pp. 1-12.

HOW TO IMPROVE INTERPRETABILITY OF THE LOGISTIC REGRESSION MODEL

Ana Bilandžić, Marina Jeger and Nataša Šarlija
Faculty of Economics in Osijek, J.J. Strossmayer University of Osijek
Trg Ljudevita Gaja 7, 31000 Osijek, Croatia
{anag, marina, natasa}@efos.hr

Abstract: Developing and interpreting a prediction model can be quite a difficult task especially when dealing with highly correlated variables. The aim of this paper is to demonstrate the application of factor analysis for the purpose of increasing overall interpretability of the logistic regression model. The results show that by conducting factor analysis prior to building regression model researchers can overcome some model interpretability issues, such as misalignment with theoretical background, and create a model that is easier to understand, explain and apply in real-life business situations.

Keywords: factor analysis, logistic regression, interpretability, growth prediction model

1 INTRODUCTION

Logistic regression modeling is widely used for the analysis of multivariate data involving binary responses we deal with in our research. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses. However, when working with highly correlated variables, logistic regression may provide results that are, from the theoretical perspective, very hard to interpret. The aim of this paper is to demonstrate the application of factor analysis prior to building logistic regression model in order to boost interpretability without compensating on predictive power of the model. The goal of the study is to develop a model for estimating growth potential of small and medium-sized enterprises (SMEs) in Croatia based on predominantly financial data and some nonfinancial data noted in their financial statements.

Previous studies used various approaches to tackle the issue of interpretability and accuracy of the prediction model. Schielzeth (2010) suggests some simple methods, such as centering and standardization of input variables or thoughtful removal of intercepts or main effects, to improve interpretability of regression coefficient in linear regression models. Furthermore, Li (2014) used a combination of principal component and logistic regression to distinguish accounting information distortion and achieve higher model accuracy. Similar approach was used in two other studies (Shengyuan, 2009; Kehong and Zhansheng, 2006) that combined principal component and logistic regression but in context of corporate financial distress prediction. Zhu and Li (2010) used principal component as a preprocessing method before applying logistic regression and discriminant analysis for credit risk estimation. Suleiman et al. (2014) used principal component as input for predicting applicants' creditworthiness in order to improve the predictive power of linear discriminant and logistic regression models. Results showed that the use of principal component as input improved models prediction by reducing their complexity and eliminating data co-linearity.

Methodological steps undertaken in this paper are the following: (i) development of the logistic regression model where independent variables are financial ratios defined as observed variables; (ii) application of factor analysis on independent variables in order to create factors; (iii) development of the logistic regression model where independent variables are factors; (iv) comparison of both logistic regression models in light of their interpretability and predictability.

The structure of the paper follows the methodological steps and ends with the discussion on advantages and limitations of this approach, possibilities of application and suggestions for further research.

2 DATA AND VARIABLES

The sample used in this research consists of 1492 privately-owned small and medium-sized companies in Croatia. They were chosen from the Financial Agency (FINA) data set that include 53434 SMEs which existed over the period from 2008 to 2013. An enterprise is defined as high growth if it has average annualized growth in assets greater than 20% a year, over a three-year period, from 2010 to 2013 (OECD, 2010). Out of total number of SMEs, 746 enterprises met this criterion. Development sample included 650 high growth SMEs, while validation sample consists of 96 high growth SMEs. The other 746 SMEs, which are not high growth, were selected randomly from the whole data set. They were divided in the same way as high growth enterprises.

Independent variables for growth prediction model are created for every enterprise in the data set for the period from 2008 to 2010. Total of 101 variables were created. They are grouped in 5 groups: liquidity (15 ratios), turnover (30 ratios), leverage (15 ratios) and profitability (15 ratios). In general, it could be noticed that most of the financial ratios from all of the groups are higher in growing SMEs compared to not growing SMEs although there are some ratios which show different behavior. The fifth group consists of 26 variables which include industry sector, non-tangible assets and percentage change in number of employees, assets, profitability, sales and some other performance indicators.

3 METHODS

In the process of developing a model with the binary dependent variable Y – probability for a firm to become high growing or not – logistic regression was used. In simple terms, for one independent variable *x* the logistic function would be:

$$y = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1}$$

The goal is to obtain β_0 and β_1. Because the above formula is not linear, through logistic transformation it becomes:

$$logit(y) = \beta_0 + \beta_1 x \tag{2}$$

where $logit(y) = \frac{y}{1-y}$, often called 'log odds'. Intuitively for more variables $logit(y)$ becomes

$$logit(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{3}$$

Selection procedures forward and backward were used, and the selected variables were used with R built in function `glm()` to obtain our first model (Agresti, 2002). To address the difficulties with interpretation of regression coefficients, factor analysis was conducted.

Factor analysis is a procedure to get a model with the following structure:

$$
\begin{aligned}
X_1 - \mu_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1r}F_r + U_1 \\
X_2 - \mu_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2r}F_r + U_2 \\
&\vdots \\
X_p - \mu_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pr}F_r + U_p
\end{aligned} \tag{4}
$$

where X_i, $i = 1,2,\dots,p$ are observed variables, and F_j and U_i , $i = 1,2,\dots,p$; $j = 1,2,\dots,r$ are unobserved. Equivalently, the set of equations can be written as

$$(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{Af} + \boldsymbol{u} \tag{5}$$

where A is the factor pattern matrix consisting of its elements a_{ij} which are called factor loadings, $\boldsymbol{x}$ is the $p \times 1$ vector of elements X_i, $i = 1,2,\dots,p$ and $\boldsymbol{\mu}$ is vector of their means. While $\boldsymbol{f}$ is the $r \times 1$ vector of elements F_j, $j = 1,2,\dots,r$, they are called common factors and are assumed to have mean 0 and variance 1. U_i , $i = 1,2,\dots,p$ are unique factors and are assumed to have mean 0, but variance σ_i^2, $i = 1,2,\dots,p$, they form the $p \times 1$ vector $\boldsymbol{u}$. Addionally, it is assumed that the unique and common factors are uncorrelated. So, by marking the covariance matrix of x with Σ, the previous equation turns into:

$$
\begin{aligned}
\Sigma &= \mathrm{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] \\
&= \mathrm{E}[(\boldsymbol{Af} + \boldsymbol{u})(\boldsymbol{Af} + \boldsymbol{u})'] \\
&= \mathrm{E}[(\boldsymbol{Af} + \boldsymbol{u})(\boldsymbol{f'A'} + \boldsymbol{u'})] \\
&= E[\boldsymbol{Aff'A'} + \boldsymbol{uf'A'} + \boldsymbol{Afu'} + \boldsymbol{uu'}] \\
&= E[\boldsymbol{Aff'A'}] + E[\boldsymbol{uf'A'}] + \boldsymbol{E}[\boldsymbol{Afu'}] + \boldsymbol{E}[\boldsymbol{uu'}] \\
&= \boldsymbol{A}E[\boldsymbol{ff'}]\boldsymbol{A'} + E[\boldsymbol{uf'} + \boldsymbol{AE}[\boldsymbol{fu'}] + \boldsymbol{E}[\boldsymbol{uu'}] \\
&= \mathbf{AA'} + \boldsymbol{\Psi},
\end{aligned} \tag{6}
$$

where Ψ is the vector of variances of U_i. Whereas the right side of the equation consists only of unobserved data, this process is not unique, and different factors can be obtained (Jobson, 2012).

To get the factor scores, F_j, $j = 1,2,\dots,r$ R was used and the function `fa()`, from the package `psych`. Not all variables showed correlation with at least one factor, so one by one was excluded from the factor analysis, until the desired result was reached. With the factor loadings from function `fa()` and some additional variables that covered profitability once again `glm()` was used to develop new logistic regression model.

4 RESULTS

For the purpose of this study, data analysis and model development procedures can be divided into four steps. First, the standard logistic regression model was built based on predefined set of financial ratios. The model was juxtaposed to the underlying theoretical background and inconsistencies in model results were identified. Second, factor analysis was applied on independent variables and three factors were singled out. Third, a new prediction model was developed by using factors as independent variables. And finally, a comparison of a prediction model without factor analysis and a prediction model with factor analysis was given and the results were evaluated in the context of both theoretical framework and methodological approach.

The first step covered development of logistic regression prediction model in a standard way with financial ratios set as observed variables. The model consists of 15 variables and has satisfactory performance measures (total hit rate 64.65%, hit rate for growth firms

64.1%, hit rate for non-growth firms 65%; AUC=0.7, KS=40.26%; AIC = 828.32). To reach an adequate level of prediction, the authors developed several prediction models that showed average performance. Still, interpretation of the model was challenging as the results collide with the underlying theory and, in some cases, even with the common sense. In that context, several inconsistencies should be noted. First, some predictor variables have negative coefficients while the theory and sound reasoning suggest the opposite. For example, Fixed-asset turnover has a negative regression coefficient suggesting that the increase of the turnover leads to decrease of potential to growth. Second, while the descriptive analysis shows that higher values of a particular indicator are characteristic for growing companies, the final model can show the negative value of respective coefficients, such as Ratio of short-term liabilities to equity. Third, sometimes the sign of the coefficients can change in the process of model development which happened with Return on equity and Ratio of retained earnings and total assets.

To account for these inconsistencies, factor analysis on independent variables was applied in the next step. Analysis of variance of eigenvalues showed that three factors could be generated. This represents 99% of total variance. The results of this procedure are shown in table 1.

Table 1: Factor loadings of three extracted factors

	Factor 1 (Turnover Factor)	***Factor 2 (Liquidity Factor)***	***Factor 3 (Leverage Factor)***
Current Ratio	-0.00012	**0.998667**	-0.0001
Leverage Ratio	3.04E-06	-0.000000172	**0.998749**
Total Equity to Total Asset Ratio	-3.00E-06	0.000000261	**-0.99875**
Quick Ratio	-6.30E-05	**0.998669**	-0.0000527
Total Asset Turnover	**0.99873**	0.000197	0.000205
Current Asset Turnover	**0.998119**	-0.00016	-0.00016
Sales to Total Asset Ratio	**0.998516**	-0.0000416	-0.0000468
Cash Ratio	0.000301	**0.847054**	0.000251

Looking at the factor loadings it could be noticed that variables are grouped in theoretically sound way assessing three groups of business performance indicators: business activity (turnover ratios), liquidity and leverage.

Furthermore, in the third step, the factors were treated as independent variables and the new logistic regression model was developed. Additionally, since profitability ratios were removed in the process of factor analysis, they were put back during the logistic regression model development together with some variable from the fifth group of variables. The logistic regression results are given in the table 2.

Table 2: Logistic regression model with factors as independent variables

Factor/Variable	***Regression coefficient***
Factor 1 (Turnover Factor)	160.3426
Factor 2 (Liquidity Factor)	8.2408
Factor 3 (Leverage Factor)	147.246
Intangible Assets/Total Assets	1.6228
Net Income/Sales	-0.0068
Net Income/Equity	0.0002
Change (%) in number of employees	-0.1437
Accuracy of the model: total hit rate 60.67%, hit rate for growth firms 56.1%, hit rate for non-growth firms 64.58%; AUC=0.64, KS=29.42%; AIC = 748.77	

The final step in the analysis was to compare logistic regression models with and without factor analysis applied prior to model building. Comparison is done according to model quality, interpretability and predictability. Based on hit rates, area under the curve and Kolmogorov-Smirnov indicators, the first model exhibits a slightly better performance in terms of model predictability. Nevertheless, according to Akaike information criterion (AIC), the second model shows a higher relative quality and therefore represents a better option when choosing a model for predicting growth. As regard to interpretability, all of the three issues introduced in the paper have been addressed. First, results obtained by combining factor analysis and logistic regression gave results that are more logical and in line with the theoretical framework (Sampagnaro, 2013; Segarra and Teruel, 2009). As it can be seen from the table 2, the potential for growth rises with an increase of liquidity, turnover, leverage and profitability, with a drop of return on sales and with a decrease in change in number or employees. Second, regression results coincide with the input data. Leverage ratios are higher in high growth SMEs and the same can be seen in the final model. Third, the results are consistent. Since one factor represents liquidity it implies that higher liquidity means higher potential to growth and it cannot happen that one liquidity coefficient is positively associated with the growth potential and the other negatively (which was the case in a regression model developed in a standard way).

5 CONCLUSION

Logistic regression is a method of choice when analyzing multivariate data involving binary responses. Still, sometimes logistic regression may result in models that are very difficult to explain, especially if derived from a large set of highly correlated variables. To overcome these issues without any great loss in information contained in data, the authors conducted factor analysis prior to regression model building. The goal was to improve interpretability of the model by identifying most important factors and reducing large number of variables, as well as lowering multicollinearity levels. Based on the empirical analysis, the model created by combining factor analysis with logistic regression exhibits variables that are simpler and easier to interpret. Those variables, when observed separately, depict the main area of business performance – business activity, liquidity, leverage and profitability. Moreover, the influence of each variable on predicting growth is unambiguous and aligned with the underlying theory.

This study has certain limitations and they are predominantly related to the data itself. Financial statements of small and medium-sized companies are not subjects to regular financial review process and therefore may have questionable quality and reliability. Furthermore, the period covered with the analysis refers to time of worldwide economic

crisis and the data may be under the influence of some macroeconomic forces that are not typical for the time of economic prosperity.

However, model for predicting company's growth can be a powerful strategic tool for managers and entrepreneurs, and is widely applicable in many areas of business decision making such as finance, management, marketing and sales. This only emphasizes the need for developing growth prediction models that are easy to understand and apply in business decision making, not only for researchers and academics, but for business people as well.

Growth prediction models can further be improved by focusing on two aspects: improving the existing data (strengthening the supervision and increasing the reliability of financial statements), and expanding the data set (by including non-financial and macroeconomic variables). All of that could lead to an increase of predictability of the model.

Acknowledgment

This research is funded by Croatian Science Foundation under grant No. 3933 "Development and application of growth potential prediction models for SMEs in Croatia".

References

[1] Agresti, A., 2002, Categorical Data Analysis Second Edition, John Wiley & Sons, Inc., Hoboken, New Jersey

[2] Jobson, J. (2012). Applied multivariate data analysis: volume II: Categorical and Multivariate Methods. Springer Science & Business Media.

[3] Kehong, G., and Zhansheng, L. (2006). Empirical Comparison among Different Predicting Models of Financial Failure for China's Listed Corporations. Journal of Shanxi Finance and Economics University, (5): 129-133.

[4] Li, Z. (2014). Research on Distinguish the Accounting Information Distortion Based on the Principle Components-logistic Regression Model. International Journal of Security and Its Applications, 8(4): 37-50.

[5] Organization for Economic Co-operation and Development. (2010). High-growth enterprises: what governments can do to make a difference. OECD Publishing.

[6] Sampagnaro, G. (2013). Predicting rapid-growth SMEs through a reversal of credit-scoring principles. International Journal of Entrepreneurship and Small Business, 18(3): 313-331.

[7] Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. Methods in Ecology and Evolution, 1(2): 103-113.

[8] Segarra, A., and Teruel, M. (2009). Small firms, growth and financial constraints. Xarxa de Referència en Economia Aplicada (XREAP)

[9] Sheng-yuan, Y. A. N. G. (2009). Study on Financial Predicament Forecast of Listed Company in China. Journal of Kunming University, (5): 94-99.

[10] Suleiman, S., Issa, S., Usman, U., & Salami, Y. O. (2014). Predicting an applicant status using principal component, discriminant and logistic regression analysis. International Journal Of Mathematics And Statistics Invention (IJMSI), 2(10): 5-15.

[11] Zhu, K. and Li, J.(2010). Studies of Discriminant Analysis and Logistic Regression Model Application in Credit Risk for China's Listed Companies. Management Science and Engineering, 4(4): 24-32.

FEATURE SUBSET SELECTION FOR B2B SALES FORECASTING

Marko Bohanec
Salvirt ltd.
Dunajska cesta 136, Ljubljana, Slovenia
E-mail: Marko.Bohanec@salvirt.com

Mirjana Kljajić Borštnar
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, Kranj, Slovenia
E-mail: mirjana.kljajic@fov.uni-mb.si

Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
E-mail: Marko.Robnik@fri.uni-lj.si

Abstract: Although the use of information technology in business to business (B2B) sales is high, and generates lots of data, sales forecasting relies more on the personal judgement and established mental models. Business domain data sets often list a high number of descriptive features, however only a few have some value for machine learning techniques. Small number of features is important for understanding and interpretation in the context of organizational learning. We present an operational guideline for selection of small number of the most relevant features while maintaining high classification accuracy. Promising results of introduced three-step operational guideline are presented in this paper.

Keywords: feature subset selection, machine learning, business case, knowledge engineering

1 INTRODUCTION

We investigate a feature subset selection problem for business data sets. As a use-case we present a data set built on insights from B2B sales experts provided in a form of sales history. In human reasoning the number of features impacts the comprehensibility and perceived complexity, therefore one has to focus only on the most important features. Definition of features relevant for human comprehension may be highly subjective. Reduction in the number of features is motivated by a need of sales teams to understand and discuss relationships between features and the sales outcome free from personal biases. In [1] author suggests the number of features should be in the range of 7 ± 2 non-redundant features.

Our research question is related to whether we can define operational steps to achieve an optimal balance between subset of features and classification accuracy (CA) of machine learning (ML) model trained on the selected subset. This optimization problem is formulated in (1).

$$min(\propto \cdot (1 - CA) + (1 - \propto) \cdot |S|/|A|) \quad 0 < \alpha < 1 \qquad (1)$$

where α is balance factor, |S| is a power of subset of features and |A| is power of all listed features. An exhaustive search through all possible feature subset is impossible due to its exponential size. We propose a sequence of operational steps which shall result in a balanced subset.

This paper is organized in three sections. In Section 2 we introduce data set, filter and classifier methods. In Section 3 proposed operational guideline is applied and in Section 4 we conclude the paper with outlined findings and suggestions.

2 METHODOLOGY OVERVIEW

The feature subset selection problem is a well-researched topic within data mining and machine learning [2, 3, 4, 5, 6]. Various techniques were proposed and thoroughly validated on different publicly available data sets, which helped us to select three techniques, suitable for our problem based on characteristics of data sets in B2B sales domain.

2.1 Building a full list of features

Our real-world domain is described by carefully selected set of features, reflecting B2B sales history. Sales opportunities are labelled with their final outcomes (yes or no). Features can be of different (numeric, nominal or ordinal) types. A data set sample is present in Table 1.

Table 1: B2B Sales history data (a sample only).

Clnt_stats	Deal_type	Scp_clrty	Strat_deal	Cross_sale	Up_sale	Imp_Client	Signed
current	project	clear	avg	yes	no	strategic	*Yes*
current	solution	clear	unimport	yes	no	normal	*No*
current	project	clear	avg	no	yes	normal	*Yes*
new	project	limited	avg	no	no	normal	*No*
current	project	clear	very_imp	no	yes	strategic	*Yes*
current	solution	clear	avg	yes	no	normal	*Yes*

This data set was obtained from a software development company, participating in the research. The data set consist of 22 features and 150 sales cases, among which for 65 cases the deal was signed and for 85 it was lost.

2.2 Feature ranking methods overview

In feature subset selection in machine learning relevant features should be detected and preserved, and irrelevant should be discarded. This typically has a positive impact on learning speed, generalization capacity and simplicity of induced models [6]. Same authors indicated other benefits, like reduced measurement costs and better understanding of a domain. These are important criteria also for our B2B sales forecasting case where the cost of data acquisition can be significant. Feature subset selection enables better understanding of the sales domain as the important features are ranked higher than rest of them. This creates a basis to effectively mitigate prior mental models based on intuition and biases of participants. A comprehensive overview of feature selection methods is available in [6], where three main techniques are listed: filters, wrappers and embedded methods.

Filter methods try to extract general characteristics of training data and rank features in a pre-processing step without interaction with classification models. This approach is promising for our research, as we would like to maintain independence from classifiers and compare CA of different classification models using top ranked features. Wrapper methods take a particular classification model as a part of feature selection process. A contribution of a feature is determined with the performance of the model for different subsets of features. This approach is computationally expensive as it requires a training re-run for each feature subset; however it has an advantage to expose low performing features and detect redundant features. For example, two features can be both relevant, however, they might be correlated and thus only one is needed in the model. Embedded methods are integrated into a specific classifier and evaluate features during the training process.

Our goal is to find a minimal non-redundant subset with acceptable CA. As most filter methods are inferior in detecting redundancy (but they are fast, tough), wrapper methods are also considered in our research.

2.3 Filter methods

Based on previous research [4,5,6], we identified a few filters, which are perceived as the best performing and are frequently used as a method of choice for ranking features. They are described in a compact form below.

ReliefF
ReliefF [4] is an extension of algorithm Relief [7] and estimates the quality of features based on how well their values separate similar cases. This means that the quality of individual feature depends on the context of other features which allows this method to detect interactions between features [4, 5].
Random Forest as a filter
Random Forest [8] is a state-of-the-art classifier, which can be applied as a feature quality estimator and selector. When used as a filter it compares classifier's performance on original and modified data, generated by random value permutations. Difference in performance on original and locally perturbed data for each tree in the forest is summarized as an importance estimate [5].
Gain Ratio
The Information Gain [12] is a classical measure based on class value impurity. It measures entropy of features conditioned upon class. To avoid preference of features with more values, Gain Ratio uses normalization with feature entropy [9].

2.4 Selected Classifiers

We selected several classifier methods, which have performed well in previous studies ([5], [10]) or exhibit good comprehensibility.

Random Forest classifier (RF)
Random Forest learning algorithm [8] is constructing a set of decision trees, called forest. Individual tree are grown with a limited set of randomly selected features. During the classification each tree is an independent "voter" for testing instances. Votes are counted and the majority class is taken as a prediction for a given instance.
Naïve Bayes (NB)
Naïve Bayes is a popular classifier due to its simplicity. Using Bayes rule is called naïve due to the assumption that features are independent.
Decision Tree classifier (DT)
Decision tree is classification technique popular due its visualization in a form of a tree. Internal nodes of trees contain features. The branches are formed based on the values of the feature. Leaves of the tree contain the predicted class.

3 APPLICATION OF METHODOLOGY

We propose application of introduced methods as an order of three steps:

1. Rank features according to their relevance using filter methods.
2. Build ML model by incrementally adding ranked features, monitoring maximal CA to detect the cut-off point.

3. Eliminate noisy and redundant features with wrapper method.

To create an initial set of features for description of sales cases, we worked with the company's sales team through guiding workshop. From the initial data set 60% of cases is randomly assigned to the training set (90 cases) and the rest are left as the testing set (60 cases). Statistical, data mining and machine learning suite R ver. 3.2.0 was used for data analysis, visualization and programming tasks. CORElearn library of machine learning techniques and methods [11] was utilized for ML model building, feature ranking and wrapper creation.

3.1 Feature ranking

We ranked features according to their relevance by applying selected filters. Comparing the obtained rankings we concluded that for all filters only between 4 and 6 features (mostly the same) rank high, the rest are less important. Figure 1 shows the distribution of rankings.

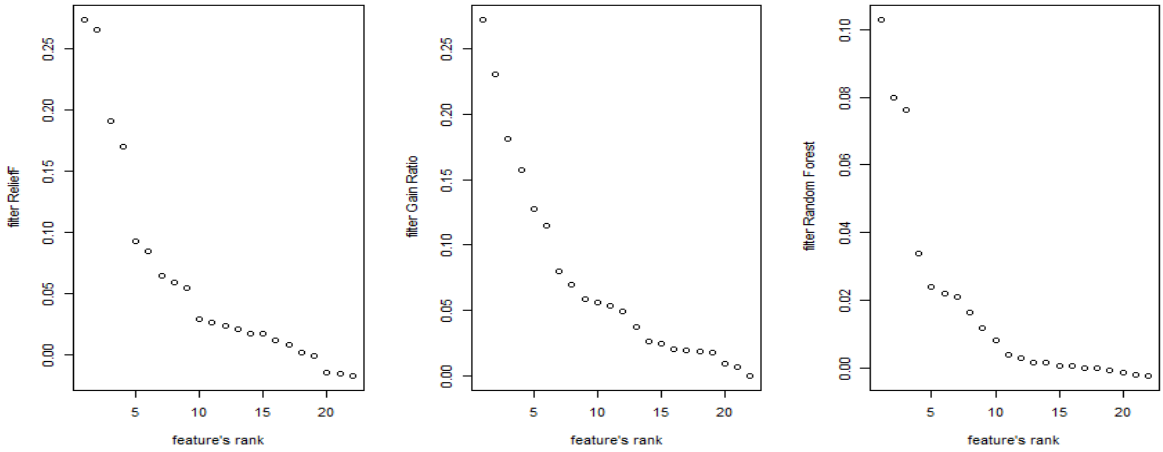


Figure 1: Feature ranking by selected filter methods (each point represent a feature).

3.2 Increasing CA with progressive addition of ranked features

Next we added ranked features (by ReliefF) one by one to training of ML model and measure their CA. As a result the maximum CA defines cut-off point of ranked features, where the rest can be discarded, as they decrease CA. Figure 2 shows growth of CA as result of adding features one by one to the model. We tested RF, NB and DT classifiers on testing data.

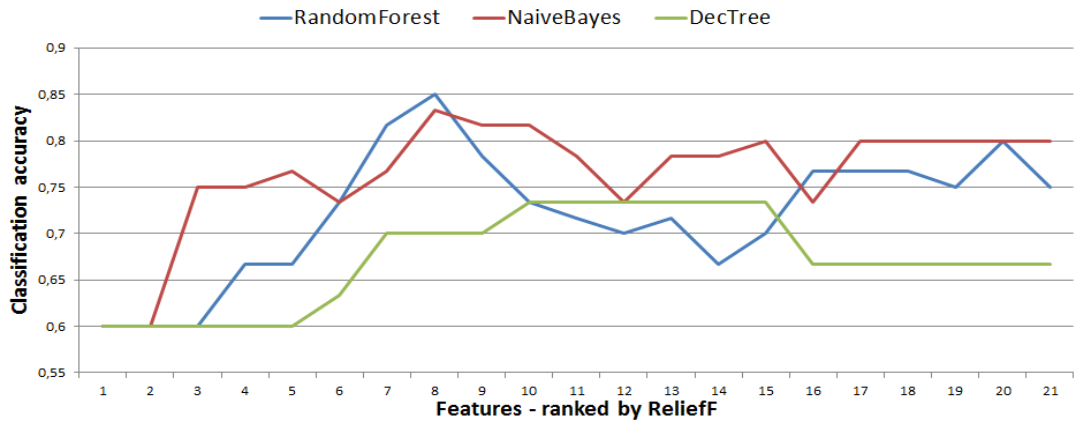


Figure 2: Incrementally adding features ranked by ReliefF to 3 classifiers – CA on testing data set.

Figure 2 reveals that RF and NB achieve their maximum CA faster than DT. After 8th feature is added to the first two models, CA starts to drop and then varies, however it never reaches maximum CA again.

3.3 Dealing with noise and redundancy

We noticed negative or no difference in CA during incremental addition of some features, which indicated noisy and redundant features. Therefore we constructed a wrapper method capable of eliminating such features. We excluded features one by one from the list of features (top down) and monitored CA. When excluded feature increased CA performance (Table 2, column *Row_miss*), we excluded it from the list. Table 2 shows remaining features after these steps. All differences in column *Diff* are negative, therefore features are contributing to CA and they can't be dropped. This list of features defines minimum feature list for our data set.

Features in Table 2 are: *Post_statmn* (does a client makes public positive statements), *Up_Sale* (does company sells more of the same service/product to existing client), *Competitors* (indicates presence of competing offerings) and *Procrmnt* (indicates involvement of a purchase department in the sale).

Table 2: The final list of the most important features for Naïve Bayesian classifier.

Feature name	All_Feat	Row_miss	Diff	Prct_Diff
Post_statmn	0,80	0,78	-0,02	-2,78%
Up_Sale	0,80	0,78	-0,02	-2,78%
Competitors	0,80	0,78	-0,02	-2,78%
Procrmnt	0,80	0,77	-0,03	-4,17%

4 CONCLUSIONS

We introduced a heuristic, but operational solution to the sales forecasting optimization problem. The proposed order of application of ML methods can be used as a general approach to balance CA for models built with selected subset of features. Reduction in number of features improves speed of execution for most ML techniques and increases parsimony of a model.

The proposed solution was applied to the real-world business data set. For this data set the minimum number of features is 4 (Table 2) for NB classifier, which represents less than 20% of the initial number of features. For different data set and different classifier different minimum number would be identified by the proposed methodology.

In the future we will test the methodology on B2B data sets from other companies and test additional methods for detection of redundancy, e.g., the wrapper approach using other classification models. It would be interesting to analyse if pairs (features, cases) create a line of equal CA for different pairs (kind of "iso-CA").

Acknowledgement
We are grateful to the Salvirt Ltd. Company for funding the research and development presented in this paper.

References

[1] Miller, G.A. (1956): "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information", Psychological Review, Vol. 63, pp. 81-97.

[2] Langley, P. (1994): "Selection of relevant features in machine learning", Proceedings of the MAl Fall Symposium on Relevance pp. 1-5.

[3] Kohavi R., John G.H. (1997): "Wrappers for feature subset selection", Artificial Intelligence 1–2, pp. 273–324.

[4] Robnik-Šikonja M., Kononenko I. (2003): "Theoretical and Empirical Analysis of ReliefF and RReliefF", Machine Learning Journal Vol 53, pp. 23-69

[5] Čehovin L., Bosnić Z. (2010): "Empirical evaluation of feature selection methods in classification", Intelligent Data Analysis 14, pp. 265-281, IOS Press

[6] Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. (2013): "A review of feature selection methods on synthetic data", Knowledge Information Systems, Vol 34, pp. 483–519

[7] Kononenko, I. (1994): 'Estimating attributes: analysis and extensions of Relief'. In: L. De Raedt and F. Bergadano (eds.): Machine Learning: ECML-94. pp. 171–182, Springer Verlag.

[8] Breiman L. (2001): "Random forests", Machine Learning Journal, Vol. 45, pp. 5–32.

[9] Quinlan, J.R. (1993): "C4.5: Programs for Machine Learning", Morgan Kaufmann, San Francisco.

[10] Kontsiantis S. B. (2007): "Supervised Machine Learning: A review of Classification Techniques", Informatica 31, pp. 249-268

[11] Robnik-Šikonja M. (2015): "CORElearn: various machine learning algorithms and techniques", R package version 0.9.46, https://cran.r-project.org/web/packages/CORElearn, 03.06.2015

[12] Witten, I.H., Eibe F., Hall M.A. (2011): "Data mining – Practical Machine Learning Tools and Techniques", Third edition, Elsevier.

INCOME CONVERGENCE OF THE EU REGIONS: SPATIAL ECONOMETRIC APPROACH

Michaela Chocholatá
University of Economics in Bratislava, Department of Operations Research and Econometrics
Dolnozemská cesta 1, 852 35 Bratislava, Slovakia
chocholatam@yahoo.com

Andrea Furková
University of Economics in Bratislava, Department of Operations Research and Econometrics
Dolnozemská cesta 1, 852 35 Bratislava, Slovakia
andrea.furkova@euba.sk

Abstract: The paper focuses on the testing of regional income convergence of NUTS2 EU regions during the 2000-2011 period using both non-spatial and spatial approaches. The first part of analysis was aimed to the testing of absolute and conditional hypotheses by using traditional *beta*-convergence concept, i.e. spatial regional dependences were not considered. The results of our analysis showed that the regions with lower per capita GDP grew at a higher speed during the period 2000-2011. Also the inclusion of dummy variable to the model was used to control different steady state levels of per capita GDP of different regions. The estimation of the spatial income convergence models confirmed that the spatial dependence among regions does matter. The inclusion of spatial effects and also the use of conditional convergence concept led to worsening of convergence characteristics. Overall, the results of our analysis imply that convergence process is not determined only by a region's initial income and other specific factors but also essentially by its neighbourhood region's growth performance.

Keywords: *beta*-convergence, spatial econometric models, NUTS2 regions.

1 INTRODUCTION

In recent years, income disparities and convergence in European Union (EU) countries and regions become an important research area for analysts and also for economic policy creators. This growing interest is mainly caused by the process of the eastward enlargement of the EU which is accompanied by the problem of regional disparities and convergence. Balanced regional development and reduction of regional disparities is one of the main objectives of the EU, as it is recognized in the EU treaty [4]. The essential argument is that the balanced regional development is necessary condition for social cohesion and increased competitiveness in the EU. Also the EU's strategic document Europe 2020 [3], new economic geography and growth theories emphasize the role of geographic spillovers in growth mechanism. Just the integration of the spatial dependence among regions and growth processes was our main motivation for spatial analysis of income convergence of EU regions.

As it has been already mentioned, regional income convergence has become more popular in recent years, but despite this interest, studies on spatial regional convergence in the enlarged EU are relatively scarce. From the studies dealing with the European regional convergence, income disparities and spillovers taking into account spatial pattern we can mention e.g. [2], [6], [7], [8].

The aim of this paper is to consider the geographical dimension of data in the estimation of the income convergence of 252 NUTS2 (Nomenclature of Units for Territorial Statistics) EU regions and to emphasize geographic spillovers in regional economic growth process. Using the 2000-2011 period we try to test hypotheses of absolute and also conditional income convergence using both non-spatial and spatial approaches. GDP per capita in euro of NUTS2 regions is used as a proxy for the income level of individual regions.

The rest of the paper is organized as follows: sections 2 and 3 deal with theoretical backgrounds of the study. Estimation results are presented and interpreted in section 4. The paper closes with concluding remarks.

2 INCOME CONVERGENCE MODELS

The issue of income convergence deals with the question whether poor economies catch-up to wealthier economies. In the literature we can distinguish the three hypotheses concerning the regional convergence: the absolute (unconditional) convergence hypothesis, the conditional convergence hypothesis and the club convergence hypothesis (for more details see e.g. [5], [7]).

In order to empirically test the convergence, the concepts of so called *beta*-convergence and *sigma*-convergence are usually used (see e.g. [7]). Furthermore it can be distinguished the unconditional and conditional *beta*-convergence [5]. Since the *beta*-convergence is concentrated on detecting of the catching-up processes, *sigma*-convergence focuses on the reduction of disparities among regions in time and gives information about the convergence "journey". In this paper we will furthermore deal only with the concept of *beta*-convergence. The analysis of the *beta*-convergence is usually based on the following cross-country/region growth regression [7]:

$$\ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \ln\left(y_{i,0}\right) + \varepsilon_i \qquad \varepsilon_i \sim i.i.d\left(0, \sigma_\varepsilon^2\right) \tag{1}$$

where $y_{i,0}$ and $y_{i,T}$ are the per capita GDP's of the region i ($i = 1,2,...,n$) in the base year 0 and in the final year T, respectively. $\ln\left(\frac{y_{i,T}}{y_{i,0}}\right)$ is the growth rate of the i-th region per capita GDP in the period $(0, T)$, T denotes also the number of periods for which we have data, n is the number of regions in the data set, α and β are unknown parameters and ε_i is an error term. The absolute convergence hypothesis can be accepted if the estimated β parameter is statistically significant and negative. Concerning the equation (1), two more indicators can be calculated – the speed of convergence and the half-life (formulas for calculation see e.g. [1], [7]).

The conditional *beta*-convergence hypothesis incorporates the variables which enable the differentiation of the regions and also enables to capture different initial conditions. Since in the further analysis only one dummy variable D_i will be used, the model (1) can be reformulated as follows:

$$\ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \ln\left(y_{i,0}\right) + \gamma D_i + \varepsilon_i \quad \varepsilon_i \sim i.i.d\left(0, \sigma_\varepsilon^2\right) \tag{2}$$

where D_i is a dummy variable indicating if the region i belongs to a postcommunist country ($D_i = 1$) or not ($D_i = 0$), γ is a parameter and all other terms were previously defined above. The conditional convergence hypothesis can be accepted if the estimated value for β is significantly negative.

Models presented above are based on the fact that each region is a geographically independent entity and no spatial interactions were considered. During the last years these models have been modified in order to incorporate the mutual regional interactions. In

spatial econometric models the spatial relationships are expressed by the spatial weight matrix **W** of dimension ($n \times n$). There are various possibilities how to specify the spatial weight matrix **W** (see e.g. [1]). One of the possibilities is to use the queen case contiguity matrix **W**.

3 SPATIAL CONVERGENCE MODELS

After estimation of the model (1) or (2) using the Ordinary Least Squares (OLS) it is necessary to calculate the spatial diagnostic statistics (e.g. the Moran's I – see e.g. [1]) which indicate whether there is spatial autocorrelation in the residuals. In case that the spatial autocorrelation is present, the Lagrange Multiplier (LM) tests can be used in order to decide whether a spatial autoregressive (SAR) or a spatial error (SEM) model of spatial dependence is the most appropriate. If both statistics are significant, robust modifications of these statistics should be used (see [1], [7]). Both SAR and SEM models can be estimated by e.g. the maximum likelihood method (ML).

3.1 Spatial Autoregressive Model

Model SAR, known also as spatial lag model, is appropriate if spatial autocorrelation among neighbouring regions exists, i.e. if the growth rate in a region is related to those of its surrounding regions conditioning on the initial level of per capita GDP. In this case the *beta*-convergence model (2) can be modified as follows:

$$\ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \ln\left(y_{i,0}\right) + \gamma D_i + \rho \sum_j w_{ij} \left(\ln\left(\frac{y_{j,T}}{y_{j,0}}\right)\right) + \varepsilon_i \ , \qquad \varepsilon_i \sim i.i.d\left(0, \sigma_\varepsilon^2\right) \tag{3}$$

where ρ is the scalar spatial autoregressive parameter, w_{ij} are the elements of matrix **W** describing the structure and intensity of spatial effects and all other terms were previously defined.

3.2 Spatial Error Model

The specification of spatial error model (SEM) is appropriate when it is supposed that the spatial autocorrelation exists in the error term. In such a case the non-spatial model (2) can be modified as follows:

$$\ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \ln\left(y_{i,0}\right) + \gamma D_i + \varepsilon_i \ , \qquad \varepsilon_i = \lambda \sum_j w_{ij} \varepsilon_j + \xi_i \ , \ \xi_i \sim i.i.d\left(0, \sigma_\xi^2\right) \tag{4}$$

where λ is a scalar spatial error coefficient expressing the intensity of spatial autocorrelation between regression residuals.

4 EMPIRICAL RESULTS

The above presented spatial econometric models were used to detect and to treat spatial dependence in the model of *beta*-convergence on per capita GDP (defined at current market prices in Euro) of 252 NUTS2 EU regions[1] over the 2000-2011 period[2]. The data were

[1] We considered NUTS2 regions of EU corresponding to actual state in 2011. Due to the possible problems with isolated regions we excluded 20 island regions of Cyprus, Malta, France, Finland, Spain, Greece, Portugal and Italy.

retrieved from the web page of Eurostat [9] and the whole analysis was carried out in the software GeoDa (Geographic Data Analysis) [11]. The corresponding shapefile (.shp) for Europe was downloaded from the web page of Eurostat [10] and thereafter 252 NUTS 2 regions were selected in GeoDa.

This paper is aimed to test absolute and conditional *beta*-convergence hypotheses using both non-spatial and spatial approaches. We begin with estimation of the traditional absolute *beta*-convergence model (1) in which the effects of spatial dependence are not considered. The estimation of this model (Model1) was done by OLS and thereafter the tests aiming at detecting the presence of spatial dependence were carried out using a spatial weight matrix **W**. The spatial weight matrix **W** was specified as a contiguity weight matrix of queen case definition of neighbours. The estimation results are summarized in Table 1.

The parameter β of Model1 associated with the initial per capita GDP is significant and negative, which confirms the hypothesis of absolute convergence for the EU regions. This means that the regions with lower per capita GDP grew at a higher speed during the period 2000-2011. The speed of convergence associated with this estimation is 3.24% per year and half-life is 21 years. The Moran's *I* test adapted to the OLS regression residuals confirms a presence of spatial dependence but does not allow testing the presence of the two possible forms of spatial dependence. For this reason we use two LM tests as well as their robust counterparts. Robust LM tests show that the absolute *beta*-convergence model is misspecified due to spatial autocorrelation and more appropriate model is model SAR because the Robust LM (lag) statistic is statistically significant while the robust LM (error) statistic is not. Therefore the model of absolute *beta*-convergence should be modified to integrate this form of spatial dependence explicitly.

We proceeded with estimation of SAR model (Model2) following the model defined in (3). The estimation results by ML for Model2 are given in Table 1. The parameters are all strongly statistically significant; β coefficient is again negative confirming absolute *beta*-convergence hypothesis. The statistical significance of spatial autoregressive parameter ρ confirms the existence of spatial effects among neighbouring regions. In spatial autoregressive model, we can notice that the convergence process appears to be weaker if spatial effects are taken into account, i.e. speed of convergence is 1.33 % per year and half-life increased to 52 years.

The concept of so far tested absolute convergence is appropriate in case that the regions are homogeneous, i.e. consideration of initial conditions is irrelevant. Concerning our analysed group of regions, this assumption is not so straightforward; we therefore proceeded with testing of conditional *beta*-convergence hypothesis based on the model (2). The estimation results of this model (Model3) are given in Table 1. The region's specific dummy variable was used to test the conditional convergence hypothesis assuming that it controls for region specific factors in our case we chose the former political orientation of the analysed region (postcommunist or not). As in previous cases, the parameter β is statistically significant and negative and the parameter γ, associated with the above defined dummy variable, is also statistically significant which confirms the hypothesis of conditional convergence hypothesis. The results of LM tests and their robust versions led to the estimation of the model SEM (Model4 – see Table 1). Since the parameter β in this model was not statistically significant and also there was not notable difference in the statistical significance of the Robust LM(lag) and Robust LM(error) statistics, we decided to prefer the SAR model with region's dummy (Model5 – see Table 1). In addition, the results of Model5 provide statistically significant estimations of all parameters.

[2] 2011 was the last year of published statistics by Eurostat.

It can be noticed that models assuming conditional convergence hypotheses yield worse values of convergence characteristics. The convergence process according to these models seems to be weaker, e.g. the half-life of convergence – the time that it takes for 50% of the initial gap to be eliminated is in case of Model5 184,811 years and the speed of convergence is only 0,38% per year. The appropriateness of the spatial aspect was checked and confirmed by Moran's *I* statistics for spatial residuals (see the last row of Table 1).

Table 1: Estimation results of *beta*-convergence models

	Model1 (Linear model)	**Model2 (SAR model)**	**Model3 (Linear model, region's dummy)**	**Model4 (SEM model, region's dummy)**	**Model5 (SAR model, region's dummy)**
Estimation	OLS	ML	OLS	ML	ML
α	3.240***	1.459***	1.585***	0.564**	0.508**
β	-0.300***	-0.136***	-0.136***	-0.032	-0.040 **
γ	-	-	0.367***	0.532***	0.252***
ρ	-	0.584***	-	-	0.524***
λ	-	-	-	0.662***	-
R^2	0.692	0.814	0.749	0.857	0.835
Convergence characteristics					
Speed of convergence	0.0324 (3.24%)	0.0133 (1.33%)	0.0132 (1.32%)	0.0030 (0.30%)	0.0038 (0.38%)
Half-life	21.417	52.148	52.346	233.051	184.811
Tests					
Moran's *I* (error)	9.265***	-	11.012***	-	-
LM (lag)	103.470***	-	99.686***	-	-
Robust LM (lag)	23.295***	-	3.727*	-	-
LM (error)	80.783***	-	112.676***	-	-
Robust LM (error)	0.607	-	16.717 ***	-	-
Moran's *I* (spatial residual)	-	-0.062	-	-0.069	0.023

Note: Symbols ***, **, * indicate statistical significance at 1%, 5% and 10% level of significance, respectively.

5 CONCLUSION

This paper tests the hypothesis of regional income convergence in NUTS2 EU regions during the 2000-2011 period using both non-spatial and spatial approaches. The estimation of absolute *beta*-convergence model was the initial point of our analysis and the absolute income convergence among regions was confirmed. Since the spatial effects were proved, also the spatial version of this model was estimated based on SAR model. The second part of our analysis was oriented on testing of conditional income convergence including region's effects using the region's dummy (postcommunist region or not) controlling for different steady states of the analysed regions. As well as absolute convergence, the conditional convergence was also confirmed. Since the spatial dependence was again revealed by the Moran's *I* and the set of LM tests, the spatial versions of this model were used.

The results show that the inclusion of spatial effects and also the use of conditional convergence concept led to worsening of convergence characteristics. Despite these facts we prefer the conditional convergence models with included spatial effects for modelling of the income convergence process. The results of our analysis imply that convergence process is not determined only by a region's initial income, but also essentially by its neighbourhood

region's growth performance. Accordingly, we found out that inclusion of region specific factors that affect income growth in particular region seems to be appropriate addition to the model of income convergence. Similar conclusions were received also by other studies dealing with this subject of area.

Acknowledgement
This work was supported by the Grant Agency of Slovak Republic – VEGA grant No. 1/0285/14 "Regional modelling of the economic growth of EU countries with concentration on spatial econometric methods".

References

[1] Arbia, G. (2006). Spatial Econometrics. Statistical Foundations and Applications to Regional Convergence. Berlin Heidelberg: Springer-Verlag.

[2] Curran, D. (2009). British Regional Growth and Sectoral Trends – Global and Local Spatial Econometric Approaches. National Institute for Regional and Spatial Analysis. Working Paper 49. http://www.researchgate.net/publication/232983243_British_regional_growth_and_sectoral_trends_global_and_local_spatial_econometric_approaches [Accessed 05/02/2015].

[3] European Commission (2010). Europe 2020: A European strategy for smart, sustainable and inclusive growth. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF[Accessed 05/02/2015].

[4] European Commission (2012). Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union. Official Journal of the European Union, 55, pp. 1-390. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:326:FULL:EN:PDF [Accessed 05/02/2015].

[5] Hančlová, J. et al. (2010). Makroekonometrické modelování České ekonomiky a vybraných ekonomik EU. Ostrava: VŠB-TU Ostrava.

[6] Lundberg, J. (2003). The Regional Growth Pattern in Sweden –A Search for Hot Spots. Centre for Regional Science, Umea University. Working Paper 68. http://www.diva-portal.org/smash/get/diva2:227098/FULLTEXT01.pdf [Accessed 05/02/2015].

[7] Paas, T. at al. (2007). Econometric Analysis of income convergence in selected EU countries and their NUTS 3 level regions. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1078863 [Accessed 05/02/2015].

[8] Paas, T. and Schlitte, F. (2009). Spatial effects of regional income disparities and growth in the EU countries and regions. http://ec.europa.eu/eurostat/documents/1001617/4398377/S3P2-SPATIAL-EFFECTS-TIIU-PAAS-FRISO-SCHLITTE.pdf [Accessed 05/02/2015].

[9] http://ec.europa.eu/eurostat/ [Accessed 05/02/2015].

[10] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units [Accessed 05/02/2015].

[11] https://geodacenter.asu.edu/software/downloads. [Accessed 05/02/2015].

HOUSING AFFORDABILITY LEVELS IN SELECTED EUROPEAN COUNTRIES: HIERARCHICAL CLUSTERING APPROACH

Ksenija Dumičić, Berislav Žmuk and Anita Čeh Časni
University of Zagreb, Faculty of Economics and Business, Department of statistics
Trg J.F.Kennedyja 6, 10000 Zagreb, HR-Croatia
{kdumicic, bzmuk, aceh}@efzg.hr

Abstract: In this paper, after data exploration, the cluster analysis is used to identify which of the selected 8 European countries have similar housing affordability, measured by seven indicators: Price to income ratio; Gross rental yield city centre; Gross rental yield outside of centre; Price to rent ratio city centre; Price to rent ratio outside of city centre; Mortgage as a percentage of income and Affordability index. The results of the analysis have shown that there is a substantial difference in the housing affordability levels between analysed groups of European Union and Western Balkan countries.

Keywords: housing affordability indicators, hierarchical clustering, Ward's linkage method, squared Euclidean distances, Western Balkan countries.

1 INTRODUCTION

Housing affordability is, in this research, defined as the measure of population's ability to afford or to purchase a particular item, such as a house, indexed to the population's income. Although, multiple indices for measuring housing affordability exist, see [1, 2, 4, 12], [8] stated that the precise definition of housing affordability is ambiguous. Anyhow, the most general measure of housing affordability is simply the relationship between housing costs and income. In that sense, affordability considers not just housing, but also the quality of housing and whether the household has enough income remaining for other necessities of life after paying the cost of housing [7, 11]. Affordability, as defined by [11], is the challenge that each household faces in balancing the cost of its actual or potential housing and its non-housing expenditures, within their available income. With the aim of accurately measuring the housing affordability, [11] made a compelling argument in favour of the residual income approach as the logical alternative to the ratio approach. Furthermore, according to [3] housing affordability is difficult to define. Different definitions of housing affordability result with different estimates of the magnitude and distribution of the housing affordability. In order to define the affordability of housing in communities, some approaches compare median housing prices and median household income. In their research [6] focus on measures that account for an individual household's ability to afford a home, providing a conceptual review of three commonly used housing affordability indices in the United States.

As housing affordability is considered to be one of the key factors that can be used to describe the socioeconomic stability and development of a state [5], in this paper an exploratory cluster analysis is used to identify which of the selected 8 European countries have similar housing affordability, measured by seven selected housing affordability indicators. This research is very interesting since it gives some insights in housing affordability levels in four different groups of countries. Namely, from the Western Balkans (WB) group, five countries were selected (Albania, Bosnia and Herzegovina, Croatia, Montenegro and Serbia); from the New Member States (NMS) Slovenia; from the EU 15 group of countries, Austria; and, finally, from the South East European countries (which is also a candidate country), Turkey is included into the analysis. We provide and discuss the results of hierarchical clustering using Ward`s linkage and squared Euclidean distances, with

two, three and four cluster solutions, where four cluster solution gives the most interesting and economically interpretable results.

The reminder of this paper is the following. After brief introduction, in section two, data and methods used in the empirical analysis are presented. In section three the results of descriptive analysis of housing affordability in eight selected countries are given. Section four presents the results of conducted cluster analysis. Finally, section five concludes.

2 DATA AND METHODS

Housing affordability is observed using following seven indicators (variables) of housing affordability: Price to income ratio (*Price income*); Gross rental yield city centre (*Rent yield centre*); Gross rental yield outside of centre (*Rent yield out centre*); Price to rent ratio city centre (*Rent ratio centre*); Price to rent ratio outside of city centre (*Rent ratio out centre*); Mortgage as a percentage of income (*Mortgage income*); and Affordability index (*Afford index*). Each of the observed indicator measures housing affordability on a slightly different way. Observing them all together enables making a reliable conclusion about housing affordability level in a country.

Variable *Price income* is given as a ratio of median house price and median family disposable yearly income. The higher *Price income* is, the lower housing affordability level will be. Variable *Rent yield centre* is defined as a ratio of average yearly rent per square meter in the city centre and buying price per square meter city centre, multiplied by 100. This variable reveals if it is better to rent or to buy a house in city centre. Opposite to this variable, variable *Rent yield out centre* observes if it is better to rent or to buy a house outside of a city centre. Variable *Rent ratio centre* is an inverse of *Rent yield centre* variable. The higher *Rent ratio centre* is, the lower housing affordability level will be. Consequently, people are considering renting instead of buying a house in city centre. Similar measure, but for houses outside of a city centre is variable *Rent ratio out centre*. Variable *Mortgage income* is a ratio of monthly payment for 20 years mortgage and median family disposable monthly income multiplied by 100. Obviously the lower this variable is, the higher housing affordability level is. Finally, variable *Afford index* is a ratio of value 100 and variable *Mortgage income* [9].

The statistical analysis is conducted for overall 8 European countries: Albania, Austria, Bosnia and Herzegovina, Croatia, Montenegro, Serbia, Slovenia and Turkey. The data for analysed variables are collected for period from 2009 to 2015. Due to the limitation of the database which was used for collecting data, the data are not available for all countries during the whole analysed period. Accordingly, the main focus is given on the most recent year: 2015. The statistical analysis is based on descriptive statistics methods. Statistical cluster analysis is conducted, as well.

3 EXPLORATORY DESCRIPTIVE ANALYSIS OF HOUSING AFFORDABILITY

According to analysed housing affordability indicators, housing affordability level varies substantially among observed countries. In order to make comparison of housing affordability level among observed countries, actual values of variables are standardized and are presented in Table 1.

Table 1: Standardized variables of housing affordability for 8 European countries in 2015 [Authors' calculations, 10].

Country	*Variable*						
	Price income	***Rent yield centre***	***Rent yield out centre***	***Rent ratio centre***	***Rent ratio out centre***	***Mortgage income***	***Afford index***
Albania	0.62	-0.01	0.87	-0.23	-1.15	0.59	-0.69
Austria	-0.83	-0.56	-0.31	0.68	0.14	-1.37	1.77
Bosnia and Herzegovina	0.01	-0.35	-0.71	0.30	0.90	0.20	-0.43
Croatia	0.01	-0.43	-0.48	0.44	0.44	-0.16	-0.13
Montenegro	1.05	-0.45	-0.61	0.47	0.70	1.44	-1.08
Serbia	1.34	-0.69	-0.67	0.96	0.82	1.04	-0.93
Slovenia	-0.53	0.10	-0.22	-0.39	0.01	-0.97	0.91
Turkey	-1.66	2.39	2.13	-2.22	-1.87	-0.78	0.58

According to Table 1 the highest and the lowest values of each variable have been specially noted. The lowest standardized value of *Price income* variable was observed for Turkey (z=-1.66) whereas Serbia had the highest value (z=1.34). These results lead to the conclusion that Turkey had the highest and Serbia the lowest housing affordability level among observed countries in 2015.

Variable *Rent yield centre* shows that in Turkey (z=2.39) it was far better to buy a house in city centre than to rent it, in comparison to the other countries in 2015. On the other hand, in Serbia (z=-0.69) it was the best to rent a house in city centre than to buy it, when compared to the other countries in 2015. According to *Rent yield out centre* variable values, it can be concluded that in Turkey (z=2.13) it was also far better to buy a house outside of city centre than to rent. The *Rent yield out centre* variable values are in favour of renting a house outside of city centre instead of buying it in Bosnia and Herzegovina (z=-0.71), Serbia (-0.67) and in Montenegro (-0.61).

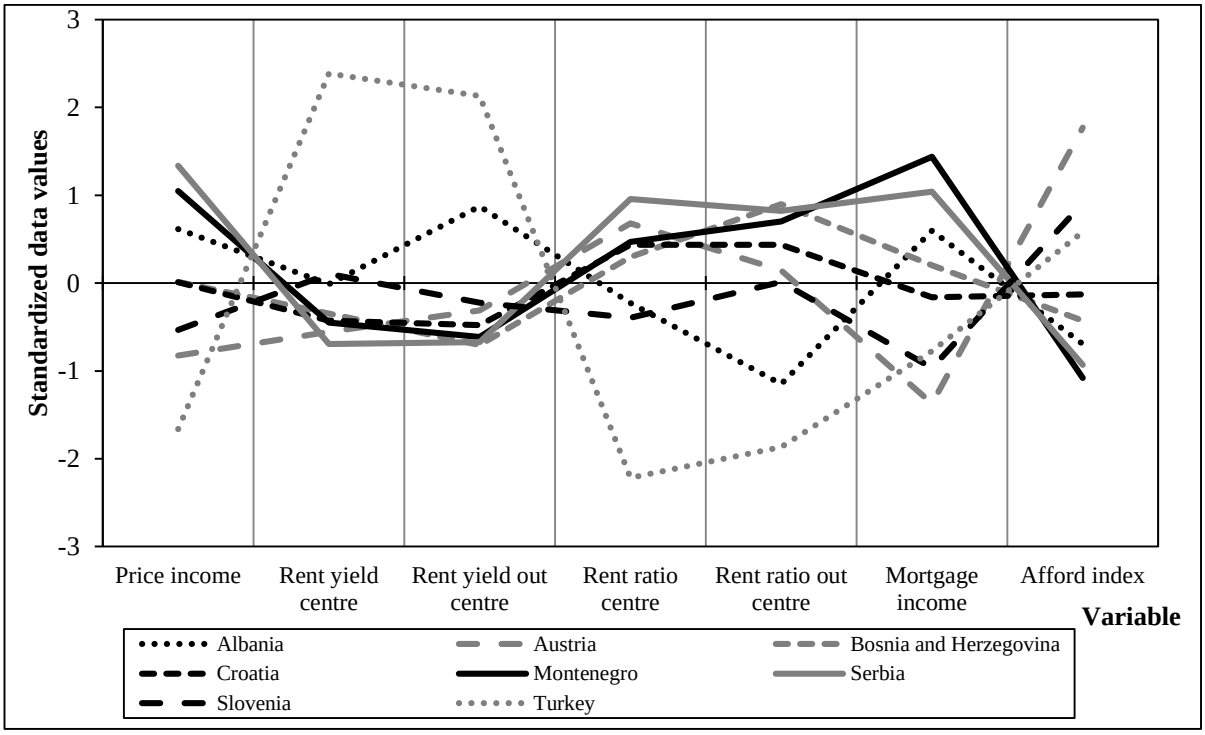


Figure 1: Profile diagram of the observed 8 countries using standardized values [Authors' calculations, 10].

According to variable *Rent ratio centre* it was far better to buy a house in city centre than to rent it in Turkey (z=-2.22) when compared to the other countries in 2015. That is also true for variable *Rent ratio out centre* (Turkey, z=-1.87). Again, the *Rent ratio centre* variable also shows that in Serbia (z=0.96) it was the best to rent a house in city centre than to buy it when compared to the other countries in 2015. Furthermore, according to variable *Rent ratio out centre* in Bosnia and Herzegovina (z=0.90), Serbia (z=0.82) and Montenegro (z=0.70) it was the highest recommended to rent a house in city centre than to buy it in compare to the other countries in 2015.

The lowest ratio of monthly mortgage payments and monthly income multiplied by 100 was in Austria (z=-1.37). This means that according to variable *Mortgage income*, the highest housing affordability level among the observed countries was right in Austria in 2015. In contrast, according to variable *Mortgage income* the lowest housing affordability level among the observed countries was right in Montenegro (z=1.44) in 2015. Exactly the same conclusions about housing affordability levels in the observed countries in 2015 can be made if the variable *Afford index* is observed.

Standardized values of all 7 observed variables considering 8 analysed countries are given in Figure 1. Figure 1 reveals that Turkey's values for 5 variables, out of 7, are quite different than they are for the other observed countries. Furthermore, it is obvious that regarding some variables, housing affordability level in Albania is very different than for other countries. The remaining six countries seem to be very "compact" with no significant and obvious differences in housing affordability levels measured by the variables of interest. However, the precise difference in housing affordability levels among the analysed countries is inspected using different approaches to statistical cluster analysis.

4 CLUSTER ANALYSIS OF HOUSING AFFORDABILITY

In order to determine which of the observed 8 European countries (Albania, Austria, Bosnia and Herzegovina, Croatia, Montenegro, Serbia, Slovenia and Turkey) have very similar housing affordability levels, the statistical cluster analysis was conducted. In the cluster analysis standardized values of 7 housing affordability indicators (variables *Price income*, *Rent yield centre*, *Rent yield out centre*, *Rent ratio centre*, *Rent ratio out centre*, *Mortgage income* and *Afford index*) for year 2015 were used. As amalgamation (linkage) rules in hierarchical clustering the following methods were used: single linkage, complete linkage, unweighted pair-group average, weighted pair-group average, unweighted pair-group centroid, weighted pair-group centroid and Ward's method. In all cases as a distance measure the squared Euclidean distances were used.

The used amalgamation (linkage) rules led to almost same results. The only difference appeared when single linkage was applied. Namely, in this case Albania appeared to be separated in cluster a step before than the same thing happened when other amalgamation rules were applied. Because there were no obvious differences between amalgamation rules applied in cluster classifications of countries, only the results provided by the Ward's linkage are given.

In Table 2 contents of clusters based on Ward's clustering method and squared Euclidean distances are shown. In the two-cluster solution Turkey was placed in the separate cluster, whereas in the second cluster all other observed countries were placed. In the three-cluster solution in the first cluster Turkey appeared, in the second one there are Albania, Bosnia and Herzegovina, Croatia, Montenegro and Serbia, whereas in the third cluster there are Austria and Slovenia. Further clustering put Albania in a separate cluster.

Consequently, it can be concluded that Turkey has different housing affordability level characteristics in comparison to all other observed countries. Furthermore, among these other

countries Austria and Slovenia appear to have similar housing affordability level characteristics than the following group of Western Balkan countries: Albania, Bosnia and Herzegovina, Croatia, Montenegro and Serbia. Also housing affordability level characteristics of Albania seem to be different of those for Bosnia and Herzegovina, Croatia, Montenegro and Serbia.

Table 2: Review of cluster memberships of the observed 8 countries, hierarchical clustering, Ward's linkage and squared Euclidean distances, with data for 7 variables, for 2015 [Authors' calculations, 10].

Number of clusters	***Clusters***			
	Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
2	Turkey	Albania, Austria, Bosnia and Herzegovina, Croatia, Montenegro, Serbia, Slovenia		
3	Turkey	Albania, Bosnia and Herzegovina, Croatia, Montenegro, Serbia	Austria, Slovenia	
4	Turkey	Bosnia and Herzegovina, Croatia, Montenegro, Serbia	Austria, Slovenia	Albania

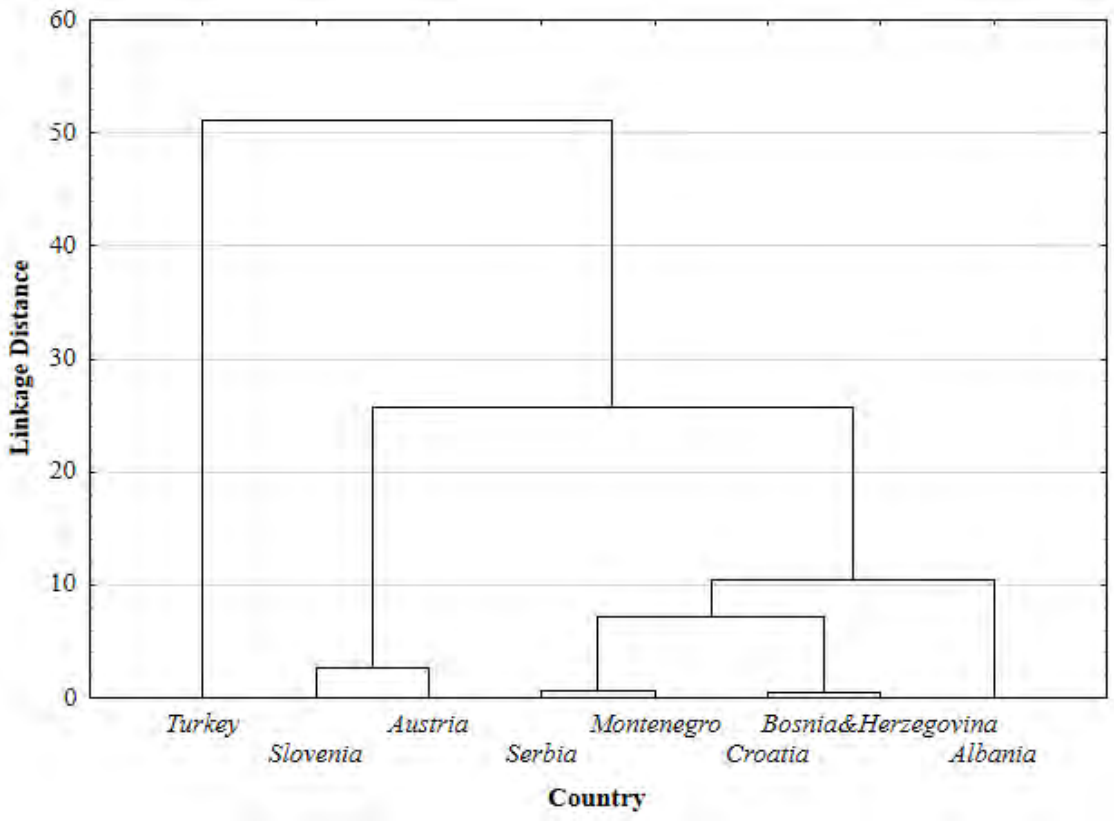


Figure 2: Dendrogram of the observed 8 countries based on the hierarchical clustering, the Ward's linkage and the squared Euclidean distances, with data for 7 variables, for 2015 [Authors' calculations, 10].

The dendrogram given in Figure 2, shows both clusters' and the linkage distances. It enables easier and faster understanding of differences in housing affordability levels in the 8 selected European countries.

5 CONCLUSION

Housing affordability is considered to be one of the key factors that can be used to describe the socioeconomic stability and development of a state. In this paper housing affordability is defined as the measure of population's ability to afford or to purchase a particular item, such

as a house, indexed to the population's income. An exploratory cluster analysis is used to identify which of the selected 8 European countries have similar housing affordability, measured by seven selected housing affordability indicators. The research gives some insights into the housing affordability levels in four different groups of countries: the WB countries, EU 15, NMS and SEE candidate country – Turkey.

The results of performed descriptive statistical analysis have shown that housing affordability level in Albania is very different than for other analysed countries. The remaining six countries had no significant and obvious differences in housing affordability levels measured by the variables of interest. According to performed hierarchical clustering using Ward`s linkage and squared Euclidean distances, the dendrogram shows that housing affordability level in Turkey is really different than it is in all other observed countries. Also, the difference in the housing affordability levels between observed European Union groups of countries: Austria (EU15) and Slovenia (NMS) and the observed Western Balkan countries (Albania, Bosnia and Herzegovina, Croatia, Montenegro and Serbia) can be considered as noticeable. Finally, the housing affordability level is quite different between Albania and observed ex-Yugoslavia countries (Bosnia and Herzegovina, Croatia, Montenegro and Serbia).

Acknowledgement

This work has been partially supported by Croatian Science Foundation under the project STRENGTHS no 9402.

References

[1] Belsky, E. S., Goodman, J. and Drew, R. (2005). Measuring the nation's rental housing affordability problems. http://www.jchs.harvard.edu/publications/rental/rd05-1_measuring_rental_affordability05.pdf [Accessed 08/06/2015].

[2] Bogdon, A. S. and Can, A. (1997). Indicators of local housing affordability: Comparative and spatial approaches. Real Estate Economics, 25(1): 43-60.

[3] Bourassa, S. C. (1996). Measuring the affordability of homeownership. Urban Studies, 33(19): 1867-1877.

[4] Combs, E. R., Combs, B. A. and Ziebarth, A. C. (1994). Housing affordability: A comparison of measures. Consumer Interests Annual, 41: 188-194.

[5] Dumičić K., Žmuk, B. and Čeh Časni, A. (2015). Forecasting Housing Affordability in Selected Post-Transition Economies. in: Jovancai Stakic, A., Kovsca, V., Bendekovic, J. (eds). Economic and Social Development Book of Proceedings, 5th Eastern European Economic and Social Development Conference on Social Responsibility, 21-22 October, 2015, Belgrade, Serbia, 134-143.

[6] Jewkes, M. D. and Delgadillo, L. M. (2010). Weaknesses of Housing Affordability Indices Used by Practitioners. Journal of Financial Counseling and Planning, 21(1): 43-52.

[7] Kutty, N. K. (2006). A new measure of housing affordability: Estimates and analytical results. Housing Policy Debate, 16(1): 113-142.

[8] Linneman, P. D. and Megbolugbe, I. F. (1992). Housing affordability: Myth or reality? Urban Studies, 29(3): 369-392.

[9] NUMBEO. (2015a). About Property Value and Investment Indicators At This Website. http://www.numbeo.com/property-investment/indicators_explained.jsp [Accessed 09/04/2015].

[10] NUMBEO. (2015b). Property Prices Index for Country 2015. http://www.numbeo.com/property-investment/rankings_by_country.jsp?title=2015 [Accessed 09/04/2015].

[11] Stone, M. E. (2006). What is housing affordability? The case for the residual income approach. Housing Policy Debate, 17(1): 151-184.

[12] Yip, N. M. and Lau, K. Y. (2002). Setting rent with reference to tenants' affordability: Public housing rent policy in Hong Kong. Jour. of Housing and the Built Environment, 17(4): 409-418.

THE PERFORMANCE OF THE MODIFIED HOLT-WINTERS METHOD: A STUDY OF 1428 MONTHLY SERIES FROM M3-COMPETITION

Liljana Ferbar Tratar
University of Ljubljana, Faculty of Economics
Kardeljeva pl. 17, 1000 Ljubljana, Slovenia
liljana.ferbar.tratar@ef.uni-lj.si

Abstract: Demand forecasting is used frequently in the world because of expedient source management and because the need for planning is becoming more important. Inventory related decisions are crucial for reduction of costs, increasing efficiency and improving customer service level. Demand forecasting and stock control are equally contributing towards such a decision making process. However, the stock control system is often examined independently of the demand forecasting process. In this paper we use 1,428 real time series from M3-Competition to evaluate the performance of the modified Holt-Winters method. From the results we show that essential reduction of supply chain costs can be achieved if we use the joint model with the modified Holt-Winters method.

Keywords: Demand forecasting, Holt-Winters method, Optimization, Inventory, M3-Competition.

1 INTRODUCTION

A fundamental aspect of supply chain management is accurate demand forecasting. Forecasting in a supply chain context has attracted a considerable amount of academic research. This is true for both fast moving items and, more recently, for slow movers or intermittent demand (see, for example, [1], [15], [10], [16]). In supply chain management, forecasts of demand are required on a regular basis for a very large number of products, so that inventory levels can be planned so as to provide an acceptable level of service to customers ([8]). The methods developed should therefore be fast, flexible, user-friendly, and able to produce results that are reliable and easy to interpret by a manager.

Exponential smoothing methods are a class of methods that produce forecasts with simple formulae, taking into account trend and seasonal effects of data (more details can be found in [3]). These procedures are widely used as forecasting techniques in inventory management and sales forecasting. Distinguished by their simplicity, their forecasts are comparable to those of more complex statistical time series models ([11]).

Unfortunately, most of the stock control studies consider demand data as an input to the model without explicitly considering that they are the results of a demand forecasting system. Even though this weakness has been highlighted in the academic literature, little empirical work has been conducted to develop understanding on the interaction between forecasting and stock control. In general, separate evaluation of the forecasting method and the stock control policy may easily lead to poorer overall performances. The choice based on measures that utilize only actual and forecasted demand data will bring worse results than the joint optimisation of total inventory costs.

Our objective in this paper is to point out that the most common procedures for evaluating a demand forecasting method, that is to compare forecast errors, are not appropriate. When the calculated forecasts are used in the other model, for instance the inventory model, they have to be evaluated on the basis of the total costs of the supply chain. We use 1,428 real time series from M3-Competition to evaluate the performance of the modified Holt-Winters method. From the results we will show that essential reduction of supply chain costs can be achieved if we use the joint model with the modified Holt-Winters method.

The remainder of the paper is organized as follows. In Section 2 we describe the classical Holt-Winters forecasting procedure, a modified Holt-Winters procedure, our model of the supply chain and we present the proposed joint model. After the description and classification of the real time series from M3-Competition (Section 3), in Section 4 a performance of the modified HW method is demonstrated and the main findings of the paper are described.

2 METHODOLOGY

2.1 The Holt-Winters forecasting procedures and an modified HW method

Exponential smoothing methods are a class of methods that produce forecasts with simple formulae, taking into account trend and seasonal effects of the data. The HW method estimates three smoothing parameters associated with level, trend and seasonal factors. We estimated smoothing and initial parameters in HW methods by minimising the mean square error (MSE).

In the multiplicative seasonal form of HW method (MHW) fundamental equations for level, trend, seasonal factors and forecast are ([12]):

$$L_t = \alpha(Y_t/S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (1)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (2)$$

$$S_t = \gamma(Y_t/L_t) + (1-\gamma)S_{t-s} \quad (3)$$

$$F_{t+m} = (L_t + b_t m)S_{t-s+m} \quad (4)$$

where m is the number of forecast ahead, s is the length of seasonality (e.g., number of months or quarters in a year) and Y_t is the observed data at time point t. There have been many suggestions for restricting the parameter space for α, β and γ ([7]). In this paper, we follow the traditional approach, requiring that all parameters lie in the interval $[0.1]$. These estimates are set to minimize the discrepancies between the in-sample one-step-ahead predictions F_{t+1} and the observed values Y_{t+1}.

Empirical study (see [2]) illustrates that the method used to designate the initial vector has very little effect on the accuracy of the predictions obtained when smoothing and the initial parameters of forecasting method are determined to minimise the forecast error measure. So, to initialize the level, we set $L_s = (Y_1 + Y_2 + \cdots + Y_n)/s$; to initialize the trend, we use $b_s = (Y_{s+1} - Y_1 + Y_{s+2} - Y_2 + \cdots + Y_{2s} - Y_s)/s^2$; and for initial seasonal indices we calculate $S_p = Y_p/L_s, p = 1,2,\cdots,s$.

The additive seasonal form of HW method (AHW) works with the following equations:

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (5)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (6)$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} \quad (7)$$

$$F_{t+m} = L_t + b_t m + S_{t-s+m} \quad (8)$$

The second of these equations is identical to (2). The only differences in the other equations are that the seasonal indices are now added and subtracted instead of relying on products and ratios. The initial values for level and trend are identical to those for the multiplicative method. To initialize the seasonal indices we use $S_p = Y_p - L_s, p = 1,2,\cdots,s$.

The only difference between the additive and modified HW method (MoHW) is in equation for level:

$$L_t = \alpha Y_t - S_{t-s} + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (9)$$

For the modified HW method in contrast to the additive HW method the smoothing parameter α occurs only at observed data Y_t and not at seasonal factor S_{t-s}. If we consider equation (9) and replace $S_t = \alpha S_t^*$, the equation (7) becomes:

$$S_t^* = \gamma^*(Y_t - L_t) + (1 - \alpha\gamma^*)S_{t-s}^*, \;\; \gamma^* = \gamma/\alpha \quad (10)$$

The other equations for the MoHW now conform to the AHW format. Thus, when we minimize forecast error with respect to the smoothing parameters, the new effect is to smooth the seasonal factors by changing them less. This effect has been noted by Miller and Williams ([13]), but not given in the form and researched as in this paper. The equations (9-10) for MoHW are similar to equations for damped additive trend ([4]), adjusted to seasonal factors. The initial values for level, trend and seasonal indices are identical to those for the additive method.

2.2 Symmetric relative efficiency measure

The efficiency of the MoHW method was measured in terms of the mean squared error (MSE) of the in-sample one-step-ahead forecasts and compared to that of AHW and MHW methods.

To compare the MoHW method with the existing AHW method, we first find their mean squared errors $\mathrm{MSE}_{\mathrm{MoHW}}$ and $\mathrm{MSE}_{\mathrm{AHW}}$ as defined above. We define the symmetric relative efficiency measure as

$$\mathrm{SREM}_{\mathrm{MoHW/AHW}} = \begin{cases} 1 - \frac{\mathrm{MSE}_{\mathrm{MoHW}}}{\mathrm{MSE}_{\mathrm{AHW}}}; & \mathrm{MSE}_{\mathrm{MoHW}} < \mathrm{MSE}_{\mathrm{AHW}} \\ \frac{\mathrm{MSE}_{\mathrm{AHW}}}{\mathrm{MSE}_{\mathrm{MoHW}}} - 1; & \mathrm{MSE}_{\mathrm{MoHW}} \geq \mathrm{MSE}_{\mathrm{AHW}} \end{cases}$$

The comparison with MHW and ETS method was done in a similar way.

Also, we use definition of SREM to compare the MoHW method with others methods regarding average costs (in these cases, SREM1 measures the percentage increase or decrease of the average costs):

$$\mathrm{SREM1}_{\mathrm{MoHW/method}} = \begin{cases} 1 - \frac{\mathrm{AC}_{\mathrm{MoHW}}}{\mathrm{AC}_{\mathrm{method}}}; & \mathrm{AC}_{\mathrm{MoHW}} < \mathrm{AC}_{\mathrm{method}} \\ \frac{\mathrm{AC}_{\mathrm{method}}}{\mathrm{AC}_{\mathrm{MoHW}}} - 1; & \mathrm{AC}_{\mathrm{MoHW}} \geq \mathrm{AC}_{\mathrm{method}} \end{cases}$$

2.3 The supply chain model and joint optimisation

In this section we show the advantage of joint optimisation over the optimisation based solely on forecasting data. The smoothing and initial parameters calculated by optimisation of the forecasting method are namely not optimal values for minimising the supply chain costs. Therefore, the joint optimisation was used and the initial and smoothing parameters were optimised to minimise the total costs of the company.

Consider a simple two-stage supply chain (with centralized demand information) consisting of one retailer (the most downstream unit of the supply chain) and one distributor. The retailer holds inventory in order to meet an external demand and places inventory replenishment orders to the distributor. Orders are placed at every time period. At time *t*, the last known value of the external demand is D_{t-1}. The retailer places an order Q_t to the distributor. We assume that the order placed one period ago is received (lead time is one period). External demand D_t is observed and filled. The unsatisfied demand is backlogged and causes penalty costs for the retailer. The distributor is able to supply any requested

quantity and delivery lead time is one period. The order placed at time t is received at time $t-1$ and is available to the retailer to fulfil external demand D_{t+1}.

Assuming that the retailer follows an order-up inventory policy, an order Q_t placed by the retailer to the distributor can be expressed as $Q_t = F_{t+1} - FS_t$, where $F_{t+1}$is the forecasted demand for the period *t+1* and FS_t is the final stock for the period *t* (if $FS_t > 0$ the retailer has on-hand inventory, if $FS_t < 0$ the unsatisfied demand occurs). When it is $Q_t < 0$, an order is not placed. The final stock is calculated as $FS_t = IS_t - D_t$, where the initial stock IS_t is obtained as $IS_t = Q_{t-1} + FS_{t-1}$. As the distributor has information about the external demand (centralized supply chain), it places the order, which is equal to the forecasted demand (less FS_t, if $FS_t > 0$). The missing amount of products supplied from the marketplace (assuming that a perfect substitute for the product exists) causes penalty costs for the distributor.

The costs of the supply chain are the sum of the holding costs and the penalty costs for all links in the supply chain. We assume the penalty costs to be higher than the holding costs, which is expressed by introducing a weight, *penalty*, that is greater than 1. In our analysis, for all calculations of total costs (average costs and minimised average costs) we assume that penalty is equal to 3 or 5.

In other words, using the common notation $X^+ = \max(X, 0)$, the supply chain costs at the time point *t* are expressed as (*n* – total number of links in the supply chain):

$$C_t = \sum_{l=1}^{n} C_t^l = \sum_{l=1}^{n} \left(\left(IS_t^l - D_t^l \right)^+ + penalty \cdot \left(D_t^l - IS_t^l \right)^+ \right)$$

Since demand data is usually considered as input to the model in stock control studies the average costs (for the period $t = T - s$) for forecasts obtained with different forecasting methods regarding minimising MSE were calculated (notation *method*). After that the smoothing and initial parameters of the joint model are estimated by minimising the average costs (notation *J_method*).

3 DATA

We used real seasonal time series from the M3-Competition to evaluate the performance of the MoHW method. The original time series data can be found in R package Mcomp ([5]). The analyses were carried out in the program R ([14]). Function sbplx from the nonlinear optimization package nloptr ([17], [9]) was used to estimate the smoothing parameters. The starting values in the minimization step were set to $\alpha_0 = \beta_0 = \gamma_0 = 0.5$ and the maximum number of iterations was set to $25{,}000$.

Table 1: Classification of monthly time series from M3-Competition.

Discipline	***Number***	***Noise***	***Trend***	***Season***	***Number***
DEMOGRAPHIC	111	A	N	N	123
FINANCE	145	A	N	A	115
INDUSTRY	334	A	A	N	167
MACRO	312	A	A	A	97
MICRO	474	M	N	N	124
OTHER	52	M	N	A	95
TOTAL	1428	M	N	M	124
		M	A	N	179
		M	A	A	56
		M	A	M	99
		M	M	N	159
		M	M	M	90
				TOTAL	1428

In our study, we analyzed 1428 monthly series. They refer to six different disciplines, as shown in Table. First we used ets function from R package forecast ([6], [7]) to classify the series by the form of their trend, seasonality and noise. Table also shows this classification. Here 'A' stands for 'additive', 'M' for 'multiplicative', and 'N' for 'none'.

We applied AHW, MHW and MoHW methods on each of the series independently of its discipline and ets classification. The estimated smoothing and initial parameters and in-sample MSE values were saved and SREM of MoHW with respect to AHW and MHW were computed.

4 RESULTS OF THE STUDY AND DISCUSSION

For each method and series, the symmetric relative efficiency measures (SREM and SREM1) of MoHW with respect to AHW, MHW and ETS were computed. Table 2 shows averages of SREM for monthly time series. We can observe that with the MoHW method the MSE can be reduced on average by more than 4% (6%) in comparison with the AHW (MHW) method. Also, the MoHW method outperforms ETS in 77% of cases, on average by almost 16%.

Table 2: Averages of the SREM and SREM1 for 1428 monthly time series from the M3-Competition.

	MSE → COST	**SREM**			**SREM1**					
					penalty = 3			**penalty = 5**		
		MoHW/ AHW	MoHW/ MHW	MoHW/ ETS	MoHW/ AHW	MoHW/ MHW	MoHW/ ETS	MoHW/ AHW	MoHW/ MHW	MoHW/ ETS
Discipline	DEMOGRAPHIC	3.3%	8.3%	13.7%	2.2%	6.4%	9.0%	2.5%	6.7%	10.1%
	FINANCE	4.1%	8.7%	18.0%	2.1%	3.5%	9.8%	2.2%	3.7%	10.5%
	INDUSTRY	2.5%	7.0%	6.7%	1.7%	4.4%	6.4%	1.8%	4.4%	7.4%
	MACRO	6.4%	4.3%	8.0%	4.1%	3.2%	6.9%	4.2%	3.5%	8.0%
	MICRO	5.3%	7.2%	26.0%	3.2%	5.0%	14.9%	3.5%	5.4%	15.5%
	OTHER	1.5%	6.6%	22.7%	1.0%	5.1%	9.6%	1.2%	5.6%	11.1%
Type	ANN	4.4%	5.3%	26.5%	1.9%	3.2%	14.3%	2.0%	3.3%	14.6%
	ANA	2.2%	7.1%	5.5%	1.5%	3.9%	6.1%	1.6%	3.9%	7.1%
	AAN	2.9%	7.1%	20.2%	1.7%	5.4%	12.4%	1.8%	5.8%	12.6%
	AAA	4.1%	6.7%	7.0%	2.0%	4.5%	5.6%	2.1%	4.8%	6.4%
	MNN	2.9%	7.8%	26.5%	1.8%	3.7%	14.4%	2.0%	4.2%	15.4%
	MNA	3.5%	8.3%	8.5%	2.2%	7.2%	6.8%	2.4%	7.4%	8.0%
	MNM	3.9%	5.2%	9.3%	3.3%	5.3%	8.3%	3.6%	5.7%	10.3%
	MAN	3.9%	6.1%	17.5%	2.2%	2.9%	10.1%	2.4%	3.1%	10.8%
	MAA	3.3%	6.2%	8.9%	1.9%	3.9%	6.5%	2.1%	4.1%	7.8%
	MAM	8.2%	5.1%	8.2%	4.6%	4.0%	6.9%	4.8%	4.5%	7.9%
	MMN	3.5%	6.6%	20.3%	2.5%	5.0%	11.9%	3.0%	5.2%	12.3%
	MMM	9.8%	7.7%	7.7%	6.5%	5.7%	5.9%	7.1%	5.9%	6.7%
	Total	**4.5%**	**6.7%**	**15.9%**	**2.7%**	**4.5%**	**10.1%**	**2.9%**	**4.7%**	**10.9%**

The MoHW method is particularly good in capturing the behavior of microeconomic time series, where the MoHW method performs better than ETS method on average by 26%. The MoHW method substantially outperforms other methods for classes with no seasonal component (xNN, xAN and xMN), irrespective of noise. Surprisingly, the fit of MoHW method is better even in xAA and xAM classes, where AHW and MHW methods are theoretically correct methods. This indicates the universality of the MoHW method regarding ETS which tries to select the most appropriate method.

Since demand data is usually considered as input to the model in stock control studies, the average costs (for the period $t = T - s$) for forecasts obtained with different forecasting methods were calculated. Table 2 also shows the averages of SREM1 (percentage of improvement of the average costs) of MoHW with respect to AHW, MHW and ETS. We can

observe that averages of the SREM1 are more than 2%, 4% and 10% (for penalty = 3 and penalty = 5) with respect to AHW, MHW and ETS. Almost the same as we observe for SREM holds for SREM1. If the MoHW substantially outperforms classical methods in some classes regarding MSE, the MoHW substantially outperforms them in the same classes regarding the average costs (as in this case the costs are calculated for forecasts considered as an input to the stock control model). We can also observe that the improvement of MoHW in comparison with other methods increases as penalty increases.

From the joint optimisation of supply chain (costs) model for 1,428 monthly series from the M3-Competition (see Table 3), we observe the following: on average JMoHW can reduce the average costs by 5.9% (7.2%) in comparison with JAHW (JMHW) for penalty =3 and by 9.2% (11.8%) for penalty = 5. We can see that the averages of the SREM1 increase as penalty increases. The JMoHW method outperforms JAHW and JMHW methods for all disciplines and it is particularly good for microeconomic and demographic time series. Also, the JMoHW method outperforms other two methods for all types and it is particularly good in MNA, MAM and MMx (with multiplicative noise and trend) classes.

Table 3: Averages of the SREM1 obtained with joint optimisation for 1428 monthly time series.

	JOINT	SREM1			
		penalty = 3		penalty = 5	
		JMOHW/ JAHW	JMOHW/ JMHW	JMOHW/ JAHW	JMOHW/ JMHW
Discipline	DEMOGRAPHIC	7.0%	8.9%	12.5%	16.1%
	FINANCE	4.0%	6.8%	7.1%	11.6%
	INDUSTRY	2.9%	5.1%	4.7%	7.1%
	MACRO	6.3%	5.5%	8.9%	8.3%
	MICRO	7.8%	9.5%	12.6%	16.9%
	OTHER	7.8%	7.3%	10.0%	11.1%
Type	ANN	5.8%	6.6%	7.2%	10.6%
	ANA	3.1%	4.9%	5.2%	7.1%
	AAN	5.7%	6.6%	9.1%	10.6%
	AAA	5.4%	6.7%	8.9%	11.8%
	MNN	5.3%	7.3%	8.0%	11.7%
	MNA	5.3%	10.0%	8.8%	15.6%
	MNM	4.7%	8.3%	8.5%	12.3%
	MAN	4.3%	5.6%	6.9%	7.9%
	MAA	6.1%	6.5%	9.9%	9.7%
	MAM	8.5%	8.7%	12.0%	14.9%
	MMN	7.1%	8.7%	13.1%	14.8%
	MMM	8.6%	8.4%	12.5%	16.6%
	Total	**5.9%**	**7.2%**	**9.2%**	**11.8%**

Finally, if we use joint optimisation with the MoHW method (JMoHW) instead of the models where forecasts are calculated with the AHW, MHW or ETS method regarding minimising MSE, we can observe the following (see Table 4): on average JMoHW can reduce the average costs by more than 24% (23% and 28%) in comparison with AHW (MHW and ETS) method for penalty = 3 and by more than 41% (40% and 43%) for penalty = 5.

The averages of the SREM1 within different discipline vary between 18.9% and 33.1% for penalty = 3 and between 33.5% and 48.9% for penalty = 5. The JMoHW substantially outperforms other methods for microeconomic time series. The averages of the SREM1 within different classes vary between 18.3% and 36.2% for penalty = 3 and between 35.5% and 51.9% for penalty = 5. The JMoHW method substantially outperforms the classical methods if a time series does not have a trend and a seasonal component. For these two

classes, ANN and MNN, the averages of the SREM1 vary between 25.9% and 36.2% for penalty = 3 and between 44.3% and 51.9% for penalty = 5.

Table 4: Averages of the SREM1 for 1428 *monthly time series from the M3-Competition.*

	JOINT/MSE	SREM1					
		penalty = 3			penalty = 5		
		JMOHW/ AHW	JMOHW/ MHW	JMOHW/ ETS	JMOHW/ AHW	JMOHW/ MHW	JMOHW/ ETS
Discipline	DEMOGRAPHIC	18.9%	19.0%	22.6%	33.5%	33.7%	36.7%
	FINANCE	21.5%	20.8%	27.0%	36.8%	36.6%	41.6%
	INDUSTRY	23.1%	23.2%	24.8%	39.6%	39.7%	41.1%
	MACRO	23.1%	22.3%	26.1%	39.2%	38.6%	42.1%
	MICRO	27.2%	25.9%	33.1%	46.2%	45.3%	48.9%
	OTHER	27.7%	28.6%	31.7%	46.2%	46.9%	48.3%
Type	ANN	27.3%	27.1%	36.2%	45.7%	45.7%	51.9%
	ANA	23.9%	24.5%	26.9%	41.0%	41.4%	43.5%
	AAN	20.1%	21.8%	27.9%	35.5%	37.1%	42.0%
	AAA	23.1%	23.4%	24.8%	40.7%	41.1%	42.2%
	MNN	27.1%	25.9%	35.2%	45.2%	44.3%	50.6%
	MNA	24.4%	26.6%	26.5%	42.3%	44.1%	42.9%
	MNM	22.6%	18.3%	21.0%	40.0%	36.7%	37.0%
	MAN	23.7%	23.1%	29.0%	40.2%	39.8%	44.4%
	MAA	23.3%	23.8%	25.3%	40.4%	40.9%	41.8%
	MAM	26.3%	23.4%	25.1%	42.8%	40.6%	41.6%
	MMN	24.6%	24.3%	30.5%	41.8%	41.7%	46.4%
	MMM	24.6%	20.6%	23.1%	40.6%	37.6%	38.7%
	Total	**24.1%**	**23.5%**	**28.1%**	**41.2%**	**40.8%**	**43.9%**

As we can see, the JMoHW method outperforms all three methods and it does not perform generally worse in none of the classes, which indicates the universality of the JMoHW method. The JMoHW method is general enough to be used as the encompassing method when the same method is applied to all time series.

5 CONCLUSION

Demand forecasting is used throughout the world more often because of proper source management and the rising need to plan. One of the most commonly used forecasting techniques is exponential smoothing, which is relatively inexpensive, fast and simple and does not demand special software.

In this paper we presented the modified Holt-Winters method and the problem of the local optimisation of forecasting methods when the calculated forecasts are used in the other models. We therefore proposed the MoHW method for a simultaneous optimisation of demand forecasting and a stock control policy. The method is computationally stable, requires little storage and produces results, which are easy to interpret.

We tested the method on 1,428 real series from M3-Competition. We developed the symmetric relative efficiency measure to compare the performance of different methods. Taking averages of these measures across several time series allowed us to indicate which method is preferable in general.

Based on the M3-Competition monthly time series we showed that the MoHW method is particularly good for microeconomic time series and for time series with multiplicative noise, trend and seasonal component. We showed that the method is general enough to be used as the encompassing method when the same method is applied to all time series.

As the method can be easily implemented in an Excel spreadsheet, we suggest the managers and supply chain decision makers to use JMoHW method to make better predictions and reduce the costs.

References

[1] Acar, Y. and Gardner, E.S. (2012). Forecasting method selection in a global supply chain. International Journal of Forecasting, Vol(28): 842–848.

[2] Bermúdez, J.D., Segura, J.V. and Vercher, E. (2006). A decision support system methodology for forecasting of time series based on soft computing. Computational Statistics & Data Analysis, Vol(51): 177 – 191.

[3] Gardner, Jr. E.S. (1985). Exponential smoothing: The state of the art. International Journal of Forecasting, Vol(4): 1–28.

[4] Gardner, Jr. E.S. (2006). Exponential smoothing: The state of the art—Part II. International Journal of Forecasting, Vol(22): 637–666.

[5] Hyndman, R.J., Akram, M. and Bergmeir, C. (2013). Mcomp: Data from the M-competitions. R package version 2.05.

[6] Hyndman, R.J., Athanasopoulos, G., Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C. and Wang, E. (2014). Forecast: Forecasting functions for time series and linear models. R package version 2.05.

[7] Hyndman, R.J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, Vol (27).

[8] Hyndman, R.J., Koehler, A.B., Snyder, R.D. and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, International Journal of Forecasting, Vol(18): 439–454.

[9] Johnson, S.G. (2013). The NLopt nonlinear-optimization package. Version 2.4.1.

[10] Liao, T.W. and Chang, P.C. (2010). Impacts of forecast, inventory policy, and lead time on supply chain inventory - A numerical study. International Journal of Production Economics, Vol(128): 527–537.

[11] Makridakis, S. and Hibon, M. (2000). The M3-competition: results, conclusions and implications. Int. J. Forecasting, Vol(16): 451–476.

[12] Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting : methods and applications. United States of America: John Wiley & Sons, Inc.

[13] Miller, D.M. and Williams, D. (2003). Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy. International Journal of Forecasting, Vol(19): 669–684.

[14] R Core Team (2014). R: A language and environment for statistical computing. Version 3.1.0. R Foundation for Statistical Computing, Vienna, Austria.

[15] Syntetos, A.A., Babai, M.Z., Davies, J. and Stephenson, D. (2010). Forecasting and stock control: A study in a wholesaling context. International Journal of Production Economics, Vol(127): 103–111.

[16] Teunter, R. and Sani, B. (2009). On the bias of Croston's method. European Journal of Operational Research, Vol(194): 177–183.

[17] Ypma, J. and Borchers, H.W. (2014). nloptr: R interface to NLopt. R package version 1.0.0.

eGOVERNMENT UTILIZATION IN EUROPEAN LANDSCAPE: HIERARCHICAL CLUSTER ANALYSIS

Mirjana Pejić Bach and Jovana Zoroja
Faculty of Economics & Business – Zagreb, University of Zagreb, Department of Informatics
Trg J.F. Kennedyja 6, HR-10000 Zagreb, Croatia
{mpejic, jzoroja}@efzg.hr

Abstract: eGovernment applications have become widely used in recent two decades with the goal to improve communication and cooperation between private and public entities. Goal of the paper is to explore the recent trends of eGovernment utilization by the enterprises in European countries with the usage of hierarchical cluster analysis. Cluster analysis revealed that European countries can be partitioned into homogenous groups regarding eGovernment usage. However, results indicate that the level of eGovernment utilization is not directly related to the economic development and competitiveness of a country, although it is related to the perceived barriers towards eGovernment.

Keywords: eGovernment, Competitiveness, Cluster Analysis, Hierarchical Cluster Analysis, Global Competitiveness Report, Barriers

1 INTRODUCTION

eGovernment provides better services for its citizens and enterprises using Internet and information and communication technologies (ICTs) [7; 15]. eGovernment offers many improvements, e.g. transformation of bureaucratic mechanisms and administration, increasing participation, openness, transparency and communication with users [20]. Stier (2015) indicate that eGovernment increases efficiency and transparency of government operations, strengthening democracy and better services to citizens and business [20].

Development of eGovernment started in OECD countries during 1990s with the increasing development and usage of Internet [5, 2]. Estonia is one of the leading countries regarding ICTs usage in public administration, which is the result of the investments in this area. In the last 10 years Estonia spend 1% of national budget on development and usage of ICTs in public services [8].

Number of authors has indicated in their research that factors which influence the most on the eGovernment usage are: economic and technical readiness of the countries, and understanding of public authorities to citizens' need and lower costs [12; 18; 11]. In order to clarify the factors that could influence eGovernment usage, our work is focused on two research goals. First, we aim to shade some light into the similarities and differences among European countries according to their recent usage of eGovernment. Second, we investigate if the European countries that are similar to the level of usage of eGovernment are, in the same period, similar according to their level of economic development and competitiveness, as well as according to the perceptions of their companies regarding the barriers in eGovnernment. In the first stage, the cluster hierarchical analysis is used to organize European countries into sensible groupings using data from Eurostat for 2013, according to their usage of eGovernment [3]. In the second stage, defined clusters are compared by the means of Anova analysis according to: (i) Global Competitiveness Indeks (GCI) and GDP per capita, and (ii) perceptions regarding the barriers in eGovernment [17].

The paper is organized as follows. The first section provides an introduction. In the second section data and the model specification are explained. Results of the cluster and the Anova analysis are described in the third section. At the end, a discussion and conclusion close the paper.

2 METHODOLOGY

2.1 Quantifying eGovernment utilisation

Data on eGovernment utilisation and data on perceived barriers towards the utilization of eGovernment were collected for 32 European countries (EU countries, Iceland, Norway, Macedonia and Turkey). Other European countries were not used in the analysis because of missing data for 2013 for the selected variables. The percentage of enterprises in each country that are using specific form of eGovernment and that encountered specific barrier is outlined in the Table 1.

Table 1: Utilization and perceived barriers of eGovernment in selected European countries in 2013

Area/ country	*Utilization of eGovernment*					*Perceived barriers towards the utilization of eGovernment*				
	eTendering	*eForms-obtaining*	*eForms-returning*	*eVAT*	*eSocial*	*Confidentiality and security*	*Complicated and time consuming*	*Requiring paper of person*	*Not being aware [d]*	*Any of the reasons*
EU-28	13	77	74	59	55	21	29	33	24	54
EU-27	13	77	74	59	55	21	29	32	24	54
EU-15	12	76	73	61	54	20	29	29	25	51
Belgium	12	72	74	65	39	26	31	24	27	56
Bulgaria	10	77	79	75	73	25	15	20	16	44
Czech Republic	19	90	81	41	30	35	46	53	26	77
Germany	7	66	61	46	48	29	19	29	25	52
Estonia	25	81	80	73	73	8	13	33	5	40
Ireland	30	88	95	84	77	18	14	25	10	39
Greece	10	92	81	78	74	24	14	22	13	44
France	19	91	87	83	80	21	46	30	41	61
Croatia	19	84	81	71	57	33	36	58	15	75
Italy	9	73	58	27	27	27	53	51	40	79
Cyprus	16	82	47	9	12	16	21	39	14	60
Latvia	18	79	88	83	84	27	20	47	26	61
Lithuania	30	99	99	95	96	8	10	26	6	33
Luxembourg	10	84	65	46	35	19	28	36	33	61
Hungary	12	82	81	73	71	8	8	33	10	39
Malta	18	79	58	15	25	20	27	36	12	57
Netherlands	13	83	85	71	57	2	2	2	1	3
Austria	15	84	77	60	51	30	20	38	17	59
Poland	24	81	86	29	70	32	36	62	34	77
Portugal	19	81	85	78	79	25	33	37	25	61
Romania	15	57	52	48	49	19	23	31	13	56
Slovenia	1	86	81	79	78	13	18	29	33	49
Slovakia	22	86	71	54	59	42	57	74	16	89
Finland	6	92	89	75	75	8	19	29	25	46
Sweden	19	92	87	64	64	7	18	23	28	43
United Kingdom	14	80	87	84	65	6	12	12	11	24
Norway	22	85	89	75	54	8	20	30	15	43
FRJ Macedonia	29	64	56	50	35	16	15	37	8	53

Source: Author's research based on Eurostat data (2014)

2.2 Hierarchical cluster analysis and Anova analysis

In this paper we use following methodological approach. Firstly, we use hierarchical cluster analysis in order to organize European countries into sensible groups for the year 2013, regarding their usage of eGovernment. Secondly, we used Anova analysis to compare identified clusters.

One of the best ways to analyse a large amount of data is to classify them into groups which members are similar among each other and are different compared to members of other groups [13]. The cluster analysis is one of a statistical classification method which is used to examine hidden data structure and to group objects into homogeneous groups based

on their similar characteristics [9; 14]. There are four main steps of cluster analysis: feature selection, algorithm selection, cluster validation and results. It is important to highlight that cluster analysis is not a one-step process, but a process of several iterations [21].

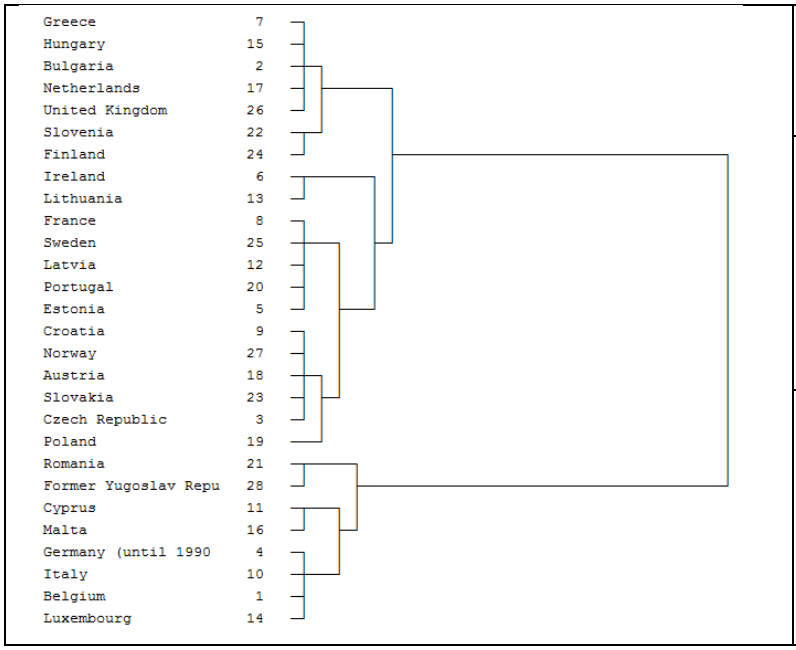
Cluster analysis can be divided into two main groups according to the algorithm usage: partitioned clustering and hierarchical clustering [4]. In this paper we used hierarchical divisive clustering, in which data can be classified with a sequence of nested partitions, which refers to the classification starting a cluster including all individuals towards the smaller clusters and backwards, in other words using a "top down" approach [10]. After the cluster analysis is conducted, the Anova analysis is often used with the goal to test differences among clusters according to the selected differences [6].

In the first stage, data on the usage of eGovernment in 32 European countries was used for cluster determination. Using the hierarchical cluster analysis, countries were grouped regarding indicators of eGovernment usage by enterprises presented in Table 1 (left columns). In this paper we used the Ward-method as the clustering method. In the second stage, we used Anova analysis to compare identified clusters according to data presented in Table 1 (right columns): (i) their competitiveness level and GDP per capita, and (ii) their perceptions regarding the barriers in eGovernment.

3 RESULTS

3.1 Hierarchical cluster analysis results

According to eGovernment usage, selected countries are divided into clusters (Figure 1). The results showed that 28 selected European countries are grouped into three clusters regarding the rank of a particular cluster.



Cluster A (N=7)	Bulgaria, Finland, Greece, Hungary, Netherlands, Slovenia, United Kingdom
Cluster B (N=13)	Austria, Croatia, Czech Republic, Estonia, France, Ireland, Latvia, Lithuania, Norway, Poland, Portugal, Slovakia, Sweden
Cluster C (N=8)	Belgium, Cyprus, FRY Macedonia, Germany, Italy, Luxembourg, Malta, Romania

Source: Author's research based on Eurostat data (2014)

Figure 1: Selected European countries grouped into specific clusters based on different characteristics and forms of eGovernment usage; 2013

In Cluster A there are seven countries: Bulgaria, Finland, Greece, Hungary, Netherlands, Slovenia and United Kingdom. It can be noted that the Cluster A includes some of the most developed European countries, e.g. Finland, as well as developing countries, e.g. Bulgaria. Cluster B consists of 13 countries: Austria, Croatia, Czech Republic, Estonia, France, Ireland,

Latvia, Lithuania, Norway, Poland, Portugal, Slovakia and Sweden. This cluster also consists of the most developed northern European countries, e.g. Norway, and developing countries, e.g. Croatia. Cluster C consists of 8 countries: Belgium, Cyprus, FRY Macedonia, Germany, Italy, Luxembourg, Malta and Romania. This cluster again consists on countries that are different according to their economic development. It can be concluded that countries within Cluster A, Cluster B or Cluster C are not similar regarding socio-economic development neither regarding geographic position. In all three clusters developed and developing European countries can be found. However, they are similar to their level of eGovernment usage by the enterprises.

Table 2 represents descriptive statistics of the eGovernment usage indicators.

Table 2: Descriptive statistics of the eGovernment usage indicators across clusters (% of enterprises)

Ward Method	eTendering	eForms-obtaining	eForms-returning	eVAT	eSocial
Cluster A (n=7)	9,429 (4,541)	83,143 (4,776)	83,286 (3,729)	**76,429 (4,315)**	**70,429 (7,161)**
Cluster B (n=13)	**21,615 (4,556)**	**86,231 (5,644)**	**85,077 (7,342)**	68,462 (18,577)	67,231 (17,210)
Cluster C (n=8)	14,500 (6,949)	72,125 (9,403)	58,875 (8,184)	38,250 (19,241)	33,750 (12,314)
Total (n=28)	16,536 (7,285)	81,429 (8,925)	77,143 (13,542)	61,821 (22,241)	58,464 (20,915)

Source: Author's research based on Eurostat data (2014)

Note: Means and standard deviations (in parentheses) are provided; bold letters indicate the highest average value

In Cluster A there is the highest percentage of enterprises according to the usage of eVAT (76,429% of enterprises) and eSocial (70,429% of enterprises). In Cluster B there is the highest percentage of enterprises according to the usage of eTendering (21,615% of enterprises), eForms obtaining (86,231% of enterprises) and eForms returning (85,077% of enterprises). Countries in Cluster C are those with the lowest average values of all eGovernment indicators.

3.2 Anova analysis results

In this section results of the Anova analysis are presented (Table 3).

Table 3: Descriptive statistics of Global Competitiveness Index, GDP per capita and perceptions regarding the barriers in eGoverment and the ANOVA analysis; 2013

	Cluster A		Cluster B		Cluster C		F-value
	Mean	St.Dev	Mean	St.Dev	Mean	St.Dev	
GCI	4,753	0,717	4,692	0,469	4,635	0,548	0,083
GDP per capita (000 EUR)	452,000	1131,062	129,408	197,004	61,513	75,701	1,009
Perceptions regarding the barriers in eGovernment							
Confidentiality and security	12,286	8,958	22,615	11,948	21,500	5,099	2,741*
Complicated and time consuming	12,571	5,940	28,385	14,936	27,125	11,667	4,042**
Requiring paper of person	21,000	10,924	41,231	16,161	35,375	8,017	5,438
Not being aware about eGovernment services	15,571	10,518	20,308	10,773	21,500	11,452	0,624**
Any of the reasons	35,571	16,521	58,308	17,698	59,250	8,548	5,934***

Source: Author's research based on Eurostat data (2014) and Schwab (2014)

Note: *** statistically significant at 1%, ** statistically significant at 5%, * statistically significant at 10%

Comparison according to Global Competitiveness Index

In order to estimate the relationship of eGovernment usage and competitiveness of selected European countries in 2013, calculated average values of GCI of each cluster are presented in Table3. It can be seen that average values of GCI are the highest in Cluster A. Cluster B (4,692) and Cluster C (4,635) have almost the same average values but less than Cluster A.

Anova analysis indicates that found differences are not statistically significant.

Comparison according to GDP per capita
In order to estimate the relationship of eGovernment usage and competitiveness of selected European countries in 2013, calculated average values of GDP per capita of each cluster are presented in Table 3. It can be seen that average values of GDP per capita are the highest in Cluster A (452,000) where are grouped one of the most developed European countries (Finland, Netherlands and United Kingdom). Cluster C (61,513) has the lowest average value for GDP per capita where the less developed European countries can be found (FRY Macedonia, Malta and Romania). Anova analysis indicates that found differences are not statistically significant.

Comparison according to perceptions regarding the barriers in eGovernment
In order to estimate the relationship of eGovernment usage and competitiveness of selected European countries in 2013, calculated average values for the barriers in eGovernment usage of each cluster are presented in Table 3. It can be seen that the highest average values are in the Cluster B for following indicators: *Confidentiality and security*, *Complicated and time consuming* and *Requiring paper of person*. While, for the other two barriers the average values are the highest in Cluster C: N*ot being aware about eGovernment services* and *Any of the reasons*). The lowest average values for all mentioned barriers in eGovernment are in Cluster A where are grouped one of the most developed European countries (Finland, Netherlands and United Kingdom). The conducted analysis showed that calculated average values for barriers in eGovernment are statistically significant for all indicators, except for the *Requiring paper of person*. Anova analysis revealed that differences among clusters are statistically significant for the following barriers: (i) *Confidentiality and security* is statistically significant at 10%, (ii) *Complicated and time consuming* and *Not being aware about eGovernment services* are statistically significant at 5%, (iii) and indicator *Any of the reasons* is statistically significant at 1%.

4 CONCLUSION

Results of our study indicated that European countries could be grouped into three clusters using hierarchical cluster analysis according to their usage of eGovernment which was our first goal of the paper. The countries grouped together in the clusters differ a lot to each other according both to the level of their economic development as well as to their geographical position. However, European countries grouped together had similar percentage of enterprises that encountered barriers regarding the utilization of the eGovernment. The conducted Anova analysis showed that calculated average values are not statistically significant for GCI, GDP per capita and for one barrier in eGovernment: *Requiring paper of person*. Nonetheless, Anova analysis revealed that differences among clusters are statistically significant for other four barriers in eGovernment. Therefore, the conclusion for our second goal could be that the level of eGovernment utilization in particular country is more related to the perceived barriers then to the level of the economic development of the country.

In this paper we collect data for the year 2013 and for the European countries. In order to expand this research, the analysis of data for the year 2015 should be conducted with the goal to compare results from the year 2013 and 2015. Expanded analysis would also show if there is any progress in eGoverment usage during the last two years, especially in the area of eTendering. Also, it will be useful to compare other non-European countries according to their level of eGovernment usage. In this paper, we have not corroborated former researchers regarding impact of socio-economic development on eGovernment usage [12; 18; 11], but we

indicate that the perceived barriers have the stronger influence. Therefore, in further research influence of national policies regarding eGovernment usage should be examined.

References

[1] Candiello, A., Albarelli, A., Cortesi, A. (2012). Quality and Impact Monitoring for Local eGovernment Services. *Transforming Government: People, Process and Policy*, 6(1): 112-125.

[2] Elliman, T., Irani, Z., Jackson, P. (2007). Establishing a Framework for eGovernment Research: Project VIEGO. *Transforming Government: People, Process and Policy*, 1(4): 364-376.

[3] European Commission Database - Eurostat (2014). http://ec.europa.eu/eurostat/web/information-society/data/main-tables [Accessed 22/05/2015]

[4] Everitt, B., Landau, S., Leese, M. (2001). Cluster Analysis, 4th Edition. London: Arnold.

[5] Guijarro, L. (2007). Interoperability Frameworks and Enterprise Architectures in E-Government Initiatives in Europe and the United States. *Government Information Quarterly*,24(1): 89-101.

[6] Hair, J.F., Black B., Babin, B., Anderson, R.E., Tatham, R.L. (2005). Multivariate Data Analysis, 6th Edition. New Jersey: Prentice Hall.

[7] Holliday, I., Yep, R. (2005). E-Government in China. *Public Administration and Development*, 25(3): 239–249.

[8] International Telecommunication Union (2012). *Measuring the Information Society*. http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012_without_Annex_4.pdf [Accessed 25/05/2015]

[9] Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8): 651-666.

[10] Jain, A., Dubes, R. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall.

[11] Janssen, M., Kuk, G., Wagenaar, R.W. (2008). A Survey of Web-based Business Models for eGovernement in The Netherlands. *Government Information Quarterly*, 25(2): 202-220.

[12] Khalil, O.E.M. (2011). E-Government Readiness: Does National Culture Matter? *Government Information Quarterly*, 28(3): 388-399.

[13] Parr Rud, O. (2000). Data Mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management. New York: John Wiley & Sons Ltd.

[14] Rodriguez-Deniz, H., Voltes-Dorta, A. (2014). A Frontier-based Hierarchical Clustering for Airport Efficiency Benchmarking. *Benchmarking: An International Journal*, 21(4): 486-508.

[15] Rorissa, A., Demissie, D., Pardo, T. (2011). Benchmarking eGovernment: A Comparison of Frameworks for Computing eGovernment Indeks and Ranking. *Government Information Quarterly*, 28(3): 354-362.

[16] Sarrayrih, M. A., Sriram, B. (2015). Major Challenges in Developing a Successful eGovernment: A Review on the Sultanate of Oman. *Journal of King Saud University-Computer and Information Sciences*, 27: 230-235.

[17] Schwab, K. (2013). The Global Competitiveness Report. Geneva: World Economic Forum. http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2012-13.pdf[Accessed 22/05/2015]

[18] Stier, S. (2015). Political determinants of e-government performance revisited: Comparing democracies and autocracies. *Government Information Quarterly*, ISSN 0740-624X, http://dx.doi.org/10.1016/j.giq.2015.05.004. [Accessed 10/06/2015]

[19] Verdegem, P., Verleye, G. (2009). User-Centred E-Government in Practice: A Comprehensive Model for Measuring User Satisfaction. *Government Information Quarterly*, 26(3): 487-497.

[20] Wong, K., Fearon, C., Philip, G. (2007). Understanding eGovernment and eGovernnance: Stakeholders, Partnerships and CSR. *International Journal of Quality and Reliability Management*, 24(9): 927-943.

[21] Xu, R., Wunsch, D.C. II (2008). Recent Advances in Cluster Analysis. *International Journal of Intelligent Computing and Cybernetics*, 1(4): 484-508.

ECONOMIC AND DEMOGRAPHIC DETERMINANTS OF DEMAND FOR LIFE INSURANCE

Snježana Pivac, Branka Marasović and Dujam Kovač
Faculty of Economics Split, University of Split
Cvite Fiskovića 5, 21000 Split, Croatia
snjezana.pivac@efst.hr, branka.marasovic@efst.hr, dkovac01@student.efst.hr

Abstract: The insurance industry, particularly its stronger component life insurance, is increasingly gaining in importance as financial intermediary that facilitate the role of the financial system in economic growth. Important position of the life insurance industry will be explained by economic and demographic determinants on the example of the transition countries of central Europe, applying a static panel model. According to the results of previous studies, hypothesis has been confirmed - a rise in income contributes to the growth of the insurance market and the demand for life insurance is function of bequest motive. The limiting factor in the development of the life insurance market is unemployment, which is becoming very disturbing in the study on a group of countries. As an additional reason for the interest of the study stands out relatively few studies comparing them with the banking sector.

Keywords: Life insurance demand, Transition countries of central Europe, Static panel analysis

1 INTRODUCTION

The growth rates of life insurance significantly exceed the growth rate of world gross domestic product (GDP), which highlights the role of the insurance sector in the economy and determines the increased importance of the insurance sector in the execution of financial intermediation.
However, the rate of growth of life insurance differed not only between countries with different levels of development, but also between countries of the same level of development [7]. Therefore, the question arises, what causes occurrence of variations and what determines the demand for life insurance. Previously conducted studies have created a clear view of a group of factors with impact on the demand for life insurance that are justifiably explored. There are different economic, demographic, social and institutional factors of demand for life insurance. Identified core factors used in previous research show often ambiguous result depending on the measure used to describe demand for life insurance. Various measures created a difficulty in presenting general conclusions about the direction of the relationship and the importance of the impact of individual factors on the life insurance industry, which is evident in the work [3] while comparing different models with different measures. Consequently, there is limited understanding of the contribution of life insurance on the real economy. This paper examines the economic and demographic determinants of demand for life insurance in the group of transition countries that are members of the European Union (EU) and which geographically belong to the region of Central Europe. Considered countries; Bulgaria, Croatia, Czech Republic, Hungary, Slovakia and Slovenia constitute a homogenous group according to the criteria of development of the life insurance market measured by the indicator of the density of life insurance or penetration of life insurance. The homogeneous group of countries allows creating model, which is adapted to the characteristics of all the countries, whose estimates reflect the actual properties of each of the observed countries [2]. The empirical confirmation of the drivers of demand is carried out by a panel method that optimizes scope of country, through the criterion of homogeneity, and the volume of data required to present a valid statistical conclusions. The insurance market of selected countries is analysed for the ten-year period from 2003 to 2012. The paper is organized as follows. The second section provides an overview of

the determinants that are included in the model. The third section presents the data and methodology involved in the analysis and presents the results. A concluding section summarizes the main results in the analysis and insights that were obtained on the basis of research.

2 DETERMINANTS OF LIFE INSURANCE

The life insurance can be either “pure insurance” products, savings products, or a combination of both [16]. Based on existing research and data availability, the following factors are abstracted.

2.1 Economic determinants of life insurance

Disposable income is considered as the central variable in the models that explain the demand for life insurance regardless of the level of economic development of the country. The possible loss due to occurrence of the insured event is more significant following the increase in income [8]. As a result of the increase in income, life insurance contract gains value. In addition, the reason for the positive relationship between the level of disposable income and demand for life insurance is the possibility to invest in life insurance product since higher income generally allows settlement necessities of life while retaining surplus funds. Disposable income, which is measured by the ratio of GDP per capita, due to the availability, is most commonly used variable and it can be considered as an appropriate measure for permanent income [3]. Research conducted [5] shows that the life insurance sector starts to grow faster in developing countries with higher income than in developing countries with lower income. **Inflation** reduces the value of the agreed life insurance contract, making a product of life insurance less attractive. The promised future payments have less value, hence it is logical less demand for life insurance. Reduced demand is achieved through two channels; the first is related to reduced consumer confidence; the second refers to the real interest rate [6]. Regardless of the type of insurance, promised payments are guaranteed in a future time period and as such are influenced by changes in the price of money. Inflation will affect the reduction in the purchasing power of future payments dependent by insured event. As a result of the process of adapting to clients' needs, it is enabled the contracting indexed life insurance which in turn can cause the effect inflation being irrelevant. Despite the existence of indexed insurance policy, whose purpose is to neutralize the impact of inflation, the data time series when there is inflation in the country, Brazil; indicate a significant reduction in demand for insurance [1]. **Unemployment** is a position where an individual is unable to compensate for work which is the basic source of all product and service consumption and therefore life insurance products. The total income that an individual gets, does not have to originate only from work, but for the majority of the population, income from work is the main source of consumption and wealth accumulation. Higher unemployment is considered to have negative impact on the demand for life insurance. The evidence on the impact of unemployment on demand are limited, [9] suggests a negative relationship between unemployment and the demand for life insurance. Studies frequently examine the link between unemployment and the rate of non-payment insurance [12]. Theoretically assumed positive relationship empirically confirms in developing countries [10].

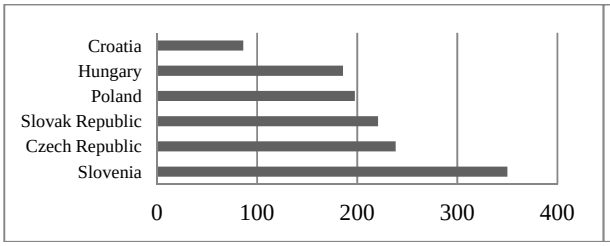
2.2 Demographic determinants of life insurance

It is assumed that a larger **number of dependents** in the population have a positive impact on the demand for life insurance. This will be explained by the fact that the recipient of income is

primarily striving to protect dependents of uncertainty, which are part of life. A sense of responsibility towards the dependents is a fundamental driver of demand for life insurance. However, the number of people dependent on the breadwinner can have the opposite effect of what was discussed in terms of income exhausting requirements of each dependent member. The result of insufficient amounts of available income is the cause of the inability of contracting life protection, so the number of dependents may have a negative effect on the demand for life insurance. The ratio of the population younger than 15 and older than 64 years to the working age population is used as a measure of dependents in society, [3] separately consider the share of the young dependent population and the older dependent population in the working age population. The number of dependents is achieving ambiguous impacts since it is manifested in different motive for hedging and the reasons for saving. Separate consideration of the young and old dependent population gains on the meaning in a situation when in the analysis is included a detailed data on demand with respect to the type of life insurance. A higher level of **education** of the individual is positively associated with the desire to protect dependents. It is considered that a higher level of education results in a stronger awareness of the uncertainties of life [4], [11]. Greater aversion to risk and a greater understanding of the risks, which are related with higher education, result in a greater demand for life insurance. The level of education is associated with the need for protection of dependent family members and their standard of living [13]. The education level of the population is usually measured by the total enrolment in tertiary education regardless of age, expressed as a percentage of the total population of the five-year age group following on from leaving secondary school.

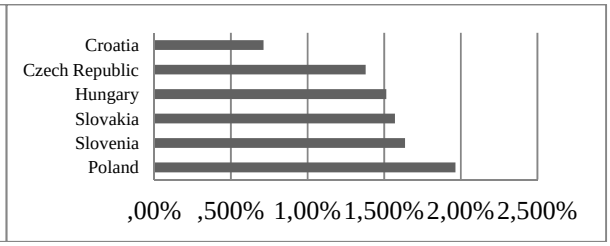
3 EMPIRICAL RESEARCH

Considered countries; Bulgaria, Croatia, Czech Republic, Hungary, Slovakia, and Slovenia constitute a homogenous group according to the criteria of development of the life insurance market. The penetration, as an indicator of the share of life insurance in the total product, and density, as the size of the premium allocation per capita, are the measures considered for six transition countries that reveal the same stage of development of life insurance market with common challenges, which is the convergence to the rest of developed Europe (EU). The indicator of the density of insurance is an absolute indicator. Penetration is a relative indicator adjusted for the change in GDP, therefore less resilient to the movement of the income [3]. Penetration reveals the position of the life insurance industry in relation to all other sectors that contribute to GDP value. Indicator of density as an absolute size indicates a change of scale that is not appointed in the context of the national economy and the position of comparison with other sectors of the national economy. Recent studies, in particular the panel included an analysis, predominantly used penetration of insurance as a measure of the demand that due to the insensitivity of the change in the price [4] and insensitivity to the level of economic development [15] created a reason to critique. The insurance market of selected countries is analysed for the ten-year period from 2003 to 2012, using the data from the annual equidistance between the members of the time series, which reduces claim for correction of possible systemic renewal phenomena within a period of one year. The data that includes both temporal and spatial component of the analysed variables are called the panel data, and the whole process is called the panel analysis.



Source: According to data http://www.insuranceeurope.eu/

Figure 1: Life insurance market size measured by average density (USD) in selected countries (2003-2012)



Source: According to data http://www.insuranceeurope.eu/

Figure 2: Life insurance market size measured by average penetration in selected countries (2003- 2012)

Dependent variable, approximated with density or penetration of insurance, is changed to the units of observation (by country) and by the time, so the evaluation of variables, which really determine the variable demand for life insurance, is considered a more precise [2]. Analytical note of the model equations is as follows:

$$LFINS_{it} = \alpha_i + \beta_1 INCOME_{it} + \beta_2 INFLN_{it} + \beta_3 UNEMP_{ti} + \beta_4 NBDEPT_{it} + \beta_5 EDUC_{it} + \varepsilon_{it} \quad (1)$$
$$i = 1,..., N; t = 1,..., T$$

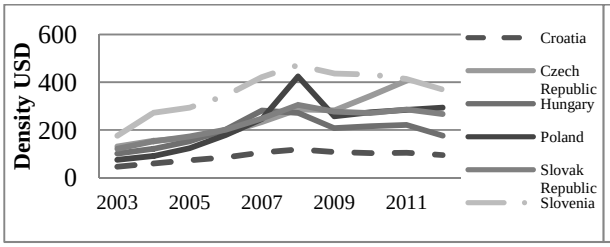
Where N is the number of units of observation, T is the number of periods. Parameter α_i is a constant member, different for each unit of observation, $\beta_1,..., \beta_5$ are parameters that are estimated. Furthermore, ε_{it} is the estimation error that is assumed to follow a white noise process.

Table 1: Description of variables and expected impact of independent variables

Label	Definition of variable	Expected impact
LFINS	Ratio of gross written premiums to total population (Density)	**Dependent variable**
LFINS2	Gross written premiums to GDP (Penetration)	**Dependent variable**
INCOME	GDP per capita (current 000 US$)	+
INFLN	Inflation, consumer prices (annual %)	-
UNEMP	Unemployment, total (% of total labor force)	-
NBDEPT	Age dependency ratio (% of working-age population)	+/-
EDUC	School enrolment, tertiary (% gross)	+

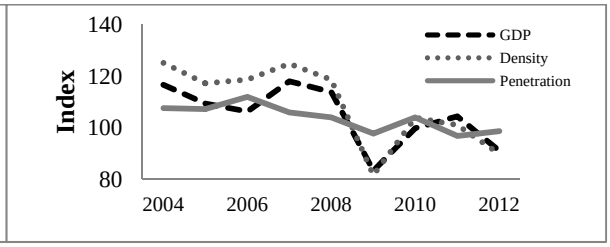
Source: Author's definition and expectations according to previous research

The movements of the dependent variable, the demand for life insurance, indicates upward trends in the performance till the financial crisis, when positive trends are replaced by a period of stagnation. Poland is the most propulsive market in the period before 2008. The period of financial and economic crisis caused a decrease in the size of the market and stagnation by the end of the period. The least developed market of life insurance among the surveyed countries is Croatia. Indicates the greatest differences compared to other observed countries. General level of the importance of the life insurance market is less influential and trends are less expressed on Croatian example.



Source: According to data http://www.insuranceeurope.eu/

Figure 3: Life insurance density in selected countries



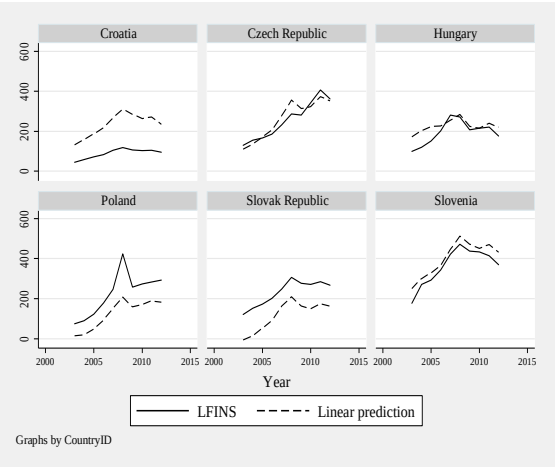
Source: According to data http://www.insuranceeurope.eu/

Figure 4: Average indices of all selected countries

Observing average indicators for six countries by years (Figure 3), it has been noted that after 2008 average gross written premium per capita declines from 312.73 USD to 260.96 USD. Hence, the dummy variable is introduced for the period of crisis according to movements of indicators of demand. However, variable crisis shows no statistical significance and does not affect the assessment parameters. The growth rate of demand for life insurance in selected countries exceeds the growth rate of GDP (Figure 4). Hausman test confirms the existence of a correlation between at least one independent variable and the individual intercept, for this reason the model with fixed effects (FE) is accepted as the appropriate model [2]. Despite the presence of a problem of correlation, the estimator of the model with fixed effect is consistent. Furthermore, noted stop of the growing trend in demand for life insurance is the reason of inclusion of the dummy variable crisis which is an adverse effect, but does not show significance in explaining the trends in demand for life insurance regardless of the measure of demand (Table 2). Wooldridge test confirms the absence of serial correlation of residuals. The usual *R*2 is valid for comparison of the pooled model estimated by OLS and the FE model, but comparison should be done only within the same class of models and estimators, [14] suggests three measures within, between and overall.

Table 2: Estimated models with graph of forecasting movements of life insurance density compared with real movements in selected countries

		Fixed effect model		
	lfins	lfins	lfins2	lfins2
income	16.38***	16.88***	0.0109	0.0128
	(2.997)	(3.381)	(0.0192)	(0.0216)
infln	-2.555	-2.544	-0.00328	-0.00323
	(3.001)	(3.029)	(0.0192)	(0.0194)
unemp	-7.862***	-7.958***	-0.0502***	-0.0506***
	(2.609)	(2.650)	(0.0167)	(0.0170)
nbdept	13.00**	12.80**	0.0442	0.0434
	(5.762)	(5.850)	(0.0369)	(0.0375)
educ	1.527	1.373	0.0119	0.0113
	(1.345)	(1.438)	(0.00861)	(0.00921)
crisis		-4.405		-0.0173
		(13.53)		(0.0866)
_cons	-582.4**	-569.8**	-0.777	-0.727
	(238.1)	(243.5)	(1.524)	(1.559)
N	59	59	59	59
Hausman	0.0011	0.0014	0.0003	0.0006
Wooldridge	0.1858	0.1817	0.1823	0.1861
R2	0.8443	0.8447	0.4767	0.2167



Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
Source: Author's estimation according to data http://www.insuranceeurope.eu/; http://www.worldbank.org/

The results of the analysis are divided depending on the measure of the demand for life insurance. Life insurance is becoming a more attractive way of investing as income rises. A positive relationship is observed in the model where demand is measured by the density of insurance. In the same model, it was confirmed that an increase of the number of dependent members of the population contributes to the development of the insurance market. The latter indicates that the motives for leaving the property in inheritance to dependent family members are expressed in the transition countries. Unemployment as a great problem of the selected countries, the average rate for observed period exceeds 10%, manifests itself as a variable that significant impedes the development of the insurance market. Indicator density of insurance is more sensitive to changes in income, than is the case with an indication of the penetration of the insurance market. Therefore it is said that the model of market penetration is less income elastic, as in the example where significance is lacking. Furthermore, the model contains only

unemployment as significant variable, thus explains the correlation connection and is not used in the assessment of the actual movement of the demand for life insurance. Fixed model best estimates demand for life insurance in the case of the Czech Republic and Slovenia. The presence of inflation shows no significance that can be justified with indexed insurance policies.

4 CONCLUSION

The variables that are estimated by the fixed effect model of demand for life insurance and that showed significance are in accordance with the theory. In conclusion, the conditions in which the economy is with low unemployment rate and households with a stable income, who care about the future of dependents, are the main drivers of economic and demographic growth of the insurance market in the transition countries of Central Europe. If the demand is assumed by indicator of penetration, the conclusion is that higher unemployment rate reduces the share of collected premiums in the total product. The inhibited trend described by variable crisis shows no significance to the assessment parameters. Future research should be consider the impact of the pension system and its long term sustainability, which is the fundamental characteristic considered by demographic indicators, on function and position of the insurance industry.

References

[1] Babbel, D.F. (1981). Inflation, Indexation, and Life Insurance Sales in Brazil. The Journal of Risk and Insurance, Vol. 48(1): 111–135.

[2] Baltagi, B.H. (2008). Econometric Analysis of Panel Data. 4th ed. Chester: John Wiley and Sons.

[3] Beck, T. and Weeb, I. (2003). Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. The World Bank Economic Review, Vol. 17(1): 51–88.

[4] Browne, M.J. and Kim, K. (1993). An international analysis of life insurance demand. The Journal of Risk and Insurance, Vol. 60(4): 616-634.

[5] Feyen, E., Lester, R. and Rocha, R. (2011). What Drives the Development of the Insurance Sector?. Policy Research Working Paper No. 5572, World Bank.

[6] Fortune, P. (1973). A theory of optimal life insurance: Development and tests. The Journal of Finance, Vol. 28(3): 587–600.

[7] Kjosevski, J. (2012). The Determinants of Life Insurance Demand in Central and Southeastern Europe. International Journal of Economics and Finance, Vol. 4(3): 237 - 247.

[8] Li, D., Moshirian, F., Nguyen, P. and Wee, T. (2007). The demand for life insurance in OECD countries. Journal of Risk and Insurance, Vol. 74(3): 637–652.

[9] Mantis, G. and Farmer, R.N. (1968). Demand for life insurance. Journal of Risk and Insurance, Vol. 35(2): 247–256.

[10] Outreville, J.F. (1990). Whole-life insurance lapse rates and the emergency fund hypothesis. Insurance: Mathematics and Economics, Vol. 9(4): 249–255.

[11] Outreville, J.F. (1996). Life insurance markets in developing countries. Journal of Risk and Insurance, Vol. 63(2): 263–278.

[12] Outreville, J.F. (2011). The Relationship between Insurance Growth and Economic Development: 80 Empirical Papers for a Review of the Literature. ICER Working Paper No. 12/2011, Montreal.

[13] Truett, D.B. and Truett, L.J. (1990). The demand for life insurance in Mexico and the United States: A comparative study. Journal of Risk and Insurance, Vol. 57(2): 321–328.

[14] Wooldridge, J.M. (2002). Econometric Analysis of Cross Section and Panel Dana. Cambridge: MIT.

[15] Zheng, W., Liu. Y. and Deng, Y. (2009). A Comparative Study of International Insurance Markets. The Geneva Papers, Vol. 34: 85-99.

[16] Insurance Europe. (2015). http://www.insuranceeurope.eu/ [Accessed May/2015].

[17] The World Bank. (2015). http://www.worldbank.org/ [Accessed May/2015].

THE COMPARISON OF HOLT-WINTERS METHOD AND MULTIPLE REGRESSION METHOD: A CASE OF MONTHLY HEAT LOAD FORECASTING IN THE COMPANY ENERGETIKA LJUBLJANA (SLOVENIA)

Ervin Strmcnik and Liljana Ferbar Tratar
Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia
strmcnikervin@guest.arnes.si; liljana.ferbar.tratar@ef.uni-lj.si

Abstract: A policy towards a low-carbon society in EU provides many measures. Appropriate heat load forecasting techniques offer opportunity for more effective schedule operations and cost minimization. This paper presents a comparison between two widely used methods. The forecasting performance of Multiple regression and Holt-Winters methods has been investigated. Based on chosen accuracy measures, Holt-Winters methods ensured much better forecasting values. The best forecasting performance was obtained by Multiplicative and Extended Holt-Winters method.
Keywords: Heat load forecasting, Holt-Winters method, Multiple regression method

1 INTRODUCTION

Current challenges facing the European Union are sustainable energy supply, reducing dependence on energy resources and energy efficiency [8]. Forecasting is one of the possible measures for the efficient energy production, and wise management of energy resources.

The Naive method has been chosen as a reference method. Forecasting values have been compared with forecasting values of other methods. The Multiple regression method belongs to the group of more advanced methods. It's based on the consideration of the external factors. Dotzauer [2] figured out that the outdoor temparature affects the heat load the most. Other potential relevant external factors are solar radiation, wind speed and humidity. The basic exponential smoothing methods have been developed by Holt [5] and Winters [10]. Extended exponentional smoothing methods have been presented by the American professor Gardner [4]. One possible way of forecasting improvement represents relaxation of space parameter restrictions. Hyndman, Akram in Archibald [6] mentioned, recommended values of space parameters are dependent on time series characteristic such as error, trend and seasonality.

The aim of this article is the comparison of monthly heat load forecasting value of the Multiple regression method and Holt-Winters method. We defined the following hypothesis: »Holt-Winters methods are more appropriate methods for heat load forecasting than the Multiple regression methods«.

The remainder of the paper is organized as follows. Section 1 presents a supply chain and case study data applied in this paper. In section 2, a methodology is explained. Section 3 consists of description of methods. Naïve method, the Multiple regression methods and Holt-Winters methods are introduced. Results are presented in section 4. In the last, section 5, conclusion and further research are discussed.

2 DATA

This study is based on heat load data *Y* and four other external quantities. These are outdoor temperature *T*, solar radiation *S*, wind speed *W*, and relative humidity H. Data was obtained from the company Energetika Ljubljana from September 2008 to February 2013. The supply chain of the company consists of three main elements: 1.) Production of heat load, 2.) A distribution network, and 3.) Customers. Company Energetika Ljubljana provides heat load (hot water and steam), natural gas and electricity. The heat load accounted for 50% of the whole sale of the company [7]. Data was gathered through a daily resolution. Time series have been divided into winter and summer seasons (Figure 1).

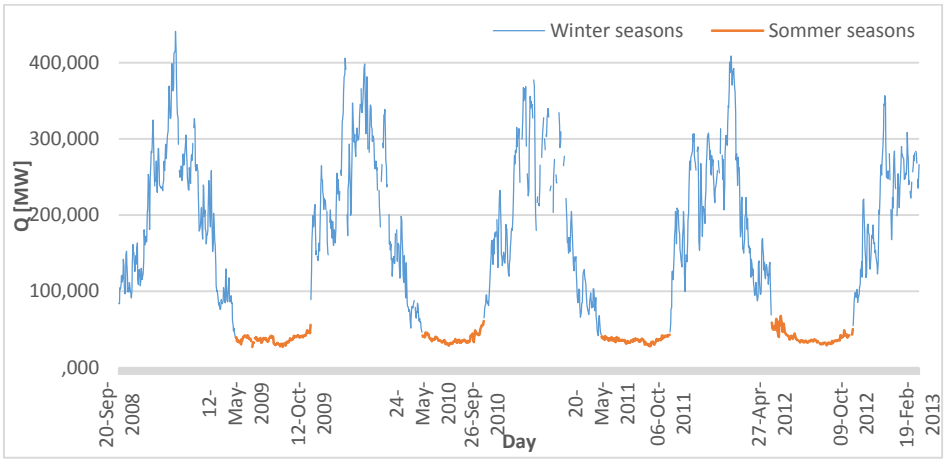


Figure 1: Heat load data *Y* in daily resolution

In this study, only winter data is analysed due to difficulty of estimation. Forecasting value was related to strong relations between heat load data and external weather factors, which caused a high dynamic of time series. Average monthly data WS1 (winter season 1), WS2, WS3, WS4, WS5 were divided into six equal periods: P1 (period 1), P2, P3, P4, P5, P6, as shown in Table 1.

Table 1: A data division

	WS1	***WS2***	***WS3***	***WS4***	***WS5***
P1	16. 10.-15. 11. 2008	16. 10.-15. 11. 2009	16. 10.-15. 11. 2010	16. 10.-15. 11. 2011	16. 10.-15. 11. 2012
P2	16. 11.-15. 12. 2008	16. 11.-15. 12. 2009	16. 11.-15. 12. 2010	16. 11.-15. 12. 2011	16. 11.-15. 12. 2012
P3	16. 12.-15. 1. 2009	16. 12.-15. 1. 2010	16. 12.-15. 1. 2011	16. 12.-15. 1. 2012	16. 12.-15. 1. 2013
P4	16. 1.-15. 2. 2009	16. 1.-15. 2. 2010	16. 1.-15. 2. 2011	16. 1.-15. 2. 2012	16. 1.-15. 2. 2013
P5	16. 2.-15. 3. 2009	16. 2.-15. 3. 2010	16. 2.-15. 3. 2011	16. 2.-14. 3. 2012	/
P6	16. 3.-15. 4. 2009	16. 3.-15. 4. 2010	16. 3.-15. 4. 2011	15. 3.-14. 4. 2012*	/

The obtained data were split into training (WS1, WS2, WS3, WS4), and testing subsets (W5). Training subset was used for method learning. With the testing subset we checked a time series learning ability. Hence, we estimated the generalization ability of the method.

Autocorrelation analysis has been used to gather information about seasonality. Due to prominent presence of seasonality, we expected better performance of Holt-Winters methods (although HW methods did not take into account external factors such as outdoor temperatures, solar radiation, wind speed and relative humidity), than methods of Multiple regression.

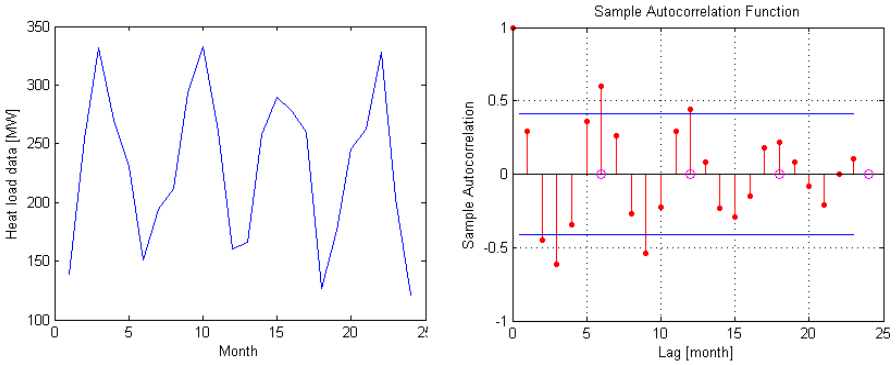


Figure 2: Autocorrelation analysis

3 METHODOLOGY

Forecasting values of heat load, which we compared to independent data (the testing subset, WS5), has been calculated in the following way:

$$F(t+m), m = 1, 2, 3, 4, \quad (1)$$

where t represents the arbitrary time, and m represents number of forecasted period in monthly resolution.

The efficiency of methods was measured in terms of the mean squared error (MSE) of the in-sample one-step-ahead forecasts. Then we used the HW methods with optimal parameters and predicted the next 4 periods. These were compared to the true values (WS5) and the mean squared forecasting error for each method was recorded. Due to gathered knowledge and experiences, we used a MARNE error. It is a relative measure depending on the size of the district heating system, and can be easily interpreted in technical or economical terms. MARNE error is calculated as the average of the absolute differences *e(t)* of forecast heat consumption, and actual heat consumption, normalized by the maximum transmission capacity of the district heating network Y_{max} [9]:

$$\text{MARNE} = 100\frac{\frac{1}{N_d}\sum_{t=1}^{N_d}|e(t)|}{Y_{max}}[\%], t = 1, 2, 3, \dots N_d, \tag{2}$$

where N_d is the number of heat load samples in a monthly resolution. MARNE error allowed us the comparison between actual and forecasted values among all samples of time series. The forecasting errors on training and testing subsets were denoted as $\text{MARNE}_{\text{train}}$ and $\text{MARNE}_{\text{test}}$. Theil's U-statistics has been used as the additional accuracy measure. It has been calculated through actual heat load values Y_t, Y_{t+1} and forecasted heat load values F_{t+1} as it is shown it the following equation:

$$U = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{F_{t+1}-Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1}-Y_t}{Y_t}\right)^2}} \tag{3}$$

If the value of Theil's U-statistics is higher than 1, it means that the forecasting technique is worse than the naïve method approach. If the value of Theil's U-statistics is lower than 1, it means that our technique is better than naïve method technique.

At this point, we would like to emphasize that lower values of the accuracy measures MSE, MARNE, and Theil's U-statistics, represent a better forecasting performance. Testing accuracy measures was accepted as the main criterion for the forecasting performance of methods.

4 FORECASTING METHODS

The Naïve method was chosen as the reference method. A detailed forecasting performance analysis has been conducted for Multiple regression methods (MR), and Holt-Winters methods (HW).

4.1 Naive Method

A Naïve method represents a benchmark method, and it is one of the most simple forecasting methods, it was used as a reference method. Predicted heat load values F_{t+1} were calculated on the basis of the previous data Y_t:

$$F_{t+1} = Y_t + e_{t+1}, \tag{4}$$

where e_{t+1} denotes noise and t denotes the arbitrary time.

4.2 Multiple regression

The motivation for Multiple regression methods was linear correlation analysis between heat load data *Q(t+1)* and external factors *T(t+1)*, *S(t+1)*, *W(t+1)*, *H(t+1)*. We also calculated the autocorrelation coefficient between *Q(t+1)* and *Q(t)*. As shown in Figure 3, there exist a high negative correlation between heat load *Q(t+1)* and outdoor temperature *T(t+1)*.

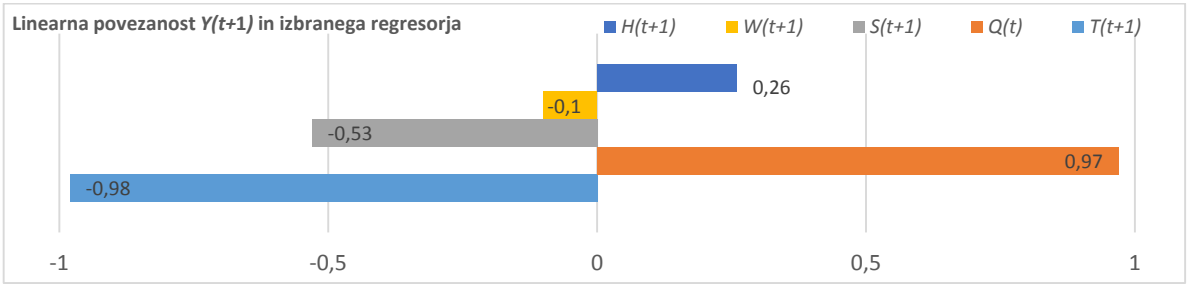


Figure 3: Linear correlation coefficient between heat load data and external factors

We defined four Multiple regression models: MR1, MR2, MR3 and MR4. We considered different number of external factors in the models. For instance, M4 denotes Multiple regression method, which took into account four external factors. These were outdoor temperature *T*, solar radiation *S*, wind speed *W* and relative humidity *H*, as is shown in Table 2. Parameters η_i, i=1:5, denote regression coefficients.

Table 2: Multiple regression (MR) methods

MR1	$F(t+m) = \eta_i + \eta_{i+1}T(t+1) + \eta_{i+2}Q(t)$
MR2	$F(t+m) = \eta_i + \eta_{i+1}T(t+1) + \eta_{i+2}Q(t) + \eta_{i+3}S(t+1)$
MR3	$F(t+m) = \eta_i + \eta_{i+1}T(t+1) + \eta_{i+2}Q(t) + \eta_{i+3}S(t+1) + \eta_{i+4}W(t+1)$
MR4	$F(t+m) = \eta_i + \eta_{i+1}T(t+1) + \eta_{i+2}Q(t) + \eta_{i+3}S(t+1) + \eta_{i+4}W(t+1) + \eta_{i+5}H(t+1)$

4.3 Holt-Winters methods

We examined six different HW methods: 1.) Additive AHW, 2.) Multiplicative MHW and 3.) Extended EHW, 4.) AHW_init, 5.) MHW_init and 6.) EHW_init. First three methods are presented in Table 3. For last three methods, the additional optimisation of the initial values (level-L_s, trend-b_s and seasonality S_i, i=1,..,s; where s denotes the number of months in a season), has been executed. AHW and MHW estimates three smoothing parameters (α, β and γ), and can incorporate additive or multiplicative seasonality depending on whether the seasonal variation is regarded as independent of the level of the local mean or as being proportional to it. EHW estimates four parameters [3]. EHW method takes into consideration the additional parameter δ. The only difference between the AHW and the EHW method is the equation for the calculation of level (look at the Table 3). All other equations – for trend (b_t), seasonality (S_t), forecast (F_{t+m}) and method initialization – remain the same as with the AHW method (Table 3). The optimal values for the smoothing parameters were obtained by minimising MSE with Solver in Microsoft Excel. The initial values of the smoothing parameters were 0.5. With the additional settings, the interval [0,1] for the smoothing parameters was accepted.

Table 3: Holt-Winters (HW) methods

	Additive HW method – AHW	Multiplicative HW method – MHW	Extended HW method – EHW
L_s	$\frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s)$	$\frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s)$	$\frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s)$
b_s	$\frac{1}{s}\left(\frac{Y_{s+1}-Y_1}{s} + \frac{Y_{s+2}-Y_2}{s} + \cdots + \frac{Y_{s+s}-Y_s}{s}\right)$	$\frac{1}{s}\left(\frac{Y_{s+1}-Y_1}{s} + \frac{Y_{s+2}-Y_2}{s} + \cdots + \frac{Y_{s+s}-Y_s}{s}\right)$	$\frac{1}{s}\left(\frac{Y_{s+1}-Y_1}{s} + \frac{Y_{s+2}-Y_2}{s} + \cdots + \frac{Y_{s+s}-Y_s}{s}\right)$
S_i	$S_1 = Y_1 - L_s, S_2 = Y_2 - L_s, \ldots, S_s = Y_s - L_s$	$S_1 = \frac{Y_1}{L_s}, S_2 = \frac{Y_2}{L_s}, \ldots, S_s = \frac{Y_s}{L_s}$	$S_1 = Y_1 - L_s, S_2 = Y_2 - L_s, \ldots, S_s = Y_s - L_s$
L_t	$\alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1})$	$\alpha\frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$	$\alpha Y_t - \delta S_{t-s} + (1-\alpha)(L_{t-1} + b_{t-1})$
b_t	$\beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$	$\beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$	$\beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$
S_t	$\gamma(Y_t - L_t) + (1-\gamma)(S_{t-s})$	$\gamma\frac{Y_t}{L_t} + (1-\gamma)(S_{t-s})$	$\gamma(Y_t - L_t) + (1-\gamma)(S_{t-s})$
F_{t+m}	$L_t + b_t m + S_{t-s+m}$	$(L_t + b_t m)S_{t-s+m}$	$L_t + b_t m + S_{t-s+m}$
	Initial values for the smoothing parameters		
	$\alpha = 0.5,\ \beta = 0.5$ in $\gamma = 0.5$	$\alpha = 0.5,\ \beta = 0.5$ in $\gamma = 0.5$	$\alpha = 0.5,\ \beta = 0.5, \gamma = 0.5$ in $\delta = 0.5$
	Optimal values for the smoothing parameters, obtained by		
	$\alpha = 0.05,\ \beta = 1.00$ in $\gamma = 0{,}50$	$\alpha = 0.04,\ \beta = 1.00$ in $\gamma = 0.58$	$\alpha = 0.13,\ \beta = 0.32, \gamma = 0.48$ in $\delta = 0.00$

5 RESULTS

The minimum $MARNE_{train}$ was obtained by Multiple regression MR4 method as it is shown in the figure. It yielded 0.94%. The best testing (generalization) results were reached with Holt-Winters methods. The best MHW method yielded testing performance $MARNE_{test}$=1.43%. It was just a little bit lower than testing performance of EHW-method ($MARNE_{test}$=1.50%). Testing Theil's U-statistics confirmed an excellent performance of MHW method. The lowest U-statistics had a value 0.08 (Figure 4). It was much lower than 1, and it meant an excellent performance in comparison to naive method.

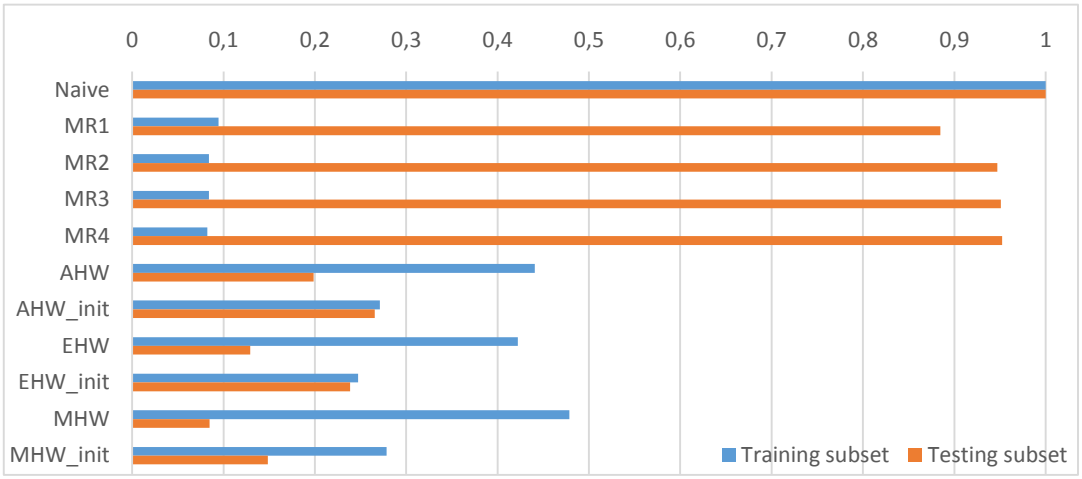


Figure 4: Theil's U-statistics measure accuracy

The additional analysis was conducted in relation to the relaxation of limitation for space parameters. We investigated the influence of extended interval on Theil's U-statistics measure accuracy. In the first step, the space parameters limitation was changed from [0,1] to [0,2]. In the second step, from [0,2] to [-2,2], positive effects were found out. As it is shown in Figure 5, training Theil's U-statistics reached the lower values after the relaxation of limitation for smoothing parameters. The biggest minimisation of training Theil's U-statistics was obtained in the second step. Regarding testing Theil's U-statistics, positive effects were noticed by four of six methods.

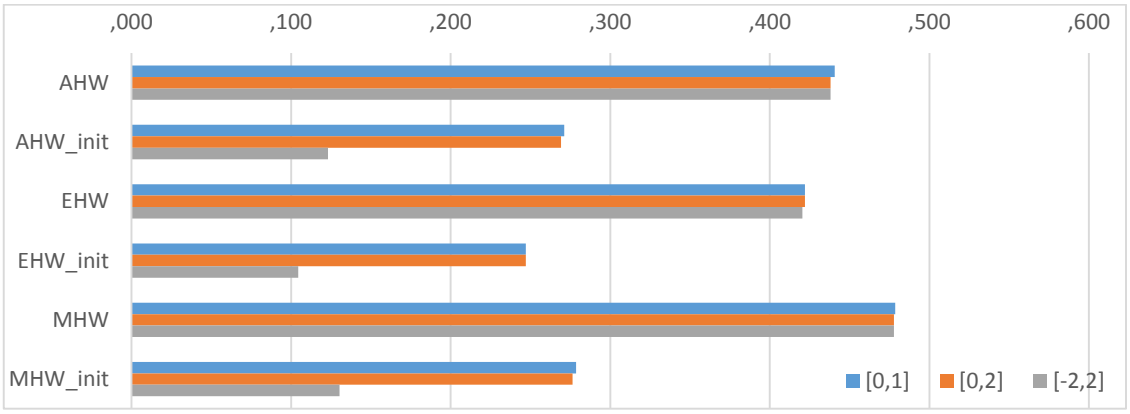


Figure 5: Influence of relaxation of limitation for space parameters on training Theil's U-statistics

Based on the results, we accepted the hypothesis »Holt-Winters methods are more appropriate methods for heat load forecasting than the Multiple regression methods«.

Accuracy measures $MARNE_{test}$ and Theil's U-statistics by Holt-Winters methods yielded much lower values than by Multiple regression methods. The best forecasting performance was obtained by multiplicative Holt-Winters method MHW.

In the end, we would like to emphasize that MHW and EHW method were identified as the most appropriate methods for the industrial forecasting implementation. Due to time series dynamics authors suggest a revision of the analysis when new data is available.

6 CONCLUSION AND FURTHER RESEARCH

We can conclude that the appropriate forecasting heat load methods offer many opportunities regarding process optimisation and appropriate strategic decisions. Monthly forecasting values were calculated with Naive, Multiple regression and Holt-Winters methods. MSE, MARNE and Theil's U-statistics have been used as statistics measures of forecasting accuracy. Within this paper our contribution to this study has been emphasized by the comparison of the forecasting performance between Multiple regression methods and Holt-Winters methods. Results showed that Holt-Winters methods are more suitable for monthly heat load forecasting than Multiple regression methods. The best performance was obtained by Multiplicative and Extended Holt-Winters method. The presented study was conducted utilizing real data, and it represents additional value within this paper. Findings and recommendations are useful immediately in the real business sector. Further research of relaxation of limitation for space parameters will be able to contribute to better forecasting performance. Another possible way for the improvement is a configuration of a new method, which would consist partly of Multiple regression method (which consider external factors), and partly of Holt-Winters method (which takes into account autoregressive elements such as level, trend, and seasonality).

References

[1] Chatfield, C., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2001). A New Look at Models for Exponential Smoothing. *Journal of the Royal Statistical Society*, *50*(2), 147-159.

[2] Dotzauer, E. (2002). Simple model for prediction of loads in district-heating systems. *Applied Energy 73*, 277-284.

[3] Ferbar Tratar, L., Mojskerc, B., & Toman, A. (2014). Demand forecasting with four-parameter exponential smoothing. *V: 17th International Symposium on Inventories, August 20-24, 2014*, Budapest, Hungary. CHIKÁN, Attila (ur.). *Book of abstracts*. Budapest: ISIR, 2014, str. 114. [COBISS.SI-ID 22182374].

[4] Gardner, E.S. Jr. (2005). Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting, 22*(4), 637-666.

[5] Holt, C.C. (1957). *Forecasting seasonals and trends by exponentially weighted moving averages 52*. Office of Naval Research, Arlington, VA;ponovno natisnjen:Holt, C.C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting, 20*(1), 5-10.

[6] Hyndman, R.J., Akram, Md., & Archibald, B. (2008). The admissible parameter space for exponential smoothing models. *Annals of Statistical Mathematics, 60*(2), 407–426.

[7] Letno poročilo 2013. *Podjetje Energetika Ljubljana*. http://www.energetika-lj.si/o-druzbi/poslovanje-2 [Accessed 4/3/2015].

[8] Roadmap 2050 – Technical & Economic Analysis: Full Report. *European Climate Foundation.* http://www.roadmap2050.eu/reports/ [Accessed 4/3/2015].

[9] Strmčnik E., Potočnik P., & Govekar E. (2015). Raziskava in razvoj empiričnih modelov za napovedovanje odjema toplote v vročevodnem sistemu. *Ventil, 21*(1), 50-54.

[10] Winters, P.R. (1960). *Forecasting Sales by Exponentially Weighted Moving Averages*. http://dx.doi.org/10.1287/mnsc.6.3.324 [Accessed 2/4/2015].

THE IMPACT OF INTRACLUSTER HOMOGENEITY ON THE SURVEY COSTS: A CROATIAN BUSINESS SURVEY

Berislav Žmuk
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J.F.Kennedyja 6, 10000 Zagreb, HR-Croatia
bzmuk@efzg.hr

Abstract: The intracluster homogeneity is one of the most challenging issue in cluster sampling. The main goal of the research is to study to what extent are the survey costs influenced by different homogeneity rates. In order to calculate rates of homogeneity, simple random sampling and two-stage cluster sampling designs in estimating proportions were used. The research has shown that lower survey costs are achieved when the intracluster homogeneity is lower.

Keywords: rate of intracluster homogeneity, cluster sampling, survey sampling design, design effect, survey costs, business survey.

1 INTRODUCTION

Depending on the used method of survey data collection, the collection cost could range from enormous to rather small. In order to reduce survey costs, new survey modes are developed. Consequently, some of the most common methods of data collection in use now are: computer-assisted personal interviewing (CAPI), audio computer-assisted self-interviewing (ACASI), computer-assisted telephone interviewing (CATI), interactive voice response (IVR) and web survey [6]. Also, mixed modes or the use of two or more methods of data collection are also possible and recommended in some cases [4]. Furthermore, in order to reach satisfactory response rates, researchers are nowadays forced to give incentives or even to pay to the respondents [2].

Survey costs could be reduced through using either the different data collection modes, or different sampling designs [5]. A cluster design is appropriate to those who want to keep survey costs at their minimum [3]. In order to successfully achieve minimum survey costs with the desired survey precision, it is necessary to find an optimal balance of the number of clusters and their size. The intracluster homogeneity (rate of homogeneity – *roh*) measures the tendency of elements, according to their values of a variable, within a cluster to be correlated among themselves in comparison to the values of a variable for elements outside the cluster [6]. In that way the intracluster homogeneity has an important role in cluster design.

Because of its importance, the paper analyses the impact of the intracluster homogeneity on the survey costs. The main problem of the intracluster homogeneity is its proper estimation. The aim of this paper is to emphasize the importance of selecting the correct intracluster homogeneity by examining its impact on survey costs. Consequently, the research hypothesis is that lower survey costs are achieved when the intracluster homogeneity, measured and estimated as the rate of homogeneity, is lower.

The paper is organized as follows. After a brief introduction of the research problem, in the second chapter important characteristics of cluster sampling are given. In the third chapter data and methods are described, whereas in the fourth chapter the analysis of the intracluster homogeneity impact on survey costs was conducted. In the fifth, final, chapter conclusions and recommendations for further research are given.

2 CLUSTER SAMPLING

According to Kish [7], in cluster sampling *N* population elements included in the sampling frame are distributed among *A* clusters. The sample size in cluster sampling is defined by the number of selected clusters (*a*) and by the average number of selected elements in selected clusters ($\bar{b}$):

$$n = a \cdot \bar{b}, \quad a < A. \tag{1}$$

In that case the mean statistics is calculated as:

$$\bar{y} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b} y_{\alpha\beta}}{a\bar{b}}, \tag{2}$$

where $\bar{y}$ is the mean of the observed variable, $\alpha = 1,2,...,a$ are clusters in the sample, $\beta = 1,2,...,b$ are elements within cluster α and $y_{\alpha\beta}$ is the variable value of the element β in cluster α.

In the case of the two-stage cluster sample, where a random sample of *a* clusters is obtained in the first stage, and a random sample of *b* elements within sampled clusters is obtained in the second stage, the sampling variance is:

$$\operatorname{var}(\bar{y}) = \frac{1-f}{a\bar{b}^2(a-1)} \cdot \left[\sum_{\alpha=1}^{a} y_{\alpha}^2 - \frac{\left(\sum_{\alpha=1}^{a} y_{\alpha}\right)^2}{a} \right], \tag{3}$$

where f is the sampling rate and s_a^2 is the between cluster variance.

Because of different approaches in the simple random sampling compared to the cluster sampling, there is a difference in variances obtained in each of these two approaches. The measure of the efficiency of two sampling designs is the Kish's design effect (*deff*) given as [7]:

$$deff = \frac{\operatorname{var}(\bar{y})_{CLUSTER}}{\operatorname{var}(\bar{y})_{SRS}}, \tag{4}$$

where $\operatorname{var}(\bar{y})_{CLUSTER}$ is the sampling variance of the cluster sample statistics (mean) and $\operatorname{var}(\bar{y})_{SRS}$ is the sampling variance of the simple random sample statistics (mean).

The intracluster homogeneity is a measure of the homogeneity of the elements within clusters and it is estimated as the rate of homogeneity (*roh*):

$$roh = \frac{deff - 1}{\bar{b} - 1}. \tag{5}$$

The *roh* can take values from $-1/(b-1)$ to 1. In most cases the *roh* is between 0 and 1. If the *roh* equals to 1, the complete homogeneity within clusters is achieved. On the other hand, if the *roh* is $-1/(b-1)$, the maximum heterogeneity within clusters exists.

The total costs of cluster sampling (*C*) are equal to:

$$C = C_0 + a \cdot c_a + a \cdot \bar{b} \cdot c_b, \tag{6}$$

where C_0 are fixed costs, c_a is the cost per cluster and c_b is the cost per element within a cluster. The cost per cluster is mainly determined by travel and preparation costs whereas the cost per element is defined by interview costs. The optimal number of clusters and the cluster (subsample) size for the given budget can be found using the Lagrange multiplier or the Cauchy-Schwartz inequality.

3 DATA AND METHODS

In order to inspect the research hypothesis, results of a survey on statistical methods use in Croatian enterprises are used [9]. The main aim of the survey was to provide an insight into the position of statistical methods in Croatian small enterprises and into the attitude of employees to their use. A web survey was used as the method of data collection. The survey was conducted in the period from October 2012 to February 2013. According to the used sampling frame, at the beginning of the survey the sampling population consisted of 24,618 Croatian small enterprises. A stratified sampling design with proportionate allocation was used as a sampling design. According to their main activity the enterprises were stratified into four strata. Overall 631 Croatian small enterprises took part in the survey, which resulted in the Response rate 1 [1] of 1.02%.

For the purpose of the analysis in this paper, the following two questions are observed: "Are statistical methods used in your enterprise?" (*Question A*) and "Does your enterprise invest in statistical methods use knowledge of the employees?" (*Question B*). Overall 215 (34%) Croatian small enterprises provided a positive answer to *Question A*, whereas the rest of them (66%) stated that they do not use statistical methods in their business. Only those enterprises which had previously stated that they use statistical methods answered to *Question B* as well. Accordingly, out of 215 Croatian small enterprises that use statistical methods 115 (53%) do not invest in statistical methods use knowledge of the employees whereas 76 (35%) enterprises do. It has to be emphasized that 24 (11%) respondents were not familiar with this topic or were not sure about the level of statistical methods use in their enterprise. Consequently, they have selected the "*I do not know*" answer.

For the purpose of this paper it will be assumed that instead of a stratified sampling design, which has been used in reality in the observed survey, a pure simple random sampling design was used. Therefore, sampling variances are going to be calculated for *Question A* and for *Question B* assuming a simple random sampling design.

In the next step, the cluster sampling design using the data from the observed previously conducted survey is going to be simulated. In the simulated cluster sampling design, the counties of the Republic of Croatia have the role of clusters. So, 21 clusters (20 counties plus the City of Zagreb) are formed. In order to obey the main rule in the cluster sampling design, it is assumed that there are more than 21 clusters. The number of selected elements (here enterprises) in a cluster (here a county) is going to be different among selected clusters. Because of that, a two-stage cluster design with probabilities proportionate to the size is going to be assumed. In that way, an equal probability selection method (*epsem*) is used. Therefore, the sampling rate is equal to the ratio of the sample size and the population size.

Based on calculated sampling variances of the two-stage cluster design for *Question A* and for *Question B* the design effects are going to be estimated. So, all elements for estimating the intracluster homogeneity through the rate of homogeneity are going to be available. Finally, the total survey costs are analysed for the given cost per cluster and the cost per element within a cluster.

4 INTRACLUSTER HOMOGENEITY ESTIMATION AND IMPACT ANALYSIS

In the conducted survey, 215 or 34% out of 631 small Croatian enterprises confirmed that they use statistical methods in their businesses. Consequently, the simple random sampling design variance for *Question A* is 0.000357. Similarly, out of 191 small Croatian enterprises (excluding enterprises that do not use statistical methods and 24 enterprises that did not know the answer to *Question B*) 76 or 40% confirmed that they invest in statistical methods use knowledge of their employees. Accordingly, it can be estimated that only 12% ($76/631 \cdot 100$)

of small Croatian enterprises invest in statistical methods use knowledge of their employees. Therefore, the simple random sampling design variance for *Question B* is 0.000168. In the calculations of the random sampling design variance, the sampling rate was considered as negligible, so it was not included in the calculations.

Table 1: Cluster sampling results for *Question A* and for *Question B*

Counties (clusters)	***Cluster size***	***Question A***		***Question B***	
		Cluster proportion	*Cluster variance*	*Cluster proportion*	*Cluster variance*
City of Zagreb	239	0.2929	0.0009	0.1004	0.0000
Zagreb	41	0.5366	0.0062	0.1707	0.0007
Bjelovar-Bilogora	10	0.3000	0.0233	0.0000	-
Slavonski Brod-Posavina	9	0.4444	0.0309	0.2222	0.0062
Dubrovnik-Neretva	11	0.1818	0.0149	0.0909	0.0008
Istria	39	0.3077	0.0056	0.1538	0.0006
Karlovac	6	0.3333	0.0444	0.0000	-
Koprivnica-Križevci	15	0.2667	0.0140	0.1333	0.0013
Krapina-Zagorje	16	0.2500	0.0125	0.0000	-
Lika-Senj	3	0.6667	0.1111	0.3333	0.0556
Medimurje	17	0.4118	0.0151	0.1176	0.0009
Osijek-Baranja	25	0.4400	0.0103	0.1600	0.0011
Požega-Slavonia	5	0.2000	0.0400	0.0000	-
Primorje-Gorski kotar	61	0.3934	0.0040	0.1475	0.0004
Sisak-Moslavina	12	0.4167	0.0221	0.1667	0.0025
Split-Dalmatia	49	0.3265	0.0046	0.1429	0.0004
Šibenik-Knin	16	0.3750	0.0156	0.1250	0.0010
Varaždin	31	0.4194	0.0081	0.1613	0.0009
Virovitica-Podravina	8	0.1250	0.0156	0.0000	-
Vukovar-Sirmium	11	0.2727	0.0198	0.1818	0.0033
Zadar	7	0.4286	0.0408	0.0000	-
Total	631	-	-	-	-

Table 1 provides cluster sampling results for the both observed questions. Here, the number of clusters is 21. For both questions the cluster sizes are considered to be the same. Obviously, the number of enterprises with an observed characteristic in the same cluster is quite different when *Question A* and *Question B* are considered. Consequently, the cluster proportions and cluster variances are quite different between the questions. In relation to *Question B* there were some clusters in which the number of enterprises with an observed characteristic was equal to 0. Because of that, cluster variances for counties of Bjelovar-Bilogora, Karlovac, Lika-Senja, Primorje-Gorski kotar, Požega-Slavonia and Zadar could not be calculated. The average cluster size for both questions is 30.05.

Because the clusters are of unequal sizes, the ratio approach is used to calculate the cluster sampling variance by using the following equation [7]:

$$\operatorname{var}(r) = \frac{1-f}{\left(\sum_{\alpha=1}^{a} b_\alpha\right)^2} \cdot \frac{a}{a-1} \cdot \left[\sum_{\alpha=1}^{a} y_\alpha^2 + r^2 \cdot \sum_{\alpha=1}^{a} b_\alpha^2 - 2 \cdot r \cdot \sum_{\alpha=1}^{a} y_\alpha b_\alpha \right], \tag{7}$$

where r is the ratio (proportion), b_α is the number of selected elements in the selected cluster α and y_α is the number of elements with the chosen characteristic in cluster α. The results for the considered questions are provided in Table 2.

Table 2: Rates of homogeneity for *Question A* and for *Question B*

Statistics	*Question A*	*Question B*
Number of clusters	21	21
Average number of elements in clusters	30.05	30.05
Total sample size	631	631
Simple random sampling variance	0.000357	0.000168
Cluster sampling variance	0.000604	0.000116
Design effect	1.695	0.691
Rate of homogeneity (roh)	0.0239	-0.0106

According to the results in Table 2, the *roh* for *Question A* is 0.0239 whereas for *Question B* it is -0.0106. So, the level of homogeneity within clusters is higher for *Question A* than for *Question B*. But also the level of heterogeneity among clusters is higher for *Question A* than for *Question B* too.

Based on calculated *roh* indicators, the total survey costs are going to be simulated. The cost per cluster is set to be €500 and the cost per element within a cluster is set to be €50. These costs already include fixed costs. The number of clusters is going to be estimated using the following equation [8]:

$$a = \frac{p \cdot (1-p) \cdot z^2}{\bar{b} \cdot e^2} \cdot \left[roh \cdot (\bar{b} - 1) + 1 \right], \tag{8}$$

where p is the expected proportion used from the previous research, z is the value from the normal distribution, based on the desired level of confidence, and e is the absolute value of the tolerated sampling variance which is based on the required precision.

In order to calculate a as an expected proportion, proportions for *Question A* (0.40) and *Question B* (0.12) were used. The used 95% confidence level resulted with z value of 1.96. The absolute value of the tolerated sampling variance which is based on the required precision e was set at 0.05. The *roh* at *Question A* is set as 0.0239 and at *Question B* it is -0.0106. The only changeable element was the $\bar{b}$ (the average number of selected elements in selected clusters). In the first case $\bar{b}$ was equal to 10, whereas in the second case $\bar{b}$ was set to 50. Based on the clusters' sizes and the number of clusters, the total survey costs are calculated and the results are given in Table 3.

Table 3: Survey costs for *Question A* and for *Question B*

Statistics	*Case 1*		*Case 2*	
	Question A	*Question B*	*Question A*	*Question B*
Expected proportion (p)	0.3407	0.1204	0.3407	0.1204
Normal distribution value (z)	1.96	1.96	1.96	1.96
Rate of homogeneity (roh)	0.0239	-0.0106	0.0239	-0.0106
Average number of elements per cluster ($\bar{b}$)	10	10	50	50
Tolerated sampling variance (e)	0.05	0.05	0.05	0.05
Number of clusters (a)	41.95	14.72	15.00	1.56
Sample size (n)	419.51	147.20	749.84	77.95
Final sample size (n_{final})	420	148	750	78
Final number of clusters (a_{final})	42	15	15	2
Final average number of elements per cluster ($\bar{b}_{final}$)	10	10	50	39
Cost per cluster (c_a)	500	500	500	500
Cost per element within a cluster (c_b)	50	50	50	50
Total survey costs (C)	41,951	14,720	44,991	4,677

The results provided in Table 3 show that in the *Case 1* the survey costs when considering *Question A* are 1.85 times higher than at *Question B*. In the *Case 2* the difference in the survey costs is even higher. So, in the *Case 2* the survey costs for *Question A* are 8.62 times higher than for *Question B*. This difference arises from the difference in expected proportions and rates of homogeneity. However, the additional analysis has shown that even for the same value of expected proportion (either 0.3407 or 0.1204) the survey costs in case of *Question A* are higher than for *Question B*. In that way the research hypothesis that lower survey costs are achieved when the intracluster homogeneity is lower can be accepted.

5 CONCLUSIONS

In order to obtain quality and valid data, attention must be paid to both, data collection, and to sampling method. In terms of the limited research budget, this process becomes particularly important.

The analysis has shown that the intracluster homogeneity has a significant impact on the total survey costs in cluster sampling. In order to reach a desired level of precision with a restricted survey budget this should be taken into account in the further surveys based on cluster sampling. So, the main scientific contribution of the paper is to raise awareness of importance of units homogeneity in and between clusters on the survey costs.

Still, the question how to choose an appropriate value of the intracluster homogeneity for a survey when they differ importantly among target survey questions stays open. Therefore, in their future research, researchers should deal with the optimal intracluster homogeneity problem.

Acknowledgement
This work has been partially supported by the Croatian Science Foundation under the project STRENGTHS (project no. 9402, Project period: 2014-2018).

References

[1] American Association for Public Opinion Research. (2015). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/publications/Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf [Accessed 01/06/2015].
[2] Bricker, J. (2014). Survey Incentives, Survey Effort, and Survey Costs. FEDS Working Paper No. 2014-74, Washington.
[3] Daniel, J. (2012). Sampling Essentials: Practical Guidelines for Making Sampling Choices. Thousand Oaks, California: SAGE Publications.
[4] de Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. The Journal of Official Statistics, 21(2): 233-255.
[5] Groves, R. M. (1989). Survey Errors and Survey Costs. New York: John Wiley & Sons.
[6] Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2004). Survey Methodology. Hoboken, New Jersey: John Wiley & Sons.
[7] Kish, L. (1995). Survey Sampling. New York: John Wiley & Sons.
[8] Leon, E. A. et al. (2014). Effectiveness of systematic foot and mouth disease mass vaccination campaigns in Argentina. Scientific and Technical Review, 33(3): 917-926.
[9] Žmuk, B. (2015). Business Sample Survey Measurement on Statistical Thinking and Methods Adoption: the Case of Croatian Small Enterprises. Interdisciplinary description of complex systems, 13(1): 154-166.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 5:

Production

THE STATIONARITY OF PER CAPITA ELECTRICITY CONSUMPTION IN CROATIA ALLOWING FOR STRUCTURAL BREAK(S)

Đula Borozan
J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek
31000 Osijek, Croatia
E-mail: borozan@efos.hr

Luka Borozan
J. J. Strossmayer University of Osijek, Department for Mathematics
31000 Osijek, Croatia
E-mail: lborozan@mathos.hr

Abstract: Understanding the stationarity properties of electricity consumption provides valuable insights for energy policy-makers and practitioners. The paper examines the unit root properties of per capita electricity consumption for Croatian counties using the panel unit root tests with structural break(s) during the period 2001-2013. The results indicate that the series of most counties are non-stationary processes, and that statistically significant structural break(s) happened only in a few of them. Hence, the impacts of shocks on per capita electricity consumption are permanent and have a long memory for a majority of them. Moreover, their behaviors are path-dependent.

Keywords: electricity consumption, panel unit root test, stationarity, structural break, Croatia.

1 INTRODUCTION

Policy authorities are very interested in finding whether movements in electricity consumption have a transitory or permanent effect, i.e., whether there is an energy convergence across regions. If electricity consumption converges to a stationary rate in the long-run, deviations therefrom are transitory and will diminish eventually. On the contrary, if the movement of electricity consumption has a characteristic of hysteresis, deviations therefrom caused by exogenous shock/innovation in energy markets will have a permanent effect thereon (see [10]). The convergence hypothesis is empirically testable by employing the unit root tests on per capita electricity consumption. Finding evidence of a unit root provides evidence in favor of the unit root (also known hysteresis or no-convergence) hypothesis, while rejecting a unit root of the stationarity or convergence hypothesis.

The majority of the preceding studies considered the issue of energy consumption convergence in the context of developed or Asia and Pacific region countries. Their empirical results are mixed; some of them indicate energy convergence, while the others point to energy divergence for some countries or regions (for a summary review, see, e.g., [14], [15]). However, only very little attention in the literature has been devoted to this issue at the sub-national level regardless of the group of economies considered and no attention in Croatia. Hence, the main aim of this paper is to test the stationarity (convergence) hypothesis in per capita electricity consumption in Croatian counties in the period 2001-2013 by employing the panel unit root tests with structural break(s). We expect to find evidence on structural break(s) since Croatia faced structural transformations in the time period considered. They appeared mainly as a result of economic crisis and institutional, economic and policy changes (including energy reforms) caused by the accession process to the European Union. They have not been equally spilled over through the country.

The remainder of the paper is organized as follows. Section 2 outlines the methodology, whereas Section 3 gives the empirical results. The conclusions are given in the last section.

2 DATA AND METHOD

The annual data used in the paper refer to 20 Croatian counties and the City of Zagreb for the period 2001-2013. They were collected from the Hrvatska elektroprivreda (HEP) database. HEP is a leading Croatian electricity company engaged in energy business. Its electricity market share amounted to 85.1% and was stable during the whole time period considered. Final electricity consumption covers electricity supplied to the final consumers for all energy uses – households and enterprises. It is expressed in kWh. The electricity consumption data are divided by the population estimate to eliminate the biasing effect of different population distribution. Since data on the population estimates for the time period 2000-2002 were not available for the public, the estimate for 2000 and 2002 is made in this paper by correcting the Census 2001 data with the data on the natural increase and total net migration. To mitigate the implications of possible cross-sectional shocks, for each county i, the natural logarithm of the ratio of per capita electricity consumption (EC) relative to the annual average of all 21 Croatian counties in the sample is given by the expression: $REC_{it} = \ln(EC_{it}/$ average $EC_t)$, where REC_{it} denotes relative per capita electricity consumption in the county i in the year t. Now, as noted by Meng et al. [9], the transformed series measure relative per capita electricity consumption, meaning that any positive shock to energy consumption across all the counties that will increase the average by the same proportion will leave the relative energy consumption series unchanged. Consequently, any structural break identified in the transformed series would be county specific [11].

Average per capita electricity consumption in Croatia (registered by HEP) was 3,239.98 kWh in the period 2001-2013. It increased annually at the rate of 1.9% at the same time, or 4.08% in the period 2001-2008, primarily due to a post-war renewal of industrial production [6]. Figure 1 shows the development of the logarithm of the relative per capita electricity consumption series for 20 counties and the City of Zagreb in the period 2001-2013.

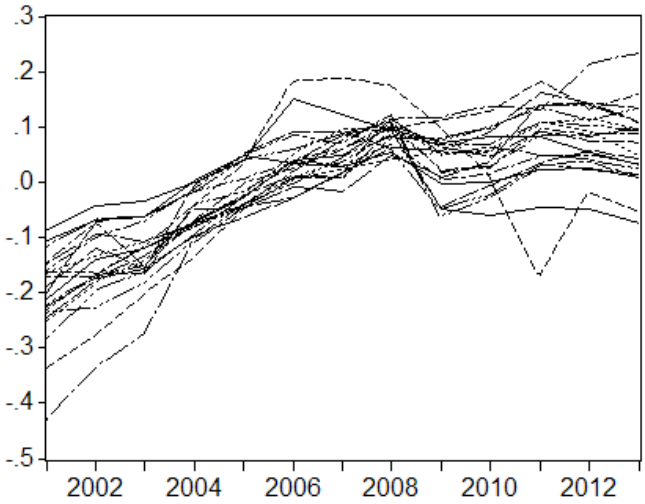


Figure 1: Development of the logarithm of the relative per capita electricity consumption series by counties

While there is an increasing trend and a similar pattern in per capita electricity consumption of the 20 Croatian counties and the City of Zagreb, the evolution in some of the counties shows strong heterogeneity in the series and across counties, and also a possible existence of significant structural break(s). All variables are given in logs.

To allow for the existence of break(s) being endogenously determined and different for different counties within the panel data, the Zivot-Andrews ([16]; ZA) test and the Clemente-Montañés-Reyes ([4]; CMR) test were employed in this paper.

The ZA test has a null hypothesis that the given series (y_t) contains a unit root with drift that excludes any structural break (equation 1). Then, depending on the model variant, the alternative hypothesis is a presence of a one-time structural break occurring at an unknown point in time. The ZA test considers each point as a potential break-date and runs a regression for them sequentially. Then the procedure selects the break-point amongst all possible break-points that minimizes the one-sided t-statistic, i.e., that supports the alternative hypothesis the most. Small values of the statistic lead to rejection of the null. Zivot and Andrews [16] developed three variants of the model to test for a unit root: model A, which permits a one-time change in the level of the series; model B, which allows for a one-time change in the slope of the trend function, and model C, which combines a one-time change in the level and the slope of the trend function of the series. Following Perron's suggestion [12] that most economic time series can be adequately modeled using either model A or model C, both models are primarily applied in the literature. However, since Sen [13] shows that model C is superior to model A, model C is chosen for the analysis of unit roots in this paper. A visual inspection of our data corroborates this choice. The alternative hypothesis for model C is that the series (y_t) is a trend-stationary process with a one-time break in both the intercept and the trend occurring at an unknown point in time (equation 2)

$$H_0: y_t = \mu + y_{t-1} + e_t \tag{1}$$

$$H_1: y_t = \mu + \theta DU_t(\lambda) + \beta t + \gamma DT_t(\lambda) + \alpha y_{t-1} + \sum_{j=1}^{k} c_j \Delta y_{t-j} + e_t, \tag{2}$$

where DU_t denotes a dummy variable that captures a shift in the intercept (mean) and DT_t is another dummy variable that represents a break in the trend occurring at a break time (TB). λ represents the break fraction (λ = TB/T) which is estimated by minimizing t-statistics for α = 1. μ, θ, β, γ and α are parameters, while k is the number of extra regressors determined by a test of significance of the estimated coefficients c_j. The definitions of DU_t and DT_t are following: $DU_t(\lambda) = 1$ if t > TB, zero otherwise; $DT_t(\lambda) = t - T\lambda$ if t > TB, zero otherwise.

The CMR test relaxes the assumption about only one structural break and allows for two breaks in the mean of the series. Bearing in mind the time period considered in this paper, i.e., 2001-2013, allowing for two structural breaks may be a reasonable expectation. In this test, the null hypothesis of a unit root with two structural breaks in the mean (equation 3) is tested against the alternative of the stationary series with two structural breaks (equation 4).

$$H_0: y_t = y_{t-1} + \delta_1 DTB_{1t} + \delta_2 DTB_{2t} + \mu_t \tag{3}$$

$$H_1: y_t = \mu + d_1 DU_{1t} + d_2 DTB_{2t} + e_t, \tag{4}$$

where DTB_{it} is a dummy variable with the value 1 if $t = TB_i + 1$ (i = 1, 2) and zero otherwise, and $DU_{it} = 1$ if $t > TB_i$ (i = 1, 2) and zero otherwise. TB_1 and TB_2 are the time periods when the mean is being identified. The authors [4] stress that $TB_i = \lambda_i T$ (i = 1, 2) with $0 < \lambda_i < 1$, and also that $\lambda_2 > \lambda_1$ can be supposed for the sake of simplicity.

Clemente, Montanes and Reyes [4] developed two variants of the model: an innovational outliers (IO) model, which allows for gradual shifts in the mean of the series, and an additive outliers (AO) model, which captures sudden changes in its mean. In this paper, both models will be used. In the case of the IO model, the unit root hypothesis is tested by first estimating equation (5), and, subsequently, by obtaining the minimum value of the pseudo *t*-ratio for testing whether the autoregressive parameter is 1 for all break-point combinations.

$$y_t = \mu + \rho y_{t-1} + \delta DTB_{1t} + \delta_2 DTB_{2t} + d_1 DU_{1t} + d_2 DU_{2t} + \sum_{i=1}^{k} c_i \Delta y_{t-i} + e_t. \tag{5}$$

In the case of the AO model, to test for the unit root hypothesis, the authors [4] suggest an application of a two-step procedure: first, the deterministic part of the variable has to be removed by estimating the following equation

$$y_t = \mu + d_1 DU_{1t} + d_2 DU_{2t} + \tilde{y}_t, \tag{6}$$

and, subsequently, the test that searches for the minimal t-ratio for the $\rho = 1$ hypothesis has to be performed for the following model

$$\tilde{y}_t = \sum_{i=0}^{k} \omega_{1i} DTB_{1t-1} + \sum_{i=0}^{k} \omega_{2i} DTB_{2t-i} + \rho \tilde{y}_{t-1} + \sum_{i=1}^{k} c_i \Delta \tilde{y}_{t-i} + e_t . \tag{7}$$

In model (7), DTB_{it} dummy variables are included in order to assure that a minimum t-statistic for the $\alpha = 1$ hypothesis is found.

3 RESULTS

The ZA test was used to test for a unit root allowing for one endogenously determined structural break. Table 1 presents the results of the tests. They show that only five counties (Istria, Karlovac, Slavonski Brod, Vukovar-Srijem and Split-Dalmatia) are sensitive to one-time structural break in both the intercept and trend. According to the ZA tests, shocks refer mostly to the beginning of the recession. This is expected since pessimistic expectations related to recession usually dampen the rate of electricity demand growth.

Certainly, the inability to reject the unit root null for the majority of the counties does not mean that this hypothesis is accepted. Moreover, Im et al. [5] observed that the results of the ZA test may be questionable since the test does not allow for a break under the null and, thus, the rejection of the null does not also guarantee that the series is stationary. Additionally, these findings may be biased due to the loss of information if actually more than one break exists (see [3], [7]).

To test for a unit root in the presence of double structural breaks, the CMR test with double mean shifts in the mean of the variable was also employed. The results for the IO model and the AO model, which allow for two breaks, are also presented in Table 1. The former indicates that relative per capita electricity consumption is not integrated across all series, i.e., that it constitutes a panel which is a mixture of trend-stationary and non-stationary processes. Namely, the unit root null is rejected for three counties at the 5% significance level. This means that a significant shock did not permanently change the growth path of the variables in three counties (Primorje-Gorski Kotar, Šibenik-Knin and Zagreb), while for the other counties we can not reject the null, i.e., that shocks had a permanent effect on electricity consumption. The results for the AO model show no evidence against the unit root hypothesis.

Comparing the results with no-break, one break or two breaks, Madsen and Smith [8] stress that as a rule of thumb, the one break case should be preferred to the no-break case if the break is statistically significant and the two break case should be preferred if the second break is statistically significant. Consequently, the IO model seems to be more appropriate for the Primorje-Gorski Kotar, Šibenik-Knin and Zagreb county series. Namely, these series can be considered as stationary around a mean which gradually changes twice in the period considered. The estimated changes are mostly related to the years 2002-2004 and 2007-2008 during which the energy legislative framework (including the real and expected increases in the price of electricity) changed. On the contrary, the counties of Istria, Karlovac, Slavonski Brod, Vukovar-Srijem and Split-Dalmatia had faced one estimated statistically significant

structural break caused primarily by the domestic recession that started in 2009 on a yearly basis.

Table 1: The results of the panel unit root test with structural break(s)

	ZA test		***CMR test – IO model***				***CMR test – AO model***			
	TB	***t-stat***	***MAIC***	***t-stat***	***TB1***	***TB2***	***MAIC***	***t-stat***	***TB1***	***TB2***
City Zagreb	2008	-2.007	1	-3.458	2004	2011	1	-3.174	2005	2008
Bjelovar-Bilogora	2007	-2.867	1	-3.524	2002	2004	1	-4.183	2002	2005
Dubrovnik-Neretva	2009	-1.206	1	-2.478	2002	2012	1	-3.358	2004	2007
Istria	2009	-5.681*	1	-3.728	2003	2009	1	-4.808	2004	2006
Karlovac	2009	-4.439*	1	-1.908	2008	2010	1	-2.698	2005	2007
Koprivnica-Križevci	2007	-3.957	1	-2.94	2003	2005	1	-4.451	2002	2005
Krapina-Zagorje	2008	-4.153	1	-2.421	2003	2009	1	-3.386	2004	2007
Lika-Senj	2004	-4.32	1	-3.921	2002	2011	1	-2.789	2005	2009
Međimurje	2011	-3.46	1	-1.963	2002	2010	1	-3.171	2005	2009
Osijek-Baranja	2009	-4.272	1	-4.264	2003	2008	1	-2.026	2006	2008
Pimorje-Gorski kotar	2005	-3.878	1	-6.306*	2003	2007	1	-2.692	2005	2008
Požega-Slavonia	2006	-2.929	1	-3.351	2004	2009	1	-2.962	2005	2008
Sisak-Moslavina	2006	-3.257	1	-4.447	2004	2009	1	-1.466	2005	2010
Slavonski Brod-Posavina	2009	-4.777*	2	-3.433	2003	2009	2	-2.825	2002	2005
Split-Dalmatia	2009	-5.985*	1	-5.06	2003	2007	1	-4.227	2004	2006
Šibenik-Knin	2009	-4.017	1	-10.015*	2002	2008	1	-2.536	2005	2008
Varaždin	2008	-4.05	1	-3.226	2003	2005	1	-3.562	2004	2007
Virovitica-Podravina	2007	-3.634	1	-3.596	2004	2012	1	-3.559	2005	2008
Vukovar-Srijem	2011	-5.976*	1	-4.688	2003	2009	1	-4.263	2005	2009
Zadar	2009	-2.237	3	-2.438	2002	2012	3	-3.218	2005	2009
Zagreb	2009	-3.342	1	-8.504*	2004	2007	1	-2.295	2005	2008

Note: TB = break point. * denotes statistical significance at the 5% level. The Stata routines zandrews for the ZA test and clemao2 and clemio2 for the CMR test (implemented by C. F. Baum [1] and [2] in Stata) were used to obtain estimates of t-statistics.
Notes for the ZA test: The optimal lag length was selected via a TT test. The critical value at the 5% significance level is -4.42.
Notes for the CMR test: The optimal lag length was chosen using the Modified Akaike Information Criterion (MAIC). The critical value at the 5% significance level is -5.490.

4 CONCLUSION

The results of the tests for a unit root for each panel member and in the presence of structural break(s) show that the per capita electricity consumption series is not integrated across all counties, i.e., that it constitutes a panel which is a mixture of trend-stationary and non-stationary processes. Moreover, some of them are sensitive to structural shock(s). The ZA tests highlight the impact of economic recession on per capita electricity consumption behavior, while the CMR test gradual changes that are provoked by changes in the energy legislative framework. However, it seems that these changes had a significant but transitory impact on the level of electricity consumption only in several Croatian counties. For the majority of Croatian counties, the changes had a permanent effect, implying the occurrence of the divergence process. Obviously, current energy policy could be improved by

addressing the causes of the divergence process. However, this may be the agenda for further research.

Moreover, different forms of energy consumption might exhibit different types of unit root behavior, and therefore testing for convergence in the total energy consumption series may be misleading. Therefore, further research should also analyze the unit root properties of each energy consumption form. Moreover, new tests more fitted to a small sample size and not restricted to one or two breaks might be used to check the robustness of the results.

Acknowledgment

This paper was supported by the Croatian Science Foundation under Grant number IP-2013-11-2203.

References

[1] Baum, C. F. (2004). ZANDREWS: Stata module to calculate Zivot-Andrews unit root test in presence of structural break. Boston College, Department of Economics. Statistical Software Components S437301, (revised 29 Nov 2009) Boston.

[2] Baum, C. F. (2005). CLEMAO_IO: Stata module to perform unit root tests with one or two structural breaks. Boston College, Department of Economics. Statistical Software Components S444302, (revised 17 Jun 2005) Boston.

[3] Ben-David, D., Lumsdaine, R. and Papell, D.H. (2003). Unit root, postwar slowdowns and long-run growth: Evidence from two structural breaks. Empirical Economics, 28 (2): 303-319.

[4] Clemente, J., Montañes, A. and Reyes, M. (1998): Testing for a unit root in variables with a double change in the mean, Economics Letters, 59(2): 175-182

[5] Im, K.S., Lee, J. and Tieslau, M. (2010). Stationarity of inflation: Evidence from panel unit root tests with trend shifts, 20th Annual Meetings of the Midwest Econometrics Group, Oct 1-2, 2010. Available at: http://apps.olin.wustl.edu/MEG Conference/Files/pdf/2010/70.pdf

[6] Ivanović, M., Blažević, D. and Glavaš, H. (2010). The structure of electricity consumption and its utilizations efficiency in the European transition countries. International Journal of Electrical and Computer Engineering Systems, 1(2): 19-34.

[7] Maddala, G.S. and Kim, I.M. (2003). Unit root, cointegration and structural change. Cambridge and New York: Cambridge University Press.

[8] Madsen, J.B. and Smith, R. (2008). Is the output-capital ratio constant in the very long time? Monash University. Discussion paper number 10., Copenhagen.

[9] Meng, M., Payne, J.E. and Lee, J. (2013). Convergence in per capita energy use among OECD countries. Energy Economics, 36(C): 536-545.

[10] Mishra V., Sharma S. and Smyth R. (2009). Are fluctuations in energy consumption per capita transitory? Evidence from a panel of Pacific Island countries. Energy Policy: 37(6): 2318-2326.

[11] Mishra, V. and Smyth, R. (2014). Convergence in energy consumption per capita among ASEAN countries. Energy policy, 73: 180-185.

[12] Perron, P. (1989). The great crash, the oil price shock and the unit root hypothesis. Econometrica, 57: 1361-1401.

[13] Sen, A. (2003). On unit root tests when the alternative is a trend break stationary process. Journal of Business and Economic Statistics, 21(1): 174-184.

[14] Shahbaz, M., Tiwari, A. and Khan, S. (2012). Is energy consumption per capita stationary? Evidence from first and second generation panel unit root tests, MPRA Paper from University Library. Number 41607, Munich.

[15] Smyth, R. (2013). Are fluctuations in energy variables permanent or transitory? A survey of the literature on the integration properties of energy consumption and production. Applied Energy, 104: 371-378.

[16] Zivot, E. and Andrews, W. K. (1992). Further evidence on the Great crash, the oil price chock, and the unit-root hypothesis. Journal of Business and Economic Statistics, 10(3): 251-270.

INNOVATIVE PRODUCTS AND NEWSVENDOR PROBLEM UNDER UNCERTAINTY WITHOUT PROBABILITIES

Helena Gaspars-Wieloch
Poznań University of Economics, Department of Operations Research
Al. Niepodległości 10, 61-875, Poznań, Poland
helena.gaspars@ue.poznan.pl

Abstract: The paper presents a new scenario-based decision rule for the classic version of the newsvendor problem (NP) under complete uncertainty. So far, NP has been analyzed under uncertainty with probabilities or with partial information. The procedure described in the contribution is based on a hybrid of Hurwicz and Bayes decision rules. It takes into account the decision maker's attitude towards risk and the whole distribution of payoffs connected with particular order quantities. It does not require any information about the probability distribution since it is designed for the sale of new, innovative products, i.e. for one-shot decisions.

Keywords: newsvendor problem, uncertainty without probabilities, one-shot decisions, scenario-based decision rule, coefficient of optimism, standard deviation, innovative products

1 INTRODUCTION

The newsvendor problem (NP), also known as the single-period problem (SPP) or the newsboy problem, consists in finding the order quantity which maximizes the expected profit (or minimizes the expected loss) in a single period probabilistic demand framework. This topic has attracted a great deal of attention and played a central role at the conceptual foundations of stochastic inventory theory. It was originally related to DMSU (decision making under stochastic uncertainty) or DMR (decision making under risk) where the demand is presented as a random variable. NP has been already analyzed on diverse assumptions and with various extensions [7]. Additionally, this problem has been also recently discussed in the context of DMPI – decision making with partial information [24, 25, 26], where the decision maker (DM) is able to subjectively define possibility degrees and satisfaction levels (the probability distribution is not known completely).

Nevertheless, according to [4], newsvendor theory should not assume that the DM faces a known distribution, since in real-life situations, demand distribution is not always known. Furthermore, the authors demonstrate that knowing the demand distribution does not necessarily lead the subject closer (than that one who is unaware of the underlying demand distribution) to the optimal solution or to improve profits (see also [5]). Therefore, the investigation of NP under complete uncertainty (i.e. without probabilities [33, 8, 42, 43, 45, 48]) is much desired. Especially in the case of new (innovative) product development where it is quite complicated to define probabilities or even probability-like quantities, because there are no data available for forecasting the upcoming demand via statistical analysis.

Newsvendor models are usually based upon the assumption of risk neutrality. Meanwhile, recently there is a growing body of literature that attempts to use alternative risk preferences rather than risk neutrality to describe the newsvendor decision-making behavior [1, 26, 30, 35, 49, 50, 51, 54]. In this contribution, we also take into account the DM's attitude towards risk. That means that the solution recommended for a particular DM depends on two factors: the target objectively defined (profit maximization) and the DM's nature. The paper is organized as follows. Section 2 deals with the main features of the traditionally understood NP. Section 3 defines the newsboy problem under complete uncertainty. Section 4 presents a 2-criteria decision rule that may be used as a tool in searching an optimal solution for the

problem aforementioned. Section 5 provides a case study. Conclusions are gathered in the last part. The contribution deals with innovative and small life cycles products.

2 NEWSVENDOR PROBLEM

The classic newsvendor problem constitutes a production/procurement problem of a retailer who sells a product under random demand without keeping inventory. There are many situations in practice where keeping a product in inventory for future use is either impossible or impractical. This is the case for products such as newspapers and perishable food. A similar situation arises when an apparel retailer makes orders in the beginning of the season for a fashion item. Such orders are made for one season (sales time window) only, and any unsold (leftover) items are not kept in inventory to be sold next year. They are rather sold at deep discounts at the end of the season. Thus, in NP the retailer places an order for a product to his own supplier at the beginning of each period and the quantity procured is used solely to satisfy the demand during the current period [6]. The demand for this product during the current period is not known in advance, but it is represented by a nonnegative random variable D. The cumulative distribution function of D is F, i.e. $P(D \le x) = F(x)$. The distribution may be continuous (e.g. the endpoints of the interval with possible values of the demand are equal to D_{min}=10 and D_{max}=25), or discrete (the demand takes a finite and countable number of values, e.g. $P(D=1)=0.3$; $P(D=2)=0.4$; $P(D=3)=0.2$; $P(D=4)=0.1$). The goal of NP may consist in expected profit maximization or expected loss minimization. Here we focus on problems with discrete demand distributions and expected profit maximization, which are described for example in [42]. We assume that c_1 is the unit production/purchase cost of the product. Symbol c_2 denotes the selling price (full retail price) of this product and c_3 stands for the discount price (price of leftover items/salvage value), where $c_3 < c_1 < c_2$. Symbol q signifies the order quantity. Values of c_1, c_2, c_3 allow one to calculate the unit profit from selling the product at price c_2: $b=c_2-c_1$, and the unit loss from selling it at price c_3: $s=c_1-c_3$. Equation (1) enables one to compute profit $g(q,D)$ gained by the retailer when the supply equals q and the demand is equal to D. Expected profit $p(q)$ is given by Equation (2), where $D_{min}=q_{min}$ and $D_{max}=q_{max}$ are the minimal and maximal quantity of demand considered by the retailer and $P(D)$ is the probability that the demand will be equal to D. NP consists in determining q^* which satisfies Equation (3). In the case of discrete demand distributions, this problem may be solved by means of a profit matrix (Table 1), a recurrence equation or a critical ratio [42].

$$g(q, D) = \begin{cases} b \cdot q, & \text{if } q \le D, \\ b \cdot D - s(q - D), & \text{if } q > D. \end{cases} \tag{1}$$

$$p(q) = \sum_{D=D_{\min}}^{D_{\max}} g(q, D) \cdot P(D) \tag{2}$$

$$q^* = \arg\max_q p(q) \tag{3}$$

Table 1. Profit matrix for the classic newsvendor problem (general case)

$q \setminus D$	$D_{min}=q_{min}$	$D_{min}+1$	…	$D_{max}-1$	$D_{max}=q_{max}$
q_{min}	$b \cdot q$	$b \cdot q$	$b \cdot q$	$b \cdot q$	$b \cdot q$
$q_{min}+1$	$b \cdot D - s(q-D)$	$b \cdot q$	$b \cdot q$	$b \cdot q$	$b \cdot q$
…	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot q$	$b \cdot q$	$b \cdot q$
$q_{max}-1$	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot q$	$b \cdot q$
q_{max}	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot D - s(q-D)$	$b \cdot q$

Extended newsvendor models are variations of the classic newsvendor model involving different objectives, utility functions, supplier pricing policies, newsvendor pricing policies, discounting structures, states of information about demand and supply, constrained multi-products, multiple-products with substitution, random yields, multi-location models and different multiple production cycles, e.g. [1, 3, 7, 9, 12, 30, 31, 32, 34, 38, 40, 50, 52, 53, 54].

3 NEWSVENDOR PROBLEM UNDER COMPLETE UNCERTAINTY

We have already mentioned that in the case of new products, no relevant historical data are available for statistical demand analysis. In such circumstances, one can refer to decision making with partial information and characterize the demand by possibility distributions [26]. Another approach consists in assuming that the decision is made under complete uncertainty, which facilitates the decision making process, since this time it is not necessary to estimate probability-like quantities. The only parameter that should be declared (if we intend to take into consideration DM's preferences) is the coefficient of optimism (or pessimism). The second benefit of applying DMCU (decision making under complete uncertainty) to NP is the possibility to combine NP with scenario planning [39, 46], because within the framework of the newsboy problem the profit matrix can be computed in a very precise way (see Table 2). The result of the choice made under uncertainty with scenario planning depends on two factors: which decision will be selected and which scenario will occur. Thus, NP under complete uncertainty may be defined by means of a scenario-based decision model with m states of nature (scenarios, events): $S_1, \ldots, S_i, \ldots, S_m$, n possible alternatives (decisions, strategies, order quantities): $A=\{A_1, \ldots, A_j, \ldots, A_n\}$, and $n \times m$ profits (a_{ij} – profit gained by the retailer if state S_i occurs and alternative A_j is selected) calculated according to formula (4). The distributions of payoffs are discrete. Symbols $a_{j,min}$ and $a_{j,max}$ denote the minimal and maximal profit connected with decision A_j (see Section 4).

Table 2. Profit matrix for NP under complete uncertainty (general case)

Scenarios \ alternatives	A_1 ($q=q_{min}$)	A_2 ($q=q_{min}+1$)	A_j	A_{n-1} ($q=q_{max}-1$)	A_n ($q=q_{max}$)
S_1 ($D=D_{min}$)	$a_{1,1}$	$a_{1,2}$	$a_{1,j}$	$a_{1,n-1}$	$a_{1,n}$
S_2 ($D=D_{min}+1$)	$a_{2,1}$	$a_{2,2}$	$a_{2,j}$	$a_{2,\,n-1}$	$a_{2,\,n}$
S_i	$a_{i,1}$	$a_{i,2}$	$a_{i,j}$	$a_{i,\,n-1}$	$a_{i,\,n}$
S_{m-1} ($D=D_{max}-1$)	$a_{m-1,1}$	$a_{m-1,2}$	$a_{m-1,j}$	$a_{m-1,\,n-1}$	$a_{m-1,\,n}$
S_m ($D=D_{max}$)	$a_{m,1}$	$a_{m,2}$	$a_{m,j}$	$a_{m,\,n-1}$	$a_{m,\,n}$

$$a_{ij} = \begin{cases} b \cdot q, & \text{if } \; j \le i, \\ b \cdot D - s(q-D), & \text{if } \; j > i. \end{cases} \tag{4}$$

We assume that NP for new products may be an uncertain problem where an optimal pure (not mixed) strategy has to be found (the DM completely executes only one alternative).

There are many classical and extended decision rules designed for DMU [2, 10, 11, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 36, 37, 41, 44, 47]. Nevertheless, the majority of the extended rules refer to the probability calculus, which is rather characteristic of DMR – decision making under risk, or DMU with probabilities. Let us remind that in this paper we focus on the Knight's definition, according to which uncertainty occurs when we do not know the probabilities of particular scenarios[1] (see complete uncertainty).

[1] Of course, we are aware of the fact that many researchers apply the alternative approach according to which each non-deterministic (with and without probabilities) decision problem is treated as an uncertain problem, meanwhile risk is understood as the possibility that some bad circumstances might happen.

It is worth emphasizing that some rules find application when the DM intends to perform the selected strategy only once (one-shot decisions). Others are recommended for people contemplating realization of the chosen variant many times (multi-shot decisions). In the NP case with innovative and small life cycles products we certainly deal with one-shot decisions.

In the next section we make an attempt to find an appropriate decision rule for the classic version of the newsvendor problem under complete uncertainty (NPCU).

4 TWO-CRITERIA (H+B) RULE FOR NPCU

When selecting a suitable procedure for NPCU (the likelihood of particular states is unknown) we must remember that the method should consider the DM's nature in the decision making process. Furthermore, it is recommended to take into account the specific structure of the profit matrix of the classic NP (see Table 3):

a) for b sufficiently bigger than s and for equal likelihood, $p(q)$ is the highest for $q=q_{max}$,
b) for b sufficiently smaller than s and for equal likelihood, $p(q)$ is the highest for $q=q_{min}$,
c) for $q=q_{min}$, particular profits a_{ij} are always the same (regardless of the state) – hence, the value of the profit is certain for decision A_1,
d) almost all alternatives have more than one profit equal to $a_{j,max}$ and the number of such profits increases for q close to q_{min} – the higher q is, the less certain value $a_{j,max}$ is,
e) for each decision particular profits $a_{1,j}, \ldots, a_{i,j}, \ldots, a_{m,j}$ are always ordered in the form of a non-decreasing sequence, which means that the sets of payoffs achievable across the states do not overlap in the top right-hand corner of the matrix,
f) the range between $a_{j,min}$ and $a_{j,max}$ is an increasing function $f(q)$.

Table 3. Examples of profit matrix for NP under complete uncertainty ($q_{min}=D_{min}=1$, $q_{max}=D_{max}=4$)

Examples	**I. b=5, s=1**				**II. b=3, s=3**				**III. b=1, s=5**			
Sc. \ Alt.	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4
S_1	5	4	3	2	3	0	-3	-6	1	-4	-9	-14
S_2	5	10	9	8	3	6	3	0	1	2	-3	-8
S_3	5	10	15	14	3	6	9	6	1	2	3	-2
S_4	5	10	15	20	3	6	9	12	1	2	3	4

In connection with all those factors, we can conclude that decision rules formulated by Wald, Hurwicz, Savage, Bayes and Hayashi can not be applied to NPCU (due to the lack of possibility to consider DM's nature, the lack of application to one-shot decisions, irrational solutions in the case of asymmetric distributions of payoffs or not overlapping sets of payoffs for particular scenarios, etc., compare with the justification presented in [13, 15, 17, 18]).

Therefore, we use in this paper a less-known extended procedure devoted to DMU, i.e. a hybrid of Hurwicz's and Bayes' rules (H+B rule), which is described in detail in [15, 19]. This method, thanks to parameters $\alpha \in [0,1]$ (the coefficient of pessimism is close to 1 for extreme pessimists) and $\beta = (1-\alpha) \in [0,1]$ (the coefficient of optimism is close to 1 for radical optimists), takes into account DM's preferences. Both coefficients are not measures of probability – they just subjectively present someone's behaviour. In the H+B rule, in contradiction to the Hurwicz's, Wald's, Hayashi's, Savage's approaches, all outcomes have an influence on the value of the final measure, which is quite advantageous for cases where alternatives contain many profits equal to extreme values. The general idea of H+B is to assign, for a pessimist, α to the last term of the non-increasing sequence of all payoffs related to a given decision and β to the remaining terms of that sequence. For an optimist, weights are set in a different way: β is connected with the first term of the sequence and α with the remaining ones. Due to the fact that the ranges of profits related to particular alternatives vary

rather significantly, we will support the H+B rule for NPCU with an additional auxiliary decision tool, which analyzes the deviations between outcomes.

The suggested 2-criteria (H+B) rule for NPCU may consist of the following steps:

1) Determine α and β. If $\alpha \in [0,0.5)$, then $\alpha = \alpha_o, \beta = \beta_o$ (α_o and β_o are optimist's coefficients). If $\alpha \in (0.5,1]$, then $\alpha = \alpha_p, \beta = \beta_p$ (α_p and β_p are pessimist's coefficients).
2) Define $q_{min}=D_{min}$, $q_{max}=D_{max}$, m, n and the set of alternatives (A).
3) Estimate prices c_1, c_2, c_3, compute b, s and generate the profit matrix.
4) Find a non-increasing sequence of gains $Sq_j = (a_{1j},...,a_{tj},...,a_{zj})$ for each order: $a_{t,j} \geq a_{t+1,j}$ (t=1,…,z-1), z – number of terms in the sequence, t – number of the term in the sequence.
5) Calculate, for each decision, index hb_j (hb_j^p, hb_j^o or $hb_j^{0.5}$ depending on parameter α). If $\alpha \in (0.5,1]$, calculate hb_j^p (index for pessimists) according to Equation (5). If $\alpha \in [0,0.5)$, compute hb_j^o (index for optimists) following formula (6). If α = 0.5, calculate hb_j^{05} using Equation (7), where b_j denotes the Bayes criterion, i.e. the average of all payoffs.

$$hb_j^p = \frac{\alpha_p \cdot a_{zj} + \beta_p \cdot \sum_{t=1}^{z-1} a_{tj}}{\alpha_p + (z-1) \cdot \beta_p} \tag{5}$$

$$hb_j^o = \frac{\alpha_o \cdot \sum_{t=2}^{z} a_{tj} + \beta_o \cdot a_{1j}}{(z-1) \cdot \alpha_o + \beta_o} \tag{6}$$

$$hb_j^{0.5} = hb_j^p = hb_j^o = b_j = \frac{1}{m} \cdot \sum_{i=1}^{m} a_{ij} \tag{7}$$

The denominators in Equations (5)-(6) are introduced so that the final value of particular indices belongs to interval $[w_j,M_j]$, where w_j and M_j are the values of the Wald's criterion and the maximax criterion, respectively, i.e. the last (a_{zj}) and the first (a_{1j}) term of Sq_j. Denominators are not crucial – they can be omitted when preparing the ranking.

6) Choose alternative A^*_j fulfilling condition (8).

$$A_j^* = \arg\max_j (hb_j) \tag{8}$$

7) If set A^* containing decisions A^*_j is a singleton set, stop the procedure (A^*_j is the final optimal solution). Otherwise, select decision A^{**}_j which satisfies Equations (8)-(9).

$$A_j^{**} = \arg\min_{j^*} (\sigma_{j^*}) \tag{9}$$

where σ_{j^*} is the standard deviation calculated for all decisions A^*_j.

The assignment of parameters α and β to particular payoffs, depending on the level of optimism, is justified in [15]. Briefly, the H+B rule recommends for a pessimist an alternative with a relatively high payoff $a_{j,min}$ or with quite frequent payoffs (almost) equal to $a_{j,max}$. On the other hand, that rule suggests for an optimist an alternative with the highest payoff $a_{j,max}$, but its highest payoffs do not have to be frequent. The second criterion (standard deviation) is introduced in the final step of the procedure in order to find a relatively safe strategy (i.e. an alternative with as few small payoffs as possible), which is especially crucial in the case of cautious DMs. The deviation criterion is only applied to decisions with the highest index hb_j. The use of the standard deviation as a supplementary tool in the scenario-based decision making process has been also proposed by [29].

5 ILLUSTRATIVE EXAMPLE

In this Section we are going to solve Example I presented in Table 3 by means of the 2-criteria (H+B) rule. We assume that the DM is almost an extreme pessimist:

1) α=0.9, β=0.1. Thus $\alpha = \alpha_p, \beta = \beta_p$.

2) $q_{min}=D_{min}=1$, $q_{max}=D_{max}=4$, $m=n=4$, $A=\{A_1, A_2, A_3, A_4\}$.

3) $c_1=6$, $c_2=11$, $c_3=5$. Hence $b=5$, $s=1$. The profit matrix is given in Table 3 (Example I).

4) $Sq_1=(5,5,5,5)$, $Sq_2=(10,10,10,4)$, $Sq_3=(15,15,9,3)$, $Sq_4=(20,14,8,2)$.

5) Indices hb_j^p are calculated in the following way:

$$hb_1^p = \frac{0.9 \cdot 5 + 0.1 \cdot (5+5+5)}{0.9 + 3 \cdot 0.1} = 5.0; hb_2^p = \frac{0.9 \cdot 4 + 0.1 \cdot (10+10+10)}{0.9 + 3 \cdot 0.1} = 5.5; hb_3^p = 5.5; hb_4^p = 5.0$$

6) There are two alternatives A^*_j: $A^*=\{A_2, A_3\}$ (see Equation 8).

7) Set A^* is not a singleton set. Thus, it is necessary to compute σ_{j*}: σ_2=3.0, σ_3=5.74. Decision A_2 is the final optimal alternative (A^{**}_j) since it has the highest index hb_j^p (in set A) and the lowest standard deviation (in set A^*). The order quantity should be equal to 2.

Note that if the DM was almost a radical optimist (α=0.1), then indices hb_j^o would be equal to 5, 9.5, 13.5 and 17, respectively. In that case, decision A_4 would be the best one.

6 CONCLUSIONS

The 2-criteria (H+B) rule, proposed and described in the paper, may be a quite comfortable and comprehensive decision tool in the newsvendor problem under complete uncertainty (NPCU) with new and small life cycles products since it takes into account DM's preferences and the very specific distribution of payoffs (asymmetry, range, frequency of extreme values). It does not require any information about the likelihood (which is justifiable in the case of innovative products) and it is rather simple to use. The 2-criteria (H+B) rule may be helpful in any uncertain decision problem (not merely NPCU), especially in the case of a considerable extreme payoffs differential. In the future it would be desirable to analyze extended newsvendor models in the context of complete uncertainty.

Acknowledgement

This research is financed by the National Science Center, Poland (project registration number: 2014/15/D/HS4/00771)

References

[1] Agrawal, V. and Seshadri, S. (2000). Impact of uncertainty and risk aversion on price and order quantity in the newsvendor problem. Manufacturing and Service Operations Management, 2(4): 410–423.

[2] Basili, M. (2006). A rational decision rule with extreme events. Risk Analysis, 26: 1721–1728.

[3] Behret, H. and Kahraman, C. (2010). A multi-period newsvendor problem with pre-season extensions under fuzzy demand. Journal of Business Economics and Management, 11(4): 613–629.

[4] Benzion, U., Cohen, Y. and Shavit, T. (2010). The newsvendor problem with unknown distribution. Journal of the Operational Research Society, 6(6): 1022-1031.

[5] Besbes, O. and Muharremoglu, A. (2013). On implications of censoring demand in the newsvendor problem. Management Science, 59(6): 1407-1424.

[6] Burnetas, A.N., Gilbert, S.M. and Smith, C.E. (2007). Quantity discounts in single-period supply contracts with asymmetric demand information. IIE Transactions, 39: 465-479.

[7] Choi, T.-M. (2012). Handbook of newsvendor problems: models, extensions and applications. Springer, International Series in Operations Research and Management Science 176.
[8] Courtney, H., Kirkland, J. and Viquerie, P. (1997). Strategy under uncertainty. Harvard Business Review, 75(6): 66–79.
[9] Dogru, M. K., de Kok A. G., and van Houtum G.J. (2013). Newsvendor characterizations for one-warehouse multi-retailer inventory systems with discrete demand under the balance assumption. Central European Journal of Operations Research, 21: 541-559.
[10] Ellsberg, D. (2001). Risk, ambiguity and decision. New York: Garland Publishing.
[11] Etner, J., Jeleva, M. and Tallon, J.-M. (2012). Decision theory under ambiguity. Journal of Economic Surveys, 26(2): 234–270.
[12] Gallego, G. and Moon, I. (1993). The distribution free newsboy problem: review and extensions. Journal of the Operational Research Society, 44: 825–834.
[13] Gaspars-Wieloch, H. (2012). Ograniczona skuteczność metod optymalizacyjnych w rozwiązywaniu ekonomicznych problemów decyzyjnych [Limited efficiency of optimization methods in solving economic decision problems]. Ekonomista 2012/3: 303-324.
[14] Gaspars-Wieloch, H. (2013). On a decision rule supported by a forecasting stage based on the decision maker's risk aversion. In: Zadnik Stirn L, Zerovnik J, Povh J, Drobne S, Lisec A (Eds.), Proceedings of the 12th International Symposium of Operational Research, 53–59.
[15] Gaspars-Wieloch, H. (2014a). Propozycja hybrydy reguł Hurwicza i Bayesa w podejmowaniu decyzji w warunkach niepewności [A hybrid of the Hurwicz and Bayes rules in decision making under uncertainty], In: Modelowanie preferencji a ryzyko'14, ed. T. Trzaskalik, Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 178, 74–92.
[16] Gaspars-Wieloch, H. (2014b). On a decision rule for mixed strategy searching under uncertainty on the basis of the coefficient of optimism. Procedia - Social and Behavioral Sciences, Elsevier, 110(2014): 923 – 931.
[17] Gaspars-Wieloch, H. (2014c). Modifications of the Hurwicz's decision rules. Central European Journal of Operations Research, 22(4): 779–794.
[18] Gaspars-Wieloch, H. (2014d). Modification of the maximin joy criterion for decision making under uncertainty. Quantitative Methods in Economics, XV(2): 84-93.
[19] Gaspars-Wieloch, H. (2015a). Modifications of the Omega ratio in decision making under uncertainty. Croatian Operational Research Review, 6(1): 181-194.
[20] Gaspars-Wieloch, H. (2015b). On a decision rule supported by a forecasting stage based on the decision maker's coefficient of optimism. Central European Journal of Operations Research 23(3): 579-594.
[21] Ghirardato, P., Maccheroni, F. and Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude. Journal of Economic Theory, 118: 133–173.
[22] Gilboa, I. (2009). Theory of decision under uncertainty. Cambridge University Press. Cambridge, New York.
[23] Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. Journal of Mathematical Economics, 18: 141–153.
[24] Guo, P. (2011). One-shot decision theory. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 41(5): 917–926.
[25] Guo, P. (2013). One-Shot decision making with regret. (ICIST) International Conference on Information Science and Technology, 493-495. http://dx.doi.org/10.1109/ICIST.2013.6747596
[26] Guo, P. and Ma, X. (2014). Newsvendor models for innovative products with one-shot decision theory. European Journal of Operational Research, 239: 523-536.
[27] Hayashi, T. (2008). Regret aversion and opportunity dependence. Journal of Economic Theory, 139(1): 242–268.
[28] Hurwicz, L. (1952). A criterion for decision making under uncertainty. Technical Report, 355. Cowles Commission.
[29] Ioan, C. and Ioan, G. (2011). A method of choice of the best alternative in the multiple solutions case in the games theory. The Journal of Accounting and Management, 1(1): 5–8.

[30] Kamburowski, J. (2014). The distribution-free newsboy problem under the worst-case and best-case scenarios. European Journal of Operational Research, 237: 106–112.
[31] Khouja, M. (1999). The single period (news-vendor) problem: literature review and suggestions for future research. Omega, The International Journal of Management Science, 27: 537–542.
[32] King, A.J. and Wallace, S.W. (2012). Modelling with stochastic programming. Springer.
[33] Knight, F. H. (1921). Risk, uncertainty, profit. Hart. Boston MA, Schaffner & Marx, Houghton Mifflin Co.
[34] Kocabıyıkoğlu, A., and Popescu, I. (2011). The newsvendor with pricing: a stochastic elasticity perspective. Operations Research, 59(2): 301-312.
[35] Lee, C.-Y., Li, X. and Yu, M. (2014). The loss-averse newsvendor problem with supply options. Naval Research Logistics, 12/2014, 62(1), http://dx.doi.org/10.1002/nav.21613
[36] Marinacci, M. (2002). Probabilistic sophistication and multiple priors. Econometrica, 70: 755–764.
[37] Piasecki, K. (1990). Decisions and reliable forecasts [Decyzje i wiarygodne prognozy]. Poznań: Akademia Ekonomiczna w Poznaniu.
[38] Petruzzi, N. C., Wee, K. E. and Dada, M. (2009). The newsvendor model with consumer search costs. Production and Operations Management, 18(6): 693-704.
[39] Pomerol, J. C. (2001). Scenario development and practical decision making under uncertainty. Decision Support Systems, 31(2): 197–204.
[40] Qin, Y., Wang, R. Vakharia, A. J., Chen, Y. and Seref, M. M. H. (2011). The newsvendor problem: review and directions for future research, European Journal of Operational Research, 213: 361–374.
[41] Savage, L. (1961). The foundations of statistics reconsidered. Studies in Subjective Probability. New York: Wiley, 173–188.
[42] Sikora, W. (Ed.) (2008). Badania Operacyjne [Operations research]. Warsaw: Polskie Wydawnictwo Ekonomiczne.
[43] Trzaskalik, T. (2008). Wprowadzenie do badań operacyjnych z komputerem [Introduction to operations research with computer] (2nd ed.). Warsaw: Polskie Wydawnictwo Ekonomiczne.
[44] versky, A. and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5: 297–323.
[45] von Neumann, J. and Morgenstern, O. (1944). Theory of games and economic behavior. Princeton, New York: Princeton University Press.
[46] Van der Heijden, K. (1996). Scenarios: the art of strategic conversation. Chichester: John Wiley and Sons.
[47] Wald, A. (1950). Statistical decision functions. New York: Wiley.
[48] Walliser, B. (2008). Cognitive economics. Springer. Berlin-Heidelberg.
[49] Wang, C.X. and Webster, S. (2009). The loss-averse newsvendor problem. Omega, 37: 93-105.
[50] Wang, C. X., Webster, S. and Suresh, N. C. (2009). Would a risk-averse newsvendor order less at a higher selling price?. European Journal of Operational Research, 196: 544–553.
[51] Wang, C. X., Webster, S., Zhang, S. (2012). Newsvendor models with alternative risk preferences within expected utility theory and prospect theory frameworks, in: Choi T.-M (ed.). Handbook of Newsvendor Problems: Models, Extensions and Applications. Springer, International Series in Operations Research and Management Science 176
[52] Wang, K., Sun J., Liang, L. and Li, X. (2015). Optimal contracts and the manufacturer's pricing strategies in a supply chain with an inequity-averse retailer. Central European Journal of Operations Research. http://dx.doi.org/10.1007/s10100-013-0335-2
[53] Wu, M., Zhu, S. X. and Teunter, R. H. (2013a). Newsvendor problem with random shortage cost under a risk criterion. International Journal of Production Economics, 145(2): 790–798.
[54] Wu, M., Zhu, S. X. and Teunter, R. H. (2013b). The risk-averse newsvendor problem with random capacity. European Journal of Operational Research, 231: 328–336.

THE ROBUST BINOMIAL APPROACH TO THE MACHINE INTERFERENCE PROBLEM WITH DIFFERENT GROUPS OF IDENTICAL MACHINES AND PREEMTIVE PRIORITY

Gregory Gurevich
Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering,
Bialik Sts. 56, P.O. Box 950, Beer Sheva 84100, Israel
gregoryg@sce.ac.il

Baruch Keren
Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering,
Bialik Sts. 56, P.O. Box 950, Beer Sheva 84100, Israel
baruchke@sce.ac.il

Yossi Hadad
Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering,
Bialik Sts. 56, P.O. Box 950, Beer Sheva 84100, Israel
yossi@sce.ac.il

Abstract: This paper presents a binomial model for a special case of the machine interference problem (MIP), where a production system consists of different groups of identical machines. All the machines produce the same product independently of each other. Each machine randomly requests a service that is provided by a group of operators. The model assumes that each group of machines has a different priority. Queue discipline serves the machines according to their priority (preemptive priority). The model enables calculation of the interference proportion for each machine, depending on the number of operators and the priority.

Keywords: machine interference problem (MIP), priority, binomial model, queuing.

1 INTRODUCTION

The latest version of the binomial model was recently developed by Hadad et al. [1], Keren at al. [2] and Gurevich et al. [3]. This model is applicable for a production system with several machines that produce the same product in parallel and independently of each other. The machines are classified into J $(1 \le J < \infty)$ different groups by types. Each group j, $j = 1,..,J$ includes N_j identical machines. Note that a production system with several types of heterogeneous machines is common in many industries because new machines are frequently introduced either through expansion or replacement. Because the complete elimination of old machines takes years, enterprises are quite likely to have a heterogeneous mixture of both old and new machines (see e.g. [4]). These machines can be expected to differ in run time for production of one unit of a product, failure rate, service time, operation cost, proportion of defects, and similar metrics. Each machine in group j needs a determinable production time T_j (run time) to produce one unit of a product. During the production process each machine may request a service for loading or unloading and for troubleshooting, where the requests are independent. The time that an operator invests in service throughout a cycle is a random variable with an average value of t_j. The assumption is that the value t_j is identical for all the machines in group j. The service is given by the group of K certified operators $\left(1 \le K \le \sum_{j=1}^{J} N_j\right)$ who are qualified to repair all machines.

If the number of operators is less than the number of machines, $K < \sum_{j=1}^{J} N_j$, then some machines may wait for a service while the operators give service to other machines (machine interference). The average time that a machine in group j produces one unit of a product is called the cycle time $H_j, j = 1,..,J$. Thus, the cycle time H_j is the sum of three components: the run-time T_j, the average service time t_j, and the average interference time t_{I_j}, i.e., $H_j = T_j + t_j + t_{I_j}$. During waiting time and service time the machines are idle. Thus, in the steady state situation each machine in group j $(j = 1,..,J)$ can be in one of the following states: running (producing items) with some probability or idle with some probability. These probabilities depend on the average interference time. If the values of these probabilities were known, one could calculate the probability that n_j machines in group j are getting or requiring a service according to the binomial distribution where $n_j = 0,1,...,N_J$. The presented model allows calculation of the average interference time t_{I_j} $(j = 1,..,J)$ for the machines in each group j where the number of operators K and the priority of each group are given. The value of t_{I_j} depends on the number of operators K, the number of machines in each group N_j, and on queue discipline. The values of T_j, t_j do not depend on K, N_j or on queue discipline. Therefore, T_j, t_j can be evaluated by work measurement (see e.g. [5], [6]). The average interference time enables one to calculate the cycle time H_j, $j = 1,..,J$, and steady state probabilities. Using the cycle time H_j, $j = 1,..,J$, one can determine an optimal number of operators in the context of different objective functions, for example, to minimize of total manufacturing cost per unit of a product.

2 MODEL DESCRIPTION

This section presents the model assumptions, notations and application.

2.1 Model assumptions

1) There are J $(1 \le J < \infty)$ groups of machines, each group j, $j = 1,..,J$ includes N_j identical machines.
2) Each machine of the group j, $j = 1,..,J$ can be in one of the following positions, where the probabilities for each position are constant in a steady state situation and identical for all machines of the group:
    a. running (producing items),
    b. having a service,
    c. waiting for a service (interference).
3) Machine failures are independent.
4) Service time is random and each service request transfers immediately to operators.
5) An available operator handles a service request immediately.
6) Each service request is handled by only one operator.
7) Walking time from one machine to another is negligible.
8) A machine is idle while waiting for a service or while getting a service.

9) Each group of machines has a different priority and groups of machines are ranked according to their priorities. Queue discipline serves the machines according to their priority. If an operator must select which of several machines from the same group must be served, that selection is made randomly.
10) Machines are served according to the absolute priority policy (a preemptive priority). When all the operators are busy and an additional machine with a higher priority requests a service, the service of one of the machines with the lowest priority ceases immediately and its operator serves the machine with the higher priority. When a previously interrupted service is resumed, this service is resumed from the point where it was preempted, without loss of the prior work.

2.2 Notations

N_j -Number of identical machines in the group j, $j=1,...,J$.

T_j - Runtime. The length of time needed for a machine in the group j, $j=1,...,J$, to process one unit of a product. The run time T_j is a pre-given deterministic value.

t_j - Average time of the service that operators invest in a machine of the group j, $j=1,...,J$, during its cycle time (H_j).

t_{I_j} - Interference time. The average time during a cycle time (H_j) during which a machine in the group j, $j=1,...,J$, is idle because it is waiting for an operator.

i_j - Interference proportion. The ratio between the interference time t_{I_j} and the cycle time H_j, that is, $i_j = t_{I_j} / H_j$, $j=1,...,J$.

Thus, the cycle time H_j is calculated as follows:

$$H_j = T_j + t_j + t_{I_j} = T_j + t_j + i_j \times H_j = \frac{T_j + t_j}{\left(1 - i_j\right)}.$$

S_j - Service proportion. The ratio between the average time of the service for a machine in group j, t_j, and the cycle time H_j, i.e., $S_j = t_j / H_j$.

p_{j0} - Probability that a machine in group j is running, $j=1,...,J$. This probability is calculated as follows:

$$p_{j0} = \frac{T_j}{H_j} = \frac{T_j\left(1 - i_j\right)}{T_j + t_j}.$$

p_j - Probability that a machine in group j is getting or requiring a service (idle), $j=1,...,J$. This probability is calculated as follows:

$p_j = S_j + i_j$, $j=1,...,J$.

Each machine in each group can be in one of two states - running or idle. A machine in the idle state can be in one of two positions - getting the service or waiting for the service. Because these states are mutually exclusive, it is clear that for any machine in group j the follows equality holds: $p_{j0} + p_j = 1$, $j=1,...,J$.

X_{j0} - Number of running machines in group j (a random variable), $j=1,...,J$.

X_j - Number of machines in group j that are getting or requiring a service (a random variable), $j=1,...,J$.

It is clear that $X_{j0}+X_j=N_j$, i.e., the sum of the number of running machines and the number of idle machines of group j, is equal to the number of machines in the group j, $j=1,...,J$. Because each group has N_j identical machines, the probability p_j is equal in the steady state for all machines in group j. Therefore, one can calculate the probability that $X_j=x_j$, i.e., x_j ($x_j=0,1,...,N_j$) machines in group j are getting or requiring a service according to the binomial distribution, $X_j \sim Bin\left(p_j,N_j\right)$, $j=1,...,J$.

K - Number of certified operators, $1\le K\le \sum_{j=1}^{J} N_j$.

3 CALCULATION OF THE INTERFERENCE PROPORTION FOR EACH GROUP OF MACHINES

Denote the group with the highest priority as "group 1" ($j=1$), and so on. Let us L_j be a number of machines in group j that are waiting for the service, $j=1,...,J$. If the value of expectation $E\left(L_j\right)$ was known, one could calculate the interference proportion as $i_j=E\left(L_j\right)/N_j$, $j=1,...,J$ (see e.g. [1], [7]). The value of $E\left(L_j\right)$ $j=1,...,J$, is calculated as follows: for the group with highest priority $(j=1)$:

$$E\left(L_1\right)=\sum_{m=K+1}^{N_1} P\left(X_1=m\right)\times\left(m-K\right)=\sum_{m=K+1}^{N_1}\binom{N_1}{m}\left(p_1\right)^m\left(1-p_1\right)^{N_1-m}\times\left(m-K\right), \qquad (1)$$

and for the other groups:

$$E\left(L_j\right)=E\left(E\left(L_j\left|\sum_{i=1}^{j-1}X_i\right.\right)\right),\quad j=2,...,J. \qquad (2)$$

The random variables X_j and $\sum_{i=1}^{j-1}X_i$, $j=2,...,J$, are independent. Therefore,

$$E\left(L_j\left|\sum_{i=1}^{j-1}X_i\right.\right)=\sum_{m=max\left\{0,K+1-\sum_{i=1}^{j-1}X_i\right\}}^{N_j} P\left(X_j=m\right)\times\left(m-max\left\{0,K-\sum_{i=1}^{j-1}X_i\right\}\right)$$
$$=\sum_{m=max\left\{0,K+1-\sum_{i=1}^{j-1}X_i\right\}}^{N_j}\binom{N_j}{m}\left(p_j\right)^m\left(1-p_j\right)^{N_j-m}\times\left(m-max\left\{0,K-\sum_{i=1}^{j-1}X_i\right\}\right). \qquad (3)$$

By (2) and (3),

$$E\left(L_j\right)=\sum_{n=0}^{\sum_{i=1}^{j-1}N_i}\ \sum_{m=max\{0,K+1-n\}}^{N_j}\left(\begin{array}{l}\binom{N_j}{m}\left(p_j\right)^m\left(1-p_j\right)^{N_j-m}\\ \times\left(m-max\left\{0,K-n\right\}\right)\times P\left(\sum_{i=1}^{j-1}X_i=n\right)\end{array}\right),\ j=2,...,J. \qquad (4)$$

Note that $\sum_{i=1}^{j-1} X_i$ is a sum of independent binomial random variables, i.e., $X_i \sim Bin(N_i, p_i)$, $i = 1,...,j-1$. For a special case where $J = 2$ the probability is:

$$P\left(\sum_{i=1}^{j-1} X_i = n\right) = P(X_1 = n) = \binom{N_1}{n}(p_1)^n (1-p_1)^{N_1-n} \quad n = 0,1,...,N_1, \tag{5}$$

for $J = 3$ the probability of the sum of two independent binomial random variables is:

$$P\left(\sum_{i=1}^{2} X_i = n\right) = P(X_1 + X_2 = n) = \sum_{i=0}^{N_1} P(X_1 = i) \times P(X_2 = n-i) \tag{6}$$

$$= \sum_{i=0}^{N_1} \binom{N_1}{i}(p_1)^i (1-p_1)^{N_1-i} \times \binom{N_2}{n-i}(p_2)^{n-i} (1-p_2)^{N_2-n+i} \times I(0 \le n-i \le N_2),$$

$n = 0,1,...,N_1 + N_2$, where $I(\cdot)$ is the indicator function. Similarly for $J > 3$ one can calculate the distribution of $\sum_{i=1}^{j-1} X_i$ recursively, by finding the distribution of $\sum_{i=1}^{j-2} X_i$ and then adding the remaining X_{j-1} as is presented by following equation (7):

$$P\left(\sum_{i=1}^{j-1} X_i = n\right) = P\left(\sum_{i=1}^{j-2} X_i + X_{j-1} = n\right) = \sum_{k=0}^{\sum_{i=1}^{j-2} N_i} P\left(\sum_{i=1}^{j-2} X_i = k\right) \times P(X_{j-1} = n-k). \tag{7}$$

Thus, with modern computing tools, given the values of N_i and p_i, $i = 1,...,j-1$, it is possible to calculate the exact distribution of $\sum_{i=1}^{j-1} X_i$, $j = 2,...,J$. The interference proportion i_1 of machines of the group 1 ($j = 1$) is calculated as the solution of the following equation:

$$i_1 = \frac{E(L_1)}{N_1}. \tag{8}$$

By substituting $E(L_1)$ from equation (1) into equation (8) and setting $p_1 = i_1 + S_1$, equation (8) has the following form:

$$i_1 = \frac{1}{N_1} \sum_{m=K+1}^{N_1} \binom{N_1}{m}(i_1 + S_1)^m \left(1-(i_1 + S_1)\right)^{N_1-m} \times (m-K). \tag{9}$$

Hadad et al. [1] showed that equation (9) has a unique feasible solution. Similarly, sequentially for $j = 2,...,J$, i_j is the solution of the follows equation:

$$i_j = \frac{E(L_j)}{N_j}, \tag{10}$$

where $p_r = i_r + S_r$, $r = 1,...,j-1$. By substituting $E(L_j)$ from equation (4) into equation (10) and setting $p_j = i_j + S_j$, equation (10) has the form of:

$$i_j = \frac{1}{N_j} \times \sum_{n=0}^{\sum_{i=1}^{j-1} N_i} \sum_{m=max\{0,K+1-n\}}^{N_j} \left(\begin{array}{l} \binom{N_j}{m}(i_j + S_j)^m \left(1-(i_j + S_j)\right)^{N_j-m} \\ \times (m - max\{0, K-n\}) \times P\left(\sum_{i=1}^{j-1} X_i = n\right) \end{array} \right). \tag{11}$$

Because $\sum_{i=1}^{j-1} X_i$ is a sum of independent binomial random variables, where $X_i \sim Bin(N_i, p_i)$, the probability $P\left(\sum_{i=1}^{j-1} X_i = n\right)$, $n = 0,...,\sum_{i=1}^{j-1} N_i$, is completely defined by the probabilities p_r, $r = 1,..., j-1$. Therefore, equation (11) has a single variable i_j. By a similar way as was presented in Hadad et al. [1], equation (11) has a unique feasible solution. The solution can be obtained numerically (e.g., by the interval halving method) or by software tools such as Excel-Solver.

4 CONCLUSION

This paper deals with a special case of the machine interference problem where a production system has several types of machines, and where all the machines work in parallel and produce the same product. The heterogeneous machines are divided into several groups, where each group that includes several identical machines has a different priority. The machines request only one type of service that is provided by a group of operators. The presented model enables calculation of the expected number of machines that are waiting for service in each group of machines and the interference proportion for each group via the binomial probability function.

References

[1] Hadad, Y., Keren, B. and Gurevich, G. (2013). A multinomial model for the machine interference problem with different service types and multiple operators. Computers & Industrial Engineering, 66(2): 264-273.

[2] Keren, B., Gurevich, G. and Hadad, Y. (2015). Machines interference problem with several operators and several service types that have different priority. International Journal of Operational Research, in press.

[3] Gurevich, G., Hadad, Y. and Keren, B. (2015). Calculation of the steady state waiting time distribution in the context of the machine interference problem. International Journal of Operational Research, in press.

[4] Prsafalvi, A. and Sztrik, J. (1987). On the heterogeneous machine interference with limited server's availability. European Journal of Operational Research, 28: 321-328.

[5] Hadad, Y., Keren, B. and Hanani, Z.M. (2014). The number of measurements to be performed for time study analysis. European Journal of Industrial Engineering, 8(2): 222-243.

[6] Niebel, B. and Freivalds, A. (2003). Methods, standards, and work design. McGraw-Hill: New York, 5-50.

[7] Jones, D.W. (1971). Work measurement of multi machine assignments, 3.92-3.112 of Maynard HB., Industrial Engineering Handbook, McGraw-Hill: New York.

A LEAN PRODUCTION PROCESS - TODAY'S DESTINATION OF COMPANIES

Eva Jordan
Kekon d.o.o
Grajski trg 15, 8360 Žužemberk, Slovenia
Eva.e.jordan@gmail.com

Tomaž Berlec
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, 1000 Ljubljana, Slovenia
Tomaz.berlec@fs.uni-lj.si

Marko Starbek
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, 1000 Ljubljana, Slovenia
Marko.starbek@fs.uni-lj.si

Janez Kušar
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, 1000 Ljubljana, Slovenia
Janez.kusar@fs.uni-lj.si

Abstract: Customers require short delivery times and lower price of products (lean production). These requirements of customers can only be met if a company switches from a classically organized to a lean production. This paper presents a value stream analysis that shows a size of the lead time (LT_{pp}). Overall equipment effectiveness (OEE) and efficiency index (PE) allow us to determine a leanness index of a production process (LI). The value LI in fact show the leanness of a production process. Portfolio analyses of leanness of production process is shown on an example of production of cooling covers.

Keywords: lean production, lead time, equipment effectiveness, OEE, leanness index

1 INTRODUCTION

Lean production is a practice of managing production processes that was developed in Japan. Eiji Toyoda, Taiichi Ohno and Shaotro Kamiya have developed new concepts for managing production processes and have gradually created the Toyota Production System TPS [5]. In 1990, James Womack joined their management concepts and thus founded lean production, the goal of which is reduction of wastes [9]. Goals of a lean production process are: shorter lead time, lower tied-up capital, better flow of material and information, motivation of workers and satisfaction of customers [3, 4].

The goal of the research presented is a procedure of portfolio analyses for leanness of production.

2 LEAN PRODUCTION

The lean production concept is a customer-oriented production process with a low level of wastes and a short lead time.

Figure 1 shows a common goal of both concepts for the improvement of the production process, yet a different path to a lean process.

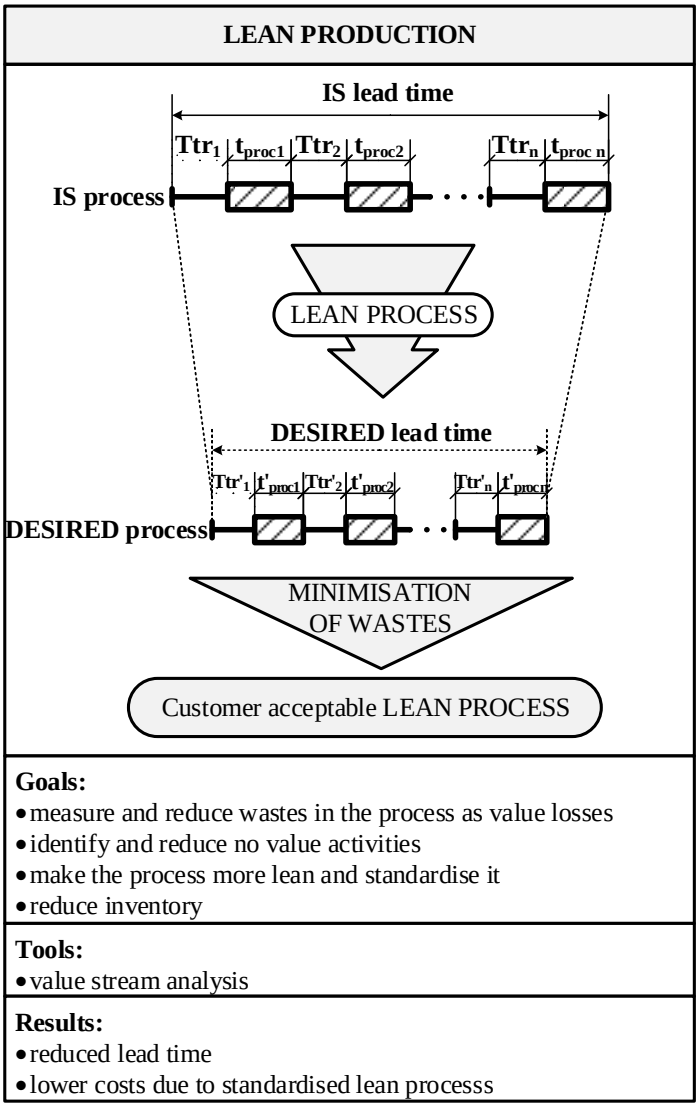


Figure 1: Common goal, yet a different path to a lean Six Sigma process

Taiichi Ohno defined the seven wastes [6] that are a burden to production processes:

- overproduction,
- waiting times,
- ineffective production,
- transport,
- scrap and processing,
- inventory, and
- redundant motion.

Today, there are more forms of waste in addition to the seven: waste related to brain drain, i. e. individuals with technical skills leave a company with their skills, and non-utilized talent of employees in a company.

2.1 Value stream analysis

The value stream analysis [8] is derived from the current situation of a value stream in a production process that is created in three steps:

Step 1: Selection of a product or a family of products

Step 2: Analysis of customers' needs

Step 3: Assessing the current value stream situation.

A stream flow analysis in large-series production of a small number of product types is carried out for each product type separately. In a small-series production of a large number of product types, a value stream analysis is performed only for a representative of an individual family of products; it can be created in two ways [7], namely by creating a matrix of families of products or by creating a flow chart of the production and similarity of the families.

An analysis of customers' needs allows assessment of a customer's tact time:

$$CT = \frac{T_d}{Q_d} \tag{1}$$

CT - customer's tact time [sec/pcs]

T_d – daily available work time [sec/Wd]

Q_d – daily needed quantity of products [pcs/Wd]

Selection of a product or a representative from a family of products and analysis of customers' needs allow assessment of the current value stream.

The lead time of a production process with »n« subprocesses represents a time interval between the beginning of the lead time of the first subprocess and the end of the last subprocess, in fact, it equals the sum of lead times of subprocesses [8]:

$$LT_{pp} = \sum_{i=1}^{n}\left(Ttr_i + t_{proc_i}\right) \tag{2}$$

LT_{pp} – lead time of production process [min]

Ttr_i – time of transition of i^{th} subprocess [min]

t_{proc_i} – processing time per piece of i^{th} subprocess [min]

n – number of subprocess

A short lead time of a production process is achieved by reducing the times of wastes. The best index for assessing reduction in the lead time of a production process is the process efficiency index (PE):

$$PE = \frac{\sum_{i=1}^{n} t_{proc_i}}{LT_{pp}} \tag{3}$$

PE – process efficiency index [-]

The equation (3) clearly shows that the PE index increases by reduced times of wastes.

The following guidelines are suggested to reach a lean process:

- introduction of continuous production,
- introduction of FIFO connections,
- KANBAN system for regulating the situation.

The overall equipment effectiveness (OEE) [6] and the process efficiency index (PE) are used to calculate the leanness index of a production process (LI).

$$LI = 100 \times OEE \times PE \tag{4}$$

LI – leanness index of a production process [-]

OEE – overall equipment effectiveness [-]

The leanness index of a production process (LI) indicates how reliable individual steps of the production process (OEE) are and how lean the production process is (PE).

By positioning the production process with respect to the achieved OEE and PE, the achieved leanness index of a production process is determined (see Fig. 2).

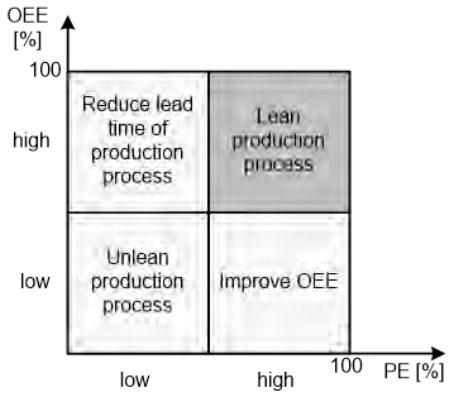


Figure 2: Portfolio analysis of leanness of a production process

We should continue improving the process and the overall equipment effectiveness until a production process falls into the square »Lean production process«.

3 PORTFOLIO ANALYSIS OF A LEAN IN A PRODUCTION OF COOLING COVERS

The company desires to assess lean indices of production processes.

A team consisting of five co-workers of the company was organised in order to assess the leanness of the process. Based on collected data on implementation of subprocesses i (processing times $t_{proc,i}$ and preparation times Tp_i) the team made an assessment of the current value stream of the process of producing cooling covers and determined scopes of the situation S, lead time LT_{pp} and process efficiency PE (see Fig. 3):

$$S = \frac{Q}{N_d} \tag{5}$$

S – scope of situation [Wd]
Q – quantity of situation [pcs]
N_d – daily need [pcs/Wd]

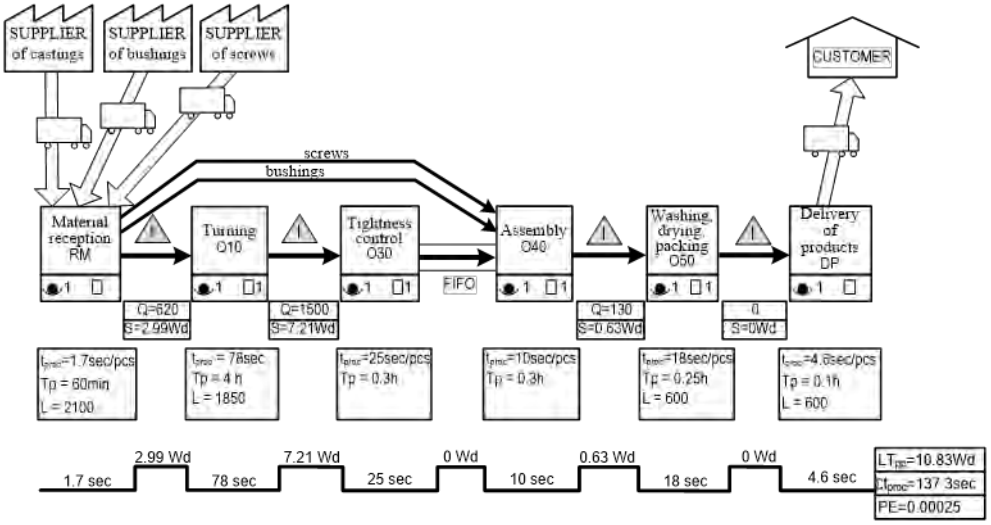


Figure 3: Current value stream of production of cooling covers

The team members have analysed the value stream chart and came to the following conclusions:

- lead time of the production process of cooling covers amounts (figure 3): LT_{pp} = 10.83 Wd (1Wd = 2 shifts of 7 effective hours)
- sum of processing times (figure 3): Σt_{proc} = 137.3 sec
- process efficiency amounts to PE:

$$PE = \frac{\sum_{i=1}^{n} t_{proc_i}}{LT_{pp}} = 0.00025 = 0.025\%$$

The team, in co-operation with the head of production, evaluated the wastes:

- time of equipment failure and time of preparation of equipment 25 %, the estimated availability rate is therefore A = 0.75
- time of idle time of equipment 10 %, the estimated efficiency rate is E = 0.90
- quality of products 20 %, the evaluated quality rate of equipment is Qr = 0.80

Evaluated overall equipment effectiveness OEE [7] is:

$$OEE = 100 \cdot (A \cdot E \cdot Qr) = 100 \cdot (0.75 \cdot 0.90 \cdot 0.80) = 54\ \%$$

The evaluated overall equipment effectiveness OEE value and the reached process efficiency PE allow us to calculate the leanness index of a production process:

$$LI = 100 \cdot (OEE \cdot PE) = 100 \cdot (0.54 \cdot 0.00025) = 0.0135\ \%$$

Figure 4 shows the achieved process leanness of the production of cooling covers.

Figure 4 indicates that the process of the production of cooling covers is unlean, improvement potentials are seen especially in terms of reducing the lead time LT_{pp}.

The team wanted to assess the situation in the lean production field, so they decided to assess the success of sub processes and of the entire process of production of cooling covers.

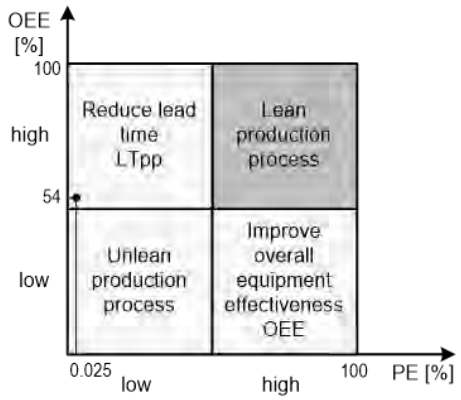


Figure 4: Portfolio analysis of process leanness of the production of cooling covers

4 CONCLUSIONS

This paper showed a path that leads to lean production processes. The past research has shown that production process become lean if wastes are eliminated or reduced and if optionally a transition from workshop production to continuous production is made [1, 2].

We propose that the index of lean processes is calculated by two parameters: OEE and PE.

The results of assessing the process leanness of the production of cooling covers have shown that the process is very unlean. These indicate the need for performing continuous improvements with a focus on reducing the lead time of the process.

Further research will be directed especially towards increasing process leanness and increasing process profitability.

For competitiveness on the global market, the companies will need to quickly switch to a lean production.

References

[1] Buchmeister, B., Friščić, D., Lalić, B. and Palčič, I. (2012). Analysis of a Three-Stage Supply Chain with Level Constraints. International Journal of Simulation Modelling, 11(4): 196-210.

[2] Debevec, M., Šimic, M. and Herakovič, N. (2014). Virtual Factory as an Advanced Approach for Production Process Optimization. International Journal of Simulation Modelling, 13(1): 66-78.

[3] Hines, P., Found, P. and Harrison, R. (2008). Staying Lean: thriving, not just surviving. Lean Enterprise Research Centre, Cardiff University, Cardiff, ISBN 0902810111.

[4] Holweg, M. (2007). The genealogy of Lean production. Journal of Operations Management, 26(2): 420-437.

[5] Monden, Y. (1998). Toyota Production System; An Integrated Approach to Just-in-Time, Third edition. Narcross, Ga: Engineering and management press

[6] Ohno, T. and Bodek, N. (1988). Toyota Production System: Beyond Large-Scale Production. Portland, Oregon: Productivity Press.

[7] Potočnik, P., Berlec, T., Starbek, M. and Govekar, E. (2012). Self-organizing neural network-based clustering and organization of production cells. Neural computing and applications. http://dx.doi.org/10.1007/s00521-012-0938-x [Accessed 2.6.2015]

[8] Tapping, D., Luyster, T. and Shuker, T. (2002). Value Stream Management, Eight Steps to Planning, Mapping, and Sustaining Lean Improvements. New York: Productivity Press.

[9] Womack, J.P. and Jones, D.T. (2003). Lean Thinking, Banish Waste and Create Wealth in Your Corporation. New York.

ON MODELLING OF SPATIAL LIGHT DISTRIBUTION OF LED WITH ATTACHED SECONDARY OPTICS

David Kaljun
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
david.kaljun@fs.uni-lj.si
Janez Žerovnik
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
janez.zerovnik@fs.uni-lj.si

Abstract: In design of optical systems based on LED technology, a crucial task is to handle the unstructured data describing properties of optical elements in standard formats. This leads to the problem of data fitting within an appropriate model. We overview our recent work on the problem and discuss plans for future work.

Keywords: least squares function fiting, Newton method, discrete optimization, local search, light distribution, LED

1 INTRODUCTION

The LED industry has been evolving rapidly in the past several years. The fast pace of research and development in the field had some expected impact. One of the results is a massive use and implementation od LED elements in all kind of luminaires. While some of these luminaires are designed for ambient illumination the majority are technical luminaires that have to conform not only to electrical and mechanical safety regulations but also to regulations that define and restrict the photo-metrics of a certain luminaire. This means that the photometry of a luminaire has to be defined prior to production. In order to do that efficiently and with minimal error the design engineer must virtually test the luminaires performance. Tools that can be used (OpticsWorks, LigthTools, TracePRO) [16, 12, 10] do exist and they offer a vast repository of sub-modules to develop and design custom lenses, reflectors, light guides, etc. These universal tools however do not completely exploit the luminaire design possibilities that were introduced by the transition from conventional light source technologies to LED. One of the possibilities which is also the main goal of a bigger study that incorporates the research presented here is to have an expert or intelligent system which would be capable of suggesting a secondary lens combination that would result in a user defined end photometry. In other words, the system would take some stock secondary LED lenses from different manufactures, place them on a defined LED array and search for the optimal combination of lenses so that the resulting photometry would be as close as possible to the user defined one. The method could enable the luminaire designer to custom design the light engine to a specific area of illumination, while keeping the mechanical and electrical parts of a luminaire untouched. This would in turn provide a customer with a tailored solution that would guarantee a maximum efficiency, lower prices, fewer light pollution and the possibility to individualize the illumination effect while maintaining a consistent visual appearance of the luminaries.

There are several optimization tasks related to development of the above idea. Here we focus on the approximation of spatial light distribution with a moderate number of suitable

basic functions [5, 15]. The problem that is defined formally in the next section is motivated by the following. The data describing the properties of the lenses and/or of the desired light distribution is nowadays usually given in some standard format files that correspond to the measured (or desired) values at a number of points in space. This results in relatively large data files of unstructured data. Clearly, if the data can be well enough approximated c.f. as a linear combination of certain basic functions, this may enable faster computations using less computer storage. Recent experiments showed that sufficiently good approximations can be obtained by some basic optimization algorithms, including local search algorithms and genetic algorithms [6, 7, 8].

However, when using predefined lenses to design a luminarire that closely approximates a desired light distribution, it may be essential that the approximation error is much lower.

The same task can also be seen as solving a problem of data compression, replacing a long unstructured data file with a much shorter one, in this case a sequence of parameters. It makes sense to aim at 0% approximation when considering the data compression task.

The rest of the paper is organized as follows. In section two we discus the problem and present the mathematical model. Section three overviews recent results on the topic and Section 4 suggests the experiment setup and algorithms for the future.

2 THE MODEL

It is known that the spatial light distribution of some LED lenses can be approximated by a sum of a small number of certain basic functions [5]. Provided the approximation is sufficiently good, it may be possible to provide designs combining several lenses with controlled error rate.

This naturally opens several research avenues. For example, it is important to have error free or at least very good approximations of the basic lenses, and to have methods that are stable in the sense that they are not too sensitive to the noise in the presentation of basic elements.

Here we focus on the first above mentioned task, approximation of the unstructured spatial light distribution data. We search for an approximation of the Luminous intensity $I(\Phi; \mathbf{a}, \mathbf{b}, \mathbf{c})$ at the polar angle of Φ in the form

$$I(\Phi; \mathbf{a}, \mathbf{b}, \mathbf{c}) = I_{max} \sum_{k=1}^{K} a_k * \cos(\Phi - b_k)^{c_k} \tag{1}$$

where K is the number of functions to sum and a_k, b_k, c_k are the function coefficients that we search for. Here we need to note two restrictions on the model. First restriction emerges from the LEDs physical design. The LED can not emit any light to the back side which is the upper hemisphere in our case. That is why all intermediate values that are calculated at the combined angle $(\Phi - b_k)$ greater than 90 equal 0. The second restriction deals with the slightly unusual description of the light distribution in standard files such ad Elumdat .ldt [18] and Iesna .ies [4]. These files present measured candela values per angle Φ on so called C planes which can be observed on Figure 1. One C plane is actually only one half of the corresponding cross-section and does not describe the other half. But from a physical point of view we need to consider the impact from the other half of the cross-section. Because all lenses used here are symmetric, we can simplify the calculation of the intermediate values and

incorporate the impact of the other half by mirroring (multiplying the angles by -1) all values that are calculated with the combined angle $(\Phi - b_k)$ less than 0. Note however that this only works with symmetrical distributions, and should be reconsidered carefully when the method is to be applied to asymmetrical distributions.

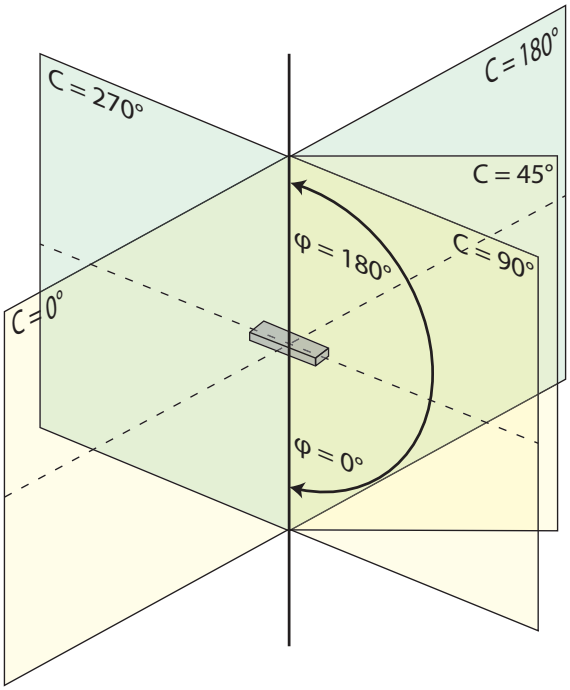


Figure 1: C-planes according to standard. C-planes angles : 0°- 360°— Φ angles : 0°to 180°

The goodness of fit is defined as the root mean square error (RMS), formally defined by the expression:

$$RMS\,(\mathbf{a},\mathbf{b},\mathbf{c}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[I_m(\Phi_i) - I(\Phi_i,\mathbf{a},\mathbf{b},\mathbf{c})\right]^2} \tag{2}$$

where RMS represents the error of the approximation in %, N the number of measured points from the input data, $I_m(\Phi_i)$ the measured Luminous intensity value at the polar angle Φ from the input data, and $I(\Phi_i,\mathbf{a},\mathbf{b},\mathbf{c})$ the calculated Luminous intensity value at the given polar angle Φ.

3 OVERVIEW OF RECENT RESULTS

The model was successfully applied to LED's with attached secondary optics and symmetric light distribution [5] showing that sufficiently good approximations (RMS error below 5%) can be obtained using a sum of only three functions ($K = 3$). Previously, approximation of spatial light distribution of a LED with uniform distribution and without a secondary lens using this

type of functions was proposed in [15]. The model was slightly modified in [5] where a new normalizing parameter was introduced, and consequently, all other parameters have values in fixed intervals known in advance. It should be noted that the modified model is equivalent to the original, only the number of parameters and their meaning differ. It may be interesting to note that due to symmetries of the examples, $K = 3$ is sufficient for both applications [5, 15]. In general case, we expect that $K > 3$ functions will be needed for sufficiently good approximations, and in view of optimization of the design of a luminaire it is interesting to have an idea how large the parameter K can grow to assure that the light distribution fits the desired (and/or standard) sufficiently well.

In previous work [6, 7], the model described above was applied in conjunction with several custom build algorithms that are based on local search heuristics and some metaheuristics. The implemented algorithms include a steepest descend algorithm, two iterative improvement algorithms with different neighborhoods and two genetic algorithms, a standard one and a hybrid one in which the best individuals of every generation are optimized with the iterative improvement algorithm. The experiment compared three local search algorithms (iterative improvement based on two different neighborhoods, and a steepest descent using one of the neighborhoods) and a genetic algorithm with several variants that had different population sizes and different number of generations. The third algorithm was called a hybrid algorithm because it combined the genetic algorithm with local search so that the best members of population were optimized by a short local search. The dataset consisted of a set of real available lenses to be approximated. The set was taken from the online catalogue of one of the biggest and most present manufacturer in the world Ledil Oy Finland [11]. The selection from the broad spectrum of lenses in the catalogue was based on the decision that the used LED is of the XP-E product line from the manufacturer Cree [2]. For more detailed description of the algorithms we refer to [6, 7]. (In [6], a version of genetic algorithm was used that some specialists argued to be non-standard, however the results did not differ much seemingly.) The results of the experiments showed that all of the applied algorithms are capable of providing satisfactory results on all tested instances, and differred mainly in the needed computational time. The average RMS values obtained on real lenses were around $RMS = 2\%$. It may be interesting to note that the hybrid algorithm performed best, but on the other hand the performance was only slightly better than that of the best local search algorithm (the differences were not statistically significant). Hence, the results mentioned proved that the model is accurate and that sufficiently good approximations can be found with a variety of algorithms for sufficiently good description of lenses.

However, recall that the model can also be used for data compression task. Zero or very low RMS error is also essential in the foreseen application, in which the premanufactured lenses are to be combined into a more complex luminaire with prescribed light distribution.

We are interested first in minimizing the approximation error, and second, in computational time of the methods.

4 CONCLUSIONS AND FUTURE WORK

When applying the model to the data compression problem, the target RMS error is 0%. Therefore, we aim to improve the approximation results that were obtained previously [6, 7] and restrict attention to symmetric light distributions. Also, we fix $K = 3$ functions in the

model. Besides the dataset of 14 realistic lenses that was used in previous studies, we will also generate an artificial dataset in which a sample is simply a sum of three basic functions with randomly chosen parameters. This assures that zero error approximation is possible for the instances of the artificial dataset.

In the model we use a sum of functions that are smooth and hence the first and second derivatives can be calculated allowing application of continuous optimization methods in addition to the general discrete optimization metaheuristics that were used before. We have chosen to use the Newton (also known as the Newton–Raphson) iterative method [17] to find the solution that we seek. It is well known that convergence of the Newton method largely depends on the initial solution. Therefore we will apply the method in two ways. First, we will use the Newton method as an optimizer which will pinpoint the local minimum of the solutions found by heuristic algorithms. In a sense this implementation of the Newton method will be an extension of the discrete optimization algorithm, used to finalize the search to end in a local minimum. (Note that the local minima may be missed by the discrete optimization algorithms due to predefined length of the discrete moves.) Second, we will use the Newton method as a standalone algorithm that will on initialization generate a number of random (initial) solutions that are uniformly scattered over the whole search space and then it will use the Newton method on them to find the local minimums. Of course, for both implementations to be comparable the iteration count has to be controlled so that the overall maximum amount of computation time will be roughly the same. Preliminary results show that we can expect significant quality and time improvements in both implementations [9].

5 Acknowledgements

This work was supported in part by ARRS, the Research agency of Slovenia, grants P1-0285 and ARRS-1000-15-0510.

References

[1] Aarts, E. and Lenstra, J. K. (2003). Local Search in Combinatorial Optimization. Town: Princeton University Press, 41 William Street, Princeton, New Jersey 08540.

[2] Cree Inc. 4600 silicon drive, north carolina 27703, USA, Technical report. http://www.cree.com/led-components-and-modules/products/xlamp/discrete-directional/xlamp-xpe [Accessed 15/06/2015].

[3] Herceg, D. (2013). Means based modifications of Newtons method for solving nonlinear equations. Applied Mathematics and Computation, Vol(No):219 6126 – 6133.

[4] IESNA. (2002). Standard File Format for the Electronic Transfare of Photometric Data and Related Information LM-63-02. ANSI.

[5] Kaljun, D. and Žerovnik, J.(2014). Function fitting the symmetric radiation pattern of a LED with attached secondary optic. Optics Express, Vol(No):22 29587-29593.

[6] Kaljun, D. and Žerovnik, J. (2014). On local search based heuristics for optimization problems. Croatian Operational Research Review, Vol(No):5(2) 317-327.

[7] Kaljun, D. and Žerovnik, J. (2014). Local search based optimization of an analytical model of symmetric spatial light distribution, in Bioinspired optimization methods and their applications : proceedings of the Student Workshop on Bioinspired Optimization Methods and their Applications - BIOMA 2014, J. Šilc and A. Zamuda, ed. Jožef Stefan Institute, pp. 81-92. (full paper will appear in journal Informatica) ,

[8] Kaljun, D. and Žerovnik J. (2015) Developing LED ilumination optics design. Papa G. (Eds.). Advances in Evolutionary Research (Announced). Town:Nova Science Publishers, Inc., 400 Oser Ave Suite 1600, Hauppauge NY 11788-3619, United States of America.

[9] Kaljun, D., Petrišič, J., and Žerovnik J. (2015) manuscript in preparation.

[10] Lambdares. (2015). TracePRO.
http://www.lambdares.com/ [Accessed 15/06/2015].

[11] Ledil Oy - Technical resources . (2015). Ledil Oy.
http://www.ledil.com/products/?y [Accessed 15/06/2015].

[12] LightTools - LightTools Illumination Design Software. (2015). Synopsis.
http://optics.synopsys.com/lighttools/ [Accessed 15/06/2015].

[13] Mladenović, N. and Hansen, P. (1997). Variable neighborhood search. Computers & OR, Vol(No):24(11) 1097-1100.

[14] Molina, D. Lozano, M. Sánchez, A. M. and Herrera, F. (2011). Memetic algorithms based on local search chains for large scale continuous optimization problems: MA-SSW-Chains. Soft Computing, Vol(No);15(11), 2201-2220.

[15] Moreno, I. and Sun, C.-C. (2008). Modeling the radiation pattern of leds. Opt. Express Vol(No):16(3) 1808-1819.

[16] OPTIS - OptisWorks. (2015). OPTIS.
http://www.optis-world.com [Accessed 15/06/2015].

[17] Quarteroni, A. Sacco, R. and Saleri. F. (2015). Nonlinear System and Numerical Optimization. Numerical Mathematics (Submitted). Town:Nova Science Publishers, Inc., 400 Oser Ave Suite 1600, Hauppauge NY 11788-3619, United States of America.

[18] SIST. (2012). Light and lighting - Measurement and presentation of photometric data of lamps and luminaries - Part 1: Measurement and file format EN 13032-1:2004+A1:2012. SIST.

[19] Thukral R. (2008). Introduction to a Newton-type method for solving nonlinear equations. Applied Mathematics and Computation, Vol(No):195 663-668.

EVALUATION OF PERFORMANCE AMONG EU28 NUTS 2 REGIONS – CONTEXT-DEPENDENT DEA TO CATCH-UP/DOWN EFFECT

Michaela Staníčková
VŠB-Technical University of Ostrava, Faculty of Economics, Department of European Integration
Sokolská třída 33, 701 21 Ostrava, Czech Republic
michaela.stanickova@vsb.cz

Abstract: Over the past half century, the European Union has been successful in securing high and rising living standards for their citizens. It is currently facing critical challenges due to the financial and economic crisis having impact on its territorial development and require policy responses. Paper deals with Context-Dependent Data Envelopment Analysis for evaluation the EU28 NUTS 2 regions performance. This topic is actual because performance of regions may be influenced by the context – processes as production, regionalization and subsidiarity. Results can provide policy makers with insights into competitive dis/advantages and can help with decision-making in regional policy.

Keywords: Competitive Dis/Advantages, Context Dependent DEA, EU Regions, Performance

1 INTRODUCTIN – TOPICALITY OF COMPETITIVENESS AND PERFORMANCE

The European Union (EU) is going through one of the most difficult periods since its establishment. Recent years have seen a myriad of economic and social difficulties, i.e. stagnating economic growth, rising unemployment, financial troubles and debt crises. Roots of this prolonged crisis lie in the lack of competitiveness. The EU competitiveness depends on contributions from all its areas. An asset for the EU is its rich regional diversity representing a unique set of potentials and challenges for development and targeted policy mix. Regional diversity represented by specific territorial endowment is considered as a competitive advantage of each region. Increasing competitiveness is one of the EU main aims.

Competitiveness is monitored characteristic of economies which is increasingly appearing in evaluating their performance and prosperity, welfare and living standards. The exact definition of competitiveness is difficult because of the lack of mainstream view for its understanding. Competitiveness is distinguished at different levels – micro, macro and regional, i.e. paper orientation. In the global economy regions are increasingly becoming the drivers of the economy [6]. Current economic fundamentals are threatened by shifting of production activities to places with better conditions. Regional competitiveness is affected by regionalization of public policy due to shifting of decision-making and coordination of activities at regional level. Regions play important role in the economic development of states.

Territories need highly performing units in order to meet their goals. Performance (and its two dimensions – efficiency/productivity and effectiveness, see Figure 1 [5]) is a major prerequisite for future economic and social development and success. Concept of competitiveness is closely linked with understanding of efficiency and effectiveness, competitiveness measures ‘‘how a nation manages the totality of its resources and competencies to increase the prosperity of its people’’ [4, p. 502].

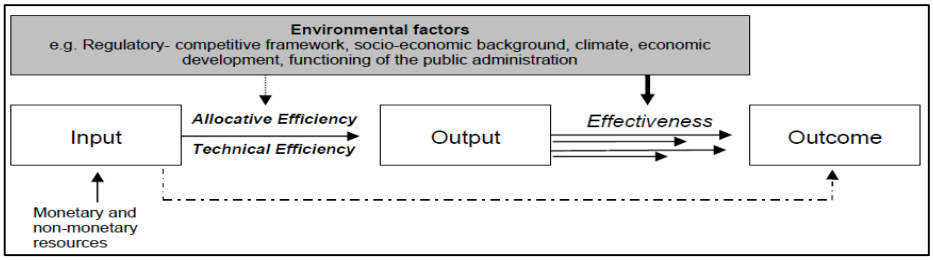


Figure 1: Relationship between efficiency and effectiveness

Policy makers are faced with evaluating relative performance of different politics, teams, units, settings of political aims and especially instruments etc. In cases where each of these performers has a set of common inputs that they utilize to produce a set of common outputs, Data Envelopment Analysis (DEA) is one of tools for efficiency analysis. DEA is often the comparison method of choice. Purpose of these comparisons is to determine the best performers along with guidelines for improving the rest. The paper' focus is in selecting the appropriate benchmarking set and understanding the implications of this choice for improving efficiency of evaluated units, i.e. the EU NUTS 2 regions based on their factor endowment identified by the EU Regional Competitiveness Index (RCI). The idea laying behind this approach is that inefficient regions can learn more easily from those, which are more similar. The contribution of this paper is thus to determine efficiency scores based on the efficient projection to the frontier along both the input and output spaces simultaneously.

2 BACKGROUND FOR EFFICIENCY ANALYSIS BY RCI AND DEA APPROACH

Efficiency analysis is based on RCI2013, resp. on its pillars (initial variables for empirical analysis) which are grouped according to different dimensions (input versus output aspects) of competitiveness they describe. Inputs and outputs describe driving forces of competitiveness, and are direct or indirect outcomes of a competitive economy [2], see Table 1. RCI seems to be convenient with respect to using DEA and its division to input and output nature of database.

Table 1: Input and output dimensions of RCI2013 pillars

Inputs	(1) Institutions, (2) Macroeconomic stability, (3) Infrastructure, (4) Health, (5) Basic education, (6) Higher education and lifelong learning, (7) Technological readiness
Outputs	(1) Labour market efficiency, (2) Market size, (3) Business sophistication, (4) Innovation

DEA originating from Farrell's work and popularized by Charnes, Cooper and Rhodes (CCR model), evaluates efficiency of a set of homogenous group (DMUs). The aim of DEA is to examine DMU into two categories – efficient and inefficient. Efficient DMUs have equivalent efficiency score, but they don't have necessarily the same performance. DMU is efficient if the observed data correspond to testing DMU on the imaginary efficient frontier. Intent of frontier estimation is to deduce the production function in form of efficient frontier. If DMUs are plotted in their input/output space, then efficient frontier that provides a tight envelope around all of DMUs can be determined. The main function of this envelope is to get as close as possible to each DMU without passing by any others. Evaluated DMUs can be divided into groups/levels according to all efficient frontiers via Context-Dependent DEA (CD-DEA). By this stratification, into efficiency analysis will enter more homogenous groups of DMUs, which will be evaluated separately according to closer features for finding relevant efficiency and inefficiency scores [6].

The first step in efficiency analysis is Returns to Scale (RTS) estimation. Why is necessary to decide RTS orientation? Various types of DEA models can be used, depending upon the problem at hand. Used DEA model can be distinguished by scale and orientation of model. If one cannot assume that economies of scale do not change, then a variable returns to scale (VRS) type of DEA model, is an appropriate choice (as opposed to a constant returns to scale, (CRS) model). If in order to achieve better efficiency, governments' priorities are to adjust their outputs (before inputs), then an output oriented (OO) DEA model, rather than an input oriented (IO) model, is appropriate. Here, OO DEA model is considered as suitable for measuring regional efficiency in links between competitiveness and efficiency. Based on RTS specification, it was possible to obtain efficient frontier, see Table 2. Calculations were made separately for group of EU15 (old) and EU13 (new) regions with respect to their integration links within the EU. EU15 were in the second step calculated by CD-DEA and efficiency scores

were obtained based on efficient levels comparison. EU13 are more homogenous group, but the distinction is not so high, because there was identified only one efficient level.

Table 2: RTS Estimation to EU NUTS 2 Regions for RCI2013 and DEA Model Results

NUTS 2	***RTS***	***Efficiency Frontier***	***DEA Model***
199 EU15	Constant	Level1 (CRS) – 199 NUTS 2 Level2 (CRS) – 6 NUTS 2	Context Dependent DEA
57 EU13	Constant	Level1 (CRS) – 57 NUTS 2	OO CCR CRS, OO APM CRS

Suppose there are *n* DMUs which consume *m* inputs to produce *s* outputs. There is a rough rule of thumb [3] which expresses the relation between the number of DMUs and the number of performance measures. Toloo et al. checked more than 40 papers that contain practical applications and statistically, they found out that in nearly all of the cases the number of inputs and outputs do not exceed 6 [7]. A simple calculation shows that when $m \leq 6$ and $s \leq 6$, then $3(m + s) \geq m \times s$. As a result, in this paper following formula (1) is applied:

$$n \geq 3(m+s). \tag{1}$$

In the paper, the rule of thumb is met – EU15: 199 ≥ 3 (7 + 4), 199 ≥ 3 (11), 199 ≥ 33. Also in the case of EU13 is rule of thumbmet, i.e. 57 ≥ 3 (7 + 4), 57 ≥ 3 (11), 57 ≥ 33.

For calculations of EU13 NUTS 2 efficiency, OO CCR CRS model is used (2) [3]:

$$\max \ \mathbf{g} = \phi_{\mathbf{q}} + \varepsilon(\mathbf{e^T s^+} + \mathbf{e^T s^-}), \tag{2}$$

subject to

$$\mathbf{X}\lambda + \mathbf{s^-} = \mathbf{x_q},$$

$$\mathbf{Y}\lambda - \mathbf{s^+} = \phi_{\mathbf{q}}\mathbf{y_q},$$

$$\lambda, \mathbf{s^+}, \mathbf{s^-} \geq \mathbf{0},$$

where g is the coefficient of efficiency of unit U_q; $\phi_{\mathbf{q}}$ is radial variable indicates required rate of increase of output; ε is infinitesimal constant; T is monotonicity which means that all inputs and outputs are freely (or strongly) disposable; e denotes the convex hull; e^T means convexity what is equivalent to decreasing marginal rates of substitution (between inputs, between outputs and between inputs and outputs); e^T is convexity condition, in the case of CRS: $e^T = (1, 1, \ldots, 1)$; s^+, and s^- are vectors of slack variables for inputs and outputs; λ represents vector of weights assigned to individual units; x_q means vector of input of unit U_q; y_q means vector of output of unit U_q; $\boldsymbol{\lambda}$ represents vector of weights assigned to variables, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, $\lambda \geq 0$; **X** is input matrix; **Y** is output matrix. In CCR model aimed at outputs coefficient of efficient DMU equals 1, but the coefficient of inefficient DMU is greater than 1.

For possibility of efficient units' classification, Andersen-Petersen's model (APM) of super efficiency is used. Following CRS model is output oriented dual version of APM (3) [1]:

$$\max \ \mathbf{g} = \phi_{\mathbf{k}} + \varepsilon\left(\sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{s_i^-} + \sum_{\mathbf{r}=1}^{\mathbf{s}} \mathbf{s_r^+}\right), \tag{3}$$

subject to

$$\sum_{\substack{\mathbf{j}=1 \\ \mathbf{j}\neq\mathbf{k}}}^{\mathbf{n}} \lambda_{\mathbf{j}}\mathbf{x_{ij}} + \mathbf{s_i^-} = \mathbf{x_{ik}},$$

$$\sum_{\substack{\mathbf{j}=1 \\ \mathbf{j}\neq\mathbf{k}}}^{\mathbf{n}} \lambda_{\mathbf{j}}\mathbf{y_{rj}} - \mathbf{s_r^+} = \phi_{\mathbf{k}}\mathbf{y_{rk}},$$

$$\lambda_{\mathbf{j}}, \mathbf{s_r^+}, \mathbf{s_i^-} \geq \mathbf{0},$$

where x_{ij} and y_{rj} are i-th inputs and r-th outputs of DMU_j; ϕ_k is efficiency index (intensity factor) of observed DMU_k; λ_j is dual weight which show DMU_j significance in definition of input-

output mix of hypothetical composite unit, DMU_k directly comparing with. The rate of efficiency of inefficient units ($\phi_k > 1$) is identical to model (1); for units identified as efficient in model (1), provides OO APM (2) the rate of super efficiency lower than 1, i.e. $\phi_k \leq 1$.

In CD-DEA approach, CRS function is used to obtain all efficient frontiers. Via Obtain Levels function, a continuous calculation, EU15 NUTS 2 regions are divided in two groups.

3 CONCLUSION

In conclusion, the main results of DEA efficiency evaluation for EU NUTS 2 regions are specified. The best results are traditionally achieved by economically powerful regions (in most cases) old and also new EU States. In Annex 1 and Annex 2, results are highlighted by traffic light method. Colour scale divides the relevant group of coefficients using three colours (shadows of grey colour), the middle colour scheme corresponds to 50 percentile, the other two colours are above (dark grey – the best results) and below (light grey – the worst results) percentile value of 50. In CD-DEA is meaning of colour the opposite.

Heterogeneity of the EU brings also differences in socio-economic position of countries and regions. It´s thus necessary to recognize specific characteristics of areas, dis/advantages for more efficient cooperation. The purpose of this paper was recognize how to determine the appropriate group exhibiting similar features for subsequent efficiency evaluation, and to identify the differences in efficiency of EU NUTS 2 regions. Obtained results by DEA approach show that targets are the coordinates of the efficient projection point on the frontier and thus represents levels of operation of inputs and outputs which would make the corresponding inefficient DMU perform efficiently. Based on results, it´s possible in the future paper to calculate best-practice units as a benchmark what improve. By calculating peer units, it´ll get results where is convenient to decrease inputs and increase outputs and what will be much more efficient combination of inputs and outputs that the region will increase the efficient frontier. Each region should know were lying its competitive dis/advantages, because there is no one-size-fits-all solution. Each region must make its own decisions about the right combination of policy objectives. DEA should be convenient tool for identifying strengths and weakness for creating these strategies and this will be orientation of future research.

Acknowledgement
The paper is supported by SGS project (SP2015/106) of Faculty of Economics, VŠB-TUO and Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296.

References

[1] Andersen, P., and N.C. Petersen. (1993). A Procedure for Ranking Efficient Units In Data Envelopment Analysis. Management Science, 39(10): 1261-1264.
[2] Annoni, P., and Dijkstra, L. (2013). EU Regional Competitiveness Index 2013. Luxembourg: Publication Office of the European Union.
[3] Cook, W.D., and Seiford, L.M. (2009). Data envelopment analysis (DEA) – Thirty years on. European Journal of Operation Research, 192(1): 1-17.
[4] IMD. (2012). IMD World Competitiveness Yearbook 2012. Lausanne: IMD.
[5] Mandl, U., Dierx, A., and Ilzkovitz, F. (2008). The effectiveness and efficiency of public spending. Brussels: EC-Directorate General for Economic and Financial Affairs.
[6] Melecký, L. and Staníčková, M. (2014). NUTS 2 regions classification: comparison of cluster analysis and DEA method. In Klímová, V. and Žítek, V. (Eds.). XVII. Mezinárodní kolokvium o regionálních vědách. (pp. 45-53). Brno: Masarykova univerzita.
[7] Toloo, M. (2012). Alternative solutions for classifying inputs and outputs in data envelopment analysis. Computers and Mathematics with Applications, 63(6): 1104-1110.

Annexes

Annex 1: Application of RTS Estimation to EU13 NUTS 2 Regions for RCI2013 and DEA Model Results

NUTS 2	Σλ	OO RTS	OO CCR CRS	OO APM CRS	Reordered NUTS 2		Rank
BG31	1,000	Constant	1,000	0,706	CZ01	1,000	1
BG32	1,000	Constant	1,000	0,836	CZ02	1,000	2
BG33	1,000	Constant	1,000	0,829	PL32	0,992	3
BG34	1,000	Constant	1,000	0,202	PL61	0,984	4
BG41	1,000	Constant	1,000	0,790	PL63	0,976	5
BG42	1,000	Constant	1,000	0,954	PL52	0,973	6
CY00	1,000	Constant	1,000	0,837	HU22	0,963	7
CZ01	1,000	Constant	1,000	1,000	CZ05	0,957	8
CZ02	1,000	Constant	1,000	1,000	PL21	0,957	9
CZ03	1,000	Constant	1,000	0,889	BG42	0,954	10
CZ04	1,000	Constant	1,000	0,843	PL51	0,954	11
CZ05	1,000	Constant	1,000	0,957	CZ06	0,950	12
CZ06	1,000	Constant	1,000	0,950	PL42	0,950	13
CZ07	1,000	Constant	1,000	0,923	PL41	0,946	14
CZ08	1,000	Constant	1,000	0,932	PL33	0,945	15
EE00	1,000	Constant	1,000	0,834	RO12	0,944	16
HR03	1,000	Constant	1,000	0,926	SK02	0,944	17
HR04	1,000	Constant	1,000	0,808	SK03	0,942	18
HU10	1,000	Constant	1,000	0,819	SI01	0,941	19
HU21	1,000	Constant	1,000	0,893	PL62	0,937	20
HU22	1,000	Constant	1,000	0,963	RO41	0,935	21
HU23	1,000	Constant	1,000	0,853	CZ08	0,932	22
HU31	1,000	Constant	1,000	0,824	HR03	0,926	23
HU32	1,000	Constant	1,000	0,775	CZ07	0,923	24
HU33	1,000	Constant	1,000	0,690	PL31	0,921	25
LV00	1,000	Constant	1,000	0,814	RO22	0,921	26
MT00	1,000	Constant	1,000	0,713	PL43	0,918	27
PL11	1,000	Constant	1,000	0,911	PL34	0,916	28
PL12	1,000	Constant	1,000	0,781	SK04	0,916	29
PL21	1,000	Constant	1,000	0,957	PL11	0,911	30
PL22	1,000	Constant	1,000	0,898	PL22	0,898	31
PL31	1,000	Constant	1,000	0,921	HU21	0,893	32
PL32	1,000	Constant	1,000	0,992	CZ03	0,889	33
PL33	1,000	Constant	1,000	0,945	RO11	0,873	34
PL34	1,000	Constant	1,000	0,916	SI02	0,867	35
PL41	1,000	Constant	1,000	0,946	RO31	0,856	36
PL42	1,000	Constant	1,000	0,950	HU23	0,853	37
PL43	1,000	Constant	1,000	0,918	CZ04	0,843	38
PL51	1,000	Constant	1,000	0,954	CY00	0,837	39
PL52	1,000	Constant	1,000	0,973	BG32	0,836	40
PL61	1,000	Constant	1,000	0,984	EE00	0,834	41
PL62	1,000	Constant	1,000	0,937	BG33	0,829	42
PL63	1,000	Constant	1,000	0,976	HU31	0,824	43
RO11	1,000	Constant	1,000	0,873	HU10	0,819	44
RO12	1,000	Constant	1,000	0,944	LV00	0,814	45
RO21	1,000	Constant	1,000	0,758	HR04	0,808	46
RO22	1,000	Constant	1,000	0,921	RO42	0,802	47
RO31	1,000	Constant	1,000	0,856	BG41	0,790	48
RO32	1,000	Constant	1,000	0,578	PL12	0,781	49
RO41	1,000	Constant	1,000	0,935	SK01	0,776	50
RO42	1,000	Constant	1,000	0,802	HU32	0,775	51
SI01	1,000	Constant	1,000	0,941	RO21	0,758	52
SI02	1,000	Constant	1,000	0,867	MT00	0,713	53
SK01	1,000	Constant	1,000	0,776	BG31	0,706	54
SK02	1,000	Constant	1,000	0,944	HU33	0,690	55
SK03	1,000	Constant	1,000	0,942	RO32	0,578	56
SK04	1,000	Constant	1,000	0,916	BG34	0,202	57

Annex 2: Application of RTS Estimation to EU13 NUTS 2 Regions for RCI2013 and DEA Model Results

NUTS 2	CDS[1]	NUTS 2	CDS[1]	NUTS 2	CDS[1]	NUTS 2	CDS[1]	NUTS 2	CDS[1]	NUTS 2	CDS[1]	NUTS 2	CDS[2]
AT11	0,670	DE72	0,829	FR22	0,727	IE01	0,539	PT15	0,607	UKJ2	0,760	ES13	1,000
AT12	0,724	DE73	0,801	FR23	0,755	IE02	0,555	PT16	0,651	UKJ3	0,751	ES23	1,003
AT13	0,724	DE80	0,791	FR24	0,717	ITC1	0,697	PT17	0,567	UKJ4	0,773	ES42	1,002
AT21	0,658	DE91	0,613	FR25	0,589	ITC2	0,522	PT18	0,442	UKK1	0,801	ES61	1,011
AT22	0,640	DE92	0,795	FR26	0,676	ITC3	0,715	PT20	0,529	UKK2	0,786	FR42	1,015
AT31	0,691	DE93	0,788	FR30	0,705	ITC4	0,707	PT30	0,482	UKK3	0,618	UKL1	1,015
AT32	0,605	DE94	0,654	FR41	0,770	ITD1	0,549	SE11	0,625	UKK4	0,741		
AT33	0,628	DK01	0,627	FR42	0,718	ITD2	0,711	SE12	0,650	UKL2	0,731		
AT34	0,680	DK02	0,695	FR43	0,739	ITD3	0,699	SE21	0,711	UKM2	0,699		
BE10	0,663	DK03	0,697	FR51	0,512	ITD4	0,697	SE22	0,633	UKM3	0,805		
BE21	0,715	DK04	0,647	FR52	0,672	ITD5	0,735	SE23	0,673	UKM5	0,578		
BE22	0,743	DK05	0,717	FR53	0,666	ITE1	0,739	SE31	0,646	UKM6	0,618		
BE23	0,664	ES11	0,811	FR61	0,646	ITE2	0,758	SE32	0,652	UKN0	0,659		
BE24	0,663	ES12	0,805	FR62	0,631	ITE3	0,758	SE33	0,609				
BE25	0,589	ES13	0,796	FR63	0,758	ITE4	0,550	UKC1	0,700				
BE31	0,663	ES21	0,819	FR71	0,715	ITF1	0,784	UKC2	0,819				
BE32	0,667	ES22	0,778	FR72	0,723	ITF2	0,759	UKD1	0,628				
BE33	0,584	ES23	0,697	FR81	0,744	ITF3	0,575	UKD2	0,684				
BE34	0,342	ES24	0,837	FR82	0,559	ITF4	0,667	UKD3	0,735				
BE35	0,487	ES30	0,821	FR83	0,647	ITF5	0,735	UKD4	0,755				
DE11	0,702	ES41	0,792	FR91	0,588	ITF6	0,589	UKD5	0,801				
DE12	0,715	ES42	0,831	FR92	0,427	ITG1	0,651	UKE1	0,809				
DE13	0,758	ES43	0,769	FR93	0,669	ITG2	0,724	UKE2	0,714				
DE14	0,701	ES51	0,754	FR94	0,634	LU00	0,597	UKE3	0,789				
DE21	0,683	ES52	0,638	GR11	0,724	NL11	0,620	UKE4	0,705				
DE22	0,688	ES53	0,632	GR12	0,614	NL12	0,755	UKF1	0,737				
DE23	0,671	ES61	0,596	GR13	0,750	NL13	0,780	UKF2	0,836				
DE24	0,746	ES62	0,540	GR14	0,552	NL21	0,739	UKF3	0,838				
DE25	0,663	ES63	0,603	GR21	0,618	NL22	0,812	UKG1	0,858				
DE26	0,818	ES64	0,621	GR22	0,418	NL23	0,761	UKG2	0,758				
DE27	0,755	ES70	0,461	GR23	0,417	NL31	0,726	UKG3	0,680				
DE30	0,792	FI13	0,696	GR24	0,360	NL32	0,761	UKH1	0,719				
DE41	0,792	FI18	0,732	GR25	0,692	NL33	0,803	UKH2	0,680				
DE42	0,792	FI19	0,829	GR30	0,716	NL34	0,739	UKH3	0,680				
DE50	0,762	FI20	0,801	GR41	0,563	NL41	0,762	UKI1	0,680				
DE60	0,692	FR10	0,791	GR42	0,586	NL42	0,723	UKI2	0,680				
DE71	0,696	FR21	0,613	GR43	0,727	PT11	0,693	UKJ1	0,657				

Note: CDS[1] = Context-Dependent Score for OO Level2 (CRS) Attractiveness; CDS[2] = Context-Dependent Score for OO Level1 (CRS) Attractiveness

A HYBRID METAHEURISTIC APPROACH FOR PRODUCTION SCHEDULING

Ozen Yavas and Ayca Altay
Istanbul Technical University, Industrial Engineering Department
ITU Isletme Fakultesi, Macka, Istanbul, 34357
{oznyavas, altaya}@gmail.com

Abstract: Production scheduling is crucial means of contributing a company to meet the demands in a timely manner, given the resources and constraining circumstances of the manufacturing environment. This problem has been dealt with numerous methods for approximately five decades. Since solving production scheduling problems with exact methods is computationally expensive, metaheuristics are frequently used where an exploration-exploitation balance has been the main objective. This study introduces a novel hybrid approach that utilizes Genetic Algorithms (GA) and Simulated Annealing (SA). A real-world application is utilized as a test problem for comparing these algorithms with the novel hybrid approach. Results indicate that the algorithm is promising for scheduling problems; however, being computationally expensive compared to classical approaches.

Keywords: Production Scheduling, Genetic Algorithm, Simulated Annealing

1 INTRODUCTION AND RELATED LITERATURE

Production scheduling is defined as the assignment and allocation of available production related resources in a way that satisfies predefined production related criteria set [1]. The aim of a production schedule is to meet all possible deliveries with respect to economic, time-related or other factor considerations [2]. It is a crucial problem for industries as it is one of the main costly activities in manufacturing industries. The production scheduling problem is generally known to be NP-hard depending on the problem structure [3]. Early production scheduling literature focuses on mathematical programming techniques and heuristics, whereas latter and recent approaches lean towards hybrid methods of heuristics and metaheuristics[4].

For this study, the literature was examined from 1980's up to 2015. During the years between 1980 and 2000, it is observed that the local search and global search algorithms were in use for solving the production scheduling problems, and the solutions found to the similar scheduling problems were compared with each other [5-8].Ulusoy et al. [5] introduced a Genetic Algorithm (GA) for the problem of simultaneous scheduling of machines and Automated Guided Vehicles(AGVs). The originality of this study provided the machines and the AGVs to be combined and randomly generalized using the same chromosome. Murata et al. [6] applied GA for multi objective production scheduling problem, whose objectives include the makespan, the total tardiness, and the total flowtime minimization. Eiben et al. [7] applied GA to the classical job shop scheduling problem by introducing more than 2 parents for the crossover procedure and proved that multi parent recombination gives more robust results than conventional crossover procedure. Shi et al. [8] used a GA to a classical job shop scheduling problem which utilised 2/4 selection method for the selection procedure.

During the years between 2000 and 2010, it is observed that hybridisation of global and local search algorithms were commonly applied as well as Matheuristics which yielded more efficient solutions. For the mathematical modelling of the problems, the scheduling rules such as Shortest Setup Time (SST), Shortest Processing Time (SPT), and Earliest Release Date (ERD) were in use. Ip et al. [9] considered the problem of minimum earliness and tardiness in production scheduling and applied GA for the solution. Guo et al. [10] solved the job shop scheduling problem by using GA. They utilised tournament selection method for selection and fitness value of the objective functions. Li et al. [11] introduced a hybrid algorithm which was

the combination of Genetic Algorithm and Tabu Search Algorithm applied to a multi objective flexible job shop scheduling problem.

During the years between 2010 and 2015, it is found that the use of hybrid algorithms were very common and the new operators and methods are being developed for different stages of GA. Lin et al. [12] considered an unrelated parallel machine scheduling problem. They inserted different heuristics into GA and compared the best solutions found. They encoded the chromosome by using the method developed by Cheng et al in 1995. In the population initialisation stage, they combined eight different heuristics. Balin [13] considered the problem of non-permutational parallel machine scheduling, and applied GA. Liu et al. [14] depicted the scheduling problem of a multi product, two stages production environment and applied GA with a Greedy 3PM Crossover operator. Dalfard and Mohammadi [15] used the hybrid algorithm of GA and SA for the flexible job shop scheduling problems. Zhang et al. [16] solved a job-shop scheduling problem with the GA and the Tabu Search (TS) hybrid.

In this study, the production scheduling of a small size company in Turkey, which consists of a multi product multi machine production line, is solved with a hybrid GA-SA algorithm. The originality of this study lies in the hybridization structure of GA and SA. The results of the hybrid algorithm are compared to algorithms themselves. The main objective of the hybridisation is to improve the results obtained by applying the GA and SA itself.

The organization of this paper is as follows: Section 2 introduces the methodology and the algorithms. Section 3 gives the mathematical model of this problem and the experimental results of the proposed algorithms. Section 4 concludes the paper.

2 PROBLEM STATEMENT AND THE MATHEMATICAL MODEL

2.1 Production Environment

In this work, in order to investigate the effectiveness of the proposed hybrid algorithm, the trials were done based on the real-world production data of a circuit board producing company in Turkey which produces 40 products. Two different technologies are used for the assembly process and they are named as Line1 and Line 2. Line1 includes two identical assembly sub-lines which are Line1.1 and Line1.2 namely. Each of these lines (Line1.1 and Line1.2) consists of an assembly machine of type 1, two different string machines (namely, machine 1 and machine 2) and an oven. On the other hand, Line2 consists of only assembly machines of type 2. The production process starts with any of the sub-line of Line1 and then it follows to Line2. After being processed consecutively in Line1 and Line2, the production process is complete. The production line diagram of this company is given below.

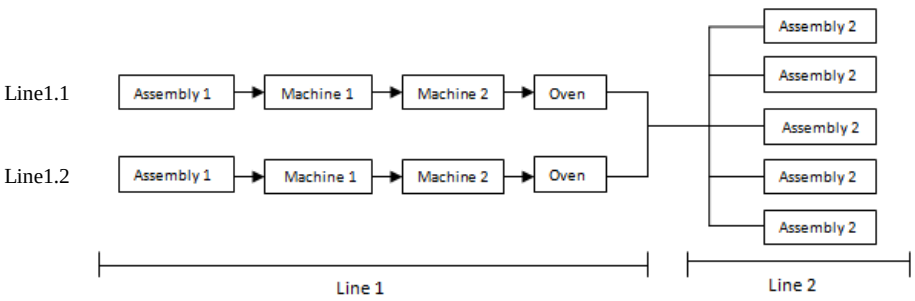


*Figure 1.*The production line diagram of the described production system.

2.2 Mathematical Model

The mathematical model of this production system can be constructed as a Mixed Integer Linear Programming (MILP) as follows.

$N: the\ number\ of\ lines\ (N = 2)$
$i: index\ of\ the\ lines\ (i = 1, N)$
$M: the\ number\ of\ machines\ in\ line\ 1\ (for\ i = 1\ , M = 2)$
$L: the\ number\ of\ machines\ in\ line\ 2\ (for\ i = 2\ , L = 5)$
$j: the\ number\ of\ machines\ in\ a\ line\ (for\ i = 1\ , j = 1, M)(for\ i = 2\ , j = 1, 2, \dots, L)$
$T: total\ number\ of\ jobs\ (circuit\ boards)\ (T = 40)$
$k: indice\ of\ the\ jobs\ (circuit\ boards)\ (k = 1, 2, \dots, T)$
$x_{ijk}: the\ k^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$x_{ijk} = \begin{cases} 1, & if\ k^{th}\ job\ is\ being\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line \\ 0, & otherwise \end{cases}$
$t_{ijk}: the\ starting\ time\ of\ the\ k^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$p_{ijk}: the\ \ processing\ time\ of\ the\ k^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$s_{ijk}: the\ machine\ setup\ time\ of\ \ the\ k^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$C_{ijk}: the\ completion\ time\ of\ the\ k^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$t_{ijk'}: the\ starting\ time\ of\ the\ k'^{th}\ job\ processed\ on\ the\ j^{th}\ machine\ of\ i^{th}\ line$
$y_{ijkk'} = \begin{cases} 1, & if\ the\ \ k'^{th}\ job\ is\ being\ processed\ on\ the\ j^{th}\ machine\ after\ the\ \ k^{th}\ job \\ 0, & otherwise \end{cases}$

The objective function is given below:

$$min(maxC_{ijk}) \tag{0}$$

Each job can be produced on only one machine.

$$\sum_{j=1}^{2} \mathrm{x_{1jk}} = 1\ ,\ \forall \mathrm{k}\ \ \text{and}\ \sum_{j=1}^{5} \mathrm{x_{2jk}} = 1\ ,\ \forall \mathrm{k} \tag{1}$$

The sum of starting, machine setup, and processing times cannot exceed the completion time.

$$t_{ijk} + p_{ijk} + s_{ijk} \leq c_{ijk}\ ,\quad \forall i, j, k \tag{2}$$

Two jobs cannot be processed on a single machine at the same time.

$$t_{ijk} + p_{ijk} + s_{ijk} \leq t_{ijk'} + M.(1 - y_{ijkk'})\ ,\quad \forall i, j, k \tag{3}$$
$$x_{ijk'} + p_{ijk'} + s_{ijk'} \leq t_{ijk} + M.y_{ijkk'}\ ,\quad \forall i, j, k$$

A job is processed on the machine that it was assigned.

$$x_{ijk} = 1\ ise\ t_{ijk} \geq 0 \tag{4}$$
$$x_{ijk} = 0\ ise\ t_{ijk} = 0$$
$$M.x_{ijk} \geq t_{ijk}$$

The values of the variables and parameters must be more than or equal to zero.

$$x_{ijk} \geq 0, t_{ijk} \geq 0, s_{ijk} \geq 0, p_{ijk} \geq 0, y_{ijk} \geq 0 \tag{5}$$

3 HYBRID GA-SA ALGORITHM

The algorithm utilised in this study is a hybrid GA-SA algorithm whose steps are explained below and described in Figure 1.

Step 1. The chromosomes are randomly generated by permutation and the initial population is generated randomly.
Step 2. The objective functions of all the chromosomes are calculated one by one.
Step 3. Improvement on the least fittest chromosomes: *n*% of the chromosomes with the worst objective function values are selected and subjected to an SA algorithm, where *n* is a predetermined algorithm parameter. In this way, fitness values of the population are improved and highly fitted chromosomes are subjected to selection procedure in the next step. With this novel step, chromosome diversity is provided and local optima are avoided at a higher rate than a classical GA. This step does not eliminate the worst chromosomes; it conducts a local search which still saves the diversity among solutions.
Step 3.1. The initial solution is provided by the GA and it is one of the worst *n*% chromosomes.
Step 3.2. A new random neighbour is selected as in the mutation form of the GA.
Step 3.3. The objective function value of the neighbour is calculated.
Step 3.4. If the objective function value of the neighbour has a lower value, the neighbour is accepted. Otherwise, the probability of acceptance is calculated as in

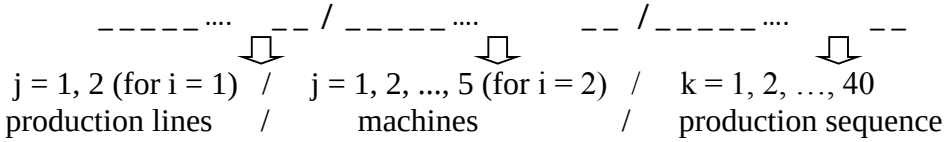
$$p_a = e^{-(neighbobj - objective)/t_{iteration}}$$

where p_a is the probability of acceptance, *neighbobj* is the objective function of the neighbourhood solution, *objective* is the objective function of the current solution and $t_{iteration}$ is the temperature of the related iteration.
Step 3.5. A random number *rn* is generated. If $rn < p_a$, then the neighbourhood solution is accepted as the new solution. If $rn > p_a$, then the current solution is revisited.
Step 3.6. The temperature is dropped as in the formula

$$t_{iteration+1} = \alpha \cdot t_{iteration}$$

Where α is the cooling rate.
Step 3.7. Steps 3.2 – 3.6 are returned until the temperature of the iteration is lower than a predetermined final temperature value.
Step 4. *Selection:* Then, the roulette wheel selection method is applied and selection is made. Since this is a minimization problem, the inverses of the objective function values (1/*f*(*x*)) are generating.
Step 5. *Crossover:* After the selection procedure is completed, crossover phase is applied.
Step 6. *Mutation:* A random mutation is achieved on genes with a predetermined mutation probability. In this process, two different mutation methods were used.
Step 7. Steps 2- 6 are returned until a stopping criterion is reached.

The adaptation of GA and SA hybrid algorithm is achieved as follows: the chromosomes are encoded in three parts. The first part indicates the index of the sub-line of the jobs that are assigned to Line1, while the second part indicates the index of the sub-line of jobs that are assigned to Line2 after being processed in Line1. The third part indicates the production sequence of the products.



Assume that one of the chromosomes has the values given below;

1 2 2 1 2 … 2 1 / 3 5 2 4 2 … 1 5 / 8 3 2 1 7 1 2 9 6 … 7 1 0

from the third part, it is observed that the 8th job is processed first. It is processed on the 1st machine of the first production line as the first gene of the first part of the chromosome states, and then it is processed on the 3rdmachine in the second production line as the first gene of the second part of the chromosome states.

4 APPLICATION & RESULTS

This structure also affects the crossover and mutation scheme of the GA. For the first and second groups of the chromosome, one point crossover operator was used since the genes can repeat themselves. For the third group of the chromosome, circular crossover operator was used where repetition of genes are not allowed. For mutation, the first two parts of the chromosome utilise the gene exchange method whereas gene conversion method is used for the last part. For detailed explanations of these methods, the reader may refer to [17].

The parameters of the algorithm that provide the best result are given below. These combinations are obtained by trial-and-error.

Table 1: The parameters of the algorithm.

Parameter	***Value***
Population size	10
Crossover probability	0.95
Mutation probability	0.05
Initial temperature	1000
Cooling rate	0.90
Final temperature	0.01
n%	50

4.1 Improvements & Trials

The investigated scheduling problem of the production system was solved by SA algorithm, GA, and the Hybrid GA/SA Algorithm on MATLAB program with aforementioned parameter values. Each of the algorithms was solved for 30 trials. Then, the best results were taken, and the mean value, standard deviation, best result, worst result, and CPU time were calculated and given in a table form. The finishing criterion is selected as a predetermined number of iterations by the algorithm.

Table 2: The results of the SA, GA and the hybrid GA-SA algorithm.

30 trials	***SA***	***GA***	***Hybrid GA-SA***
Mean value	11378,19	9840,893	9817,272
Std. Deviation	944,2655	4,849342	5,504509
Best Result	9912,033	9830,033	9803,3
Worst Result	14494,73	9849,333	9828
CPU time	0,529	1,926	204,759

As can be observed from the results, as a local search algorithm, SA alone does a local search but its exploration ability is not good enough as GA. A simple t-test between GA and the hybrid GA-SA suggests that the results of the hybrid algorithm are extremely statistically lower, thus, better than the GA algorithm alone. However, one drawback of this algorithm is that it takes more computational time.

5 CONCLUSIONS

In this study, a novel hybrid GA-SA algorithm is proposed for production scheduling. The results suggest that the algorithms results are statistically better, yet, it requires more execution time. Further studies include different hybridization of heuristics and metaheuristics in order to achieve more exploration, yet, mainly for achieving a much lower execution time.

References

[1] Graves, S. C., (1981). A review of production scheduling. Operations Research, 29(4):646-675.

[2] Herrmann, J. W. (2006). Improving production scheduling: integrating organizational, decision-making, and problem-solving perspectives. IIE Annual Conference. Proceedings, 1-6.

[3] El Raheb, K., Kiranoudis, C. T., Repoussis, P. P., and Tarantilis, C. D. (2005). Production scheduling with complex precedence constraints in parallel machines. Computing and Informatics, 24(3): 297-319.

[4] Jones, B. (2013). Identifying Real Cost Saving in LeanManufacturing. Purdue University,Aviation Technology Graduate StudentPublications. Paper 30, California.

[5] Ulusoy, G., Sivrikaya-Serifoglu, F. and Bilge, U. (1997). A genetic algorithm approach to the simultaneous scheduling of machines and automated guided vehicles. Computers & Operations Research, 24(4):335-351.

[6] Murata, T., Ishibuchi, H. And Tanaka, H. (1996). Multi-objective genetic algorithm and its applications to flowshop scheduling. Computers & Industrial Engineering, 30(4): 957–968.

[7] Eiben, A. E., Raue, P-E., Ruttkay, Z. (1996). Genetic algorithms with multi-parent recombination. Parallel Problem Solving from Nature — PPSN III, Lecture Notes in Computer Science, 886: 78-87.

[8] Shi, G. (1997). A genetic algorithm applied to a classic job-shop scheduling problem. International Journal of Systems Science, 28(1): 25–32.

[9] Ip, W.H., Li, Y., Man, K.F. and Tang,K.S. (2000). Multi-product planning and scheduling using genetic algorithm approach. Computers & Industrial Engineering, 38(2): 283–296.

[10] Guo, Z.X., Wong, W.K., Leung, S.Y.S., Fan, J.T., Chan, S.F.(2006). Mathematical model and genetic optimization for the job shop scheduling problem in a mixed- andmulti-product assembly environment: A case study based on the apparel industry. Computers & Industrial Engineering, 50(3): 202-219.

[11] Li, J., Pan, Q., Liang, Y-C. (2010). An effective hybrid tabu search algorithm for multi-objective flexible job-shopscheduling problems. Computers & Industrial Engineering, 59(4): 647–662.

[12] Lin, Y.K., Pfund, M.E., Fowler, J.W. (2011). Heuristics for minimizing regular performance measures in unrelated parallel machine scheduling problems. Computers & Operations Research, 38(6): 901-916.

[13] Balin, S. (2011). Non-identical parallel machine scheduling using genetic algorithm. Expert Systems with Applications, 38(6): 6814-6821.

[14] Liu, Y., Chai, T., Qin, S. J., Pan, Q., and Yang, S. (2012). Improved genetic algorithm for magnetic material two-stage multi-product production scheduling: a case study.51st IEEE Conference on Decision and Control. 2521-2526,Maui, Hawaii, USA.

[15] Dalfard, V. M., Mohammadi, G. (2012). Two meta-heuristic algorithms for solving multi-objective flexible job-shop scheduling with parallel machine and maintenance constraints. Computers and Mathematics with Applications, 64(6): 2111-2117.

[16] Zhang, Q., Manier, H., Manier, M-A. (2012). A genetic algorithm with Tabu Search procedure for flexible job shop scheduling with transportation constraints and bounded processing times. Computers & Operations Research, 39(7): 1713-1723.

[17] Engelbrecht, A.P.(2007) Computational Intelligence: An Introduction. Wiley Press.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 6:

Finance and Investments

ANALYSIS OF THE LEVERAGE EFFECT – EVIDENCE FROM LJUBLJANA STOCK EXCHANGE

Josip Arnerić
University of Zagreb, Faculty of Economics and Business Zagreb
Trg J. F. Kennedyja 6, Zagreb, Croatia
jarneric@efzg.hr

Mirjana Čižmešija
University of Zagreb, Faculty of Economics and Business Zagreb
Trg J. F. Kennedyja 6, Zagreb, Croatia
mcizmesija@efzg.hr

Petar Sorić
University of Zagreb, Faculty of Economics and Business Zagreb
Trg J. F. Kennedyja 6, Zagreb, Croatia
psoric@efzg.hr

Abstract: This paper studies GARCH models that can absorb the leverage effect, i.e. impacts of negative and positive shocks to conditional variance. Different functional forms of asymmetric GARCH models are analyzed on the daily Slovenian stock market index returns. The goal of this paper is to examine how different GARCH models reflect on asymmetric responses to bad and good news, i.e. different shapes of news impact curve (NIC) can be determined. The rotation of NIC is examined in the crisis and post-crisis period. The type of a GARCH model that fits data the best is chosen according to various information criteria.

Keywords: news impact curve, leverage effect, asymmetric GARCH model, rotation parameter, SBI stock market index, LR test.

1 INTRODUCTION

Because of their ability to capture the main stylized facts observed in stock market data (autocorrelation of the squared returns (time varying conditional variance, i.e. heteroscedasticity), fat-tails (leptokurtosis), asymmetric effects of high frequency data (leverage effects) and volatility clustering); GARCH type models are the most commonly used volatility-forecasting tool [5],[2]. These models were continuously improved during time. Two empirical findings determined the improvement of GARCH models: stock returns are strongly asymmetric and fat-tailed. However, standard ARCH/GARCH type models usually overestimate the true volatility when good news accrue and underestimate the true volatility when bad news accrue. It is assumed that good news are followed by significantly lower variance of returns, while bad news are followed by significantly higher variance of returns. Different functional forms of asymmetric GARCH type models determine different shapes of the news impact curve (NIC). According to [1], bad news tend to drive down the stock price, thus increasing the stock leverage (i.e. debt-equity ratio) and causing the stock to be more volatile. Based on this conjecture, the asymmetric news impact is usually referred to as the leverage effect. NIC usually follows a U-shaped pattern but it is not clear which of the proposed asymmetric GARCH type models (such as TGARCH, GJR-GARCH, EGARCH and APARCH) best describe the leverage effect. According to [6], NIC is defined as the functional relationship between conditional variance at time t and the shock term (error term or innovation) at time $t-1$, holding constant the information dated $t-2$ and earlier and with all lagged variances evaluated at the level of the unconditional variance.

Therefore, this paper studies various functional forms of asymmetric GARCH type models which arise from a more general family of GARCH models [8]. This model allows

different shifts and rotations in the NIC. The main objective of this paper is to examine if the rotation is the main source of asymmetry and which asymmetric GARCH type model best describes the leverage effect. In addition, we compare NIC in the crisis and post-crisis period to conclude if the rotation parameter changes between two periods. The present paper contributes to the existing literature as it covers the most recent period and offers a novel new approach to estimating asymmetric GARCH type models.

2 LITERATURE REVIEW

One of the very useful tools in analyzing and discussing stock returns asymmetry is the NIC, introduced by [10] and named by [6]. This curve measures the effect of past return shocks on stock return volatility. It is of interest here to explore how new information are included in volatility estimates. The finding that positive and negative shocks have different impacts on dynamic volatility has led to asymmetric specifications of the GARCH model. The most popular asymmetric GARCH models are EGARCH [9], Quadratic GARCH, TGARCH model [12] and the Glosten, Jagannathan and Runkle (GJR) model [7]. All of these asymmetric GARCH models are based on the premise that negative shocks introduce more volatility than the positive ones. Also, all these models permit shifts in the NIC or the rotation of the curve. Later on new diagnostic tests were presented that emphasize the asymmetry of the volatility response to news [6]. Several empirical studies have emphasized the relevance of the signs of the shocks on conditional volatility, but no particular attention was given to the sizes of these shocks. [3] provide a different solution, including a size effect in a special type of a threshold GARCH model which generalizes GJR to allow for multiple thresholds. In the class of GARCH models, they used the Dynamic Asymmetric GARCH model, which enables an analysis of the persistence of the size and sign effects of shocks by adding dynamics to volatility asymmetry. One of the major contributions in that area was also given by [4], who find that moderately good news reduces volatility, while both very good news and bad news increase volatility, with the latter having a more severe impact. Recent studies specifically focus on the impact of financial crisis on modeling the leverage effect. For example, [11] obtained somewhat surprising results for the pre-crisis data, finding that the NIC monotonically decreases for positive lagged residuals.

3 METHODOLOGY AND DATA

According to [8], the family of GARCH models is most easily derived from the asymmetric absolute value GARCH model. He applies a Box-Cox transformation to the conditional standard deviation and allows different powers of the transformed shocks (innovations):

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = u_t \sigma_t \ , \quad u_t \sim i.i.d. \tag{1}$$

$$\sigma_t^{\lambda} = \omega + \alpha_1 \sigma_{t-1}^{\lambda} \left(\left| \varepsilon_{t-1} - a \right| - b\left(\varepsilon_{t-1} - a \right) \right)^{\delta} + \beta_1 \sigma_{t-1}^{\lambda} \tag{2}$$

Relation (1) is the mean equation that describes the observed returns r_t as a function of other variables plus a shock (innovation ε_t). It is assumed that innovations ε_t have multiplicative structure of *i.i.d* random variables u_t. Relation (2) is the "variance" equation which specifies the evaluation of the Box-Cox transformation for the conditional standard deviation σ_t^{λ}, determined by parameter λ. It easy to notice that, when $\lambda = 2$, the equation (2) specifies the evaluation of the conditional variance σ_t^2. The mean equation will include only the constant

term μ and the same mean equation will be used in the comparison of different asymmetric GARCH models that arise from equation (2). It is also necessary to emphasize that only one time lag is included in equation (2) due to simplicity and sufficiency of GARCH(1,1) to capture all statistical properties of the innovation process.

The asymmetry in model (1-2) is introduced in equation (2) by shifting and rotating transformed absolute function of innovations:

$$f^{\delta}(\varepsilon_t) = \left(|\varepsilon_{t-1} - a| - b(\varepsilon_{t-1} - a)\right)^{\delta} \tag{3}$$

where δ is the power parameter of absolute value function $f(\cdot)$, a is the shift parameter and b is the rotation parameter. In this paper, the shift parameter is fixed to zero ($a = 0$) while the rotation parameter b is to be estimated because we believe that the main source of asymmetry is the rotation of the NIC. Moreover, by fixing the value of the shift parameter to zero, the absolute value function of shocks is centered on zero. Negative values of shocks (innovations) represent bad news from the previous day while positive values of shocks (innovations) represent good news from the previous day.

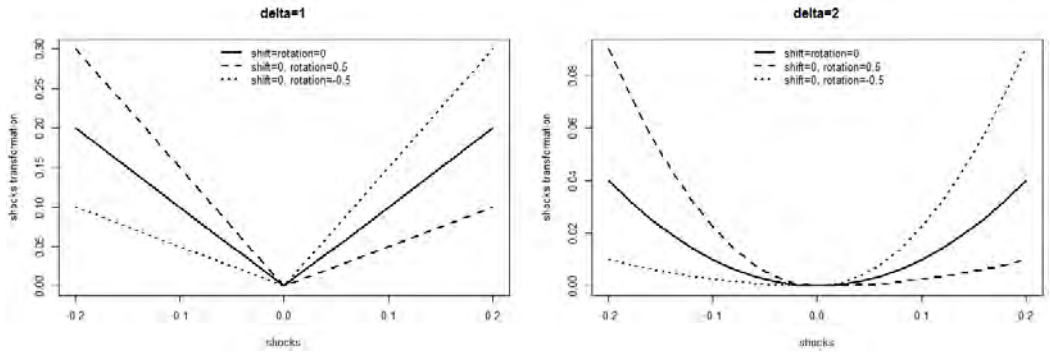


Figure 1: The effect of shift and rotation parameters on the absolute value function

Figure 1 shows that the absolute value function of shocks (innovations) is a convex function if $\delta \geq 1$, as it should be according to its expected U-shape. In both cases ($\delta = 1$ and $\delta = 2$), the absolute value function rotates to the right if the shape parameter is positive ($b > 0$) (indicating the presence of the leverage effect) and it rotates to the left if the shape parameter is negative ($b < 0$) (indicating no presence of the leverage effect). When both the shift and rotation parameters are equal to zero, the absolute value function is symmetric, i.e. past positive and negative innovations have the same effect on the transformed conditional standard deviation. Some or all of the model parameters (μ, ω, λ, δ ,α_1, b and β_1) are usually estimated by maximizing the log-likelihood function computed as the natural logarithm of the product of the conditional densities of the innovations. However, special cases of the existing asymmetric GARCH type models can be specified by appropriate restrictions on the above-mentioned parameters.

The dataset analysed in this paper consists of SBI stock market index returns. The sample period is from January 01, 2007 to September 13, 2013, i.e. total of 1731 daily observations according to the number of trading days (left panel of Figure 2). On the right panel of Figure 2, adjusted returns are presented, obtained by the Hodrick-Prescott (H-P) filter (the size of the penalty parameter is adjusted according to the number of trading days, i.e. 5 days of the week). H-P filtered returns are used to divide the entire sample into the crisis period and post-crisis period due to the stability/instability of the market movements on Ljubljana Stock Exchange. It is easy to notice that the market became stable after February 02, 2010.

Therefore, the crisis period includes 813 daily observations, while the post-crisis period includes 918 daily observations.

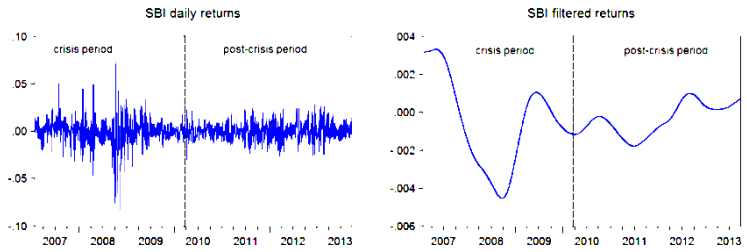


Figure 2: SBI daily returns

4 EMPIRICAL RESULTS

Four types of asymmetric GARCH(1,1) models are estimated in the crisis period and post-crisis period. The quasi-maximum likelihood QML estimates are obtained using the Broyden, Fletcher, Goldfarb and Shanno (BFGS) numerical optimization technique in R software. Additionally, innovations are assumed to follow the skewed-Student distribution due to the existence of fat tails and asymmetry of the distribution of returns. Two parameters are estimated together with other unknowns: the scale parameter, i.e. degrees of freedom df , and the shape parameter s . News impact curves are given in Figure 3 and Figure 4.

Table 1: Asymmetric sub-models estimated assuming skewed t-distribution in the crisis period

	TGARCH (1,1)		GJR-GARCH (1,1)		APARCH (1,1)		EGARCH (1,1)	
μ	-0.001*	-0.001	-0.000	-0.000	-0.001	-0.000	-0.000***	-0.001*
ω	0.001***	0.001***	0.000***	0.000***	0.000	0.000	-0.710***	-0.727***
λ	1 [a]	1 [a]	2 [a]	2 [a]	1.576***	1.870***	0 [a]	0 [a]
δ	1 [a]	1 [a]	2 [a]	2 [a]	λ	λ	1 [a]	1 [a]
α_1	0.257***	0.276***	0.276***	0.298***	0.275***	0.299***	0.460***	0.507***
β_1	0.716***	0.711***	0.670***	0.670***	0.691	0.676***	0.919***	0.918***
a	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]
b	0.223***	0 [a]	0.150**	0 [a]	0.175**	0 [a]	0.081**	0 [a]
df	11.030***	8.653***	10.201***	8.575***	10.648***	8.624***	9.577***	8.076***
s	0.883***	0.849***	0.890***	0.865***	0.888***	0.863***	0.869***	0.843***
AIC	-6.007	-5.998	-6.009	-6.004	-6.008	-6.001	-6.008	-6.001
BIC	-5.966	-5.963	-5.969	-5.969	-5.961	-5.961	-5.967	-5.967
LR-	0.002		0.011		0.007		0.008	

Note: * (**, ***) represent significance at the 1% (5%, 10%) level; *a* denotes pre-fixed parameters; LR test entries represent p-values.

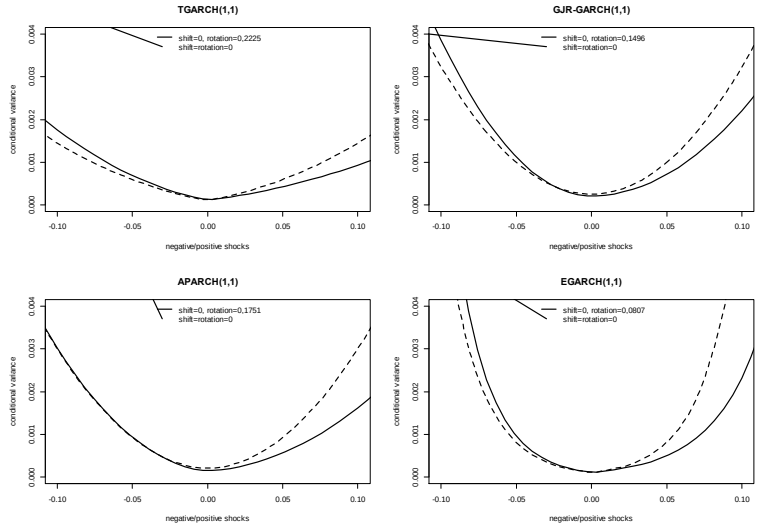


Figure 3: News impact curves for the crisis period

Table 2: Asymmetric sub-models estimated assuming skewed t-distribution in the post-crisis period

	TGARCH (1,1)		GJR-GARCH (1,1)		APARCH (1,1)		EGARCH (1,1)	
μ	-0.000	-0.001*	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000*
ω	0.001***	0.001***	0.000***	0.000***	0.000	0.000	-0.917***	-0.912***
λ	1 [a]	1 [a]	2 [a]	2 [a]	2.862***	1.264*	0 [a]	0 [a]
δ	1 [a]	1 [a]	2 [a]	2 [a]	λ	λ	1 [a]	1 [a]
α_1	0.171***	0.171***	0.174***	0.173***	0.142***	0.176***	0.329***	0.327***
β_1	0.779***	0.780***	0.761***	0.760***	0.724***	0.777***	0.900***	0.901***
a	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]	0 [a]
b	-0.110	0 [a]	-0.064	0 [a]	-0.055	0 [a]	-0.028	0 [a]
df	5.602***	5.599***	5.436***	5.416***	5.567***	5.553***	5.552***	5.548***
s	1.023***	1.022***	1.036***	1.034***	1.041***	1.026***	1.026***	1.025***
AIC	-6.463	-6.464	-6.461	-6.463	-6.455	-6.462	-6.462	-6.463
BIC	-6.426	-6.433	-6.425	-6.431	-6.413	-6.422	-6.425	-6.432
LR-test	0.3297		0.4386		0.2341		0.410	

Note: * (**, ***) represent significance at the 1% (5%, 10%) level; *a* denotes pre-fixed parameters; LR test entries represent p-values.

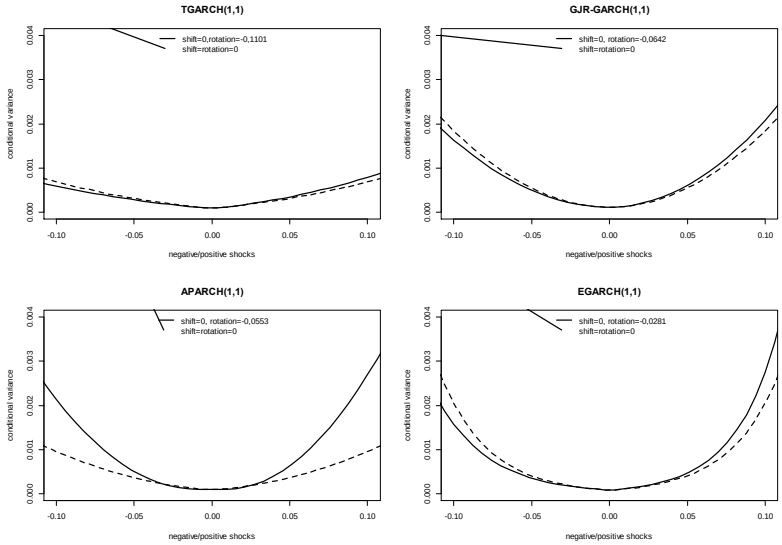


Figure 4: News impact curves for the post-crisis period

5 CONCLUSION

The empirical analysis demonstrates that the asymmetrical reaction of stock market volatility to news shocks is highly pronounced only in the crisis period. Based on the LR test results, the rotation parameter b is positive and highly significant during the crisis in all model specifications, resulting in rightward rotation of NIC. Therefore the validity of the leverage effect is strongly confirmed for the crisis data. The information criteria suggest that the asymmetrical response to news shocks can be best described by the GJR-GARCH(1,1) model. However, the post-crisis period offers just the opposite conclusions: the rotation parameter is negative, but not statistically significant. In other words, the leverage effect is found not to be significant after the recession. All of these conclusions are valid under the assumption of skewed t-distribution and with the shift parameter being equal to zero. The stated findings can be valuable to market participants for their investment decisions.

Acknowledgement
This work has been fully supported by Croatian Science Foundation under the project "Volatility measurement, modeling and forecasting" (5199).

References

[1] Black, F. (1976). Studies of Stock Price Volatility Changes. Proceedings of the Business and Economics Section of the American Statistical Association, 177–181.
[2] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31: 307-328.
[3] Caporin, M. and McAller, M. (2006). Dinamic Asymmetric GARCH. *Journal of Financial Econometrics*, 4(3): 385-412.
[4] Chen, X. and Ghysels, E. (2007) News - Good or Bad - and its Impact Over Multiple Horizons. http://www.unc.edu/~eghysels/papers/CG_news.pdf, [Accessed on June 6, 2015].
[5] Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987-1007.
[6] Engle, R. and Ng, K. V. (1993). Measuring and Testing the Impact of News on Volatility. *The Journal of Finance*, 48(5): 1749-1778.
[7] Glosten, L., Jagannathan, R. and Runkle, D. (1993). On the Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 48(5): 1779 – 1801.
[8] Hentschel, L. (1995). All in the family: Nesting symmetric and asymetric GARCH models. *Journal of financial Economics* , 39: 71-104.
[9] Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, 59(2): 347 – 370.
[10] Pagan, A. R. and Schwert, W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics* 45(1-2): 267-290.
[11] Vyrost, T. and Baumoh, E. (2009). Asymmetric GARCH and the fnancial crisis: a preliminary study, http://mpra.ub.uni-muenchen.de/27939/, [Accessed on June 6, 2015].
[12] Zakoian, J. M. (1994). Threshold Heteroskedastic Models. *Journal of Economic Dynamics and Control*, 18(5): 931 – 955.

THE ANALYSIS OF THE FINANCIAL EDUCATION IMPACT ON INVESTMENT BEHAVIOR IN CROATIA: ORDERED LOGISTIC REGRESSION APPROACH

Dajana Cvrlje
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
dcvrlje@efzg.hr

Vlasta Bahovec
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
vbahovec@efzg.hr

Irena Palić
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
ipalic@efzg.hr

Abstract: This study explores the financial education of Croatian citizens as an important factor which affects the investment behaviour using ordered logistic regression approach. There is considerable evidence on positive effects of financial education on financial behavior. The survey was conducted using the random sample. The results point out that higher financial education is related to better investment behavior, what is in line with economic theory and relevant empirical research and emphasizes the need for financial education in Croatia.

Keywords: financial education, investment behaviour, survey research, multivariate analysis, logistic regression, Croatia

1 INTRODUCTION

Financial education is a type of education on financial concepts undertaken with the explicit purpose of increasing financial knowledge and the skills, confidence, and motivation to use it Financial education is manly conducted through classroom teaching, self-study materials, informational websites, interactive games etc. Educational programs usually vary in content, audience, and methodology, but they all aim to achieve financially welfare-enhancing behavior engaged in as the result of acquired financial literacy [21]. Therefore, the effectiveness of financial education should be measured by the level of acquired financial literacy of individuals and their subsequent financial behavior i.e. financial decisions and actions they take on the marketplace.

Active discussion on the role of financial education and financial literacy continues. Research and policy interest is increasingly focused on the links between financial education, financial literacy and household saving and investments, seeking to explain why financial consumers undersave for retirement, take on too much debt, make poor mortgage and investment decisions, and experience other problems in the modern financial environment [9]. Numerous academic studies have investigated the importance of both financial education and financial literacy for various aspects of household's stability and prosperity [2, 5, 6, 12 and 19]. It has been shown that financial education has significant impact on individual's and household's wealth accumulation.

The findings of [20] and [5] emphasized the proportion of financial illiteracy problem found among Croatian consumers which suggested that many citizens are ill-equipped to make sound financial actions and decisions and be financially successful. Taking into account current situation in Croatia regarding the level of registered financial illiteracy; it is unrealistic to expect that the average person will become wiser and more active investor

unless he/she is adequately financially educated, what is confirmed by [20], who examines the proportion of respondents living in Croatia who are completely unfamiliar with certain financial products and services. The results showed that the worst results were found in the investment segment where 48.9% of retired respondents and 23,4% of active respondents were completely unfamiliar with investment funds, followed with 24,3% of active respondents who were unfamiliar with voluntary retirement funds.

2 LITERATURE REVIEW

There is considerable evidence on positive effects of financial education on financial behavior whereat the importance of financial education is particularly emphasized in matters of savings, investment and insurance pensions. Most of the research done on this subject is conducted in the developed countries while the empirical evidence for developing countries is rather limited. The research of [22] found an evidence of positive correlation between financial education and financial behavior. Findings of her research suggest that individuals who possess an adequate level of financial education should be better planners, more moderate borrowers and efficiently better in accumulating wealth. In their work, [3] found a positive link between financial education and individuals' participation in pension plans. Financial literacy was also found to be correlated with accumulated retirement saving though it appeared to be a weaker predictor of sensitivity to framing in investment decisions [9]. Furthermore, [5] confirmed positive relationship between the level of financial knowledge and skills with the successfulness in managing personal finances and [14] showed that less financially literate individuals are less likely to save for retirement. This argument was supported by the research done by [13] and [5], who observe that financial education is an important determinant of saving behavior. The recent research of [10] and [13] are among the few researches which specifically addressed the relationship between financial education and investment behavior. They argue that financial literacy is an applicable instrument for predicting individual's investment behavior.

3 THE EMPIRICAL ANALYISIS OF THE FINANCIAL EDUCATION IMPACT ON INVESTMENT BEHAVIOR IN CROATIA

3.1 The survey design and data

Prior to conducting multivariate logistic regression analysis, the survey was conducted using the random sample of 494 respondents living in Croatia. Before conducting the main survey, the questionnaire was pretested on 100 respondents living in Croatia. Feedback from both samples was used in finalizing the structured questionnaire which was conducted during July 2014. The questionnaire was administered in the form of telephone survey using the base of randomly selected telephone numbers of fixed and mobile telephony. The survey was carried out from 08:00 am 08:00 pm in order to ensure a representative sample of all required groups [17,18]. We tested both measures (financial education and investment behavior) for validity and reliability. Content validity of the study was achieved through Cronbach's internal consistency (alpha). Alpha coefficients for both measures were above 0.70, indicating a satisfactory level of internal consistency among items in each factor [16].

The questionnaire was composed out of 3 parts; I. Financial literacy, II. Financial behavior and III. Socio-demographic characteristics. The sample includes respondents living in Croatia, 20-69 years old. After conducting the survey, the logistic regression modelling is used to assess the relationship between investment behaviour as dependent variable and

financial education as explanatory variable. Both analysed variables are categorical ordinal variables and Table 1 shows the variable ranking.

Table 1: Categorical variables used in the empirical analysis (source: authors)

Variable	***Category***	***Rank***
Investment	I do not invest in voluntary pension insurance (third pillar) nor other types of financial assets (investment funds, stocks, bonds etc.)	0
	I invest my money **either** in voluntary pension insurance (third pillar) **or** in other types of financial assets (investment funds, stocks, bonds etc.)	1
	I invest my money **both** in voluntary pension insurance (third pillar) **and** in other types of financial assets (investment funds, stocks, bonds etc.)	2
Financial education	I do not have **any** financial education	0
	I have **at least** some type of financial education (either formal, no-formal or informal)	1
	I have **both** formal/no-formal and informal financial education	2

Investment behavior is ordinal variable, with three categories which are defined in Table 1. Formal financial education corresponds to a systematic, organized education model, presenting a rather rigid curriculum as regards objectives, content and methodology. It corresponds to the education process normally adopted by schools and universities. Non-formal education is rather similar to formal education, but, compared to formal education, usually lacks some features like regular student attendance. It has a rather flexible curricula and methodology and it usually includes most activities taking place outside the institution. Informal education does not correspond to an organized and systematic view of education. Informal education for instance comprises the following activities: visits to museums or to scientific and other fairs and exhibits, etc.; listening to radio broadcasting or watching TV programs on educational or scientific themes; reading texts on sciences, education, technology, etc. in journals and magazines; participating in scientific contests, etc.; attending lectures and conferences etc. [6].

3.2 The logistic regression model

Since the variable investment behavior (INV) is categorical ordinal variable, the ordinal logistic regression model is used to investigate the effects of financial education (FE) on investment (INV) [see 11]. In a model with the ordinal dependent variable, initial assumptions of a linear regression model are violated and standard least square estimator (LS estimator) is no longer the best linear estimator. Therefore, ordinal logistic model is estimated using maximum likelihood estimation (ML) method. ML estimator finds the maximum likelihood parameter estimates by maximizing the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters. The ML procedure is used in an iterative manner, to find the most likely estimates for parameters [11,15].

In the ordered logit model, there is an observed ordinal variable *y*, which is a function of another variable *y**, which is unobservable latent variable. The variable *y** has various threshold points [15]. For the purpose of empirical analysis, ordered logit model specifies that an unobservable measure of investment behavior is given by:

$$y^* = \alpha + \beta x + \varepsilon \tag{1}$$

Since investment behavior in this research has three categories (low, medium, high), in line with [11] and [15] we observe:

$$y_i = 1 \text{ if } y_i^* \le \delta_1, \tag{2}$$

$$y_i = 2 \text{ if } \delta_1 \le y_i^* \le \delta_2, \quad (3)$$

$$y_i = 3 \text{ if } y_i^* > \delta_2. \quad (4)$$

The ordered logit model results from assuming that the cumulative density of ε is the logistic function [11] The detailed explanation of ordered logit model is given in [4] and [15] The key assumption in ordinal regression is that the explanatory variable has the same effect on the odds regardless of the threshold. That means that the relationship between each pair of outcome groups is the same, and because of that, there is only one set of coefficients (only one model). This is usually termed the assumption of proportional odds. If the assumption of proportional odds is not met, instead of ordinal logistic regression, the generalised nonlinear model should be applied [1]. Once the logistic regression model is estimated, some measures of goodness of fit are calculated, the test of proportional odds, the respective tests of significance of independent variable are performed, and the obtained parameter estimates are interpreted.

3.3 The results of empirical analysis

All 494 observations in analysed data set obtained using random sampling were used in the ordinal logistic regression analysis. Analysis was performed using the statistical software SAS Enterprise Guide. The estimated model includes financial education (FE) as classification variable, and the convergence criterion is satisfied, what indicates that the estimated model converges. Furthermore, test for the proportional odds assumption for estimated model is conducted. The chi-square statistic χ^2 for testing the proportional odds assumption equals 0.8228, with empirical significance level *p- value* equal to 0.6627, which is not significant with respect to a chi-square distribution with 2 degrees of freedom. This indicates that the proportional odds assumption is reasonable.

Table 2 provides the analysis of Maximum Likelihood estimates. Note that the response categories are arranged from lowest to highest in the order that they arise in the dataset, where *FE* = 0 for no financial education and *FE* = 1 for some type of financial education. In this case, a positive coefficient β means that increasing the value of *x* tends to the higher response categories (i.e. more investment).

Table 2: Maximum Likelihood Estimates (source: authors' calculation)

Parameter	***DF***	***Estimate***	***Standard Error***	***Wald Chi-Square***	***Pr > ChiSq***
Intercept 01:0	1	0.9437	0.1099	73.7498	<.0001
Intercept 02:1	1	2.8854	0.2114	186.2248	<.0001
FE 0	1	0.4679	0.1468	10.1660	0.0014
FE 1	1	0.0379	0.1419	0.0713	0.7895

The parameter estimate at zero level financial education (FE 0) estimates the effect of the financial education level zero (FE 0) compared to the average effects of levels FE 1 and FE 2. The positive value (0.4679) for the parameter estimate is the ordered log-odds estimate of comparing the effect of FE 0 in relation to FE 1 and FE 2 on expected investment. As one goes from no financial education FE0 to FE1 or there is expected a 0,4679 unit increase in the expected value of investment in the ordered logit scale. The proportional odds, $e^{0,4679} = 1,597$ show that as one goes from FE 0 to higher categories of FE (FE 1 or FE 2) expected investment is 1,597 time greater. In other words, the respondents with some type of financial education (FE 1) and both formal/no-formal and informal financial education (FE 2)

show better investment behavior in comparison with the respondents in FE 0 category (without financial education).

Table 3 shows Odds Ratio Estimates and corresponding Wald Confidence Intervals. The odds ratio estimate of level FE 0 versus the FE 2, which is equal to 2.648, shows the difference in effect of FE 0 and FE 2. The FE 0 level is highly effective over FE 2 in reducing investment (INV). Similarly, the FE 1 level is effective over FE 2 in reducing investment (INV). That implies following ordering of levels of financial education according to favourable impact on investment: FE 2, FE 1, FE 0.

Table 3: Odds Ratio Estimates and Wald Confidence Intervals (source: authors' calculation)

Effect	***Unit***	***Estimate***	***95% Confidence Limits***	
FE 0 vs 2	1.0000	2.648	1.510	4.643
FE 1 vs 2	1.0000	1.722	0.997	2.976

Table 4 shows the various tests of the overall model: Likelihood Ratio, Score and Wald test with corresponding chi-square statistics and empirical significance levels.

Table 4: Overall model significance tests (source: authors' calculation)

Test	***Chi-Square***	***DF***	***Pr > ChiSq***
Likelihood Ratio	11.6939	2	0.0029
Score	12.1760	2	0.0023
Wald	11.7204	2	0.0029

The conducted significance tests indicate that the model is statistically significant, since *p-values* are lower than any reasonable significance level, so the stated model is appropriate.

4 CONCLUSIONS

Financial education is aimed at improving individuals' financial knowledge, skills and subsequently their financial behavior. There is considerable evidence on positive effects of financial education on financial behavior whereat the importance of financial education is particularly emphasized in matters of savings, investment and insurance pensions. This research discusses the impact of financial education on investment behavior of Croatian financial consumers using ordered logistic regression approach. Prior to conducting multivariate logistic regression analysis, the survey was conducted using the random sample of 494 respondents living in Croatia. Before conducting the main survey, the questionnaire was pretested on 100 respondents living in Croatia. Ordinal logistic model is estimated using maximum likelihood estimation (ML) method.

The results of the conducted empirical analysis indicate that the investment behavior of Croatian financial consumers depends on the level of financial education. The estimated ordinal regression model is significant and the assumption of proportional odds is satisfied. It is concluded that respondents with higher financial education show better investment behavior, what is in line with economic theory and relevant empirical research.

Regarding the low level of investment activities and investment knowledge in Croatia, the need for financial education in Croatia is obvious. Without the needed minimum level of financial education, creating a nation of responsible and wise investors in Croatia will remain a distant dream and utopia.

Acknowledgement
This work has been fully supported by Croatian Science Foundation under the project STRENGTHS no 9402.

References

[1] Agresti. A. (2013). Categorical Data Analysis, 3rd ed., John Wiley & Sons.

[2] Beckmann, E. (2013). Financial Literacy and Household Savings in Romania. Numeracy,Vol 6 (No 2).

[3] Bernheim, B. D. and Garrett, D. M. (1996). The Determinants and Consequences of Financial Education in the Workplace: Evidence from a Survey of Households. National Bureau of Economic Research. Working Paper No. 5667, Cambridge.

[4] Borooah, V.K. (2001). Logit and Probit: Ordered and Multinomial Models. Thousand Oaks: Sage.

[5] Cvrlje, D. (2014). Linking the concept of financial literacy with the successfulness in managing personal finances (Doctoral thesis). Zagreb: Faculty of Economics and Business.

[6] Cvrlje, D., Bahovec, V. and Palić, I. (2015). Exploring The Effects of Financial Education: Analysis of Saving Behavior of Croatian Consumers. In Jovancai Stakic, A., Kovsca, V. i Bendekovic, J. (Eds.) Book of Proceedings of the Economic and Social Development (ESD), 5th Eastern European Economic and Social Development Conference on Social Responsibility (pp. 241–250.) Varazdin: Varazdin Development and Entrepreneurship Agency.

[7] Field, A. (2009). Discovering statistics using SPSS, 3rd ed., Thousand Oaks: Sage.

[8] Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R.E. (2010). Multivariate Data Analysis, 7th Ed.. Pearson Education, Inc. New Jersey: Upper Saddle River.

[9] Hastings, J.S. and Mitchell, O.S. (2011). Financial literacy and impatience shape retirement wealth and investment behaviors. National Bureau of Economic Research. Working Paper No. 16740, Cambridge

[10] Jappelli, T. and Padula, M. (2013). Investment in financial literacy and saving decisions, Journal of Banking and Finance, Elsevier, Vol 37 (No 8): 2779–2792.

[11] Kennedy, P. (2008). A guide to Econometrics, 6th ed.. Blackwell Publishing.

[12] Lusardi, A. (2012). Numeracy, Financial Literacy, and Financial Decision-Making, Numeracy, Vol. 5 (No. 2).

[13] Lusardi, A. and Mitchell, O. (2009). How ordinary consumers make complex economic decisions: financial literacy and retirement readiness. National Bureau of Economic Research. Working Paper No. 15350, Cambridge

[14] Lusardi, A. and Mitchell, O. S. (2008). Planning and Financial Literacy: How Women Fare? American Economic Review, Vol 98 (No 2): 413–417.

[15] Menard, S.W. (2002). Applied Logistic Regression analysis, 2nd ed.. Thousand Oaks: Sage.

[16] Milas, G. (2005). Istraživačke metode u psihologiji i drugim društvenim znanostima. Naklada Slap: Jastrebarsko.

[17] Rea, L. M. and Parker, R. A. (2005). Designing and Conducting Survey Research: A Comprehensive Guide. Jossey-Bass.

[18] Salant, P. and Dillman, D. A. (1994). How to Conduct Your Own Survey. Wiley.

[19] Van Rooij, M. C. J., Lusardi, A. and Alessie, R. J. M. (2011). Financial literacy and retirement planning in the Netherlands. Journal of Economic Psychology, Vol 32: 593–608.

[20] Vehovec, M. (2012). Financijska i mirovinska pismenost: međunarodna iskustva i prijedlozi za Hrvatsku. Privredna kretanja i ekonomska politika, Vol 21(No 129): 65–85.

[21] Willis, L. E. (2008). Against Financial Literacy Education. University of Pennsylvania.

[22] Yoong, J., (2010). Making financial education more effective: lessons from behavioral economics. Rand Corporation Center for Financial and Economic Decision Making. http://www.oecd.org/dataoecd/ 47/62/45485586.pdf [Accessed 10/05/2015].

FLEXIBILITY OF REAL AND NOMINAL WAGES IN SLOVENIA AND CROATIA: VAR APPROACH

Anita Čeh Časni, Ksenija Dumičić and Irena Palić
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
{aceh, kdumicic, ipalic}@efzg.hr

Abstract: In this paper the flexibility of real and nominal wages in Slovenia and Croatia is investigated. The impact of macroeconomic shocks in real and nominal wages in aforementioned countries is analysed using two unrestricted vector autoregressive models. It is explored whether they show a different behaviour with the respect to the response of nominal and real wages shocks. The initial impacts of the changes in nominal wages and real wages in both coutries are similar and the results of the analysis conducted for Slovenia and Croatia are comparable.

Keywords: macroeconomic shocks**,** real wages, small open economies, vector autoregression model, Western Balkans.

1 INTRODUCTION

Adjustment of wages is seen as a mechanism to accommodate shocks in labour supply, demand, or productivity [7]. The aim of this research is to model the impact of macroeconomic shocks on small open economies. For this purpose two European Union (EU) countries are selected: Croatia, from the Western Balkan (WB) region [7] and Slovenia, which is the Central-East European (CEE) country belonging to the New EU Member States (NMS).

According to previous research, results exploring the responsiveness of real and nominal wages to shocks in some of the new member countries of European Union are ambiguous with a substantial heterogeneity across countries, regarding the responsiveness of earnings to local labour market conditions [10]. Furthermore, the economies in the WB countries have been facing complex and interrelated political and economic problems when compared to the EU countries belonging to the NMS.

Thus, the aim of this research is to investigate flexibility of real and nominal wages in Slovenia (NMS) and Croatia (WB) using an unrestricted vector autoregressive model. Namely, it is investigated whether those countries show a different behaviour with the respect to the response of nominal and real wages shocks.

The rest of the paper is organized as follows. Section 2 briefly summarizes the existing empirical literature on flexibility of nominal and real wages in the NMS (including Slovenia) and the WB countries (including Croatia). Section 3 presents research data, research methods and estimation results. The final section provides an overview of the main findings of this study.

2 LITERATURE REVIEW

According to relevant economic literature, see for instance [4,5] wage flexibility is considered to be a key determinant of labour market flexibility. In Central and Eastern European transition countries, wage flexibility has been analysed mainly using time series methods. According to research [15] where a two-equation Structural Vector Autoregression (SVAR) framework was used in the study of the responsiveness of nominal and real wages to shocks, among the new EU members, only Hungary and the Czech Republic are well-prepared for EMU membership in terms of labour market adjustment. Furthermore, [1]

compares eight new EU member states as well as three Eurozone members to explore how far wages can flexibly accommodate macroeconomic shocks, but this study did not result in supportive evidence for a higher adjustment capability of labour markets in the analysed countries. In [4] a framework for investigating the flexibility of wages called the wage curve was popularized. Country-specific wage curves for six new EU member states, Romania, Bulgaria as well as for five EU-15 members were estimated in [6]. In the group of the CEE countries, [6] reports elasticities of pay ranging between 0.01 in Slovenia and -0.09 in Hungary. This research also provides quantitative evidence that wages are more responsive to local labour market conditions in the CEE countries when compared to the old EU members.

According to previously mentioned studies, results exploring the responsiveness of real and nominal wages to shocks in some of the new member countries are ambiguous with a substantial heterogeneity across countries. Generally, wages in the new member states are more responsive to local labour market conditions than in the old member states [10].

On the other hand, according to European Commission report [7], output recovery has been much slower in South-East Europe (WB countries) than in the Central European countries as a result of which labour markets began to improve with some delay (when compared to the NMS). In most countries of the WB region, gross domestic product started to grow by the end of 1999, resulting in the increase of employment everywhere, but in Serbia and Montenegro. Consequently, increased productivity, not the creation of new jobs, has been the driving force behind this growth, since the entire WB region is characterized by extremely low employment rates [10]

Also, when analysing WB countries, an important issue to tackle are data limitations, impeding the analyses of the labour markets with, in some cases, controversial outcome, depending on the data source. So, the adequate statistical tool for modelling the responsiveness of real and nominal wages to macroeconomic shocks would be the unrestricted VAR model, since it has useful statistical properties, see [9, 14] for technical details. This is in line with the critique on identifying restrictions in general discussed in [17].

3 THE EMPIRICAL ANALYISIS OF THE RELATIONSHIP OF REAL AND NOMINAL WAGES IN SLOVENIA AND CROATIA

3.1 Data and methods

Since the main objective of this paper is to analyse the relationship of nominal and real wages in Slovenia and Croatia, firstly real and nominal wages in Croatia and Slovenia are explained in brief and their dynamics is graphically analysed. Furthermore, stationarity analysis which is prior step to VAR modelling is conducted.

The data we use are taken from the Vienna Institute for International Economic Studies (WIIW) Monthly Database, providing data on average gross monthly wages in EUR. In order to calculate real gross wages for Croatia and Slovenia, nominal gross wages in EUR are deflated by Harmonised Indices of Consumer Prices (HICPs), 2005=100 HICPs for Slovenia and Croatia are provided by [11] and serve to calculate the comparable measures of inflation within European Union. They measure the change of the prices of consumer goods and services acquired by households, according to a harmonised approach at the European Union level. Therefore, HICPs provide the official measure of consumer price inflation in the euro area for the assessment of inflation, [11].

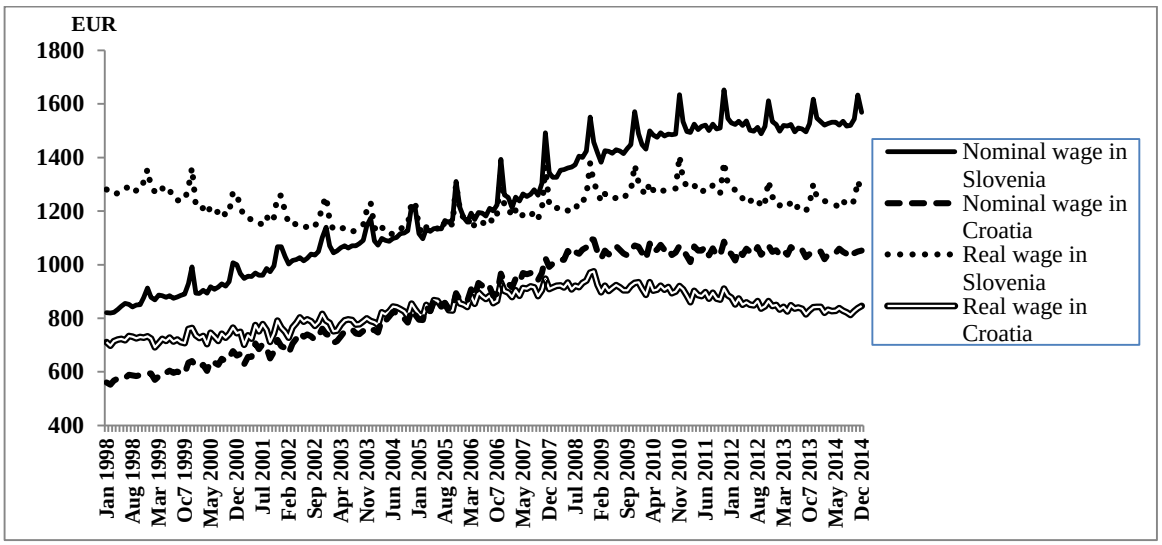


Figure 1: Nominal and real wages in Croatia and Slovenia from January 1998 to December 2014
Source: WIIW, Eurostat, authors' calculation

Nominal wages in Slovenia exhibit an increasing trend, while real wages oscillate in the observed period, whereas decrease in real wages refers to an increase in prices measured by HICPs, i.e. to inflationary pressures. With the financial crisis and recession, nominal wages stagnated in Croatia after 2008, while real wages decreased in Croatia and are stagnating since 2013.

In order to assess the relationship between nominal and real wages in Slovenia and Croatia, Vector Autoregression methodology is used. For analysis of the relationship of economic variables in [17] Vector Autoregression model (VAR) was first proposed. The basic VAR model treats all variables symmetrically, regardless of whether they are endogenous or exogenous. Namely, in the VAR model, the time path of each variable is influenced by current and past realization of the variables in the model [9]. The basic form of the VAR model in the case of *n* variables and lag *k* is given by:

$$Z_t = \mu + A_1 Z_{t-1} + ... + A_k Z_{t-k} + \psi D_t + e_t, \quad (1)$$

where Z_t is *n*-dimensional vector of variables of order (*n x 1*), $A_1,...,A_k$ are matrices of autoregressive parameters of order *(n x n)*, D_t is the vector of non stohastic exogenous variables with matrix of parameters , μ is the vector of constant terms for each variable and e_t is the vector of innovations. In estimating VAR models, question of time series stationarity arises. Time series is considered to be stationary in a broader sense if the expected value and variance of the population does not depend on time *t* and if the covariance of two members of the series Y*t* and Y*t+s* which are separated by one time period depends on the distance *s*, but not on time *t* [8]. The most common stationarity test is Dickey-Fuller test which can be modified as Augmented Dickey Fuller test by the inclusion of an additional shift of the dependent variables in order to eliminate autocorrelation of error terms [2].

3.2 The results from VAR analysis

In order to empirically assess mutual relationship of nominal and real wages in Slovenia and Croatia, previously explained bivariate VAR model is used. In estimating VAR models, it is common to take the logarithmic values of the original series to remove heteroskedasticity, i.e. the volatility of the variance of error terms [16]. Moreover, Figure 1 shows that all the observed time series show the presence of seasonal component and therefore variables are

seasonally adjusted. Seasonal adjustment is conducted using TRAMO-SEATS quarterly seasonal adjustment method developed in the research [12].

In order to test the stationarity of the selected time series, the Augmented Dickey-Fuller (ADF) unit root test is conducted. All selected variables are shown to be integrated of order 1, at 1% significance[1]. Therefore, series are first-differenced and the stationary time series are obtained. The following variables are used in the empirical analysis: DNW_SLO and DNW_CRO, which refer to percentage change in nominal gross monthly wages for Slovenia and Croatia; and DRW_SLO and DRW_CRO, which refer to percentage change in real gross monthly wages for Slovenia and Croatia, 2005=100, whereat all mentioned are seasonally adjusted. The time span for all variables is from January 1998 to December 2014.

The existence of cointegration among nominal and real wages both in Croatia and Slovenia is tested. In both cases trace test and maximum eigenvalue test indicate no cointegration at the 0.05 level. Therefore, two bivariate VAR models are estimated. The first VAR model is estimated for Croatia using variables DNW_CRO and DRW_CRO and the second VAR model is estimated for Slovenia using variables DNW_SLO and DRW_SLO.

The model diagnostics tests are conducted for both estimated VAR models. In order to eliminate residual heteroskedasticity and autocorrelation problems, both models are estimated with 27 lags. For both models, White test has shown that the problem of heteroskedasticity is not present at 1% significance level. Furthermore, the LM test of autocorrelation is conducted. For both models, at 1% significance level, null hypothesis of no autocorrelation of residuals cannot be rejected up to lag length k=12, since all corresponding empirical significance levels are higher than 0.01.

The stability of both VAR models is checked by calculating the inverse roots of characteristic AR polynomial using EViews8[2]. For detailed explanation of problems of heteroskedasticity, autocorrelation as well as AR roots calculation see, for example [9]. The analysis has shown that no inverse root of AR characteristic polynomial lies outside the unit circle and therefore both VAR models satisfy the stability condition. However, inverse roots are very close to one, and therefore impulse response functions do not die out to zero after one year.

The impulse response functions of VAR model for Croatia are presented in Figure 2, while Figure 3 shows the impulse response functions of VAR model for Slovenia. The initial effects shown by impulse response functions are similar for Croatia and Slovenia. The impact of the percentage change in nominal wages on the variable itself is initially positive and statistically significant, as well as the impact of the percentage change in real wages on the variable itself. In both countries the impact of the percentage change in nominal wages on the percentage change in real wages is unclear and statistically insignificant. On the other hand, the impact of the percentage change in real wages on the percentage change in nominal wages is initially statistically significant and positive. Therefore, the results confirm that Croatia and Slovenia, which have similar political legacy, are also similar regarding the flexibility of nominal and real wages.

[1] The results of ADF test conducted in EViews 8 software are available on request.

[2] The calculated roots are available on request.

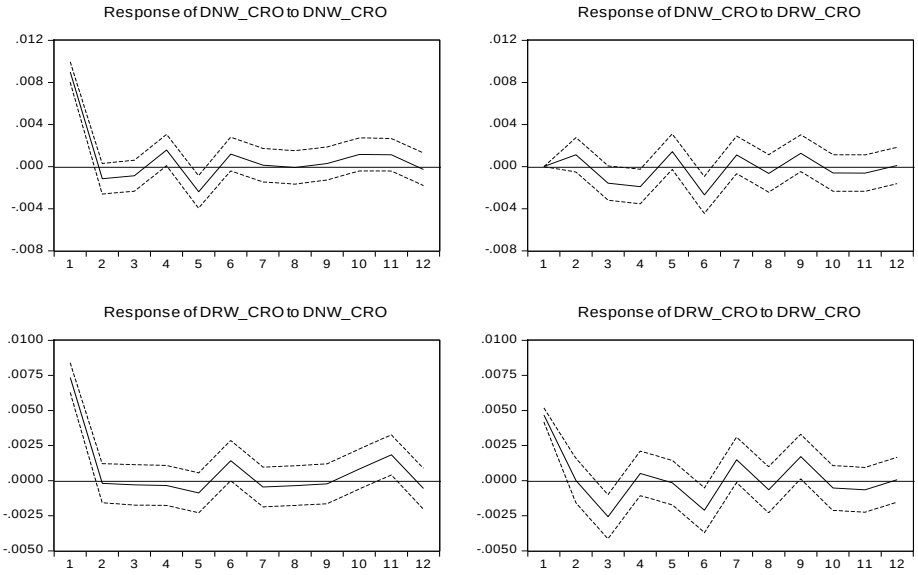


Figure 2: Orthogonalized impulse response functions for Croatia
Source: authors' calculation

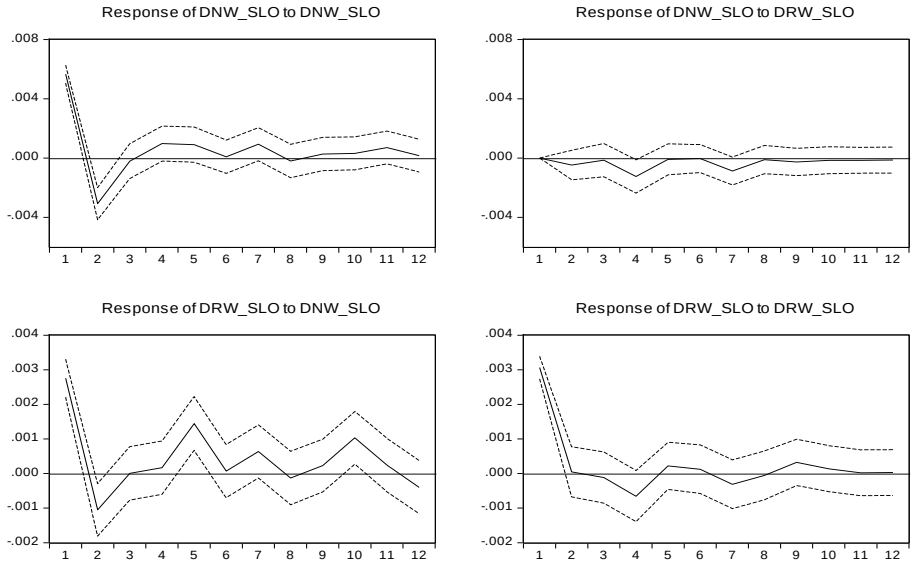


Figure 3: Orthogonalized impulse response functions for Slovenia
Source: authors' calculation

4 CONCLUSIONS

In this research the flexibility of real and nominal wages in Slovenia and Croatia using an unrestricted vector autoregressive model is investigated. The quality of the data regarding number of observations and potential structural breaks is a matter of concern in this analysis, since that can result in a bias of the estimated parameters. However, due to short time series and the general economic situation in analysed countries, it is impossible to deal with the related problems properly.

The results of the conducted empirical analysis suggest that the impacts shown by impulse response functions of two estimated VAR models are similar for Croatia and Slovenia, confirming that Croatia and Slovenia, which have similar political legacy, are also similar regarding the flexibility of nominal and real wages.

However, one should be aware of potential data problems mentioned above. Second, it is not straightforward to use this kind of time series analysis for proper cross-country comparisons without the knowledge of the data generating process. Nonetheless, we can gain some insights on the flexibility of wages and differences across the two countries in a qualitative manner. In further research we will focus on wage responses to shocks in all of the Western Balkan countries and comparison with the group of peer countries (NMS) which should shed some more light on wage setting behaviour in the aforementioned groups of countries.

Acknowledgement: This work has been partially supported by Croatian Science Foundation under the project STRENGTHS (Project no. 9402).

References

[1] Babetskii, I. (2006). 'Aggregate Wage Flexibility in Selected New EU Member States', Czech National Bank, Working Paper No. 1/2006.

[2] Bahovec, V. and Erjavec, N. (2009). Uvod u ekonometrijsku analizu. Zagreb: Element.

[3] Blanchflower, D. (2001). 'Unemployment, Well-Being and Wage Curves in Eastern and Central Europe', Journal of the Japanese and International Economies, Vol. 15, pp. 364-402.

[4] Blanchflower, D. and A. Oswald (1994). The Wage Curve, The MIT Press, Cambridge, Mass.

[5] Boeri, T., M. Burda and J. Köllö (1998). 'Mediating the Transition: Labour markets in Central and Eastern Europe', Centre for Economic Policy Research, London.

[6] Buettner, T. (2007). 'Unemployment Disparities and Regional Wage Flexibility. Comparing EU Members and EU-Accession Countries', Empirica, Vol. 34, No. 4, pp. 287-97.

[7] Commission of the European Communities (2008). Communication from the Commission to the European Parliament and the Council, Western Balkans: Enhancing the European perspective, Brussels, 5.3.2008,COM(2008) 127 final.

[8] Dougherty, C. (2006). Introduction to Econometrics (3rd ed.). London: Oxford University Press

[9] Enders, W. (2010). Applied Econometric Time Series (3rd ed.). London: John Wiley & Sons Inc.

[10] European Commission (2004), 'Adjustment capacity of labour markets of the Western Balkan countries', Study coordinated by the Vienna Institute for International Economic Studies (WIIW), DG Economic and Financial Affairs, Brussels.

[11] Eurostat HICP Database. (2015). Eurostat
http://ec.europa.eu/eurostat/cache/metadata/en/prc_hicp_esms.htm [Accessed 19/5/2015].

[12] Gomez. V. and Maravall. A. (1996). Programs TRAMO and SEATS: Instruction for Users. Banco de Espana. Working Paper No. 9628, Madrid.

[13] Kennedy, P. (2003). A guide to Econometrics. Wiley Blackwell.

[14] Lutkepohl, H. (1991). Introduction to multiple time series. Springer, New York.

[15] Moore, T. and E. J. Pentecost (2006), 'An Investigation into the Sources of Fluctuation in Real and Nominal Wage Rates in Eight EU Countries: A Structural VAR Approach', Journal of Comparative Economics, Vol. 34, No. 2, pp. 357-376.

[16] Palić, I. (2014). The relationship of private savings and economic growth: Case of Croatia. International Journal of Social, Human Science and Engineering, Vol. 8 (No. 5): 47–53.

[17] Sims, C.A. (1980). Macroeconomics and Reality. Econometrica, Vol. 48: 1–18.

HIGHER-LEVEL INDICES IN THE CONSUMER PRICE INDEX

Draženka Čizmić
Faculty of Economics & Business, University of Zagreb
Trg J.F. Kennedyja 6, Zagreb 10 000, Croatia
dcizmic@efzg.hr

Abstract: Consumer price indices are compiled at the higher level using Laspeyres-type arithmetic averages. These Laspeyres-type indices include the Young and the Lowe indices, both of which have serious shortcomings. There may be practical advantages and financial savings from continuing to make repeated use over many years of the same fixed set of quantities to calculate a CPI. However, the amount by which such a CPI exceeds some conceptually preferred target index is likely to get steadily larger the longer the same set of reference quantities is used.

Keywords: consumer price index, Laspeyres-type indices, bias

1 INTRODUCTION

Consumer price indices (CPIs) are economic indicators constructed to measure the changes over time in the prices of consumer goods and services acquired, used or paid for by households.

The calculation of a CPI proceeds in stages. In the first stage, elementary price indices are estimated for the elementary expenditure aggregates of a CPI.[1] In the second stage, these elementary indices are aggregated, or averaged, to obtain higher-level indices using the elementary expenditure aggregates as weights.

CPIs are used for a wide variety of purposes, including: as a guide for monetary policy; for the indexation of commercial contracts, wages, social protection benefits or financial instruments; as a tool for deflating the national accounts[2] or calculating changes in national consumption or living standards.

Various mathematical formulas for constructing these indices are discussed including problems for prices statisticians in selecting the most appropriate methodology. The advantages and disadvantages of the various formulas are discussed, along with criteria to guide decisions on the most appropriate formula.

2 ARITHMETIC FORMULAS

Consumer price indices are compiled at the higher level using Laspeyres-type arithmetic averages. These Laspeyres-type indices include the Young and the Lowe indices, both of which have serious shortcomings.

A Laspeyres price index formula is used in price statistics to measure the difference in the cost of a fixed basket of goods and services between the price reference period (0) and the current period (t). A price reference period is the period with which current period prices are compared. The Laspeyres price index can thus be written as an arithmetic average of the price relatives weighted by the price reference period expenditures shares.

[1] An elementary aggregate consists of the expenditures on a small and relatively homogeneous set of products defined within the consumption classification used in the CPI. Research on elementary indices has mostly been restricted to unweighted formulae (arithmetic and geometric means) because detailed quantity data are not available.

[2] Lowe price indices may be used to deflate time series of consumption expenditures at current prices in order to obtain the implicit quantity indices.

In principle, any set of quantities could serve as the basket. The basket does not have to be restricted to the quantities purchased in one or other of the two periods compared. For practical reasons, the basket of quantities used for CPI purposes usually has to be based on a survey of household consumption expenditures conducted in an earlier period than either of the two periods whose prices are compared (b). This class index is called a Lowe index.[3] Most statistical offices make use of some kind of Lowe index in practice. The Laspeyres price index is the Lowe index in which the reference quantities are those of the price reference period 0 (period b coincides with period 0). The Lowe index satisfies 12 axioms with time reversibility and transitivity among them.

$$P_{Lo} = \frac{\sum_{i=1}^{n} p_i^{\,t} q_i^{\,b}}{\sum_{i=1}^{n} p_i^{\,0} q_i^{\,b}} = \sum_{i=1}^{n} \left(p_i^{\,t} / p_i^{\,0} \right) s_i^{\,0b} \quad \text{where} \quad s_i^{\,0b} = \frac{p_i^{\,0} q_i^{\,b}}{\sum_{i=1}^{n} p_i^{\,0} q_i^{\,b}}$$

Instead of holding constant the quantities of period b, a statistical office may calculate a CPI as a weighted arithmetic average of the individual price relatives, holding constant the revenue shares of period b. The resulting index is called a Young index.[4]

$$P_{Yo} = \sum_{i=1}^{n} s_i^{\,b} \left(\frac{p_i^{\,t}}{p_i^{\,0}} \right) \quad \text{where} \quad s_i^{\,b} = \frac{p_i^{\,b} q_i^{\,b}}{\sum_{i=1}^{n} p_i^{\,b} q_i^{\,b}}$$

(p_i – prices, q_i – quantities, 0 – price reference period, t – current period, b – the period whose quantities are actually used in CPI, s_i^b – revenue shares of period b)

The production of a price index by reference to a fixed basket of goods and services has several advantages. Firstly, the concept is easy to understand. Secondly, by fixing both the items within the basket and their quantities, the resulting values provide a measure of pure price change that is free from compositional change.

The households' expenditures on all consumer goods and services in the CPI basket are mainly sourced from information derived from the Household Expenditure Survey (HES). The compilation of national accounts (NA) estimates within National Statistical Offices will usually be given the highest priority in terms of resources and effort. Household Final Consumption Expenditure (HFCE) estimates are usually compiled as part of the NA suite of statistics. These data are very often based on the HES but, unlike the HES, estimates may be provided on a more frequent basis, annually or quarterly for example.

The Harmonised Index of Consumer Prices (the "European CPI")[5] uses HF(M)CE to derive its weights. In general, the NA HFCE estimates include elements of imputed expenditures and HF(M)CE is a subset of HFCE that is based only on monetary expenditures. The weights used for computing HICPs within a country may relate to a period up to seven years prior to the current year. However, to minimise any incomparability this might cause, adjustments must be made each year for any especially large changes in expenditure patterns.

[3] After the index number pioneer who first proposed it in 1823. The name "Lowe Index" was introduced in the international Consumer Price Index Manual: Theory and Practice (2004) and in the paper by Balk and Diewert (2003).

[4] This type of index was first defined by the English economist Arthur Young (1812). The Lowe index may be preferred to the Young index because the Young index has some undesirable properties.

[5] The differences between HICPs and individual national CPIs can sometimes be significant in practice. The differences have in general been diminishing, although national CPIs use their own national methodologies.

3 GEOMETRIC FORMULAS

Using a geometric formula at the higher level would be compatible with the currently widely used Jevons index at the lower level and would have the benefit of maintaining consistency in aggregation. The advantages of the geometric mean are: 1) not as sensitive as arithmetic means to the extreme values, 2) it is circular[6] and 3) is more likely to lie between the Laspeyres and Paasche bounds. There are some disadvantages to the geometric mean approach. The main disadvantage of geometric aggregation will be in its complexity.

Balk (2010) demonstrates that the substitution bias of the geometric Young index is less than the substitution bias of the currently widely-used Lowe index. The CPI Manual considers the geometric Young index to be a serious practical possibility for CPI compilation. With unitary elasticity of substitution, the geometric Young can be shown to lie within Laspeyres-Paasche interval. The geometric Young is easily explained as a weighted geometric average of price changes, using the survey period expenditure shares as weights.

$$P_{GYo} = \prod_{i=1}^{n} \left(\frac{p_i^{\,t}}{p_i^{\,0}} \right)^{s_i^{\,b}}$$

Unlike the arithmetic Lowe index, the geometric Lowe indices have no fixed quantity basket definition. The price updating of the weights has no rationale for the geometric Lowe. The geometric Lowe index has no meaningful interpretation.

$$P_{GLo} = \prod_{i=1}^{n} \left(\frac{p_i^{\,t}}{p_i^{\,0}} \right)^{s_i^{\,b0}}$$

A 2012 IMF Working Paper proposes geometric alternatives to the traditional arithmetic Laspeyres-type indexes currently used by most National Statistics Offices in the compilation of price indices. These geometrically-based indexes share the advantage of their arithmetic counterparts of being able to be computed in real time and are thus practical alternatives to arithmetic versions.

4 HIGHER-LEVEL SUBSTITUTION BIAS

The Boskin Commission[7] concluded that the change in the Consumer Price Index overstates the change in the cost of living by about 1.1 percentage points per year (higher-level substitution 0.15, lower-level substitution 0.25, new products/quality change 0.60, new outlets 0.10). The Commission's estimate of 0.15 points greatly understated the significance of higher-level substitution bias. This is especially true, given the much more rapid updating of higher-level weights in the CPI for the 2000-06 period from which the 0.38 percent number is calculated. It is possible that, with this new information, the Commission's estimate of higher-level substitution bias for the 1995-96 interval might be 0.45 to 0.50 points.

[6] Fulfils a multi-period transitivity property that the product of index change going from a period 1 to a period 2 times the price index change going from period 2 to a period 3 equals the price index going directly from period 1 to 3.

[7] The previous report was the famous Stigler (1961) Commission Report. There were several important differences between the two Commissions. The Boskin report concerned only the CPI, while the Stigler report also covered the Producer Price Index and agricultural price indices. The Stigler report did not produce any numerical bias estimates. The Stigler Commission had a substantial budget to commission new research studies, whereas the Boskin Commission had no research budget at all.

The Australian Bureau of Statistics has constructed a retrospective superlative-type index to provide an estimation of potential higher-level substitution bias in the fixed-weighted Australian CPI. The analysis found the total higher-level substitution bias of the All groups CPI[8] was 3.6 percentage points after 11 years due to the inability of the fixed-base index to take account of the item substitution effect.

Higher level substitution bias is expected to occur when households exhibit cost minimising behaviour, specifically, as their preferences shift from products with relatively higher levels of inflation to products with relatively lower levels of inflation. An element of higher-level substitution bias is inevitable in the compilation of a fixed basket index and it is generally accepted that the magnitude of the bias is likely to increase with the "age" of the basket.

The different index formulas produce different index numbers, and thus different estimates of the price movements. Typically the Laspeyres formula will produce a higher index number than the Paasche formula in periods after the base period, with the Fisher Ideal and the Törnqvist of similar magnitude falling between the index numbers produced by the other two formulas.

Under the assumptions that there are long-term trends in prices and normal consumer substitution responses, the Lowe index will normally be greater than the corresponding Laspeyres index.

Lowe ≥ Laspeyres ≥ Fisher ≥ Paasche

A price index may be described as biased if it produces estimates which depart from a notionally true or correct measure. In the case of consumer price indices, the true measure is usually taken to be the cost of living index, as it allows for the substitutions in consumption that consumers make in response to changes in relative prices. As it is impractical to construct a true cost of living index, official agencies are forced into second-best solutions.

5 A COST OF LIVING INDEX

The Boskin Commission's first and overarching recommendation is that the BLS[9] should establish a cost of living index as its objective in measuring consumer prices. All of the other specific recommendations are aimed toward achieving this goal.

A cost of living index (COLI) may be defined as the ratio of minimum expenditures needed to attain the same level of utility in two time periods. COLIs cannot be calculated exactly because the second set of expenditures cannot be observed. However, a COLI may be approximated by means of superlative index.[10] Superlative indices treat both periods symmetrically, the two most widely used examples of superlative indices being the Fisher index and the Törnqvist index.

A well-known result in index number theory is that the Laspeyres price index places an upper bound on the COLI bases on the first period, while the Paasche index places a lower bound on the COLI based on the second period.

Balk and Diewert (2003) conclude that, the difference between a Lowe index and a COLI may be reduced to a negligible amount if: 1) the lag in obtaining the base year quantity weights is minimized, and 2) the base year is changed as frequently as possible.

[8] as measured by the difference between the Laspeyres-type index and the Fisher-type index

[9] Bureau of Labor Statistics is a unit of the United States Department of Labor. It is the principal fact-finding agency for the U.S. government in the broad field of labor economics and statistics and serves as a principal agency of the U.S. Federal Statistical System.

[10] The concept of a superlative index was introduced by Erwin Diewert (1976).

6 USING A WEIGHTED AVERAGE OF BASE PERIOD PRICE INDICES TO APPROXIMATE A SUPERLATIVE INDEX

A body of theory suggests that certain superlative index formulas give a good approximation to a COLI. These formulas have been difficult to implement because they require information on consumer expenditure patterns for both of the two reference periods. Such information on the more recent period is usually unavailable at the time of index production.

Shapiro and Wilcox advocated the Lloyd-Moulton price index as a timely approximation to a superlative index. Instead of second period expenditure information, the Lloyd-Moulton index relies on a parameter representing "elasticity of substitution" and uses past years' data as a basis for evaluating the parameter estimate.

An alternative approach to estimating a COLI is to compute a weighted average of the base-weighted arithmetic and geometric (Laspeyres) indexes. Lent and Dorfman (2009) show through Taylor series expansions that the arithmetic-geometric average (or "AG Mean") index closely approximates the Lloyd-Moulton index and hence the superlative indices. The weight applied to the geometric index in the AG Mean may be estimated from prior data through a simple formula and then systematically updated with more recent consumer expenditure data.

7 CROATIAN CPI

The consumer price indices have been calculated and published since January 2004. The publishing of consumer price indices resulted in termination of calculation and publishing of retail price indices, cost of living indices and indices of hotel and restaurant services.

The classification of products is based on the Classification of Individual Consumption by Purpose (COICOP). This classification breaks consumer expenditure into twelve different divisions of consumer goods and services.

Since January 2013 the year 2010 has been taken as the index base period and, for comparability purposes, indices that were calculated on the 2005 base year have been recalculated on the 2010 base year.

Since January 2013 the compilation of consumer price indices has been based on the weights derived from annual average of household expenditures obtained from the 2011 Household Budget Survey recalculated into price in the December 2012.

The calculation of the consumer price indices begins with the computation of elementary aggregate indices at the level of geographical locations as a ratio of geometric means of the current and the reference price period. The consumer price index is then compiled by using the formula for the weighted arithmetic mean of indices at the lowest aggregation levels (modified Laspeyres formula).

8 CONCLUSION

Price index theory provides prices statisticians with guidance to employ the best practices and formulas in compiling price indices in order to produce reliable price measures. However, the highly desirable methods must be balanced against the practical ones. It would be highly desirable to use a superlative index formula such as the Fisher ideal for all price indices, but timing issues and data availability preclude this.

A Laspeyres price index is one in which the quantities that make up the basket are the actual quantities of the price reference period. CPIs are not Laspeyres indices, even though they may officially be described as Laspeyres type indices. The expenditures and quantities used as weights for CPIs typically come from household budget surveys undertaken some

years before the price reference period for the CPI. Lowe indices are popular for several reasons. They are conceptually simple and meaningful. They enable statistical agencies to economize by continuing to make use of the same set of quantities over many years. They are transitive and additive.

There may be practical advantages and financial savings from continuing to make repeated use over many years of the same fixed set of quantities to calculate a CPI. However, the amount by which such a CPI exceeds some conceptually preferred target index is likely to get steadily larger the longer the same set of reference quantities is used.

This overstatement will have important unintended consequences, including overindexing government outlays and tax brackets and increasing the deficit and debt.

References

[1] Akem A., Bradley T.; (2014); ABS study into the use of HFCE estimates for CPI weighting purposes; UN Economic Commission for Europe, Geneva

[2] Akem A., Opryshko A.; (2014), Why "Lowe" when "Young" and "Laspeyres" are available?, United Nations Economic Commission for Europe, Geneva

[3] Allen R.G.D., (1975), Index Numbers in Theory and Practice, The Macmillan Press Ltd., London and Basingstoke

[4] Armknecht P., Silver M.; (2012); Post-Laspeyres: The Case for a New Formula for Compiling Consumer Price Indexes; IMF Working Paper WP/12/105, International Monetary Fund, Washington, D.C.

[5] Balk B.M., (2010), Lowe and Cobb-Douglas Consumer Price Indices and its Substitution Bias, Jahrbűcher fűr Nationalökonomie und Statistik, Vol. 230. No.6

[6] Balk B.M. and Diewert W.E., (2003), The Lowe Consumer Price Index and its Substitution Bias, Discussion Paper No. 04-07, Department of Economics, University of British Columbia, Vancouver

[7] Bishop G., (2013), Geometric Indexes and Substitution Bias in the CPI, Ottawa Group Meeting

[8] Boskin M.J., Dulberger E.R., Gordon R.J., Griliches Z., Jorgenson D.W.; (1996); Toward a More Accurate Measure of the Cost of Living; Final Report to the Senate Finance Committee, December 4.

[9] Central Bureau of Statistics of the Republic of Croatia, (2014), Statistical Yearbook of the Republic of Croatia 2014, Zagreb

[10] Diewert W.E., (1998), Index Number Issues in the Consumer Price Index, Journal of Economic Perspectives 12 (1)

[11] Diewert W.E., Nakamura A.O.; (1993); Essays in Index Number Theory, Volume 1; North-Holland, Amsterdam

[12] Eurostat, (2004), Harmonized Indices of Consumer Prices (HICPs), A Short Guide for Users, Luxembourg

[13] Gordon R.J., (2006), The Boskin Commission Report: A Retrospective One Decade Later, Working Paper 12311, NBER, Cambridge

[14] Hill P., (2008), Lowe Indices, The 2008 World Congress on National Accounts and Economic Performance Measures for Nations, Washington, D.C.

[15] International Labour Office (ILO/IMF/OECD/IMECE/Eurostat/World Bank), (2004), Consumer Price Index Manual: Theory and Practice, Geneva

[16] Kints van M., Bradley T.; (2014); The calculation of higher-level indices in the CPI – a practical guide, UN Economic Commission for Europe, Geneva

[17] Lent J., Dorfman A.H.; (2009); Using a Weighted Average of Base period Price Indexes to Approximate a Superlative Index; Journal of Official Statistics; Vol. 25, No.1

CAUSALITY BETWEEN STATE AID AND GDP IN EUROPEAN COUNTRIES

Renata Kožul Blaževski
University of Split / University Department of Professional Studies
Kopilica 5, 21000 Split, Croatia
rkozulb@oss.unist.hr

Jelena Vidović
University of Split / University Department of Professional Studies
Kopilica 5, 21000 Split, Croatia
jvidovic@oss.unist.hr

Abstract: The connection between state aid expenditures and economic growth in economic literature was never confirmed. In this paper we explore connection between state aid expenditures for horizontal and sectoral purposes and GDP using Granger causality test on panel data set. Panel data set consists of ten EU member states in period from 2000 to 2011. Our results suggest that the changes in GDP will create changes in horizontal aid but changes in horizontal aid will not create changes in GDP. Sectoral aid appears to cause changes in GDP but this connection remains unexplained due size of sectoral aid expenditures and political decision to reduce share of sectoral aid in total state aid.

Keywords: state aid, GDP, EU member states, panel data, Granger causality.

1 INTRODUCTION

The main objective of state aid is to encourage economic growth, particularly in sectors that without state aid could hardly evolve. Number of studies dealing with the effectiveness of state aid is very small. In this study we investigate the effect of state aid expenditures on GDP using a balanced panel data set which covers 10 European Union countries over the period from 2000 to 2011. We observe two main categories of state aid; horizontal state aid and sectoral state aid.

The paper is organized into five parts. In the second part of the paper we define categories of state aid and analyze ratio between horizontal and sectoral state aid in EU. In the same part we give overview of previous researches. Data and methodology are presented in the third part of paper. In the fourth part of the paper the results of the impact of state aid to GDP using the Granger causality test under VAR environment are presented. Main conclusions are presented in the fifth part of paper.

2 STATE AID

All programs of state aid before their implementation must be approved by the European Commission. In addition two types of state aid measures do not need prior approval, that are aid covered by the general provision of block exemption (General Block Exemption Rules - GBER) and de minimis aid (de minimis). According to the category, state aid can be horizontal and sectorial, regional and support at the level of local and regional governments. Horizontal aid does not distort competition in contrast to the sectorial aid provided to certain sectors or individual entrepreneurs. Horizontal aid has smaller impact on distortion of market competition, it is usually comprised of the aid earmarked for research and development and innovation, environmental protection and energy saving, small and medium enterprises, training, employment, culture and the similar. Share of state aid in GDP (Figure 1) in period from 1992 to 2011 decreased. The dominant sectoral aid is declining in favor of horizontal

support. Sectoral aid in Croatia in recent years has been most awarded in the sectors of metallurgy, railways, shipbuilding and transport. Such an unfavorable ratio of the sectoral aid in contrast to the horizontal support is the result of an unsuccessful privatization and poor implementation of structural reforms. The biggest criticism facing Croatia in its state aid policy is to large share of aid in GDP, but also a large share of sectoral aid in total aid. In 2012. the share of aid in GDP in Croatia was 2.7%, in 2011, and 2010., this shares were respectively 2.73% and 2.9%. The share of aid in GDP within the EU is 0.5%. In EU horizontal aid dominates over sectorial aid due its less negative impact on competition, and to a greater extent has positive effects on the economy. The fraction between the assigned horizontal and sectorial aid in the EU is 90:10 percent, and in Croatia 30: 70 percent.

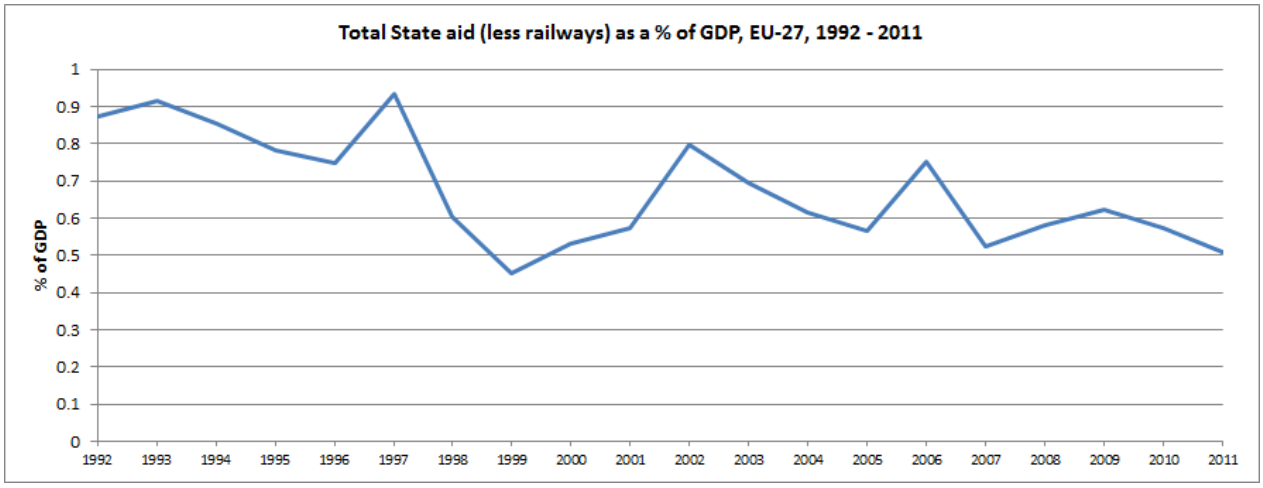


Figure 1: Trend in the overall level of non-crisis aid as a percentage of GDP, EU-27, 1992-2011. Total aid excludes railways

Tunali and Fidrmuc [7] have explored the impact of state aid on economic growth and investment on an unbalanced panel data set of 27 European Union countries over the period 1992-2011. In order to examine the effects of state aid on total value added growth, they used the Solow Model of economic growth and add state aid as an additional explanatory variable to this model. Although horizontal and sectoral state aid is statistically significant according to the OLS estimation results, the sign of the coefficients of these variables is negative and the magnitude of these coefficients is very close to zero. They concluded that total state aid and its subcomponents do not have an important effect on economic growth.

3 DATA AND METHODOLOGY

This study attempts to investigate the causality between Gross domestic product, Horizontal aid and Sectoral aid in a panel data context using the Granger causality method originating from the seminal work of Granger [2].

The data used in this study consist of a sample of 10 EU countries observed in the period from 2000 to 2011 or 120 annual observations. Countries studied in this paper are listed in Table 1. The data, a balanced panel, is chosen based on the availability of data. All variables are taken from web sites of EU Commission relating Competition and State aid. Variables used in the study are: Gross domestic product (GDP), Horizontal aid (HOR) and Sectoral aid (SEK). Variables are in million euros and are log-transformed and denoted as LGDP, LHOR and LSEK.

Table 1: Countries included in analysis

Czech Republic	Denmark	France	Germany	Hungary
Italy	Poland	Slovakia	Slovenia	United Kingdom

4 EMPIRICAL RESULTS

Empirical analysis starts by testing the presence of a unit root in the variables. We employ several panel unit root tests that are extensively used for panel data. These tests are Levin-Lin-Chu panel unit root test [3], Maddala-Wu panel unit test (Fisher-type test using ADF test; [4]), and Choi panel unit root test (Fisher-type test using Phillips Perron test; [1]). While the first test assumes a common unit root for all panel members, the last two tests allow for individual unit roots for panel members. Table 2 presents the results derived from panel unit root tests at levels and first differences.

Table 2: Results of panel unit root tests

		Levels			***First differences***		
	Variables	***LLC***	***MW***	***Choi***	***LLC***	***MW***	***Choi***
Intercept	***LGDP***	-2.2211**	22.5386	25.0995	-5.3585***	33.2911**	38.3848***
	LHOR	-2.3078**	29.1346*	29.1465*	-14.0667***	103.629***	129.020***
	LSEK	-4.4826***	42.9476***	44.0269***	-11.648***	86.7293***	118.597***
Intercept and trend	***LGDP***	-1.8287**	10.0539	7.3528	-4.8791***	25.2685	31.7047**
	LHOR	-8.8321***	49.0596***	69.5379***	-11.882***	71.8462***	132.904***
	LSEK	-7.3670***	48.3089***	56.0088***	-13.6796***	66,7484***	133.344***
No intercept and trend	***LGDP***	6.5592	1.0378	1.0398	-5.0717***	60.5701***	59.1036***
	LHOR	0.4406	11.8518	14.5913	-13.7540***	133.046***	136.959***
	LSEK	-1.7917**	19.2544	26.1420	-10.8395***	122.524***	144.932***

Notes: ***, **, and * denote the rejection of the null hypothesis at 1%, 5% and 10% levels, respectively. Lag lengths for all tests are selected by Schwarz information criterion (SIC).

While most of the tests lead us to accept the existence of unit root at levels for all variables, all tests indicate that the first differences of all variables are stationary. Hence, we use first differenced values of all variables in the model.

Next, the existance of a long-term equilibrium relationship among Gross domestic product, Horizontal aid and Sectoral aid is explored. Pedroni [5,6] suggested a number of panel cointegration tests. Pedroni [6] derives seven panel cointegration test statistics, four are based on within-dimension, and three are based on between-dimension. Table 4 reports the results of the cointegration tests. Based on the p-values, corresponding to the seven test statistics, the null hypothesis of no cointegration can not be excepted at the significance level of 5% in all but six (out of 21) cases. Thus, it can be concluded that no cointegration exist between the variables.

Table 3: Results of Pedroni residual cointegration tests

	no deterministic intercept and trend		***no deterministic trend***		***deterministic intercept and trend***	
	statistic	***p-value***	***statistic***	***p-value***	***statistic***	***p-value***
Panel v-statistic	-2.4066	0.9919	-0.8963	0.8150	0.1121	0.4554
Panel rho-statistic	-0.3596	0.3596	0.7414	0.7708	2.6821	0.9963
Panel PP-statistic	-1.7485**	0.0402**	-2.1257**	0.0168**	0.9370	0.8256
Panel ADF-statistic	-1.3893*	0.0824*	-2.6253***	0.0043***	-0.4496	0.3265
Group rho-statistic	0.2458	0.5971	2.0967	0.9820	3.7617	0.9999
Group PP-statistic	-2.5544***	0.0053***	-2.3990***	0.0082***	1.1979	0.8845
Group ADF-statistic	-1.5713*	0.0581*	-2.8870***	0.0019***	-0.0150	0.4940

Notes: ***, **, and * denote the rejection of the null hypothesis at 1%, 5% and 10% levels, respectively. Lag length for based on Schwarz information criterion (SIC)

Therefore, the empirical properties of the examined variables require estimation of the VAR in first differences. To test for panel causality, a time-stationary VAR model may be employed in the following form:

$$\Delta \mathrm{LGDP}_{\mathrm{it}} = \alpha_0 + \sum_{\mathrm{k}=1}^{\mathrm{p}} \alpha_{\mathrm{k}} \Delta \mathrm{LGDP}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \beta_{\mathrm{k}} \Delta \mathrm{LHOR}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \gamma_{\mathrm{k}} \Delta \mathrm{LSEK}_{\mathrm{it\text{-}k}} + \varepsilon_{\mathrm{it}}, \quad (1)$$

$$\Delta \mathrm{LHOR}_{\mathrm{it}} = \delta_0 + \sum_{\mathrm{k}=1}^{\mathrm{p}} \delta_{\mathrm{k}} \Delta \mathrm{LGDP}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \theta_{\mathrm{k}} \Delta \mathrm{LHOR}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \vartheta_{\mathrm{k}} \Delta \mathrm{LSEK}_{\mathrm{it\text{-}k}} + \epsilon_{\mathrm{it}}, \quad (2)$$

$$\Delta \mathrm{LSEK}_{\mathrm{it}} = \varphi_0 + \sum_{\mathrm{k}=1}^{\mathrm{p}} \varphi_{\mathrm{k}} \Delta \mathrm{LGDP}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \omega_{\mathrm{k}} \Delta \mathrm{LHOR}_{\mathrm{it\text{-}k}} + \sum_{\mathrm{k}=1}^{\mathrm{p}} \sigma_{\mathrm{k}} \Delta \mathrm{LSEK}_{\mathrm{it\text{-}k}} + \nu_{\mathrm{it}}, \quad (3)$$

where Δ denotes first difference operator, i = 1,..., N denotes the countries, t = 1, …, T denotes the time period, $\alpha_0, \delta_0, \varphi_0$ are constant terms, $\alpha_k, \beta_k, \gamma_k, \delta_k, \theta_k, \vartheta_k, \varphi_k, \omega_k, \sigma_k$ autoregressive parameters, $\varepsilon_{it}, \epsilon_{it}, \nu_{it}$ are error terms and p number of lags. Based on LM test, Akaike information criterion, Schwarz information criterion and Final predictor error lag length two is selected.

Granger causality test will help us conclude whether past values of one variable affect another variable in the current period. These test results also indicate the directions of causal relationships between variables. Results of Granger causality test in the context of VAR model are presented in Table 4.

Table 4: Results of Granger causality test

Null Hypothesis	***Wald statistic***	***p-value***	***Conclusion***
1. GDP and HOR			
GDP does not Granger cause HOR	13.9079***	0.0010***	GDP → HOR
HOR does not Granger cause GDP	2.5790	0.2754	HOR --- GDP
2. GDP and SEK			
GDP does not Granger cause SEK	2.6837	0.2614	GDP --- SEK
SEK does not Granger cause GDP	7.1629**	0.0278**	SEK → GDP
3. SEK and HOR			
SEK does not Granger cause HOR	4.7778*	0.0917*	SEK --- HOR
HOR does not Granger cause SEK	0.1249	0.9395	HOR --- SEK

Notes: ***, **, and * denote the rejection of the null hypothesis at 1%, 5% and 10% levels, respectively. X →Y denotes that the null of no causal effect from X to Y can be rejected and X --- Y denotes that the null of no causal effect from X to Y cannot be rejected.

From the results of Granger causality test, at the significance level of 5% it can be concluded that in case of GDP and HOR, there is causality running from GDP to HOR. In the case of GDP and SEK causality runs from SEK to GDP. When we observe SEK and HOR there is no causality found. In other words, the changes in GDP will create changes in HOR but not in SEK. On the other hand, changes in HOR will not create changes in GDP but changes in SEK will. Also, there is no causal relationship between HOR and SEK.

5 CONCLUSIONS

The results show that state aid in the EU has no impact on GDP. This conclusion is consistent with Tunali and Fidrmuc [7]. The main conclusion that we can draw from this analysis is that changes in GDP are not result of changes in horizontal state aid expenditures. On the other hand, we can derive the conclusion based on influence of GDP on changes in horizontal aid meaning that the amount of expenditures for horizontal purposes depends on changes in GDP.

Results indicate that there exists causality from sectoral state aid to GDP. As sectoral state aid is intended to particular segments of economy bringing them in such way to a privileged position, over the years European Commission persists in reducing the share of sectoral state aid in total aid and in GDP. Continuous policy of the EU is to reduce the share of sector-specific aid due its negative impact on competition. Although in this paper the direction of relationship between state aid and GDP is not examined political framework guides to conclusion that this relation should be negative.

The impact of state aid expenditures on GDP growth in this paper is not proven leaving the place for further researches. This research could be extended in several different directions. At first the direction of relationship between GDP and horizontal and sectoral state aid could be examined. Horizontal state aid consists of different categories of state aid such as research and development and innovation, environmental protection and energy

saving, small and medium enterprises, training, employment, culture. It is possible to examine if every single of these state aid expenditures meet the intended goals.

References

[1] Choi, I. (2001). Unit root tests for panel data. Journal for international money and finance, Vol (20): 249–2726.

[2] Granger, C. W. J. (1969). Investigating causal relation by econometric and cross-sectional method. Econometrica, Vol (37): 424–438.

[3] Levin, A., Lin, C. F. and Chu, C. S. J. (2002). Unit root tests in panel data asymptotic and finite-sample properties". Journal of Econometrics, Vol (108): 1–24.

[4] Maddala, G. S. and Wu, S. (1999). "A comparative study of unit root tests with panel data and a new simple test". Oxford bulletin of economics and statistics, Vol (61): 631–652.

[5] Pedroni, P. (1999). Critical values for cointegrating tests in heterogeneous panels with multiple ¸ regressors. Oxford bulletin of economics and statistics, Vol (61): 653–670.

[6] Pedroni, P. (2004). Panel cointegration; asymptotic and finite application to the PPP hypothesis, Econometric theory, Vol (20): 597-625.

[7] Tunalı, Ç. B., & Fidrmuc, J. (2014). State Aid Policy in the European Union, Working Paper, No. 14-04, available at: http://www.brunel.ac.uk/__data/assets/pdf_file/0006/362913/1404.pdf [Accessed 20/9/2014].

TIME VARYING CAPM BETAS ON ZAGREB STOCK EXCHANGE

Tihana Škrinjarić
University of Zagreb, Department of Mathematics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
tskrinjaric@efzg.hr

Abstract: This paper employs a CCC GARCH(1,1) model in order to identify the volatility dynamics of stock market and sector indices on Zagreb Stock Exchange. Time varying CAPM betas are estimated in order to test whether portfolio formation based on the results can enhance portfolio performance. Based upon data on 5 sector indices from January 2nd 2012 to May 15th 2015, the results indicate that time varying betas should be taken into account when forming portfolios.

Keywords: CCC and DCC GARCH, Zagreb Stock Exchange, CAPM, time varying beta, stocks.

1 INTRODUCTION

Capital Asset Pricing Model's (CAPM) beta is a widely known concept today in finance [26, 32]. However, in its original form, model has pitfalls which have been criticized over the decades [19]. One of the most common critiques regards the assumption of beta being constant. Research has shown that betas vary over time [9, 15, 16, 17, 18, 19]. This means that the relationship between the risk and return is not linear. In that way, original model cannot explain movements in stock return and risk properly. Estimating CAPM by using OLS, results in inconsistent estimates of conditional alphas and betas [3]. The majority of the literature today agrees that varying dynamics on financial markets should be modelled in a proper manner. In the last two decades, more research which estimates time varying betas has emerged [7, 10, 22, 27, 35, etc.]. Mostly, regime switching models are used (for a list of references see [34]). However, when using many approaches, it is often assumed that there is no feedback from stock to market return. Since market indices are constructed based upon individual stocks or sectors, this is not a realistic assumption. This is why MGARCH models are becoming popular to estimate such dynamics. They are able to capture not only the volatility dynamics on financial markets, but the co-movements of financial assets as well [1, 2, 5, 6, 30]. Croatian capital market is a relatively new one; and has not yet been explored sufficiently. There exist only few papers which deal with CAPM betas, and every paper estimated the linear model with exception of one paper. [20] gave a warning to be careful when using their results referring to changing betas, [31] concluded that it is not advised to use beta as a suitable risk measure, [12] had conclusions similar to previous two papers. Paper [34] is the only one addressing nonlinearity in risk-return relationship. Markov-switching models have been used to estimate nonlinear CAPM. Results pointed out that betas vary over time, meaning that the linear CAPM is misleading. However, mentioned studies did not assume a feedback relationship from individual stocks to the market. This paper addresses this, by estimating MGARCH models which capture the return co-movements of 5 sub-indices with market index on the Croatian capital market. The structure of this research is as follows. Second section deals with methodology, while third section presents the results from the empirical analysis. Fourth section is the final one, giving conclusions.

2 METHODOLOGY

It is well known that CAPM beta is defined as [26, 32]: $\beta_i = \mathrm{Cov}(r_i, r_M)/\sigma^2_M$, where β_i represents beta for i-th stock, r_i excess return on the i-th stock and r_M excess market return. When using MGARCH models, formula becomes: $\beta_{i,t} = \mathrm{Cov}_t(r_{i,t}, r_{M,t})/\sigma^2_{M,t}$, where index t is added, representing time period t. As previously mentioned, financial returns and their volatilities move together in the majority of cases. That is why MGARCH models have been developed in order to capture these co-movements. In that way, they become very useful in

portfolio selection [28, 29]. Different models exist within this methodology. The Dynamic and Constant Conditional Correlation model (DCC, CCC) are very popular because they successfully capture dynamics on financial markets and they are more parsimonious compared to early developed models. They have also been used in previous research regarding this study. Engle [14] developed the DCC model:

$$\left.\begin{aligned} &\boldsymbol{r}_t = \boldsymbol{\Theta}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t = \boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_t \\ &\boldsymbol{\Omega}_t = \boldsymbol{D}_t^{1/2}\boldsymbol{R}_t\boldsymbol{D}_t^{1/2} \\ &\boldsymbol{R}_t = \operatorname{diag}\left(\boldsymbol{Q}_t\right)^{-1/2}\boldsymbol{Q}_t\operatorname{diag}\left(\boldsymbol{Q}_t\right)^{-1/2} \\ &\boldsymbol{Q}_t = \left(1-\theta_1-\theta_2\right)\boldsymbol{R} + \theta_1\tilde{\boldsymbol{\varepsilon}}_{t-1}\tilde{\boldsymbol{\varepsilon}}'_{t-1} + \theta_2\boldsymbol{Q}_{t-1} \end{aligned}\right\} \quad (1)$$

$\boldsymbol{r}_t$ is the (m,1) vector of returns, $\boldsymbol{\Theta}$ (m,k) matrix of parameters, $\boldsymbol{x}_t$ (k,1) vector of independent variables, $\boldsymbol{\varepsilon}_t$ (m,1) vector of innovation processes, $\boldsymbol{\Omega}_t^{1/2}$ the Cholesky factor (m,m) matrix of the conditional covariance matrix $\boldsymbol{\Omega}_t$ (m,m), $\boldsymbol{u}_t$ (m,1) vector of normal i.i.d. innovations[1], $\boldsymbol{D}_t$ (m,m) diagonal matrix of conditional variances, $\boldsymbol{R}_t$ (m,m) matrix of conditional correlations. $\boldsymbol{Q}_t$ is (m,m) variance and covariance matrix of standardized innovations, $\tilde{\boldsymbol{\varepsilon}}_t$ (m,1) vector of standardized innovations, $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{D}^{-1}\boldsymbol{\varepsilon}_t$, and $\boldsymbol{R}$ (m,m) positive definite unconditional correlation matrix. Univariate GARCH(1,1) model is used to model each variance in matrix $\boldsymbol{D}_t$: $\sigma_{i,t}^2 = \alpha_{0,i} + \alpha_{1,i}\varepsilon_{i,t-1}^2 + \beta_{1,i}\sigma_{i,t-1}^2$. $\alpha_{0,i}>0$, $\alpha_{1,i}\geq 0$ and $\beta_{1,i}\geq 0$ must hold for conditional variances to be strictly positive; and $\alpha_{1,i}+\beta_{1,i}<1$ must hold in order for them to be finite. The dynamics of conditional correlations is defined by nonnegative scalars θ_1 and θ_2. Moreover, it must hold $\theta_1+\theta_2<1$ for the model to be stationary [14]. Standardized innovations satisfy $E(\tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}'_t)=\boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix, $\operatorname{Cov}(\tilde{\varepsilon}_{i,t}^2,\tilde{\varepsilon}_{j,t}^2)=0\ \forall\ i\neq j$, and $\operatorname{Cov}(\tilde{\varepsilon}_{i,t}^2,\tilde{\varepsilon}_{j,t-k}^2)=0$, $k>0$ [11]. However, if $\theta_1=\theta_2=0$ holds, DCC model becomes CCC model. It is less flexible than DCC model, because it assumes that correlation between two assets is constant over time [33]. It was developed by Bollerslev [5] and (1) becomes:

$$\left.\begin{aligned} &\boldsymbol{r}_t = \boldsymbol{\Theta}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t = \boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_t \\ &\boldsymbol{\Omega}_t = \boldsymbol{D}_t^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{D}_t^{\frac{1}{2}} \end{aligned}\right\} \quad (2)$$

where it can be seen that matrix $\boldsymbol{R}$ is constant, hence the name of the model. More details on these models can be found in [1], [2], [5], [6] or [30].

3 EMPIRICAL ANALYSIS

Daily data on 5 sector indices and CROBEX was downloaded from ZSE [37]. It refers to closing prices of industry, construction, foods, tourism and transport. These indices are calculated on ZSE from January 2nd 2012 and data was collected up until May 15th 2015. In order to calculate excess returns, daily data on 6 month Treasury bill interest rate was collected from Global Financial Data [21][2]. All of the analysis was done in EViews 8. Returns were calculated as continuously compounded returns and excess returns were calculated by subtracting the T-bill interest rate from original returns. There is total of 840 observations for each index. Returns[3] were filtered by an appropriate ARMA(p,q) model[4]

[1] Usually, the assumption of multivariate normal distribution of $\boldsymbol{u}_t$ are made, because correctly specifying the conditional mean and variances results with consistent estimates [6, 14].
[2] Usually, in analysis a 3 month T-bill interest rate is used, but the data was available only up until 2014.
[3] Unit root tests indicated that all excess returns are trend stationary. Thus, trend was subtracted from all series. Detailed results are available upon request.

(following [9]) before estimating MGARCH models. In that way, the main models are estimated with less parameters. A bivariate DCC GARCH(1,1) was estimated for each index excess return and excess CROBEX return. Since in all of the models the parameters θ_1 and θ_2 were found to be statistically not significant and (or) negative[5], a bivariate[6] CCC GARCH(1,1) was estimated because there is no support for the assumption of varying correlation between each index and market index. The results of estimated models are shown in table 1. It can be seen that the industry sector is highly sensitive to market shocks compared to other sectors, with an alpha coefficient greater than 0.1 (0.156). This means that among CAPM betas which will be calculated, industry beta should be most volatile.

Table 1: Estimation results of bivariate CCC(1,1) GARCH

Coefficient	Subindex				
	Industry	Construction	Food	Transportation	Tourism
Mean equations					
μ_0	-0.0003 (0.00046)	-0.0002 (0.00068)	-0.0002 (0.00032)	-0.0001 (0.00045)	-0.0004 (0.00042)
μ_{CROBEX}	-0.00015 (0.00019)	-0.00013 (0.00019)	-0.00015 (0.00019)	-0.00012 (0.00019)	-0.00011 (0.00019)
Variance equations					
α_0	0.0003* (0.00001)	0.0003* (0.00001)	10^{-6} * ($6\cdot10^{-7}$)	0.0001 (0.0001)	0.00001* ($3\cdot10^{-6}$)
α_1	0.1559* (0.00231)	0.0649* (0.0162)	0.0474* (0.0106)	0.0360*** (0.0206)	0.0804* (0.0125)
β_1	0.6817* (0.0398)	0.8585* (0.0389)	0.9366* (0.0132)	0.8769* (0.0814)	0.8746* (0.0217)
$\alpha_{0,\text{CROBEX}}$	$7\cdot10^{-7}$ * ($3\cdot10^{-7}$)	10^{-7} * ($5\cdot10^{-8}$)	8^{-7} ** ($3\cdot10^{-7}$)	10^{-6} ** ($5\cdot10^{-7}$)	10^{-6} ** ($5\cdot10^{-7}$)
$\alpha_{1,\text{CROBEX}}$	0.0589* (0.0150)	0.0609* (0.0171)	0.0524* (0.0130)	0.0632* (0.0172)	0.0641* (0.0170)
$\beta_{1,\text{CROBEX}}$	0.9207* (0.0196)	0.8990* (0.0262)	0.9217* (0.0191)	0.9024* (0.0248)	0.8995* (0.0245)
$\rho_{i,\text{CROBEX}}$	0.3906* (0.0296)	0.3267* (0.0299)	0.4527* (0.0297)	0.3761* (0.0270)	0.2168* (0.0337)
Diagnostics					
Log Likelihood	5634.401	5324.692	5959.297	5661.837	5624.67
Akaike info. criteria	-13.426	-12.671	-14.184	-13.475	-13.387
Schwartz info. criteria	-13.375	-12.621	-14.134	-13.424	-13.336
Hannan-Quin info. criteria	-13.406	-12.652	-14.165	-13.456	-13.367
Resid LM 30 (p-v)	5.11 (0.276)	4.50 (0.343)	1.29 (0.864)	4.99 (0.288)	2.82 (0.588)
Sq. Resid LM 30 (p-v)	5.49 (0.241)	2.01 (0.734)	2.50 (0.645)	3.52 (0.476)	4.25 (0.373)

Source: author's calculation

Note: Standard errors are given in parenthesis with exception of p-values in the last two rows. *, ** and *** denote statistical significance on 1%, 5% and 10%. Resid LM 30 is LM test statistic for testing multivariate autocorrelation of standardized residuals up to lag of 30 days, Sq. Resid LM 30 is LM test statistic for testing multivariate heteroskedasticity of standardized residuals up to lag of 30 days. All of the evaluated diagnostic tests show that there is no autocorrelation and no heteroskedasticity up until lag 30. Correlation coefficients between standardized residuals of all are not statistically significant. Covariances are not significant and covariances between squared residuals up to lag 30 are also not significant[7]. All of the parameters in

[4] AR(1), MA(1), AR(2), MA(2) and ARMA(1,1) were considered. Parsimony was followed in combination with information criteria, log likelihood and the statistical significance of estimated parameters. All of the series were estimated as an ARMA(1,1) process with the exception of construction excess return, which was modelled as an AR(1) process. Detailed results are available upon request.

[5] The results were omitted due to limited number of pages but are available upon request.

[6] This approach has been used because using many variables in the system can result in biased estimates [4].

[7] Detailed results are available upon request.

GARCH(1,1) equations are positive which ensures a strict positivity of variances, and the condition $\hat{\alpha}_{1,i} + \hat{\beta}_{1,i} < 1$ holds, which ensures that conditional variances are finite.

Sector time varying betas have been calculated and are shown on Figure 1. It can be seen that all of the betas vary over the observed period. Most aggressive sector was construction, followed by industry. If we observe how much time did betas surpass the value of 1, it was 32.3% for industry, 71.04% for construction and 2.74%, 21.22% and 2.86% for food, transportation and tourism respectively. This indicates that aggressive investors should have invested more in construction, and opposite is valid for food and tourism. The most volatile beta was that of industry, as expected in the estimation. Its standard deviation is equal to 7.06%, and for other sectors is equal to 4.90%, 1.46%, 2.50% and 2.86% for construction, food, transportation and tourism.

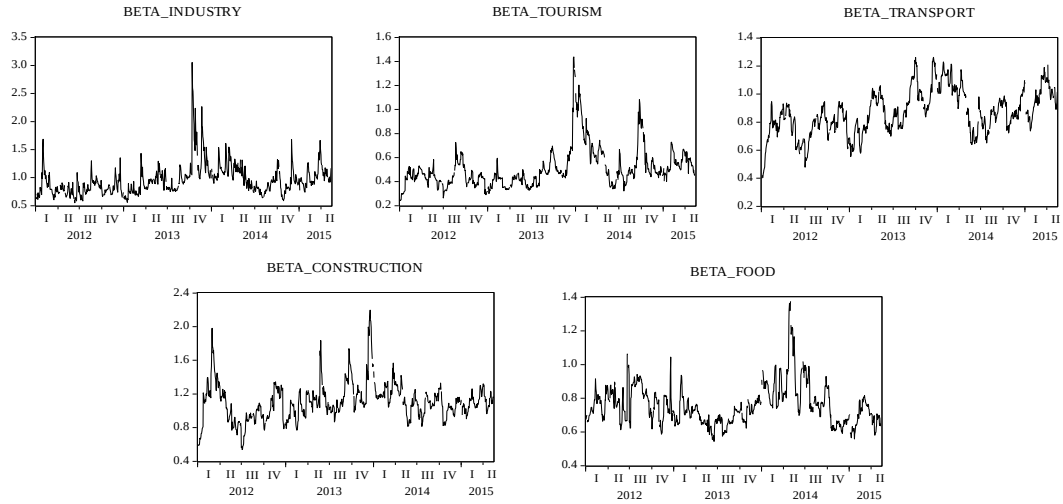


Figure 1: Sector betas calculated from CCC(1,1) GARCH results
Source: author

Next, it is useful to know whether dynamic betas can help to explain the variations of sector returns. A regression analysis was performed, in which 5 equations were estimated. The results are given in Table 2. Every beta is statistically significant with exception of food which means that the time varying beta helps to explain movements of sector returns. Although, these results should be taken with some caution because other variables influence sector returns (such as macroeconomic variables, Fama French factors, trade volume, etc.).

Table 2: Regression results, effects of betas on sector return

Sector beta:	Industry	Construction	Food	Transportation	Tourism
Estimated coefficient	0.0102	0.0100	-0.0042[a]	0.0376	0.0276
R^2	2.25%	0.95%	0.10%	10.58%	6.29%

Source: author
Note: [a] denotes that food beta is not statistically significant.

Since beta is a measure of how strongly excess returns change with respect to changes in excess market returns, it can be used as a guidance how to form portfolios. Such portfolios can benefit more when markets are bullish and suffer smaller loses when they are bearish. In order to see how results from this analysis help in forming portfolios a simulation was made. One portfolio was formed based upon holding equal weights of all indices and its beta was calculated as a simple average beta (Average portfolio). The other portfolio was formed with respect to the market movements (MGARCH portfolio). When the market index is falling, it is considered as bearish market, and when it is rising, it is considered as bullish. Assumption was made that when market is bearish, investor holds the subindex which has the lowest beta so he suffers lower loses and when it is bullish, he holds the subindex which has the highest beta in order to gain most. The results from two mentioned portfolios have been compared

and are shown in Table 3. First of all, if we compare average betas in each market state, it can be seen that MGARCH outperforms the Average portfolio in both states. When the market is bullish, it has a beta greater than other portfolio and greater than 1. Similarly, when market is bearish, it has a beta smaller than Average portfolio's beta. Average portfolio has a beta greater than 1 in bull market 13.30% of the time, whilst MGARCH portfolio has 73.89%. In bear markets, Average portfolio has beta smaller than 1 almost 85% of the time and MGARCH has it all the time. Finally, a *t*-test of the equality of two means was performed in order to see whether two portfolios performed differently. The results are not surprising that the difference is statistically significant.

Table 3: Portfolio simulations comparison

Market	Portfolio	Average beta	No >1 or < 1	% >1 or < 1	Max	Min	*t*-test (p-v)
Bull	Average	0.8343	54	13.30	1.2458	0.5158	21.57
	MGARCH	1.1457	300	73.89	2.5960	0.6309	(0.0000)
Bear	Average	0.8421	366	84.53	1.2350	0.5100	36.95
	MGARCH	0.5027	433	100	0.9628	0.2427	(0.0000)

Source: author

4 CONCLUSION

Dynamics on financial markets, which can be very turbulent, have been recognized a long time ago. That is why initial models in finance have been revised in the last couple of years. In that way, they take into account more realistic assumptions and result with more accurate predictions. This paper addresses the issue of time varying systematic risk, beta, on Zagreb Stock Exchange by using MGARCH models. The results indicate that it is advisable to assume that betas vary on the stock market. Using the output from these models can improve portfolio selection and rebalancing. In that way, investors have more control over their investments and can expect to achieve better results (in terms of risk and return); compared to times when ignoring such dynamics.

References

[1] Aielli, G. P. (2013). Dynamic Conditional Correlations: On Properties and Estimation. Journal of Business & Economic Statistics, 31(3): 282-299.

[2] Alexander, C. (2008). Practical Financial Econometrics, John Wiley & Sons, New York.

[3] Ang. A. and Chen, J. (2005). CAPM over the long run: 1926-2001. NBER Working Paper Series, National Bureau of Economic Research.

[4] Baumöhl, E. and Lyócsa, Š. (2014). Volatility and dynamic conditional correlations of worldwide emerging and frontier markets. Economic Modelling, 38: 175-183.

[5] Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. Review of Economics and Statistics, 72: 498–505.

[6] Bollerslev, T. and Wooldridge J.M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews 11: 143-172.

[7] Caporale, T. (2012). Time varying CAPM betas and banking sector risk. Economics Letters, 115: 293-295.

[8] Chen, R. (2009). Regime switching in volatilities and correlation between stock and bond markets. Discussion paper, 640. Financial Markets Group, London School of Economics and Political Science, London, UK.

[9] Chen, S. N. (1982). An examination of risk-return relationship in bull and bear markets using time-varying security betas. Journal of Financial and Quantitative Analysis, 17: 265-286.

[10] Choudhry, T., Lu, L. and Peng, K. (2010). Time-varying beta and the Asian financial crisis: Evidence from the Asian industrial sectors. Japan and the World Economy, 22: 228-234.

[11] Ding Z., and Engle R. F. (2001). Large scale conditional covariance matrix modeling, estimation and testing. Working Paper FIN-01-029, NYU Stern School of Business, New York.

[12] Džaja, J. and Aljinović, Z. (2013). Testing CAPM model on the emerging markets of the Central and Southeastern Europe. Croatian Operational Research Review (CRORR), 4: 164-175.

[13] Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics, 20: 339–350.
[14] Engle, R. F. (2009). Anticipating Correlations: A New Paradigm for Risk Management. Princeton, Princeton University Press, New York.
[15] Fabozzi, F. and Francis, J. C. (1977). Stability tests for alphas and betas over bull and bear market conditions. Journal of Finance, 32: 1093-1099.
[16] Faff, R. W.F. and Brooks, R. D. (1997). Further evidence on the relationship between beta stability and the length of estimation period. Advances in Investment Analysis and Portfolio Management, 4: 97-113.
[17] Fama E. and French, K. (1992). The cross-section of expected stock return. Journal of Finance, 47: 427-465.
[18] Fama E. and French, K. (1996). Multifactor Explanations of Asset Pricing Anomalies. Journal of Finance, 51(1): 55-84.
[19] Fama, E. and French, K. (2004). The Capital Asset Pricing Model: Theory and Evidence. Journal of Economic Perspectives, 18(3): 25-46.
[20] Fruk, M. and Huljak, I. (2004). Testiranje Sharpe-Linterova modela na Zagrebačkoj burzi. Financijska teorija i praksa, 28(1): 77-91.
[21] Global Financial Data (2015). https://www.globalfinancialdata.com/GFDPlatform/Login.aspx [Accessed 19/May/2015].
[22] Groenewold, N. and Fraser, P. (1999). Time-varying estimates of CAPM betas. Mathematics and Computers in Simulation, 48: 531-539.
[23] Jagannathan, R. and Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. The Journal of Finance, 51(1): 3-53.
[24] Kim, C. J., Morley, J. C. and Nelson, C. R. (2004). Is there a positive relationship between stock market volatility and the equity premium? Journal of Money, Credit, and Banking, 36: 339-60.
[25] Kunovac, D. (2011). Assymetric correlation on the Croatian equity market. Financial Theory and Practice, 35(1): 1-24.
[26] Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics, 47(1): 13-37.
[27] Long, L. and Tsui, A. K. (2014). Estimating time-varying currency betas with contagion: New evidence from developed and emerging financial markets. Japan and the World Economy, 30: 10-24.
[28] Longin F. and Solnik B. (1995). Is the correlation in international equity returns constant: 1960-1990?. Journal of International Money and Finance, 14: 3-26.
[29] Longin F. and Solnik B. (2001). Extreme correlation in international equity markets. Journal of Finance, 56: 649−676.
[30] Lüketpohl, H. (2006). New Introduction to Multiple Time Series Analysis, Springer, Berlin.
[31] Perković, A. (2011). Research of beta as adequate risk measure - is beta still alive? Croatian Operational Research Review (CRORR), 2: 102-111.
[32] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19 (3): 425–442.
[33] Silvennoinen, A. and Teräsvirta, T. (2009). Multivariate GARCH models. Handbook of Financial Time Series, (ed. T. G. Andersen, R. A. Davis, J.-P. Kreis, and Mikosch, T.), 201–229, Springer, Berlin.
[34] Škrinjarić, T. (2014). Testing for regime-switching CAPM on Zagreb Stock Exchange. Croatian Operational Research Reveiw 5(2): 119-133.
[35] Tai, C-S. (2000). Time-varying market, interest rate, and exchange rate risk premia in the US commercial bank stock returns. Journal of Multinational Financial Management, 10: 397-420.
[36] Tsay R. S. (2010). Analysis of Financial Time Series, Third Edition (Wiley Series in Probability and Statistics), John Wiley & Sons, New Jersey.
[37] Zagreb Stock Exchange (2015). http://www.zse.hr [Accessed 19/May/2015].

DYNAMIC MODELING OF STOCK AND BOND RETURN CORRELATION IN CROATIA

Tihana Škrinjarić and Boško Šego
University of Zagreb, Department of Mathematics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
tskrinjaric@efzg.hr, bsego@efzg.hr

Abstract: This paper employs a DCC GARCH model on Croatian stock and bond market in order to indentify the dynamic relationship between stock and bond return correlation. In that way, dynamic conditional correlations are estimated for the period June 1st 2004 to May 18th 2015 on Zagreb Stock Exchange. The results from the estimation are used in order to calculate hedge ratio and optimal portfolio weights, as well to compare the results from multiple simulated portfolios.

Keywords: DCC GARCH, Zagreb Stock Exchange, stock and bond returns, risk hedging, volatility.

1 INTRODUCTION

Since the financial crisis in 2008, a lot of focus and research of academics and practitioners has been shifted again towards volatility. Many financial markets have been affected by the mentioned crisis, such as the Croatian market. Since then, the trade is stagnating and investors are careful when investing in financial assets (see [29]). Thus, there exists a need in finding trading strategies which could try to ensure investors achieving good returns (greater than stock market return or some goal return) and successfully managing risks. The importance of risk management and portfolio diversification has been known for many decades now (beginnings range from [26]). In the last two decades, many econometric models have been developed in order to capture and describe the co-movements of financial asset returns. A famous class of models are multivariate GARCH models (MGARCH henceforth) which have been found to successfully capture the interdependence of financial assets return volatilities. Much research has been published on volatility transmissions between different markets in the last couple of years. Majority of research has focused on stock markets to explore interactions and volatility spillovers between countries. Since the number of papers in this area is growing rapidly, here we mention studies closely linked to this one. CEE counties have been examined in the last couple of years in order to identify volatility transmissions between them and developed markets [8, 15, 12, 17, 28, 24, 30, 32, etc.]. However, there is a lack of studies which focus on sector diversifications on particular markets [13, 14, 16, 27]. Therefore, the aim of this paper is to use a MGARCH model in order to estimate volatility co-movements between stock and bond market in Croatia, which has not yet been analyzed. The structure of the paper is as follows. Second section deals with methodology used in this study, while the results are shown in the third section. The final, fourth section concludes the paper.

2 METHODOLOGY

A widely accepted fact in this area of research is that financial volatilities move together across markets and time (see [5]). MGARCH models are broadly being used in research over the last decade, because they enhance portfolio selection, asset pricing models, and hedging (see [22, 23] and [20]). Many different specifications of MGARCH models have been developed depending on the topic investigated. By examining the previous literature, it can be seen that the DCC(1,1) GARCH is broadly accepted due to its power to capture the interactions between volatilities and realistic assumption that correlation between financial

asset returns changes over time. Thus, the DCC(1,1) GARCH of Engle [10] model will be applied in this study. Engle [10] proposed the following model:

$$\left.\begin{aligned} \boldsymbol{r}_t &= \boldsymbol{\Theta}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t &= \boldsymbol{\Omega}_t^{\frac{1}{2}}\boldsymbol{u}_t \\ \boldsymbol{\Omega}_t &= \boldsymbol{D}_t^{\frac{1}{2}}\boldsymbol{R}_t\boldsymbol{D}_t^{\frac{1}{2}} \\ \boldsymbol{R}_t &= \operatorname{diag}\left(\boldsymbol{Q}_t\right)^{-\frac{1}{2}}\boldsymbol{Q}_t\operatorname{diag}\left(\boldsymbol{Q}_t\right)^{-\frac{1}{2}} \\ \boldsymbol{Q}_t &= \left(1-\theta_1-\theta_2\right)\boldsymbol{R} + \theta_1\tilde{\boldsymbol{\varepsilon}}_{t-1}\tilde{\boldsymbol{\varepsilon}}'_{t-1} + \theta_2\boldsymbol{Q}_{t-1} \end{aligned}\right\} \quad (1)$$

where $\boldsymbol{r}_t$ is (m,1) vector of returns, $\boldsymbol{\Theta}$ (m,k) matrix of parameters, $\boldsymbol{x}_t$ (k,1) vector of independent variables, $\boldsymbol{\varepsilon}_t$ (m,1) vector of innovation processes, $\boldsymbol{\Omega}_t^{1/2}$ the Cholesky factor (m,m) matrix of the conditional covariance matrix $\boldsymbol{\Omega}_t$ (m,m), $\boldsymbol{u}_t$ (m,1) vector of normal i.i.d. innovations, $\boldsymbol{D}_t$ (m,m) a diagonal matrix of conditional variances, $\boldsymbol{R}_t$ (m,m) matrix of conditional correlations. $\boldsymbol{Q}_t$ is (m,m) variance and covariance matrix of standardized innovations, $\tilde{\boldsymbol{\varepsilon}}_t$ is (m,1) vector of standardized innovations, $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{D}_t^{-1}\boldsymbol{\varepsilon}_t$, and $\boldsymbol{R}$ is (m,m) positive definite unconditional correlation matrix. Variances in matrix $\boldsymbol{D}_t$ are modelled by a univariate GARCH(1,1) model: $\sigma_{i,t}^2 = \alpha_{0,i} + \alpha_{1,i}\varepsilon_{i,t-1}^2 + \beta_{1,i}\sigma_{i,t-1}^2$. Of course, in order to conditional variances to be strictly positive, it must hold $\alpha_{0,i} > 0$, $\alpha_{1,i} \geq 0$, $\beta_{1,i} \geq 0$, and in order to be finite, $\alpha_{1,i} + \beta_{1,i} < 1$. Parameters θ_1 and θ_2 are nonnegative scalars which govern the dynamics of conditional correlations. In order to model to be stationary, it must hold $\theta_1 + \theta_2 < 1$ ([10]). Standardized innovations should satisfy the following conditions: $E\left(\tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}'_t\right) = \boldsymbol{I}_m$, where $\boldsymbol{I}_m$ is the identity matrix, $\operatorname{Cov}\left(\tilde{\varepsilon}_{i,t}^2, \tilde{\varepsilon}_{j,t}^2\right) = 0, \forall i \neq j$ and $\operatorname{Cov}\left(\tilde{\varepsilon}_{i,t}^2, \tilde{\varepsilon}_{j,t-k}^2\right) = 0, k > 0$ (see [9]). The estimation of this model is done in two steps. In the first step likelihood functions of individual GARCH processes are maximized in order to obtain estimates of $\alpha_{1,i}$ and $\beta_{1,i}$. In the second step, likelihood function is maximized in order to obtain estimates of θ_1 and θ_2. Although financial series are mostly not normally distributed, the assumptions made on $\boldsymbol{u}_t$ are justified in [6] and [10]. Consistent estimator can be obtained even if the data generating process is not Gaussian, as long as the conditional mean and the conditional variance are specified correctly. For more details see [10], [11], [31], [1], [2], [6], [11] or [25].

The results from estimating MGARCH models can be used in many ways. Since investors are interested in managing risks, conditional volatility estimates from the model can be used to calculate hedge ratios. [18] reviewed hedging strategies, of which a naive one is used in this study: investor wants to minimize the exposure by hedging a long position in one asset with a short position in the second one. Thus, the hedge ratio h_t is calculated as follows (see [4, 19] for the derivation of the formula): $h_t = \sigma_{12,t} / \sigma^2_{j,t}$, where $\sigma_{12,t}$ is the conditional covariance between two assets in moment t and $\sigma^2_{j,t}$ is the conditional variance of asset j in moment t. Moreover, the results can be used to construct optimal portfolio weights of CROBEX and CROBIS in the portfolio, by using the following expression (see [18]): $w_{1,t}^* = \left(\sigma_{1,t}^2 - \sigma_{12,t}\right) / \left(\sigma_{1,t}^2 - 2\sigma_{12,t} + \sigma_{2,t}^2\right)$. Optimal portfolio weight of the other asset is simply $1 - w_{1,t}^*$. If short selling is not allowed, the following constraints are made:

$$w_{1,t}^* = \begin{cases} 0, & \text{if } w_{1,t} < 0 \\ w_{1,t}^*, & \text{if } 0 \leq w_{1,t} \leq 1 \\ 1, & \text{if } w_{1,t} > 1 \end{cases}. \quad (2)$$

3 EMPIRICAL ANALYSIS

In order to estimate the DCC(1,1) GARCH model, data on CROBEX and CROBIS indices was obtained from Zagreb Stock Exchange [33]. It consists of daily data ranging from June 1st 2004 to May 18th 2015 (2735 daily observations). The calculations and estimation were performed in EVeiws 8. Returns were calculated as continuously compounded returns. Unit root tests have resulted with conclusion that both return on CROBEX and CROBIS are stationary. Each of the return series was filtered by appropriate ARMA(*p*,*q*) model (following [7]). In that way, less parameters will be estimated in the DCC(1,1) GARCH model[1]. Mean equations were modelled as constants due to the filtered data and the individual conditional variances were modelled as GARCH(1,1) processes. The estimation was performed with the assumption of multivariate Normal distribution, and BFGS optimization algorithm was used. An additional model of asymmetric DCC(1,1) GARCH was estimated in order to look for asymmetry in the volatility, but the asymmetry parameter was shown to be statistically not significant[2]. The results are shown in Table 1. As it can be seen, the DCC(1,1) GARCH model describes the data well. CROBEX's volatility is more influenced by market shocks compared to CROBIS's volatility; it is highly sensitive to market shocks. The persistency of volatility is greater for CROBIS returns; it takes a long time for shocks to die after they occur. This can be seen on Figure 1, where conditional variances for CROBEX and CROBIS are shown. The impact of the financial crisis in 2008 has caused CROBEX's volatility to react more than CROBIS's. On the other hand, the greater persistency of CROBIS's volatility has caused the shocks to die after CROBEX's volatility (in the beginning of 2010 compared to CROBEX in the middle of 2009).

Table 1: Estimation results of DCC(1,1) GARCH model

	$\hat{\mu}_{CROBEX}$	$\hat{\mu}_{CROBIS}$
Mean equation	-0.00001	0.00001
	(-0.000139)	(0.000027)

	$\hat{\alpha}_{0,CROBEX}$	$\hat{\alpha}_{0,CROBIS}$	$\hat{\alpha}_{1,CROBEX}$	$\hat{\beta}_{1,CROBIS}$	$\hat{\alpha}_{1,CROBIS}$	$\hat{\beta}_{1,CROBIS}$
Variance equations	$9.6\cdot10^{-7}$	$4.42\cdot10^{-8}$	0.101875*	0.897715*	0.073415*	0.916906*
	$(3.52\cdot10^{-7})$	$(3.88\cdot10^{-6})$	(0.020582)	(0.016948)	(0.016076)	(0.01675)

$\hat{\theta}_1$	0.002471***
	(0.001342)
$\hat{\theta}_2$	0.996152*
	(0.002196)

Diagnostics					
Resid LM (30) (p-value)	Sq. Resid LM (30) (p-value)	Log likelihood	AIC	SIC	HQIC
8.7531 (0.0676)	4.6645 (0.3235)	22603.99	-16.5214	-16.4976	-16.5128

Source: authors' calculation

Note: *, ** and *** denote statistical significance on 1%, 5% and 10%. Estimated values are given in first rows and standard errors in parenthesis. Resid LM (30) is LM test statistic for testing multivariate autocorrelation of standardized residuals for the lag of 30 days, Sq. Resid LM (30) is LM test statistic for testing multivariate heteroskedasticity of standardized residuals for the lag of 30 days. p-values are given in parenthesis. AIC, SIC and HQIC stand for Akaike, Schwartz and Hannan-Quin information criteria. All of the evaluated diagnostic tests show that there is no autocorrelation and no heteroskedasticity up until lag 30. Following [9], the correlation coefficient between standardized residuals is equal to 0.03; it is not statistically significant (p-value 0.104). Covariance between squared residuals is equal to 0.28 and is also not significant (p-v 0.051). Covariance between squared residuals of CROBEX and lagged squared residuals of CROBIS up to lag 30 is equal to -0.11 and is also not significant (p-v 0.444). All of the parameters in GARCH(1,1) equations are positive which ensures a strict positivity of variances, as well as the condition $\hat{\alpha}_{1,i} + \hat{\beta}_{1,i} < 1$, $i \in \{\text{CROBEX}, \text{CROBIS}\}$ holds for both conditional variances which ensures that they are finite. θ_1 and θ_2 are significant which excludes CCC model; the condition $\hat{\theta}_1 + \hat{\theta}_2 < 1$ is met, which ensures the stationarity of the model.

[1] The results are omitted due to the limited number of pages of the paper but are available upon request. The following models were considered: AR(1), MA(1), AR(2), MA(2) and ARMA(1,1). Parsimony was followed, as well as was information criteria taken into account, log likelihood and the statistical significance of estimated parameters. Return on CROBEX was estimated as an ARMA(1,1) model, whilst return on CROBIS was estimated as an AR(1) model.

[2] Again, the results are omitted, but are available upon request. The assumption of *t*-distribution of DCC(1,1) GARCH model was not considered because the asymptotic properties of estimated parameters in this case are not fully investigated.

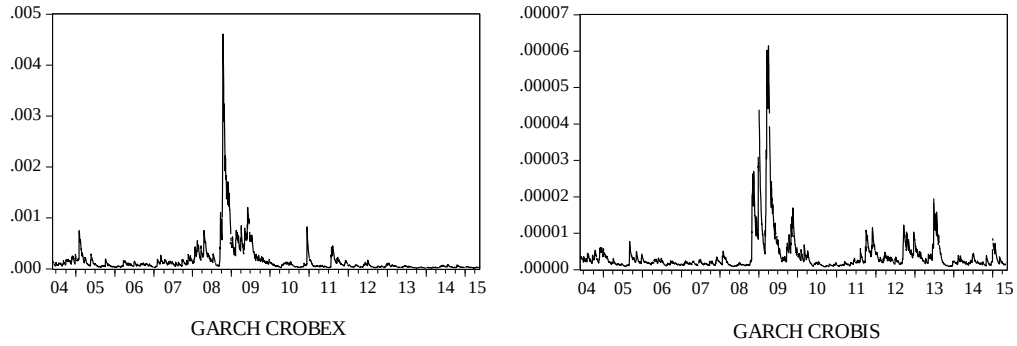
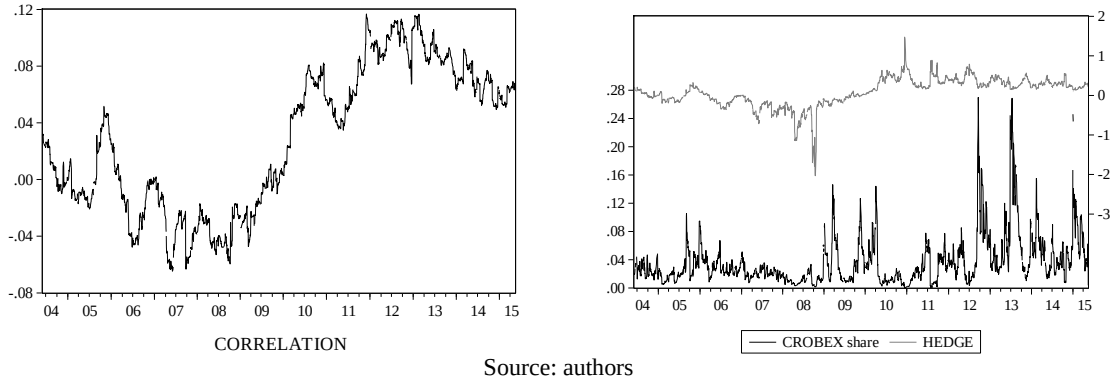


Figure 1: CROBEX and CROBIS conditional variances

Figure 2 (left panel) shows time-varying conditional correlation between CROBEX and CROBIS returns. The correlation had a downward trend before the financial crisis in 2008, which was favourable for diversification. In the crisis and afterwards there is an upward trend in the observed correlation (up until middle of 2013). This is in accordance with previous findings in the literature, where authors agree that correlations tend to rise in bear markets and decline in bull markets (see [3], [22, 23], [21]). In the last two years of the observed period, the correlation is facing a declining trend which is again favourable for investors.



Source: authors

Figure 2: CROBEX and CROBIS correlation (left panel) and CROBEX's optimal share (right panel) (black line, left axis) and hedge ratio (grey line, right axis)

Optimal hedge ratio and CROBEX's share in the optimal portfolio have been calculated and are shown on figure 2 (right panel). Looking at the share of CROBEX in the simulated portfolio, it does not exceed 30% of the portfolio in the entire observed period. Before the crisis, it should have been on average 2.3% of the portfolio, and afterwards 3.7% from 2009 to 2012, and in the period from 2013 until the end of the period 6.1%. A much greater share in the portfolio should have been of CROBIS, due to the greater standardized returns over the whole period. In that way, investors should be more drawn to achieving bigger returns over the same amount of risk. The conditional volatilities and covariance were used to construct hedge ratio with a short position in CROBIS. As it can be seen, this ratio varies from −2 to +1.5. Before the crisis the short position should have been taken in CROBEX (due to negative values) and after the crisis in CROBIS. Lastly, few portfolios have been simulated with their expected returns and risks constructed based upon the return series, conditional variances and covariance. In that way the portfolio which is a result from the analysis above can be compared to some benchmarks. First one, named MGARCH_portfolio is based upon the calculated CROBEX share on Figure 2. The second one, called EW_portfolio consists of 50% of CROBEX, third one is called R_portfolio which consists of 80% of CROBEX (risk seeking investors) and fourth one is called S_portfolio which consists of 80% of CROBIS

(risk averse investors)[3]. Bold numbers indicate the portfolio with the biggest return, smallest risk and biggest standardized return in the observed period. As expected, the portfolio simulated from the results of the MGARCH estimation is the best one, both in terms of risk and return.

Table 2: Descriptive statistics of 4 simulated portfolios

Statistics/portfolio	EW_portfolio	MGARCH_portfolio	R_portfolio	S_portfolio
Average return	0.00009	**0.00016**	0.00013	0.00006
Risk	0.000043	**0.000004**	0.000107	0.000009
Standardized return	8.192019	**27.24548**	3.913608	22.35665

Source: authors

4 CONCLUSION

In portfolio management, investors seek to earn great returns while trying to minimize risks they are exposed at. Since financial markets exhibit great dynamics in terms of return volatility, it has become almost an imperative to use MGARCH models when describing the co-movements of financial return series. This study employed DCC(1,1) GARCH model on Croatian market in order to identify dynamics between stock and bond returns. The results indicate that correlation varies over time, which has to be taken into account when managing portfolios. Portfolio formed based upon the results from the analysis is superior (with respect to return and risk) to benchmark portfolios simulated in the paper. This shows that MGARCH models are a useful tool when forming portfolios. In that way, investors could have more control over risk and try to beat the market. Some of the shortfalls of this study were fact that only stock and bond market indices were examined. This is because Croatian financial markets are underdeveloped. Thus, future research will try to implement MGARCH models on other financial assets, as well as look at sector diversification. Consequently, investors will have a clearer picture of investment opportunities on Croatian financial markets.

References

[1] Aielli, G. P. (2013). Dynamic Conditional Correlations: On Properties and Estimation. Journal of Business & Economic Statistics, 31(3): 282-299.

[2] Alexander, C. (2008). Practical Financial Econometrics, John Wiley & Sons, New York.

[3] Ang, A. and Bekaert, G. (2002). International Asset Allocation with Regime Shifts. Review of Financial Studies, 15(4): 1137-1187.

[4] Baille, R. T. and Myers, R.J. (1991). Bivariate Garch Estimation of the Optimal Commodity Futures Hedge. Jorunal of Applied Econometrics, 6(2): 109-124.

[5] Bauwens, L., Laurent, S. And Rombouts, J.V.K. (2006). Multivariate GARCH Models: A Survey. Journal of Applied Econometrics, 21: 79-109.

[6] Bollerslev T. and Wooldridge J.M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews 11: 143-172.

[7] Chen, R. (2009). Regime switching in volatilities and correlation between stock and bond markets. Discussion paper, 640. Financial Markets Group, London School of Economics and Political Science, London, UK.

[8] Dajčman, S. (2013). Dependence between Croatian and European stock markets – A copula GARCH approach. Zbornik radova Ekonomskog fakulteta u Rijeci, 31(2): 209-232.

[3] 80% was chosen arbitrary. In total, 19 different portfolios have been simulated, with CROBIS share ranging from 5% to 95% (with a 5 percentage points difference intervals). All of the simulated portfolios had returns smaller, as well as risks greater than the MGARCH_portfolio. Detailed results are available upon request.

[9] Ding Z., and Engle R. F. (2001). Large scale conditional covariance matrix modeling, estimation and testing. Working Paper FIN-01-029, NYU Stern School of Business, New York.
[10] Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics, 20: 339–350.
[11] Engle, R. F. (2009). Anticipating Correlations: A New Paradigm for Risk Management. Princeton, Princeton University Press, New York.
[12] Gelos, R.G. and Sahay, R. (2001). Financial market spillovers in transition economies. Economics of Transition, 9(1): 53-86.
[13] Hassan, S. A. and Malik, F. (2007). Multivariate GARCH modelling of sector volatility transmission. The Quarterly Review of Economics and Finance. 47: 470-480.
[14] Ho, K-Y. and Tsui, A. K. C. (2004). An Analysis of the Sectoral Indices of Tokyo Stock Exchange: A Multivariate GARCH Approach with Time Varying Correlations. Stochastic Finance, Autumn School and International Conference.
[15] Horvath, R. And Petrovski, D. (2013). International stock market integration: Central and South Easter Europe compared. Economic Systems, 37: 81-91.
[16] Katzke, N. (2013). South African Sector Return Correlations: using DCC and ADCC Multivariate GARCH techniques to uncover the underlying dynamics. Stellenbosch Economic Working Papers, No. 17/13, University of Stellenbosch.
[17] Kenourgios, D. And Samitas, A. (2011). Equity market integration in emerging Balkan markets. Research in International Business and Finance, 25: 296–307.
[18] Kroner, K. F. and Ng, V. K. (1998). Modeling asymmetric comovements of asset returns. Review of Financial Studies, 11: 817-844.
[19] Kroner, K. F. and Sultan, J. (1993). Time Varying Distributions and Dynamic Hedging with Foreign Currency Futures. Journal of Financial and Quantitative Analysis, 28: 535-551.
[20] Lien D. and Tse Y. K. (2002). Some recent developments in futures hedging. Journal of Economic Surveys, 16: 357-396.
[21] Lin, W. L., Engle, R. F., and Ito, T. (1994). Do bulls and bears move across borders? International transmission stock returns and volatility. Review of Financial Studies, 7: 507−538.
[22] Longin F. and Solnik B. (1995). Is the correlation in international equity returns constant: 1960-1990?. Journal of International Money and Finance, 14: 3-26.
[23] Longin, F., and Solnik, B. (2001). Extreme correlation in international equity markets. Journal of Finance, 56: 649−676.
[24] Lucey, B. M. and Voronkova, S. (2008) Russian equity market linkages before and after the 1998 crisis: Evidence from stochastic and regime-switching cointegration tests. Journal of International Money and Finance, 27: 1303-1324.
[25] Lüketpohl, H. (2006). New Introduction to Multiple Time Series Analysis, Springer, Berlin.
[26] Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1): 77-91.
[27] Righia, M. B. and Ceretta, P. S. (2012). Multivariate generalized autoregressive conditional heteroskedasticity (GARCH) modelling of sector volatility transmission: A dynamic conditional correlation (DCC) model approach, African Journal of Business Management, 6(27):8157-8162.
[28] Schotman, P. C. and Zalewska, A. (2006) Non-synchronous trading and testing for market integration in Central European emerging countries. Journal of Empirical Finance, 13: 462-494.
[29] Škrinjarić, T. and Besek, B. (2014). Pre and Post Crisis Performance Measurement of Croatian Stock Market. Zagreb International Review of Economics and Business, XVII(2): 97-116.
[30] Syllignakis, M. N. and Kouretas, G.P. (2011). Dynamic correlation analysis of financial contagion: Evidence from the Central and Eastern European markets. International Review of Economics and Finance, 20: 717-732.
[31] Tsay R. S. (2010). Analysis of Financial Time Series, Third Edition (Wiley Series in Probability and Statistics), John Wiley & Sons, New Jersey.
[32] Wang, P. And Moore, T. (2008). Stock market integration for the transition economies: time-varying conditional correlation approach. The Manchester School, Supplement 2008: 116-133.
[33] Zagreb Stock Exchange (2015). http://www.zse.hr [Accessed 19/May/2015].

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 7:

Location and Transport

A POLYNOMIAL ALGORITHM FOR A PARTICULAR OBNOXIOUS FACILITY LOCATION PROBLEM

Peter Czimmermann
University of Žilina, Žilina, Slovakia, peter.czimmermann@fri.uniza.sk
Štefan Peško
University of Žilina, Žilina, Slovakia, stefan.pesko@fri.uniza.sk

Abstract: The topic of obnoxious facility locations includes various types of problems and models related to the siting of environmentally hazardous facilities (such as garbage dumps, incinerator plants, water purification plants, chemical factories, nuclear reactors, etc.). One of these problems (defined by E. Minieka) is finding a discrete, weighted anti-p-center in a network. In our contribution, we present an exact polynomial algorithm for this problem and we suggest certain generalisations of the discrete weighted anti-p-center problem, which better represent real situations.

Keywords: obnoxious location, anti-p-center, polynomial algorithm

1 INTRODUCTION AND DEFINITIONS

A lot of papers deal with facility location problems, which are obviously modelled by weighted p-centers and weighted p-medians in networks ([3, 5]). It means that we need to find "the closest" place for location facilities (emergency medical stations, fire or police stations). However, over the last few years there has been much attention paid to the problem of the optimal siting of unattractive facilities, for instance, the locating environmentally hazardous facilities as far as possible from local inhabitants. These problems are called the obnoxious facility location problems. Appropriate mathematical models and algorithms for these problems has been studied by Church and Garfunkel [1], Minieka [7], Moon and Chaudry [8], Erkut and Neuman [2].

It was proved for almost all of these problems, that they are $\mathcal{NP}$-hard problems. One of the few exceptions is the discrete p-maxmin problem defined by I.D.Moon and S.S.Chaudry in [8]. C.M.Klein and R.K.Kincaid present the exact polynomial algorithm that solves this problem in [6]. In our contribution, we concern ourselves with the discrete anti-p-center problem formulated by E.Minieka in [7] (this problem is also called: the discrete p-maxmax problem). In the next section we show the exact polynomial algorithm for this problem. We also make an analysis of its efficiency and time complexity. The anti-p-center models situations in which we need to locate obnoxious facilities under the condition that the pollution or a large ecological disaster (such as a radioactivity leakage) would affect the least possible number of inhabitants in a given area - on the assumption that the pollution or contamination decreases with distance from the place of disaster.

1.1 Definitions

A network is an ordered quadruple $G = (V, E, d, w)$, where G is a connected undirected graph, $V \neq \emptyset$ is a set of vertices, E is a set of edges, c: $E \to R_o^+$ is a function which assigns to each edge a non-negative real number - its length, and w: $V \to R_o^+$ is a real non-negative function, its values are weights of vertices (the weight of vertex is the number of inhabitants in the vertex).

The distance $d(u, v)$ between vertices u and v is a length of the shortest path from u to v. The distance between a vertex v and a set of vertices $U \subset V$ is

$$d(v, U) = \min_{x \in U} d(v, x)\,.$$

The eccentricity of a set U is

$$ec(U) = \max_{v \in V} d(v, U) = \max_{v \in V} \min_{x \in U} d(v, x)\,.$$

The weighted eccentricity is

$$ecc(U) = \max_{v \in V} w(v) \cdot d(v, U)\,.$$

Let p be a natural number. Then the p-center of a network is a p-element subset of V with the minimum eccentricity. It means, we need to find the p-set object to

$$\min_{U \subset V, |U|=p} ec(U) = \min_{U \subset V, |U|=p} \max_{v \in V} \min_{x \in U} d(v, x)\,.$$

The weighted p-center is a p-subset with the minimum weighted eccentricity:

$$\min_{U \subset V, |U|=p} ecc(U) = \min_{U \subset V, |U|=p} \max_{v \in V} w(v) \cdot d(v, U)\,.$$

The anti-p-center is a p-subset with the largest eccentricity:

$$\max_{U \subset V, |U|=p} ec(U) = \max_{U \subset V, |U|=p} \max_{v \in V} \min_{x \in U} d(v, x)\,.$$

Similarly, the weighted anti-p-center is a p-subset with the largest weighted eccentricity:

$$\max_{U \subset V, |U|=p} ecc(U) = \max_{U \subset V, |U|=p} \max_{v \in V} w(v) \cdot d(v, U)\,.$$

In many practical location problems, we do not choose a set U from the whole set V, but we have prescribed the set of vertices $V_1 \subset V$. In this case, our definitions are as follows:

$$\max_{U \subset V_1, |U|=p} ec(U) = \max_{U \subset V_1, |U|=p} \max_{v \in V} \min_{x \in U} d(v, x)\,,$$

$$\max_{U \subset V_1, |U|=p} ecc(U) = \max_{U \subset V_1, |U|=p} \max_{v \in V} w(v) \cdot d(v, U)\,.$$

In such cases, we say that the anti-p-center is from the set V_1, but this notation will be omitted in the paper, because the basic set will be clear from the context.

2 P-MAXMAX ALGORITHM

In this section, the exact polynomial algorithm for weighted anti-p-center in networks will be introduced. It will be called p-maxmax algorithm.

2.1 Description of the algorithm p-maxmax

Let $G = (V, E, c, w)$ be a network, $V = \{v_1, \ldots, v_n\}$ its set of vertices and $I = \{1, \ldots, n\}$ be a set of indices. Let $X \subseteq V$ be a set of vertices which are potential obnoxious facility location sites (we suppose that $|X| = m \geq p$) and J be the set of indices of vertices from X. Let $D = (d_{i,j})_{n,m}$ be the matrix of weighted distances between the vertices of V and X. Let the elements of every row of D be sorted in a descending order. The sorted matrix is denoted by $\overline{D} = (\overline{d}_{i,j})_{n,m}$. Let the sorted i-th row be given by mapping $\pi_i : \{1, \ldots, m\} \to J$:

$$\overline{d}_{i,j} = d_{i,\pi_i(j)}\,.$$

Then the largest element in column p is chosen:

$$\overline{d}_{x,p} = \max_{i \in I} \overline{d}_{i,p}\,.$$

Then from the values $\overline{d}_{x,1} = d_{x,\pi_x(1)}$, $\overline{d}_{x,2} = d_{x,\pi_x(2)}$, $\ldots$, $\overline{d}_{x,p} = d_{x,\pi_x(p)}$, we obtain the anti-$p$-center $\{v_{\pi_x(1)}, v_{\pi_x(2)}, \ldots, v_{\pi_x(p)}\}$.

2.2 Example

Let a network G with vertex set $V = \{v_1, \ldots, v_6\}$ be given. Let $p = 2$ and the weights of vertices be equal to 1. Let the set $X = \{v_1, v_2, v_3, v_4\}$. Let the distance matrix D be given:

D	v_1	v_3	v_5	v_6
v_1	0	3	1	3
v_2	3	3	2	4
v_3	3	0	4	2
v_4	3	1	2	2
v_5	1	4	0	5
v_6	3	2	5	0

We sort the elements in every row:

$\overline{D}$	1	2	3	4
1	3	3	1	0
2	4	3	3	2
3	4	3	2	0
4	3	2	2	1
5	5	4	1	0
6	5	3	2	0

Then

$$\max_{i \in I} \{\overline{d}_{i,2}\} = 4$$

for $i = 5$. Hence $\overline{d}_{5,1} = 5 = d_{5,6}$ and $\overline{d}_{5,2} = 4 = d_{5,3}$. Then the anti-2-center of G is the set $\{v_6, v_3\}$.

2.3 Analysis of the algorithm p-maxmax

At first, we describe the algorithm in five steps.

1. Input finite weighted network $G = (V, E, c, w)$ and set $X \subseteq V$.
2. Construct weighted distance matrix $D_{n,m} = (d_{i,j})$.
3. Sort the entries of every row - matrix $\overline{D}_{n,m} = (\overline{d}_{i,j})$.
4. Find the largest entry $\overline{d}_{x,p}$ in column p of $\overline{D}_{n,m}$.
5. Construct the anti-p-center $\{v_{\pi_x(1)}, v_{\pi_x(2)}, \ldots, v_{\pi_x(p)}\}$.

The next theorem states more about the time complexity of the algorithm.

Theorem 2.1 *The complexity of the algorithm p-maxmax is $O(n^3)$ in general and for sparse networks it is $O(n^2 \log n)$.*

Proof. Since $m \leq n$, we can write $m = O(n)$. It is possible to show that Step 2 involves $O(en)$ operations, where the number of edges of the networks $e = O(n^2)$ in general, but in sparse networks we have $e = O(n)$ (transport networks are usually sparse networks). Step 3 involves $O(n^2 \log n)$ operations. Step 4 needs $O(n)$ operations and step 5 involves a constant number of operations. Therefore the complexity of the algorithm follows from Step 2 in general and from Step 3 for sparse networks. □

The next theorem states that the algorithm is exact.

Theorem 2.2 *Let a set $U = \{u_1, \ldots, u_p\}$ be an output of the algorithm p-maxmax. Then U is the weighted anti-p-center of G.*

Proof. Step 3 allows us to compute the farthest p-set of every vertex. From these sets, we choose one with the largest distance from a given vertex. It means, we find the p-set X, for which

$$\max_{v \in V} \max_{X \subset V} w(v)d(v, X) = \max_{X \subset V} \max_{v \in V} w(v)d(v, X) \ .$$

Therefore, by the definition, the set X is the weighted anti-p-center. □

2.4 Results on real data

We tested the algorithm on real data. We used the distance matrix of all settlements of Slovakia based on Euclidean metric. The network contains 2916 vertices. Weights of the vertices are numbers of inhabitants. We solved the problem of locating incinerators. The set V_1 of location possibilities contains 54 vertices. We have found anti-p-centers (where $p = 10, 20, 30$). Some results are in the following table:

p	time of computation	weighted eccentricity
10	$67.11s$	78887320
20	$68.98s$	59092240
30	$69.46s$	43134288

No optimal solution is known for Slovakia, when the problems of locating emergency medical stations are solved (weighted p-centers, weighted p-medians) [4]. We can see that it is easy to find the optimal solution, when we solve the problems based on anti-p-centers.

However, it is necessary to mention that the vertices of the anti-p-center can be close together (in our solution it is the east of Slovakia). This solution is not appropriate in real situations. In the next section, we consider some ideas that could help with this problem.

3 GENERALISATIONS

As we see, the anti-p-center model is easy for finding an optimal solution, but tests on real data show one important weakness of"Advanced microscopic modeling and complex data sources for designing spatially large public service systems" the model - vertices of the anti-p-center can be close to each other. This is the reason why we want to consider several generalisations, which can give solutions with a higher dispersion of vertices.

1) Changing the weights of all vertices. In small countries with a higher density of population (for example Slovakia), the weights of vertices in the anti-p-center problem are more important than the distances. We can consider the possibility that the weight of every vertex would be multiplied by a constant $c_1 \in (0, 1)$. New weights will be of the form $w_1(v) = \lceil c_1 w(v) \rceil$, where $\lceil x \rceil$ is the smallest integer not less than x. Determination of a convenient value of c_1 is an open question.

2) Changing the weights of vertices in V_1. The weights of vertices in V_1 will be replaced by constant $c_2 > \max\{w(v) : \forall v \in V\}$, because we need to increase their importance. The problem of determining the value of c_2 is another open question.

3) Anti-(p, q)-center. Let S be a family of q-element subsets of V. Let a p-element subset U be given. A q-eccentricity of the set U (according to S) is

$$q - ec(U) = \max_{X \in S} d(X, U) = \max_{X \in S} \min_{u \in U, v \in X} d(u, v) \, .$$

Weight of the set X is

$$w(X) = \sum_{v \in X} w(v)\,.$$

The weighted q-eccentricity is

$$q - ecc(U) = \max_{X \in S} w(X) \cdot d(X, U)\,.$$

The weighted (p, q)-center is a p-subset with the minimum weighted q-eccentricity:

$$\min_{U \subset V, |U|=p} q - ecc(U)\,.$$

The weighted anti-(p, q)-center is a p-subset with the largest weighted q-eccentricity:

$$\max_{U \subset V, |U|=p} q - ecc(U)\,.$$

If the set S does not contain all q-element subsets or there are some restrictions on the set S, then we say that the anti-(p, q)-center is with respect to the set S.

The (p, q)-centers, anti-(p, q)-centers, their mathematical programming models and the affection of optimal solution by the form of set S will be studied in our next paper which is prepared for publication.

4 CONCLUSIONS

The topic, studied in our contribution, includes various types of subproblems. One of them - the weighted anti-p-center problem - is solved in this paper. We suppose that this simple model is convenient for uncomplicated real problems. However, its generalisations (such as anti-(p, q)-centers) could be a good approximation of large real problems related to location of obnoxious facilities. The $p-$ max max algorithm seems to be a very good starting heuristic algorithm for the mentioned problems, but this will be the topic of our next paper.

5 Acknowledgements

The research of the first author is supported by the research grant VEGA 1/0339/13. The research of the second author is supported by the research grant APVV-0760-11.

References

[1] Church, R.L. and Garfinkel, R.S. (1978). Locating an obnoxious facility on a network. Transportation Science, 12(2): 107–118.

[2] Erkut, E. and Neuman, S. (1989). Analytical models for locating undesirable facilities. European Journal of Operational Research, 40: 275–291.

[3] Janáček, J., Jánošíková, Ľ. and Buzna, Ľ. (2012). Optimized design of large-scale social welfare supporting systems on complex networks. In Thai, M.T. and Pardalos, P.M. (Eds.). Handbook of optimization in complex networks: theory and applications (pp. 337–361). Springer.

[4] Janáček, J. (2014). Radial approach to the emergency public service system design with generalized system utility. International journal of applied mathematics and informatics, 8: 7–14.

[5] Jánošíková, Ľ. (2007). Emergency Medical Service Planning. Communications: Scientific Letters of the University of Žilina, 9(2): 64–68.

[6] Klein, C.M. and Kincaid, R.K. (1994). The discrete anti-p-center problem. Transportation Science, 28(1): 77–79.

[7] Minieka, E. (1983). Anti-centers and anti-medians of a network. Networks, 13: 359–364.

[8] Moon, I.D. and Chaudhry, S.S. (1984). An analysis of network location problems with distance constraints. Management Science, 30: 290–307.

INTRAMAX AND CONSTRAINTS

Samo Drobne and Mitja Lakner
University of Ljubljana, Faculty of Civil and Geodetic Engineering,
Jamova cesta 2, SI-1000 Ljubljana, Slovenia
{Samo.Drobne,Mitja.Lakner}@fgg.uni-lj.si

Abstract: In this paper, we analyse the simultaneous use of three different constraints in the original Intramax procedure, namely, the contiguity constraint, the maximum-share-of-intra-regional-flows constraint, and the minimum-coefficient-of-variation constraint. While the second constraint gives singleton regions, the last one forces the basic data unit with the highest population of flows as an individual region and up to a relatively small number of big functional regions.

Keywords: Intramax, constraints, functional region, functional regionalisation.

1 INTRODUCTION

The Intramax method was first developed by Masser and Brown [19] for the purpose of analysing the structure of flows in a square interaction matrix [2]. In such a matrix, interaction flows – like commuting or migration flows, journeys to school, shopping or recreation, traffic and passenger flows by land/sea/air, money flows, commodity flows, telephone traffic, information flows, gas/water/electricity flows, etc. – are recorded within and between a single set of areas, zones, regions or basic data units (BDUs). The Intramax procedure seeks "to maximise the proportion of the total interaction which takes place within the aggregations of basic data units that form the diagonal elements of the matrix, and thereby to minimise the proportion of cross-boundary movements in the system as a whole" [19:510].

The objective function of Masser and Brown [19] has been improved by Hirst [10] – to take into account the influence of the row and column totals on the residual values that appear in the objective function – and later by Brown and Pitfield [2] – to simplify the form:

$$\text{Max } Z = \sum_{i \neq j} \frac{a_{ij}}{a^*_{ij}} + \frac{a_{ji}}{a^*_{ji}} \tag{1}$$

where a_{ij} is the observed value of the cell entry in the *i*th row and the *j*th column in the interaction matrix, after it was standardized, so that

$$\sum_i \sum_j a_{ij} = 1 \tag{2}$$

and the expected values are calculated as follows

$$a^*_{ij} = \sum_p a_{pj} \sum_q a_{iq} \tag{3}$$

$$a^*_{ji} = \sum_p a_{pi} \sum_q a_{jq} \tag{4}$$

It should be noted that: (a) the procedure maximizes the proportion within the group interaction at each stage of the grouping process (it is a heuristic procedure which does not guarantee a globally optimal solution to the partitioning problem), and (b) the inner flows, i.e. the values on the main sub-diagonal matrices of the partitioned matrix, should be taken into account of the row and column totals all the time.

The objective function (1) is implemented in *Flowmap*, which is a software package for analysing and displaying interactions or flow data. This is the reason, beside its relative simplicity, why the Intramax method has been used so many times to analyse functional regions (FRs) of so many different kinds of interactions at very different consideration levels; e.g. for labour market areas delineation [7, 15, 17, 18, 20, 24], for housing market area delineation [1,

9, 11, 12], for commodity market delineation [2], for world's trade blocks delineation [14, 26], for functional economic regions delineation [21, 22, 23], in telecommunication analysis [8], to identify possible administrative or statistical regions [4, 25], or transport regions [16], in (allocation) analysis of services [5, 6], etc.

In this paper, we analyse the (simultaneous) use of three different constraints in the Intramax procedure using the objective function (1); namely the contiguity constraint, the maximum-share-of-intra-regional-flows constraint, and the minimum-coefficient-of-variation constraint. Here two facts should be noted, (a) the implementation of the spatial contiguity constraint can dramatically increase the complexity of the problem [13]; (b) while the spatial contiguity constraint is the often used constraint in the Intramax applications, the two other here suggested and tested constraints have not been applied in this hierarchical aggregation procedure yet.

2 METHODOLOGY

To test the simultaneous use of three constraints in the Intramax procedure, we developed a programme code in *Mathematica 10.1*. We strictly applied the constraints that could be calculated solely by interaction flows. In each stage of the grouping process, we implemented the use of the objective function (1) and a chosen combination of constraints. The procedure seeks for the maximum value of objective function (1) until the chosen combination of constraints is satisfied. The following constraints have been applied: (a) the spatial contiguity constraint, C_SC, which ensures that only adjacent regions are grouped together; (b) the higher-share-of-intra-regional-flows constraint, C_HF, which ensures that those regions are grouped together that gave a higher share of the intra-regional (inner) flows; and (c) the lower-coefficient-of-variation constraint, C_LCV, which ensures that grouping of regions gives the lower coefficient of variation of intra-regional flows.

Ad (a) Spatial contiguity is often translated into a network tree generation problem to check the validity of contiguity. Regions and their adjacency relationships are expressed as nodes and edges in terms of a graph, so that a region is verified as contiguous only if there is at least one path connecting all the spatial units within the region or if all the spatial units within the region are connected to the tree structure [13]. In our programme, the spatial contiguity is checked by the depth-first search (DFS) algorithm [3].

Ad (b) The higher-share-of-intra-regional-flows constraint, C_HF, forces seeking the maximum value of (1) until $IRFC_{new} > \left(1+\frac{1-IRFC_{old}}{10n}\right)\cdot IRFC_{old}$; where $IRFC_{new}$ is a virtual share of inner flows after aggregation of two candidate regions, $IRFC_{old}$ is a share of inner flows before aggregation, $1+\frac{1-IRFC_{old}}{10n}$ is a parameter defined heuristically that ensures that the change of $IRFC$ is significant, and n is the dimension of the interaction matrix at each stage of the grouping process.

Ad (c) The lower-coefficient-of-variation constraint, C_LCV, forces seeking the maximum value of (1) until $CV(IRFC_{new}) < CV(IRFC_{old})$; where $CV(IRFC_{new})$ is a virtual coefficient of variation of inner flows after aggregation of two candidate regions, and $CV(IRFC_{old})$ is a coefficient of variation of inner flows before aggregation.

The inclusion of constraints in the Intramax procedure was analysed by the processing time (at computer: Intel i7-4771 CPU @ 3.50GHz, RAM 16GB, WolframMark Benchmark Score: 1.78), by the share of intra-regional (inner) flows, by the numbers of singleton and isolated regions, by the number of aggregation steps where a combination of constraints was applied,

by the number of searching steps until the combination of constraints was satisfied, as well as by surveying geographic results. Before performing each aggregation procedure, we stopped the local kernel.

In the application, we analysed the inter-municipal labour commuting flows in 2011 in Slovenia. The dimension of the interaction matrix was $N^2 = 210^2$. The total number of steps to aggregate all municipalities into only one FR was ($N-1$), i.e. 209 steps. Out of a total of 44,100 cells in the matrix, there were 31,557 (71.56%) empty cells. In 2011, there were 778,776 labour commuters in total, but only 388,376 (49.87%) of them commuted between municipalities. The maximum inflow of 109,884 labour commuters (28.29% of all inter-municipal flows) terminated in the biggest employment centre of Slovenia (Ljubljana), while the outflow from Ljubljana was 16,027 labour commuters (4.13% of all inter-municipal flows).

3 RESULTS

Tab. 1 shows statistics on modelling 209 to 1 FRs using the Intramax procedure and a combination of constraints. The fastest result was obtained without using any combination of constraints, and the slowest one by using all three constraints simultaneously. In terms of single constraints, C_HF loaded the processor the most, and C_SC the least.

Using C_HF and C_SC AND C_HF (simultaneous use of both constraints, C_SC and C_HF) give singleton regions (SR), i.e. BDUs that are aggregated just at the end of the procedure; here, singleton regions are municipalities with a very small number of outgoing and ingoing flows, mostly located at the border of Slovenia.

The occurrence of isolated BDUs/regions, IR, is not desired at the latter stages of the aggregating procedure. Constraints C_LCV and C_SC AND C_LCV force Ljubljana to stay an IR even in the system of 13 FRs. It stays also as an IR after 189 steps of aggregation (system of 21 FRs) if no constraints are used or if using C_SC. Otherwise, Ljubljana aggregates together with surrounding BDUs/regions much earlier.

By registering a number of steps of the aggregation procedure where a combination of constraints has been applied, we measured the deviation of the chosen procedure from the original Intramax procedure. The lowest deviation is obtained using solely the C_SC constraint: in this case two small, adjacent municipalities are forced to aggregate together just one step earlier (in the 3rd step) than in the procedure without the constraint (in the 4th step). However, the results of the 4th step of aggregation are equal for both procedures. A small deviation from the original procedure is shown also when using C_LCV where only 16 steps of aggregation were forced no earlier than at the 190th step (moreover, later, 3 aggregation steps did not use C_LCV). The use of the C_HF constraint that should ensure higher intra-regional flows – what should be the main objective of the Intramax method – gives the maximum deviation from the original procedure. Here, in each of 203 steps of aggregation procedure, 21 searching steps had to be done on average to satisfy the constraint. Consequently, the use of C_HF influences the results/statistics in combination with other constraints.

The original Intramax procedure should seek to maximise the proportion of the total interactions which take place within the aggregations of basic data units that form the diagonal elements of the matrix [19]. For this reason, we measured the performance of chosen constraints by the share of intra-regional flows (see Fig. 1). In all cases, the use of the C_HF constraint gives the best results: it gives FRs with the highest share of inner flows. But, when C_HF is combined with C_LCV and C_SC AND C_LCV, the share of inner flows becomes lower for a small number of larger regions (from 44 FRs to 2 FRs). For a high number of small FRs (up to 21 FRs), performing the Intramax procedure without the constraint(s)

(C_no), the use of C_SC, C_LCV and C_SC AND C_LCV constraints gives the same results. For 20 FRs and bigger ones, the original objective function with no constraints as well as the use of the C_SC constraint give FRs with a higher share of intra-regional flows than if using the C_LCV constraint.

Table 1: Statistics on modelling functional regions using the Intramax procedure and constraints

	(I) C_no	(II) C_SC	(III) C_HF	(IV) C_LCV	(V) C_SC AND C_HF	(VI) C_SC AND C_LCV	(VII) C_HF AND C_LCV	(VIII) C_SC AND C_HF AND C_LCV
PT [s]	13.70	13.97	33.92	14.53	34.59	14.31	33.92	35.45
SR	0	0	6	0	6	0	0	0
IR(Lj)	21(0,0)	21(0,0)	35(8,6)	13(0,0)	35(8,6)	13(0,0)	36(14,0)	36(14,0)
TNAS	0	1	203	16	203	60	208	208
FAS	NA	3	1	190	1	(3)190	1	1
LAS	NA	3	203	208	203	206	208	208
0AS	NA	0	0	3	0	(189)3	0	0
TNSS	NA	1	4219	63	4273	60	4769	4823
Geography	Logical	Logical	Logical (but SR)	Logical (but IR(Lj) and SD)	Logical, but SR	Logical (but IR(Lj))	Non-logical	Non-logical

Legend and notes: FR – functional region; SFR – small functional region; LFR – large functional region; SD – spatial discontinuity; C_no – performing the Intramax procedure without the constraint(s); C_SC – spatial contiguity constraint; C_HF – higher-share-of-intra-regional-flows constraint; C_LCV – lower-coefficient-of-variation constraint; PT [s] – processing time in seconds (together with reading data and calculating different statistics); SR – number of singleton regions; IR(Lj) – the notion about Ljubljana as an isolated region at the highest possible step of aggregation as a number of total FRs (the number of other IRs/the number of SRs); TNAS – the total number of aggregating steps where a combination of constraints has been applied; FAS – the first step of aggregation where the combination of constraints has been applied; LAS – the last step of aggregation where the combination of constraints has been applied; 0AS – the number of aggregating steps in the group of applied constraints (between FAS and LAS) where no constraints were applied; TNSS – the total number of searching steps (of the highest values of the objective function) until the combination of constraints was satisfied; Geography – a short notation about the geographical results of functional regionalization.

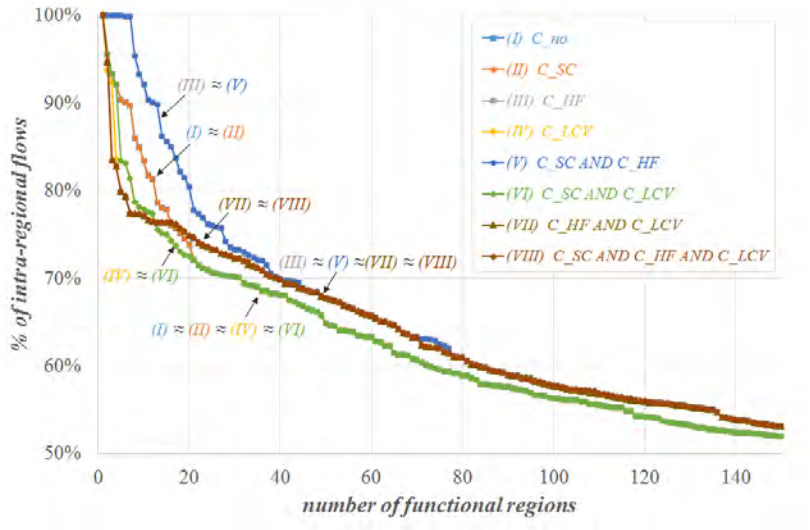


Figure 1: Share of intra-regional flows in relation to the number of functional regions

3 DISCUSSION AND CONCLUSIONS

The original Intramax objective function (1) gives reasonable results regarding FRs. When using labour commuting data, the use of the spatial continuity constraint (C_SC) is not needed while the only differences are shown, just for a while, for a very high number of small FRs. But, the original Intramax procedure lacks from delineating large urban areas that are

disaggregated into smaller urban and adjoining suburban/rural FRs, as has been already shown in [7, 9, 18, 19, 20, 21, 22, 23]. The use of the higher-share-of-intra-regional-flows constraint (C_HF) gives better results regarding the share of inner flows, but it delineates singleton regions. For larger FRs, and if ignoring singleton regions, the original Intramax procedure and the procedure with the C_HF constraint give similar results (see Fig. 2), but the geography is different for smaller FRs.

The use of the C_LCV constraint forces to aggregate FRs with a similar share of intra-regional flows. For this reason, a BDU with a much bigger population than others (in our case Ljubljana) stays isolated for a long time before enlargement of other FRs. The second important disadvantage of using C_LCV is the spatial discontinuity for 2 to 4 FRs; but, this can be solved by using the combination of C_SC AND C_LCV (see also Fig. 1). The combinations of C_HF AND C_LCV and C_SC AND C_HF AND C_LCV constraints give geographically unexpected and unacceptable results for larger FRs.

In this paper, we have demonstrated the simultaneous use of constraints in the Intramax procedure for hierarchical aggregation of basic data units according to the interaction flows. While the use of the spatial continuity constraint has been included in the procedure already from the very beginning [19], this is the first time that the other two constraints are included in the Intramax procedure.

Recently, Koo [15] suggested using a different objective function with the spatial continuity constraint. So, future research could be expanded to couple the use of different objective functions in the same procedure together with the constraints.

Figure 2: (a) Seven functional regions using the original Intramax procedure, and (b) thirteen functional regions using the Intramax procedure with the higher-share-of-intra-regional-flows constraint (C_HF, singleton regions are denoted by intensive colours), inter-municipal labour commuting flows, Slovenia, 2011.

References

[1] Brown, P.J.B. and Hincks, S. (2008). A Framework for Housing Market Area Delineation: Principles and Application, Urban Studies, 45(11): 2225–2247.

[2] Brown, P.J.B. and Pitfield, D.E. (1990). The Intramax Derivation of Commodity Market Structures from Freight Flow Data. Transportation Planning and Technology, 15(1): 59–81.

[3] Daras, K. (2005). An Information Statistics Approach to Zone Design in the Geography of Health Outcomes and Provision. PhD dissertation, University of Newcastle. Newcastle, United Kingdom.

[4] Drobne, S. and Bogataj, M. (2012). A method to define the number of functional regions: An application to NUTS 2 and NUTS 3 levels in Slovenia = Metoda opredelitve števila funkcionalnih regij: Aplikacija na ravneh NUTS 2 in NUTS 3 v Sloveniji. Geodetski vestnik, 56(1): 105–150.

[5] Drobne, S. and Bogataj, M. (2014). Regions for Servicing Old People: Case study of Slovenia. Business systems research journal, 5(3): 19–36.

[6] Drobne, S. and Bogataj, M. (2015). Optimal allocation of public service centres in the central places of functional regions. In: 15th IFAC/IEEE/IFIP/IFORS Symposium on Information Control Problems in Manufacturing (pp. 2449–2454), INCOM 2015. Ottawa: IFAC.
[7] Feldman, O., Simmonds, D.C., Troll, N. and Tsang, F. (2005). Creation of a System of Functional Areas for England and Wales and for Scotland. Proceedings of the European Transport Conference (15 pp).
[8] Fischer, M.M., Essletzbichler, J., Gassler, H. and Trichtl, G. (1993). Telephone Communication Patterns in Austria – A Comparison of the IPFP-Based Graph-Theoretic and the Intramax Approaches. Geographical Analysis, 25(3): 224–233.
[9] Goetgeluk, R. and de Jong, T. (2007). What about the Spatial Dimension of Subsidiarity in Housing Policy? ENHR 2007 International Conference on 'Sustainable Urban Areas' (pp. 1–16). Rotterdam: ENHR.
[10] Hirst, M.A. (1977). Hierarchical Aggregation Procedures for Interaction Data: A Comment. Environment and Planning A, 9(1): 99–103.
[11] Jaegal, Y. (2012). Delineating Housing Market Areas in the Seoul Metropolitan Area Using a Geo-Computational Approach, Master's Thesis, Department of Geography Education, Seoul National University.
[12] Jaegal, Y. (2013). Delineating Housing Market Areas in the Seoul Metropolitan Area Using a Geo-Computational Approach. Journal of the Association of Korean Geographers, 2(1): 7–20.
[13] Kim, H., Chun, Y. and Kim, K. (2015). Delimitation of Functional Regions Using a p-Regions Problem Approach. International Regional Science Review, 38(3): 235–263.
[14] Kohl, T. and Brouver, A.E. (2014). The Development of Trade Blocs in an Era of Globalisation. Environment and Planning A, 46(7): 1535–1553.
[15] Koo, H. (2012). Improved Hierarchical Aggregation Methods for Functional Regionalization in the Seoul Metropolitan Area. Journal of the Korean Cartographic Association, 12(2): 25–35.
[16] Krygsman, S., de Jong, T. and Nel, J. (2009). Functional Transport Regions in South Africa: An Examination of National Commuter Data. Proceedings of the 28th South African Transport Conference (pp. 144–154). Pretoria, South Africa.
[17] Landré, M. (2012). Geoprocessing Journey-to-Work Data: Delineating Commuting Regions in Dalarna, Sweden. ISPRS International Journal of Geo-Information, 1(3): 294–314.
[18] Landré, M. and Håkansson, J. (2013). Rule versus Interaction Function: Evaluating Regional Aggregations of Commuting Flows in Sweden. European Journal of Transport and Infrastructure Research, 13(1): 1–19.
[19] Masser, I. and Brown, P.J.B. (1975). Hierarchical Aggregation Procedures for Interaction Data. Environment and Planning A, 7(5): 509–523.
[20] Masser, I. and Scheurwater, J. (1980). Functional Regionalisation of Spatial Interaction Data: An Evaluation of Some Suggested Strategies. Environment and Planning A, 12(12): 1357–1382.
[21] Mitchell, W. and Stimson, R. (2010). Creating a New Geography of Functional Economic Regions to Analyse Aspects of Labour Market Performance in Australia. In Dalziel, P. (Ed.). Innovation and Regions: Theory, Practice and Policy (pp. 178–220). Lincoln, New Zealand: AERU Research Unit.
[22] Mitchell, W. and Watts, M. (2010). Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques. Geographical Research, 48(1): 24–41.
[23] Mitchell, W., Baum, S., Flanagan, M. and Hannan, M. (2013). CofFEE Functional Economic Regions. AURIN project. Centre of Full Employment and Equity. Darwin, Australia.
[24] Mitchell, W., Bill, A. and Watts, M. (2007). Identifying Functional Regions in Australia Using Hierarchical Aggregation Techniques. Centre of Full Employment and Equity. Working Paper No. 07-06. Newcastle, Australia.
[25] Nel, J.H., Krygsmany, S.C. and de Jong, T. (2008). The Identification of Possible Future Provincial Boundaries for South Africa Based on an Intramax Analysis of Journey-to-Work Data. Orion, 24(2): 131–156.
[26] Poon, J.P. (1997). The Cosmopolitanization of Trade Regions: Global Trends and Implications, 1965–1990. Economic Geography, 73(4): 390–404.

PRACTICAL LOGISTIC APPLICATIONS OF SMALL SCALE TRAVELING SALESMAN PROBLEM FOR SMALL AND MEDIUM ENTERPRISES

Eloy Hontoria
Technical University of Cartagena, Business Management Department
30202 Cartagena, Spain
E-mail: eloy.hontoria@upct.es

Darko Aleksovski
Jožef Stefan Institute, Knowledge Technologies Department
1000 Ljubljana, Slovenia
darko.aleksovski@ijs.si

Abstract: The optimal allocation of orders to a transport fleet is an important task for functioning of different types of medium-sized enterprises. This problem was analysed using Monte Carlo simulation in [3]. In this paper we deal with a related task: the route optimization for visiting the firm's customers. We model this task using the Traveling Salesman Problem (TSP). The method to solve this small scale TSP must provide near-optimal solutions in reasonable time when running on existing computers in Small and Medium Enterprises (SMEs). For this practical application we start by analysing Brute Force Search (BFS), and comparing its performance to other methods.

Keywords: TSP, Brute Force Search Algorithm, Heuristic, Practical Application.

1 INTRODUCTION

According to the Spanish Ministry of Industry, Energy and Tourism in 2014 the 99.88% of the Spanish firms were SMEs[1] (less than 249 employees). From this percentage, 95.7% were micro-enterprises (less than 9 employees). On the other hand, the European Commission reported that the transport and storage services sector accounted for about 4.9% of the total Gross Value Added (GVA) in the EU-27 in 2010[2].

Transport and associate costs also represents a crucial factor for an enterprise and according to the Establish United Logistics Group, in 2007 the logistic cost represented 9.74% of the total sales for American firms and the 8.39 % for European ones. Several attempts have been made to reduce this cost.

Among these attempts, one can find the Vehicle Routing Problem (VRP). The VRP definition states that *m* vehicles initially located at a depot are to deliver discrete quantities of goods to *n* customers. Determining the optimal route used by a group of vehicles when serving a group of users represents a VRP problem. The VRP is one of the most popular problems in combinatorial optimization and has been thoroughly studied in several papers [5, 7, 12]. The VRP is NP-hard, and it includes the Traveling Salesman Problem (TSP) as a special case when the number of vehicles equals to 1 and its capacity is = ∞ [6].

TSP has thus been applied to solve routing problems, like the vehicle routing problem for 28 Dutch cities [9]. Also, in [3] a Monte Carlo Simulation was applied to allocate orders to a fleet of transport in a typical case of a Capacitated Vehicle Routing

1 http://www.minetur.gob.es/energia/en-US

2 http://ec.europa.eu/transport/facts-fundings/statistics

Problem (CVRP). In practice, the CVRP turns out to be significantly harder to solve than the TSP, but both of them pursuit an efficient use of the enterprises' resources.

The work in this paper considers a Spanish SME with one salesman who has to visit the firm's customers located in 21 cities of Spain. This firm collects orders until Monday morning and after that the salesman has to start travelling to visit them. As a consequence this salesman needs a specific application which must be able to give the order of visiting of the customers in a maximum computational time of 1 hour (early Monday morning). Both the real distances between these cities and the needed time to drive to them can be obtained for this salesman using Google Maps. However, the order to visit them for optimizing cost (and time) is a "hard problem" for this company. Also, it will become even a harder one if in future the company grows (e.g. its customers are located in the 51 largest Spanish cities).

As discussed above, a small scale Travelling Salesman Problem [10] is presented and the enterprise has to solve it with limited computational resources. This situation is presented usually in a working day in many companies and the dilemma for most of the them is whether to choose an exact solution which requires large computational time, or an approximate solution which requires a shorter time. For this reason our aim is to design a specific method for the small scale TSP problem, and illustrate its application for SMEs.

We want to check if obtaining the exact solution of the TSP problem for an SME with 21 customers is possible in a reasonable time. In case it is not, we are interested in obtaining a near-optimal solution, using some heuristic method. To compare the results obtained with the proposed methods, a TSP solver will be used.

In the rest of this paper, we first describe the TSP problem and provide a short classification of it variants. Then, we outline and present our method, followed by an illustration of its application. Finally, we outline the conclusions and plans for future research.

2 THE TRAVELING SALESMAN PROBLEM

TSP was first proposed by [2] and is one of the most popular and most studied problems from the NP set. Its solution has a wide applicability in a variety of practical fields and therefore there have been many efforts in the past to provide efficient solutions for the problem, both exact and approximate.

2.1 Classification of TSP

Broadly, the TSP can be classified as symmetric travelling salesman problem (sTSP), asymmetric travelling salesman problem (aTSP), and multi travelling salesman problem (mTSP). In the sTSP, given an n by n symmetric matrix of distances between n cities, the task is to find a minimum-length tour that visits each city exactly once. In the sTSP the distance between cities is the same in each direction, forming an undirected graph and having a route and its reverse.

In the mTSP m salesman located at a single depot representing the starting and ending point have to visit just once all the customers. It can be generalized to a wide variety of routing and scheduling problems, i.e., the School Bus Routing Problem [11]. Different instances of the TSP are also divided into different classes based on the type of graph in concern or how to consider the distances between the cities.

With the aim of finding exact or near-optimal solutions for this practical application to a small scale symmetric TSP, an algorithm based on Brute Force Search is presented and analysed in this work.

3 DESIGN OF AN ALGORITHM BASED ON BRUTE FORCE SEARCH

The brute force search (BFS) approach systematically visits all possible routes for the traveling salesman. It does a so called exhaustive search. As the size *n* of the graph increases, the time required for the exhaustive search to finish increases exponentially [10]. This poses a limitation on the size of the graph that BFS can handle.

In order to illustrate this limitation, we have applied the BFS method on random complete graphs with different sizes (different number of nodes, *n*). The duration of the method is shown (in minutes) in Figure 1. The exponential increase of the amount of time needed is clearly visible here. In particular, for our 21 city problem, the BFS method required approximately 3 days to find the exact optimum for the TSP on a modern-day PC.

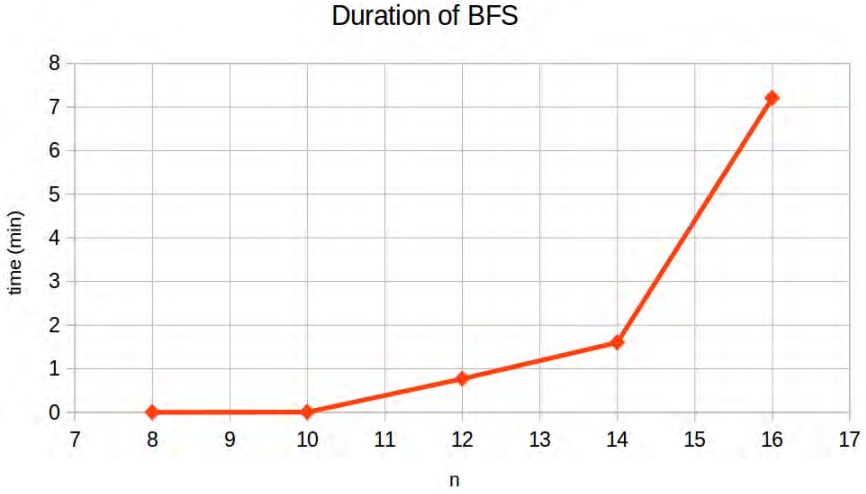


Figure 1: The duration of BFS method on random graphs of different sizes.

Having this fact in mind, our aim is to use some intermediate (near-optimal) solutions that the BFS provides. In particular, our aim is to split the searching of solutions into two phases. The first phase will use an intermediate solution provided by the BFS method, within some time limit. In the second phase the tour constructed will be improved, by using heuristic approaches such as 2-opt and 3-opt tour improvement. This two-step procedure could be understood as a global search, with local refinement of the obtained solution.

In the following text, we first discuss the tour construction, which provides the near-optimal solution, and then the tour improvement.

3.1 Tour construction

In this context we use the BFS method for the TSP problem, and we consider only the Symmetric TSP (sTSP) problem [8]. As we said, due to the exponential time requirements of BFS, we are going to use some of its intermediate solutions. However, to make sure that at least one solution will be produced within the time limit, we employ two modifications to the classical BFS method.

Adding a Greedy Heuristic. In the literature, a fast greedy heuristic is frequently used: the algorithm starts with a tour containing a chosen city and then always adds the closest unvisited city [4]. The algorithm stops when all cities are included in the tour. We have incorporated this heuristic in our BFS search. This heuristic changes the order of visiting of the not-yet-visited nodes: first visit the closest node, and then the rest. This strategy is shown in line 9 of Algorithm 1, where the pseudocode of the algorithm used can be seen.

Search pruning. Additionally, we employ the technique of pruning during the search: if the tour length is already longer that the global optimum, we discard the last visited node, and

consider a new node for visiting. This is shown in line 12 of Algorithm 1. We use the abbreviation BFScn for the BFS method with two modifications: closest neighbour heuristic and pruning. The final method is outlined in Algorithm 1.

```
1  Algorithm Run_BFScn(R = {c_1, c_2, c_3, .., c_k})
       Data: route R = {c_1, c_2, c_3, .., c_k}
2      if route R has N elements and we can return to the starting node c_1 then
3          if the length of the solution R is shorter than the length of R_glob then
4              R_glob = R                          // A new intermediate solution
5              return
6          end
7      else
8          Create a list L of unvisited nodes directly connected to c_k
9          Sort the list L according to the distance from node c_k
10         for all nodes c in L do
11             if adding c to R results in a longer route than R_glob then
12                 Do not add node c to R                                  // prune
13             else
14                 Let the new route R_1 be {c_1, c_2, c_3, .., c_k, c}    // extend route
15                 Run_BFScn(R_1)
16             end
17         end
18     end
19 Algorithm BFScn()
20     Let R_glob = {}
21     Run_BFScn({})
```

Algorithm 1: The pseudocode for the BFScn tour construction method.

3.2 Tour improvement

To refine the tour previously constructed, an improvement heuristic can be applied. In this paper we have chosen a simple k-opt heuristic for the TSP [10] which modifies an existing tour in order to reduce its length. In particular, to improve the solution a 2-opt and 3-opt exchange heuristic has been applied.

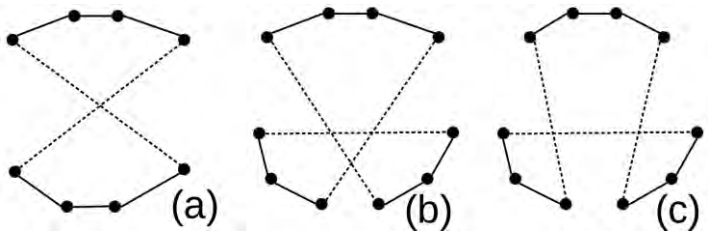


Figure 2: Tour improvement: (a) a 2-opt move and (b,c) the two variants for a 3-opt move.

The 2-opt algorithm removes two edges from the already generated tour, and reconnects the two new paths created. The reconnection is done in such a way to keep the tour valid (see Figure 2 (a)) and the move will be performed only if the new tour is shorter than the older one. The algorithm repeats this procedure until no better solution is found obtaining a 2 optimal tour. When dealing with the 3-opt algorithm the number of removed edges is three. There are two variants to perform the reconnection, as shown in Figure 2 (b) and (c). The k-opt tour improvement idea was further improved by an algorithm which

discovers moves with large values of k [10, 1]. It is worth noting that, the performance of 2-opt or 3-opt heuristic mainly depends on the tour generated by the tour construction, i.e. the first step.

We have implemented the tour improvement as an iterative method. In each iteration we systematically check every pair of nodes for the 2-opt and every triple of nodes for the 3-opt heuristic. At the end of the iteration, we select the particular 2-opt or 3-opt move which produces the lowest tour length. As noted before, the iterations are performed as long as we get shorter tours.

4 APPLICATION

Data. We test our approach using data for the distance (and duration of travel) between Spanish cities, where the customers of the SME are located. We have obtained the data using the Google Distance Matrix API[3]. The inputs to this service are the locations of the cities (or names in the case we are not interested in precision), and the outputs are the distance and duration data in XML format. For the purpose of obtaining and transforming the data to a matrix format, we use the scripting language Python. Using this procedure we obtained two sets of data, one smaller and the other larger: a) a set of 21 Spanish cities and b) a set of 51 Spanish cities.

It is also worth noting that the distances provided were not only distances of "direct roads", but also roads visiting several other cities. For instance, the distance provided by this service, for travel between Madrid and Barcelona was 619 km. However, in reality there is no direct road connecting them and the route must pass through Zaragoza. This is why we transformed the complete graph, by removing some roads, which in reality were not direct. This means that the graph representing the problem is not a complete graph, i.e. some cities are not connected directly, as illustrated in Figure 3.

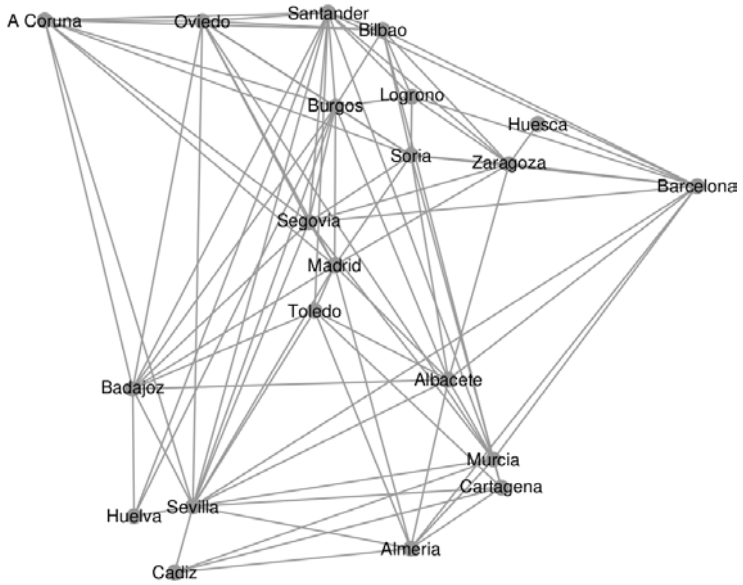


Figure 3: The dataset of 21 Spanish cities.

Methods. We implemented the BFScn and the 2-opt and 3-opt tour improvement approaches in Python and tested it on both sets of data. We reported the length of the tour found and the time required.

3 https://developers.google.com/maps/documentation/distancematrix/

We also compared the BFScn approach to a TSP solver made available within the Google's Operations Research tools (OR-TOOLS)[4]. OR-TOOLS are a freely available piece of software which uses constraint programming (CP), and have their own CP solver. Their TSP solver uses a "cheapest addition" heuristic to generate tours, and uses local search to improve them.

However, the constraint programming technique used is quite different from the standard imperative one. This means that any modification or extension of the code (for example for a different version of TSP) would require knowledge of CP and Google's CP solver. It would be hard to imagine that a typical SME employs an expert for this specialized type of programming, for potential modifications to this tool.

Results. After running the BFScn method for a maximum of 1 hour, we extract the intermediate solutions, and apply the tour improvement methods. It is visible in the second column of Table 1 that the particular solutions are obtained very fast (in less than 1 minute), and in the remainder of the 1 hour time limit BFScn cannot improve them. This is due to the closest neighbour heuristic. Also, when comparing the third and fourth columns of Table 1, we can see that the 2-opt and 3-opt tour improvements result in solutions very close or identical in length to the solutions of OR-TOOLS. This shows the usefulness of the method proposed here. The total time required to travel the 21 cities is 45.66 hours and this could potentially be handled by one salesman during one working week.

length (km)	BFScn	BFScn + impr	OR-TOOLS
51 cities	7137.29	6275.53	**6179.02**
21 cities	4347.39	**4343.58**	**4343.58**

time (min)	BFScn	BFScn + impr	OR-TOOLS
51 cities	0.93	1.36	0.05
21 cities	0.54	0.71	0.11

Table 1: The results of the application of BFScn and OR-TOOLS to the two datasets of Spanish cities.

6 CONCLUSIONS

This paper dealt with the problem of obtaining solutions for a small-scale TSP, in the context of a Small and Medium Enterprise (SME). The first conclusion was that obtaining the exact solution was not possible in reasonable time, even for a small scale problem. Because of this, a methodology based on the Brute Force Search was presented, for obtaining near-optimal solutions. It incorporates a useful closest neighbour heuristic, and uses tour improvement (2-opt and 3-opt exchange strategies) to increase the quality of the obtained solution. The methodology showed similar or identical results to a Google's OR-TOOLS solver.

Also, this paper discussed and illustrated the applicability of the methodology for SMEs, by considering two datasets, of 21 and 51 Spanish cities. Finally, this paper outlined a possible way to obtain distance and duration data for travel between towns and cities, using the Google Distance Matrix API.

Future research will further explore the link between VRP and TSP, considering the working hours of the salesman, for the small scale TSP. With this aim, future work could calculate the TSP route and split it in daily stages, taking into account that at the end of the day, the salesman has to sleep in one of the cities.

4 https://github.com/google/or-tools

References

[1] Applegate, D.L., Bixby, R.E., Chvatal, V., and Cook, W.J. (2011). The Traveling Salesman Problem: A Computational Study: A Computational Study. Princeton university press.

[2] Dantzig, G., Fulkerson, R., and Johnson, S. (1954). Solution of a large-scale traveling-salesman problem. Journal of the operations research society of America, 2(4):393–410.

[3] Hontoria, E., de la Fuente Aragon, V., Ros-McDonnell, L., and Bogataj, M. (2011). Monte Carlo simulation applied to the logistics of ceramics. In Proceedings of the 11th International Symposium on Operational Research in Slovenia, SOR 2011.

[4] Johnson, D. S. and McGeoch, L. A. (1997). The traveling salesman problem: A case study in local optimization. Local search in combinatorial optimization, 1:215–310.

[5] Laporte, G. (1992). The vehicle routing problem: An overview of exact and approximate algorithms. European Journal of Operational Research, 59(3):345–358.

[6] Laporte, G. (2007). What you should know about the vehicle routing problem. Naval Research Logistics, 54(8):811–819.

[7] Laporte, G., Gendreau, M., Potvin, J.-Y., and Semet, F. (2000). Classical and modern heuristics for the vehicle routing problem. International transactions in operational research, 7(4-5):285–300.

[8] Lawler, E.L., Lenstra, J.K., Kan, A.R., and Shmoys, D.B. (1985). The traveling salesman problem: a guided tour of combinatorial optimization. Wiley.

[9] Lenstra J. and Kan, R.A.H.G. (1975). Some simple applications of the travelling salesman problem. Operational Research Quarterly, 26(4):717–733.

[10] Lin, S. and Kernighan, B. W. (1973). An effective heuristic algorithm for the traveling-salesman problem. Operations research, 21(2):498–516.

[11] Orloff, C. (1974). Routing a fleet of m vehicles to/from a central facility. Networks, 4(2):147–162.

[12] Toth, P. and Vigo, D. (2002). Models, relaxations and exact approaches for the capacitated vehicle routing problem. Discrete Applied Mathematics, 123(1):487–512.

TIME-DISTANCE VERSUS UTILITY IN THE PUBLIC SERVICE SYSTEM DESIGN

Marta Janackova
University of Zilina, Department of Applied Mathematics
Univerzitna 1, 01026 Zilina, Slovak Republic
Marta.Janackova@fstroj.uniza.sk

Alzbeta Szendreyova
University of Zilina, Department of Mathematical Methods and Operations Research
Univerzitna 1, 01026 Zilina, Slovak Republic
Alzbeta.Szendreyova@fri.uniza.sk

Abstract: Frequently used criteria in public service system design are minimizing of cost for locating facilities or minimizing the distance between user and located facility. A utility of the public service system for the user may not be directly proportional to the distance between user and service center. There are also some services, where the utility decreases rapidly behind the given border. This paper deals with the function which expresses the utility of the public service systems for the users. We investigate how the change of the function parameter influences the properties of the function. We solve the p-median problem for input data and different values of the parameter.

Keywords: *p*-median problem, public service system, utility.

1 INTRODUCTION

In to this contribution we deal with a designing process of a public service system. Public service systems are usually established to provide public with some obligatory service. The servicing facilities are located at places called service centers and a user demand is covered either by delivering the service to the demand location or the user has to travel to the nearest service center to satisfy his demand for service. The way of demand satisfaction depends on the sort of service, which can consist of school attendance, visiting some office of a public administration system or visiting some health center.

In these cases a user usually travels to the nearest service center. In the other cases a service facility must travel from center location to the user's location. These cases involve the emergency systems as medical emergency system, fire brigade system and others.

The objective used in the public service system design is to deploy the given number of service centers or facilities in a set of possible locations so that the resulting effect of the service providing is as big as possible.

The notion of „resulting effect" depends on the criteria, which are applied on the system design. Mostly the sum of distances from user to the nearest service center is minimized. This criterion can be modified by considering travel time or cost instead of the distance. Within this contribution, the criterion is generalized and sum of general utility perceived by the individual system users is maximized.

2 THE *p*-MEDIAN PROBLEM WITH SERVICE UTILITY FUNCTION FOR USER

A user considers his access to service optimal, if his location coincides (is in proximity) with location of a service center. In this situation the user perceives the maximal utility of the system. Utility of system for given user decreases with increasing distance to the nearest service center.

Nevertheless, the same distance from the nearest center does not mean the same portion of perceived utility, if different service is considered. Service center of some service system

(e.g. an office of on public administration system) is visited by an average user one in a year's period. In such case, a longer distance to the service center does not represent a substantial decrease of system utility, contrary to the daily demanded service. There are also sorts of service, which are useful for a user only if the user lies in a given threshold radius from a service center, and the utility sharply drops to zero outside the radius.

In this contribution, we express the perceived user's utility by a decreasing function of the user's distance from the nearest service center. The shape of the function is determined by two parameters. The first one is the critical radius, which is a threshold dividing domain of the distance into parts of high and low utilities. The second parameter T is so called shaping parameter, which influences the slope of the function decrease and behavior of the function, which can change from almost linear to jump form.

In the following computational study, we evaluate the influence of the parameters on the objective function value of the optimal design of the public service system. In addition, we try to explore, which values of the shaping parameter T are suitable for the medical emergency system and which ones can be used for service systems, where a longer time of service accessibility is acceptable.

The result obtained by utility maximization will be compared to the results obtained by solving the problem of minimization the sum of the distances from individual users to the nearest service center.

The used utility function of a time – distance t follows:

$$u(t) = \frac{c_0}{1+e^{\frac{t-t_{crit}}{T}}}.$$

The variable t corresponds to the time of service delivery. The parameter t_{crit} is the above mentioned threshold or critical radius. If the delivery time t exceeds the threshold, then the utility sharply drops to the zero value. The positive parameter T influences the shape of the function $u(t)$ so that it makes the form of the function near to linear dependence if the parameter T takes a bigger value.

The function $u(t)$ obtains the maximal value at the point $t = 0$ on the range of non-negative values of t. The function is decreasing and the point $t = t_{crit}$ it changes the concave characteristic to convex one.

In the following numerical experiments, we will use a normalized form of the utility function, where the parameter c_0 is determined by the expression

$$c_0 = 1 + e^{\frac{-t_{crit}}{T}}.$$

Then, the utility function gets the value of 1 for $t = 0$.

Objective of the designing process is to locate at most p service centers so that the sum of all utility values perceived by individual users is maximal. The problem can be formalized in the following way.

Let I denote a finite set of possible center locations and J denote a finite set of user's locations. User's location $j \in J$ *is* shared by b_j users. The time-distance between locations $i \in I$ and $j \in J$ *is* denoted as t_{ij} and the utility value corresponding to this time-distance is denoted as $u_{ij} = u(t_{ij})$.

To formulate mathematical programming model of the problem, we introduce two series of decision variables. The variable $y_i \in \{0, 1\}$ for $i \in I$ models decision on locating or non-locating a service center at the location i. The variable takes the value of one if a service center is located at i and it takes the value of zero otherwise.

The second series of decision variables consist of allocation variables $z_{ij} \in \{0, 1\}$ for $i \in I$ and $j \in J$. The variable z_{ij} takes the value of one if the user's location j is assigned to the possible service center location i, and it takes the value of zero otherwise.

The model can be constituted as follows.

$$Maximize \quad U = \sum_{i \in I} \sum_{j \in J} b_j u_{ij} z_{ij} \tag{1}$$

$$Subject\ to \quad \sum_{i \in I} z_{ij} = 1 \qquad \text{for } j \in J \tag{2}$$

$$z_{ij} \leq y_i \qquad \text{for } i \in I, j \in J \tag{3}$$

$$\sum_{i \in I} y_i \leq p \tag{4}$$

$$y_i \in \{0,1\} \qquad \text{for } i \in I \tag{5}$$

$$z_{ij} \in \{0,1\} \qquad \text{for } i \in I, j \in J \tag{6}$$

The maximized expression (1) defines so called system utility *U*.

Constraints (2) ensure that each user's location *j* is assigned to one of the possible center location for service.

Link-up constraints (3) assure the implication that if user's location *j* is assigned to possible center location *i* for service, then a service center must be placed at the location.

The constraint (4) enables to locate at most *p* centers and obligatory constraints (5) and (6) allow the variables to take only the values of one or zero.

3 COMPUTATIONAL STUDY OF PUBLIC SERVICE SYSTEM DESIGNING

The computational study was performed on self-governing region of Zilina using its road network to obtain the time-distances among dowelling places of the region. The region contains 315 dwelling places, which represent the set *J* of user's locations. One hundred of most populated villages were chosen to constitute the set *I* of possible center locations. Studied instances were derived from the original emergency health care system which consists of 36 ambulance vehicles deployed in the 29 dwelling places of the region. As our model does not allow to place more than one facility at one location, the derived instances considered p=29. For this experiment, we used b_j=1, for $j \in J$. Based on the original emergency health care system, where recommended service access time varies from fifteen to twenty minutes, we suggested t_{crit} = 20 for the first series of experiments.

The second and the third series of experiments were performed for t_{crit} = 60, which corresponds with recommended access time of common health care. The above-mentioned instances were solved for different values of the shaping parameter *T*. The parameter *T* varied over the range {3, 6, 9, ..., 39} in our experiments.

For comparison, the individual instances were reformulated as the *p*-median problem, where time-distances t_{ij} weighted by the coefficients b_j were used to constitute the minimized objective function.

In the both cases, i.e. the classical *p*-median and maximal utility service system design, the solution consists of the service center locations.

Each such solution can be evaluated accordingly to the objective function (1), which determines the system utility, and it can be also evaluated as objective function of the *p*-median problem, i.e. sum of weighted time-distances from each user's location to the nearest service center. This sum of time-distances will be referred as the total of time-distances.

Table 1 shows the solutions of series for *p* = 29 and critical parameter t_{crit} = 20, t_{crit} = 40 and t_{crit} = 60.

In the column "Value of the objective" of the table, there are objective function values of the utility, which correspond to different values of the parameter *T*. In the column "Total of the time-distances", there are given the sums of time-distances from user to the nearest service centers (such as in classical *p* – median problem).

For comparison, the bottom row contains objective function values of the optimal solution of the associated classical p – median problem for $p = 29$.

Table 1: Results for the maximization of utility

$t_{crit}=20$, $p = 29$			$t_{crit}=40$, $p = 29$			$t_{crit}=60$, $p = 29$		
T	***Value of the objective***	***Total of the time distances***	***T***	***Value of the objective***	***Total of the time distances***	***T***	***Value of the objective***	***Total of the time distances***
3	305.86	2 161	***3***	314.99	2 161	***3***	315.00	2 153
6	287.87	2 143	***6***	313.79	2 151	***6***	314.96	2 151
9	280.24	2 143	***9***	310.06	2 151	***9***	314.45	2 151
12	278.94	2 143	***12***	305.92	2 143	***12***	313.16	2 143
15	280.07	2 143	***15***	302.73	2 143	***15***	311.45	2 143
18	282.01	2 143	***18***	300.65	2 143	***18***	309.71	2 143
21	284.11	2 141	***21***	299.46	2 143	***21***	308.20	2 143
24	286.16	2 141	***24***	298.87	2 143	***24***	306.99	2 143
27	288.04	2 141	***27***	298.69	2 143	***27***	306.07	2 143
30	289.75	2 141	***30***	298.76	2 143	***30***	305.42	2 143
33	291.29	2 141	***33***	298.99	2 143	***33***	304.97	2 143
36	292.67	2 141	***36***	299.33	2 141	***36***	304.68	2 143
39	293.92	2 141	***39***	299.71	2 141	***39***	304.52	2 143
Classical p-median problem, p = 29					**2 141**			

The value of the system utility decreases only to a certain value of the increasing shaping parameter T. If $t_{crit} = 20$, $U(t_{crit}, T)$ decreased for $T < 12$ and it increased for $T > 12$. If $t_{crit} = 40$, $U(t_{crit}, T)$ obtained its minimal value for $T = 27$. In our experiment, the value of $U(t_{crit}, T)$ decreased for all parameters T when $t_{crit} = 60$ (i.e. $U(t_{crit}, T)$ obtains its minimum for $T > 39$).

In the table 2, the values of the system utility $U(t_{crit}, T)$ are compared for the parameters $t_{crit} = 20$ and $t_{crit} = 60$.

Table 2: Differences between the objective function values $U(t_{crit}, T)$ for $t_{crit} = 20$ and $t_{crit} = 60$

T	U, $t_{crit}=20$	U, $t_{crit}=60$	***Differences***	***Differences in %***
3	305.86	315.00	9.14	2.99
6	287.87	314.96	27.09	9.41
9	280.24	314.45	34.21	12.21
12	278.94	313.16	34.22	12.27
15	280.07	311.45	31.38	11.20
18	282.01	309.71	27.70	9.82
21	284.11	308.20	24.09	8.48
24	286.16	306.99	20.83	7.28
27	288.04	306.07	18.03	6.26
30	289.75	305.42	15.67	5.41
33	291.29	304.97	13.68	4.70
36	292.67	304.68	12.01	4.10
39	293.92	304.52	10.60	3.61

Having analyzed the above-mentioned solutions in details, we have found that the time-distance between the worst situated user and his nearest located service centers takes the same value of 20 minutes in almost all instances. This inspired us with the idea to perform an analysis of time-distances from users to the nearest centers for all obtained optimal solutions. These integer time-distances vary from zero to the maximal value of 20 minutes and we give their frequencies for individual instances derived from the case t_{crit} = 20 in the table 3. Each row of the table corresponds to one time-distance value and each column corresponds with one instance specified by the value of shaping parameter *T*. The last column denoted as p-median contains frequencies of time-distances from the optimal solution of the classical *p*-median problem..

The table 3 shows the comparison of the values of the time-distance frequencies for the individual user in the system utility $U(t_{crit}, T)$ for the parameters *T*.

Table 3: Frequency of time-distance values, t_{crit}=20

t_{ij} \ *T*	*3*	*6*	*9*	*12*	*15*	*18*	*21*	*24*	*27*	*30*	*33*	*36*	*39*	*p-med*
1	4	4	4	4	4	4	5	5	5	5	5	5	5	5
2	13	15	15	15	15	15	17	17	17	17	17	17	17	17
3	13	13	13	13	13	13	14	14	14	14	14	14	14	14
4	23	24	24	24	24	24	25	25	25	25	25	25	25	25
5	45	45	45	45	45	45	42	42	42	42	42	42	42	42
6	30	31	31	31	31	31	32	32	32	32	32	32	32	32
7	28	28	28	28	28	28	27	27	27	27	27	27	27	27
8	24	24	24	24	24	24	23	23	23	23	23	23	23	23
9	27	25	25	25	25	25	23	23	23	23	23	23	23	23
10	19	19	19	19	19	19	16	16	16	16	16	16	16	16
11	16	13	13	13	13	13	13	13	13	13	13	13	13	13
12	6	7	7	7	7	7	8	8	8	8	8	8	8	8
13	11	10	10	10	10	10	10	10	10	10	10	10	10	10
14	8	8	8	8	8	8	9	9	9	9	9	9	9	9
15	4	4	4	4	4	4	5	5	5	5	5	5	5	5
16	6	6	6	6	6	6	6	6	6	6	6	6	6	6
17	4	4	4	4	4	4	5	5	5	5	5	5	5	5
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	2	2	2	2	2	2	2	2	2	2	2	2	2

Similar results were obtained for the instances derived for critical radius t_{crit} = 40 and t_{crit} = 60.

3 CONCLUSION

The tables show, that the highest system utility was obtained for the smallest' value of the shaping parameter *T* and the highest value for system utility $U(t_{crit}, T)$ = 315 was obtained for the parameters t_{crit} = 60 and *T* = 3, what is natural property of the utility function $u(t)$.

The value of system utility $U(t_{crit}, T)$ grows with the growing critical radius t_{crit}. The improvement of the system utility from 3% to 12% corresponds with change of t_{crit} from 20 to 60 minutes.

As concerns dependence of the system utility on parameter *T*, we have found that the dependence is a quasi-convex function. Minima of the functions differ for different t_{crit} values.

As can be seen in the table 1, we can find that the best system utility corresponds with the worst value of the total time-distance. In addition, the performance of the total time-distance does not follow the performance of the system utility. Whereas the system utility is a quasi-convex function of shaping parameter *T*, the total time-distance values monotonously decreases to the value of the classical *p*-1median with the increasing value of *T*.

The time-distance between the worst situated user and his nearest located service center takes the same value of 20 minutes for all parameters.

The range of the input data used in solved instances was not large enough to prove that the number of the worst situated users can be reduced by usage of suitable utility function. We prepare testing of larger instances in our future research.

Acknowledgement

This work was supported by the research grants VEGA 1/0518/15 and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks".

References

[1] Janacek, J. (2013). Modeling of utility in public service system design problem. Proceedings of the workshop "Optimization of region covering by public service systems", Pardubice, Czech Republic: 50-59.

[2] Janackova, M., and Szendreyova, A. (2009). First Aid Stations Location Problem and an Impact of the Set of Candidates on the Recourse of a Patient. Proceedings of 10th International Symposium on Operational Research in Slovenia-SOR´09, Nova Gorica: 357-360.

[3] Janackova, M., and Szendreyova, A. (2012). Territory decomposition parameters of distribution tasks. Proceedings of the 30th international conference: Mathematical methods in economics, Karviná, Czeck Republic: 373-379.

[4] Janackova, M., and Szendreyova, A. (2013). An Influence of the Lengths Distribution in the Distance Matrix on the Efficiency of the Exact p-Median Algorithm. Proceedings of the 1st WSEAS International Conference on Optimization Techniques in Engineering (OTENG'13), Antalya, Turkey: 192-195.

[5] Marianov, V., and Serra, D. (2002). Location problems in the public sector. Facility location - Applications and theory, Berlin, Springer, pp. 119-150.

CROATIAN MARITIME FREIGHT TRANSPORT IN THE ENVIRONMENT OF SCHENGEN AREA

Elza Jurun
University of Split/Faculty of Economics Split, Department of Quantitative Methods
21000 Split, Cvite Fiskovića 5, Croatia
Phone: ++ 38521430648; Fax: ++38521430701
E-mail: elza@efst.hr

Nada Ratković
University of Split/Faculty of Economics Split, Department of Quantitative Methods
21000 Split, Cvite Fiskovića 5, Croatia
Phone: ++ 38521824171; Fax: ++38521430701
E-mail: nada.ratkovic@efst.hr

Ivan Marijanović
University of Split/Faculty of Economics Split
21000 Split, Cvite Fiskovića 5, Croatia
Phone: ++ 38521274171; Fax: ++38521430701
E-mail: marijanovic2015@gmail.com

Abstract: Croatia has geo-traffic extremely favorable position in Pan-European transport corridors. This paper deals with Croatian maritime freight transport analysis whose current trends are far below by their economic opportunities. Comprehensive statistical analysis of Croatian maritime freight transport as well as the comparative analysis with neighboring countries have been carried out on the real data bases and covers time horizon of the last decade. This research also aims to propose priorities among activities necessary to promote Croatian sea transport indeed as an integral part of the Schengen area.

Keywords: Schengen area, Croatian maritime freight transport, Pan-European transport corridors, Comparative analysis

1 INTRODUCTION

Croatian as well as European development strategies have the same goals based on creation and implementation of transport policy as an integral part of a complex policy of sustainable development. The biggest step towards achieving this goal recently is definitely introduction of the Schengen system. From this point of view accession to the Schengen area becomes a strategic Croatian goal. Declaration of readiness to start the Schengen evaluation process has been submitted of Government of the Republic of Croatia on March 12th 2015 with the hope of adopting the European Council decision of the full application of the Schengen acquis starting from July 1st 2015. [3]. It remains already questionable will this happens. Namely, some of the Croatian comparative advantages are becoming fundamental barriers of the Schengen evaluation process implementation speed such as for example the length of the maritime border which becomes the external Schengen area border.

Croatia is situated on the west coast of the Adriatic Sea and thus its exit at the Mediterranean Sea is also proposed. Croatia has 1.880 km coast line (6.278 km including islands) and includes 718 islands and 467 sea reefs. As one of the most developed coastal regions in Europe Croatia has a long maritime tradition and maritime sector has always played a key role in the economic, trade and social development of the country.

In the focus of this paper is a new approach of Croatian maritime freight transport (CMFT) analysis. Namely, the analysis results have shown that the current trends of CMFT were far below by their economic opportunities in every aspect. National or seaports of international economic importance in Croatia are Rijeka, Zadar, Šibenik, Split, Ploče and

Dubrovnik. They are suitably geographically located so as to facilitate maritime transport between Central and Eastern Europe and further South Asia, Australia and further more Oceania with Europe (by way of Suez Channel).

This research of the current CMFT situation and past development failures also aims to propose priorities in a series of activities that Croatia necessary has to realize in order to promote Croatian sea transport indeed as an integral part of the Schengen area.

2 CROATIAN MARITIME FREIGHT TRANSPORT

2.1 Republic of Croatia in Trans-European Transport Network

Croatia has in the terms of geographical position as well as geopolitical, and therefore in terms of geo-traffic extremely favorable position. In the focus of this chapter is Croatian position in Pan-European transport corridors.

Pan-European transport corridors make the roads in the countries of Eastern, South-East and parts of Central Europe. Corridors directions serve to enable inland and coastal Europe to be connected in the fastest, the most efficient, the safest, the most convenient and the most environmentally friendly way. Croatia is one of the few countries which has the privilege that through its territory pass four corridor direction; V_B, V_C, X and X_A. These corridors pass through the Croatian territory and are an integral part of the Trans-European Transport Network (TEN-T). TEN-T Network consists of: V_B (TEN-T Mediterranean Corridor), V_C (TEN-T comprehensive network), X (TEN-T core network) and X_A (TEN-T comprehensive network). Corridor directions are the way of involving Croatia and its transport system into the European transport and economic system. Croatia is the only Adriatic and Mediterranean, Central European, Pannonian and Danube country in Central Europe whose territory is direct continuation of southeast Europe. Therefore freight transport is mainly oriented to two TEN-T transport directions: V and X.

EU and Croatia are firmly connected by mutual economic interests because EU is the most important Croatian trading partner. It is the reason of the high degree of causal relationship between Croatia and EU economic development which on the other side further contributes to the exceptional importance of European regions cohesion. All this implies high importance of interconnection between Croatia and other European regions. In this sense, closer connection between north Adriatic ports has been achieved by establishing North Adriatic Ports Association (NAPA) on March 1st 2010.

By NAPA Italian ports Trieste, Venice, Ravenna have teamed up with Slovenian port Koper and the biggest Croatian port Rijeka. These North Adriatic ports formally agreed to cooperate to create appropriate synergies when carrying out promotional activities at many different levels (regional, national, European, international) in order to eliminate the constraints and infrastructure problems that are holding back the development of the North Adriatic Port. [5]

Right from the start, NAPA has demonstrated the validity of its strategy, based primarily on the competitiveness of the ports, but also on a spirit of cooperation, developed at national, European and international level. Its many successes include the container shipping services linking the NAPA ports with the Far East, the EU recognition of the Baltic-Adriatic Corridor and the accreditation of the Association with the European Union. The ports of the Association present themselves together on the international market as one large hub by jointly organizing, managing and participating in trade fairs, conferences and events dedicated to the markets of the Far East, the Eastern Mediterranean and Central and Central-Eastern Europe, and by jointly publishing information and promotional materials as well as articles in the press. Italian port Ravenna ceased to be a member of NAPA in January 2013.

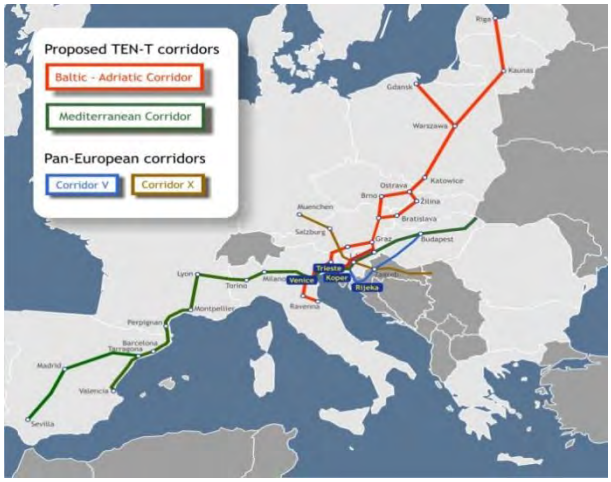


Figure 1: NAPA ports in TEN-t Network
Source: www.portsofnapa.com

Croatian membership in the NAPA significantly contributes in the process of achieving the strategic development goal –accession to the Schengen area.

2.2 Republic of Croatia in the Schengen area

The implementation of the Schengen Agreements started in 1995, initially involving seven EU States. Born as an intergovernmental initiative, the developments brought about by the Schengen Agreements have now been incorporated into the body of rules governing the EU. Today, the Schengen Area encompasses most EU States, except for Bulgaria, Croatia, Cyprus, Ireland, Romania and the United Kingdom. However, Bulgaria and Romania are currently in the process of joining the Schengen Area. Of non-EU States, Iceland, Norway, Switzerland and Liechtenstein have joined the Schengen Area. So, for these countries there are no longer any frontier controls at the borders for all kinds of goods as well as for about 400 million people over an area of 4.312.099 km^2. At the same time the Schengen area establishes effective controls at the external borders of the EU and introduces a common visa policy.

Schengen implementation for a particular Member State is a long and expensive process. Basic calculations show that the process of accession to the Schengen area for Croatia will cost approximately 120 million €. Government of the Republic of Croatia has declared the readiness to start the Schengen evaluation process on March 12th 2015.
The mechanism of evaluation and monitoring covers all aspects of the Schengen acquis: efficient and effective application of measures by Member States at the external borders, visa policy, the Schengen Information System, data protection, police cooperation, judicial cooperation in criminal matters, the absence of border control at internal limits.

The process of accessing Croatia in the Schengen system as a long and expensive process certainly has its advantages and disadvantages. For the Republic of Croatia advantages of joining the Schengen area without borders, among others are better control of borders towards third countries, more effective fight against organized crime with a safer, more open and fairer integration system. Disadvantages of entering into the Schengen area are the high cost of implementation of the acquis as well as a greater risk of the increase of illegal immigrants. Thus third-country nationals are subject to thorough checks when entering and leaving the Schengen area, while EU citizens and citizens from the four non-EU members of the Schengen system enjoy the right of free movement and subject to minimal checks identification.

2.3 Present Croatian Maritime Freight Transport Movements

Croatia fleet consists of 1 245 ships and 10% of them are sailing in international navigation. But from day to day the fleet numerically reduces with the shipping companies business decline. According to the Association of Croatian ship owners 'Mare Nostrum', in 2012 were operating 12 Croatian shipping companies with 154 vessels. In 2013 that number dropped to 10 companies with 142 ships. With the regard to reduction in the number of vessels in the fleet, it would be necessary to adopt measures that ensure the survival and modernization of the Croatian fleet which would be competitive on the world market.

According to "Statistical Report of Transport and Communications in 2013" of Croatian Bureau of Statistics from 2014, in the Croatian seaports 246.939 vessels were received in 2013. Under the domestic flag were even 98.04% of them. Compared to 2003 the overall increase was 55.842 ships. If we analyze the data for September 2014, the movement of ships in the Croatian ports amounted to 40.359 ships which is increase of 44.9% compared to September 2013.

Table 1: Total maritime throughput in Croatian port authorities from 2005 - 2013 (in 000 tonnes)

Year	Dubrovnik	Ploče	Pula	Rijeka	Senj	Split	Šibenik	Zadar
2005	288,5	2.794,6	3.151,5	14.251,6	129,1	3.124,9	1.453,8	1.007,1
2006	231,7	3.182,1	4.287,1	12.632,1	152,6	3.416,5	1.280,4	1.142,7
2007	293,3	4.211,8	4.803,6	14.935,1	159,9	3.327,9	1.292,0	1.073,4
2008	334,4	5.103,7	4.620,0	13.682,4	145,5	3.316,0	898,5	1.122,4
2009	274,2	2.759,1	3.933,3	12.195,3	128,7	2.651,6	613,8	821,7
2010	212,4	4.510,1	3.977,0	10.727,4	93,4	3.195,8	645,7	967,7
2011	360,8	4.399,5	3.259,7	9.570,7	105,5	2.653,9	785,0	726,9
2012	361,6	2.524,6	2.033,0	9.806,2	123,7	3.190,2	434,4	498,5
2013	493,1	2.558,9	2.362,9	9.714,5	111,3	3.213,5	499,2	412,6

Source: Corresponding Croatian Bureau of Statistics' yearbooks

Unlike other countries Croatia didn’t exploit its position at the Pan-European transport corridors.

- The reason for this state of Croatian sea ports has its stronghold since the 1980s when the port of Rijeka missed an important investment cycle of modernization in ports facilities and technology. In the same period neighboring competitive ports of Koper and Trieste have invested a lot in the modernization and construction of new port capacity;
- At the same time, another reason was the inadequate and outdated railway and roads, and the traffic redirected to other European ports;
- The third reason of turning traffic flows into other ports is war impact of the 1990s because of increased risk and the introduction of additional insurance. Indeed it resulted in changes in the cargo structure, reducing the market as well as reducing the gravitational field of Croatian sea ports;
- All this caused a drop in turnover due to technological obsolescence problems of port capacity with utilization at the level of 30-50%.
- During the war time and after it comes to a dramatic disparity in the development of Croatian sea ports with the strong dominance of Rijeka. This is especially true with container traffic.

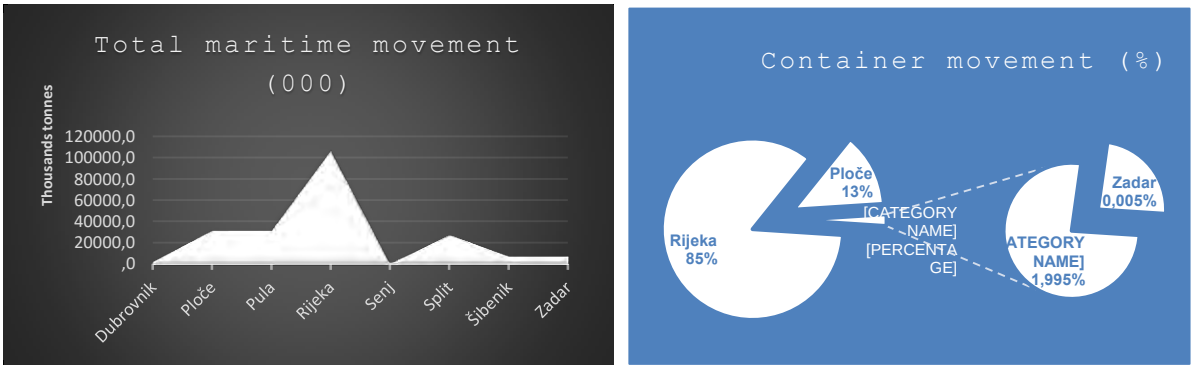


Figure 2: Total maritime and container movement in Croatian ports from 2005 - 2014 (%)
Source: Based on data in *Table 1* and traffic data of respective port authorities

Table 1 and *Figure 2* clearly illustrate the major differences between Croatian Adriatic ports. Besides, only Rijeka among the large Croatian ports is geographically part of the northern Adriatic. That's why it is the only member of NAPA and it will be the only port included in the comparative analysis with the neighboring ports.

3 COMPARATIVE ANALYSIS AND DEVELOPMENT PRIORITIES

Comparative analysis results by the NAPA ports for total and container traffic are in *Table 2*.

Table 2: Total maritime (in 000 tonnes) and container traffic (in 000 TEU) of NAPA from 2005 - 2014

Year	Rijeka		Koper		Trieste		Venice		Ravenna	
	Tonnes	*TEU*	Tonnes	*TEU*	Tonnes	*TEU*	Tonnes	*TEU*	Tonnes	*TEU*
2005	11.863,8	*76,3*	13.066,1	*179,7*	47.718,3	*198,3*	29.099,0	*289,9*	23.879,2	*168,6*
2006	10.887,0	*94,4*	14.030,7	*219,0*	48.167,7	*220,3*	30.937,0	*316,6*	26.770,2	*162,1*
2007	13.212,5	*145,0*	15.363,0	*305,6*	46.116,1	*265,9*	30.214,7	*329,5*	26.304,5	*206,8*
2008	12.391,6	*168,8*	16.050,4	*353,9*	48.279,1	*335,9*	30.247,6	*379,1*	25.904,2	*214,5*
2009	11.238,2	*130,7*	13.143,6	*343,2*	44.393,3	*277,0*	25.232,1	*369,5*	18.702,9	*185,0*
2010	10.183,3	*137,0*	15.372,0	*476,7*	47.634,2	*281,6*	26.367,9	*393,9*	21.915,0	*183,0*
2011	9.390,4	*150,7*	17.051,3	*589,3*	48.238,0	*393,2*	26.301,2	*458,4*	23.343,6	*215,3*
2012	8.554,0	*152,0*	17.880,7	*570,7*	49.206,9	*408,0*	25.395,7	*429,9*	21.460,5	*208,2*
2013	8.687,7	*169,9*	17.999,7	*600,4*	56.585,7	*458,6*	24.350,3	*446,6*	22.486,3	*226,9*
2014	9.022,8	*192,0*	18.965,4	*674,0*	57.154,0	*506,0*	21.779,1	*456,1*	24.460,2	*222,5*

Source: Based on traffic data of respective port authorities

It is evident that in the each year of the research period the largest total turnover in tonnes has achieved port of Trieste. In 2014 it had the record result with the share of 43.5% of the total selected ports turnover. In contrast, the port of Rijeka in the observed period had a decrease in turnover of 23.95% in 2014 compared to 2005. Rijeka also had the smallest individual traffic in each year through over research horizon.

Until 2010, the largest Adriatic port for container traffic (in number TEU) was Venice. From 2010 the convincing first place took over the port of Koper with an average growth rate of +6%, and in 2014 compared to 2013 its rate increased to +11%. All observed ports have experienced a marked decline in traffic at the beginning of the economic crisis of 2009. The largest decline (almost 60,000 TEU) had the port of Trieste.

In the Croatian sea ports 41% of the total turnover of goods refers to the bulk, 24% to the liquid cargo while the goods in containers constitute only 9% of the total traffic of goods. At the same time structure is more favorable for the Rijeka port where container traffic reaches 33%. However, these figures are far behind other NAPA ports where container traffic is much higher, for example in the port of Koper it is about 65% and in the port of Trieste exceed 80% intra NAPA traffic.

The results of dedicated studies for market potential of NAPA ports [6] indicate the expected continuous growth of transport in the coming years. Projections show that, for example container traffic is expected to be about 6 million TEU by 2030. So, it is expected to be 4 times greater than the traffic achieved in 2012. The study projections also predict potential traffic growth of more than 90% in Croatia, Slovenia, Hungary and Slovakia.

All it requires capacity modernization and solving the problem of "transport bottlenecks" in particular with new investments. Within the development of intermodal transport and sustainable mobility development of short sea lines, so-called "marine highways" have a special place. They are one of the priorities of the TEN-T network together with the development of port policy.

In Croatian development of intermodal transport system remains the biggest problem. Namely, although the Croatian road network is in all aspects developed above the EU average it does not provide a satisfactory solution for intermodal transport. Therefore, the most important priority for the Croatian transport development as well as for overall involvement in the TEN-T and into the Schengen area is the railway network development.

4 CONCLUSION REMARKS

Although Croatia has extremely favorable geo-traffic position by Pan-European transport corridors its comparative advantage is still far from being used. Despite the activities related to the recent started Schengen evaluation process, the comparative analysis indicators for each segment of maritime freight movements show that Croatian ports will still trot for a long time behind transportation achievements of neighboring foreign ports. Even for Rijeka, as the largest and the most developed port, intermodal transport system remains the biggest problem. Besides existing modern and fast highways, the most important priority for the Croatian transport development in this sense as well as for better integration into the Schengen area and in Pan-European transport corridors is the railway network development.

References

[1] Croatian Bureau of Statistics (http://www.dzs.hr)

[2] *Developing a Maritime Strategy and Integrated Maritime Policy*, GRC. (2014.), (http://bit.ly/1cAsESm)

[3] *Draft declaration of readiness to start the Schengen evaluation process*, GRC. (2015.)

[4] Kandžija, V., Cvečić, I., (2010.) *Economics and politics of the European Union*, Faculty of Economics-Rijeka

[5] *North Adriatic Ports Association* (http://bit.ly/1QuANFh)

[6] Regulations for Croatian sea fishing boundaries, PN. (2011.), (http://bit.ly/1M8HITt)

[7] *Revised Schengen Action Plan*, GRC. (2015.), (http://bit.ly/1Ju9Vow)

[8] *Schengen Area*, EC. (http://bit.ly/1lnN3g1)

[9] Statistical data of respective port authorities: www.portravenna.it, www.luka-ploce.hr, www.luka-kp.si, www.porto.trieste.it, www.lukarijeka.hr, www.port.ravenna.it, www.port.venice.it.

[10] Šegvić, S. (2011.) *Schengen regime for administering EU external borders*, Faculty of Law-Split

[11] *Transport Development Strategy of the Republic of Croatia (2014-2030)*, MMATI of GC. (2014.), (http://bit.ly/1di5mkR)

[12] Zerbini F. (2013.) NAPA Presentation - *Your five star route to Europe* (http://bit.ly/1MqUtJC)

SEMI-FAIR DEPLOYMENT OF THE SERVICE CENTERS IN A TRANSPORTATION NETWORK

Marek Kvet
University of Žilina, Faculty of Security Engineering
1. mája 32, 010 26 Žilina, Slovak Republic,
marek.kvet@fbi.uniza.sk

Jaroslav Janáček
University of Žilina, Faculty of Management and Informatics,
Univerzitná 1, 010 26 Žilina, Slovak Republic,
jaroslav.janacek@fri.uniza.sk

Abstract: This contribution deals with the semi-fair public service system design problem, in which the request of equal accessibility is taken into account. Within this paper, we introduce a semi-fair approach, which is based on the combination of so-called system optimal design and the lexicographic min-max optimal design. The radial formulation of the problem was used as a concept of solving technique due to previously observed excellent performance and flexibility of the approach. To evaluate the resulting system design, we use not only the price of fairness, but we have suggested a gauge of the fair deployment to evaluate the gain of the serviced community from the point of fairness. We study here mutual impact of the both objectives.

Keywords: fairness, service center deployment, lexicographical min-max, radial formulation

1 INTRODUCTION

Classical approach to the system optimal public service system design locates limited number of service centers at positions from a given finite set of possible locations to minimize the sum of distances from users' locations to the nearest located service center. The mentioned distance can represent travelling time or some cost connected with service providing. This problem can be tackled as an instance of the weighted *p*-median problem using some of the plethora developed exact and approximate methods [1], [4].

Usage of the weighted *p*-median problem for the service center deployment optimizes service accessibility of so-called average user, but it may lead to such system design, where some minority of users are caught in locations, which are inadmissibly distant from any service center. Such design is considered to be unfair, even if it is optimal from the point of the average user, i.e. it minimizes the sum of distances from users to the nearest service centers (so-called min-sum optimization). The fair designing or scheduling emerge whenever limited resources are to be fairly distributed among participants, who claim their rights to equal access to the service. The fairness has been broadly studied in [2], [8], [9] and many schemes of fairness were suggested. The lexicographic min-max criterion was denoted as the strongest one. Various approaches to the lexicographic minimization were developed [3], [10]. The effectiveness of the homogenous radial formulation led to an application of the radial approach to the lexicographic min-max location problem [5].

This paper is devoted to exploration of the idea mentioned in [10]. The idea consists in combination of the min-sum and lexicographic min-max methods. We want to use an excellent performance and flexibility of the homogenous radial method for deployment of the service centers in a transportation network. To be able to compare loss and gain from the point of the average users' disutility and the point of fair criterion, we use price of fairness [2], when fair objective is accented in the design. To be able to evaluate the gain of the serviced community from the point of fairness, we suggest a gauge of the fair deployment and study mutual impact of both objectives.

The remainder of the paper is organized as follows. Section 2 is devoted to explanation of notion of the semi-fair service center deployment. The radial approach to the semi-fair service center deployment is concisely described in Section 3 and the associated numerical experiments are performed in Section 4. The results and findings are summarized in Section 5.

2 SEMI-FAIR DEPLOYMENT OF THE SERVICE CENTERS

The optimal deployment of the service centers is defined as a task to locate at most p service centers at positions from the given finite set I so that a given objective function modelling the discomfort of users' is minimized. In the further formulations, the set J is a set of users' locations and b_j denotes the number of users sharing the location j. The distance from users' location j to the possible service center location i is denoted as d_{ij}. To model the decisions on the particular center locations, we introduce a zero-one variable $y_i \in \{0,1\}$ for each possible center location from I. The general optimal service center deployment problem can be formulated by the model (1), where m corresponds to the cardinality of I.

$$\min\{ f(\boldsymbol{y}) : \boldsymbol{y} \in \{0, 1\}^m, \sum_{i \in I} y_i \le p\} \tag{1}$$

The min-sum problem can be described by substituting function (2) for f in (1). The problem corresponds to the well-known weighted p-median problem.

$$f_s(\boldsymbol{y}) = \sum_{j \in J} b_j \min\{d_{ij} : i \in I,\ y_i = 1\} \tag{2}$$

The min-max fair deployment can be modelled by the usage of the function (3) in the model (1).

$$f_m(\boldsymbol{y}) = \max\{\min\{ d_{ij} : i \in I,\ y_i = 1\} :\ j \in J\} \tag{3}$$

If the lexicographical min-max fair problem is solved, then the distance from the worst situated user to the nearest located center is minimized first, and then the distance from the second worst situated users is minimized unless the minimal reached distance from the previously processed users gets worsened. This process is repeated until no users distance from the nearest located center can be reduced.

If the fair solution is accepted, then the min-sum objective function value computed according to (2) is worsened. Let $\boldsymbol{y}^s$ and $\boldsymbol{y}^f$ denote the system and fair optimal solutions respectively, then the relation between the system and fair optimum called the price of fairness (*POF*) [2] can be calculated according to (4).

$$POF = \frac{f_s(\boldsymbol{y}^f) - f_s(\boldsymbol{y}^s)}{f_s(\boldsymbol{y}^s)} \tag{4}$$

To express the quality of a solution of the service center deployment problem from the point of lexicographical fairness, the situation is a bit more complicated in comparison to the min-sum criterion. Given service center deployment induces differences in users' access to provided service. Disutility perceived by an individual user corresponds to the distance from the user location to the nearest located service center, and this distance gets only a value from the upper subscripted sequence $d^0 < d^1 < \ldots < d^u$ of all possible u+1 distance values, which occur in the matrix $\{d_{ij}\}$, $i \in I$, $j \in J$. Let d^w denote the highest but one member of the sequence $d^0 < d^1 < \ldots < d^u$, which limits the distance from a user location to the nearest located center. Then, the range of all disutility values can be represented by a finite set of ordered values $G_0 = d^{w+1}$, $G_1 = d^w$ … $G_w = d^1$ and $G_{w+1} = d^0$. A solution $\boldsymbol{y}$ can be characterized by the distribution vector $[B_0(\boldsymbol{y}),\ B_1(\boldsymbol{y})\ \ldots\ B_w(\boldsymbol{y})]$, where the t-th component of the vector is defined as the number of users, whose distance from the nearest located service center belongs to the semi-closed interval $(G_{t+1}, G_t]$. The lexicographic min-max problem according to [10] consists in

lexicographic minimizing of the vector $[B_0(\boldsymbol{y}), B_1(\boldsymbol{y}) \ldots B_w(\boldsymbol{y})]$ subject to $\boldsymbol{y} \in \{0, 1\}^m$ and the condition that vector $\boldsymbol{y}$ contains at most p ones.

The lexicographic ordering of the distribution vectors enables to decide on which of two different deployments is better from the point of fairness, but it does not enable to quantify the difference between them. That is why we introduce the following gauge of the min-max lexicographic fairness. First, we extend the distribution vector by the component $w+1$, which gives the number of users, whose distance from the nearest service center equals to $G_{w+1}=d^0$. After these preliminaries, the sum of the distribution vector components is equal to the number B of all users for any solution $\boldsymbol{y}$. The suggested gauge $E(\boldsymbol{B}(\boldsymbol{y}))$ of the extended distribution vector $\boldsymbol{B}(\boldsymbol{y}) = [B_0(\boldsymbol{y}), B_1(\boldsymbol{y}), \ldots, B_w(\boldsymbol{y}), B_{w+1}(\boldsymbol{y})]$ is defined by (5).

$$E(\mathbf{B}(\mathbf{y})) = \log_B \left(\left(\sum_{t=0}^{w+1} B_t(\mathbf{y}) * (B)^{w+1-t} \right)^{\frac{1}{w+2}} \right) \tag{5}$$

The suggested semi-fair deployment algorithm can be described by two phases, where the first one performs the lexicographic minimization of the distribution vector restricted to the given number of the r first components. The second phase minimizes the objective function (2) subject to additional condition that the r first components of the associated distribution vector cannot get worsened. The suggested algorithm is based on the concept of homogenous radial formulation of the service center deployment problem, which was broadly studied in [6], [7] both for the fair and system optimal deployment design.

3 TWO-PHASE METHOD FOR SEMI-FAIR DEPLOYMENT OF THE SERVICE CENTERS

The problem (1) with the objective function (3) is also known as the p-center problem, what is the task of determination of at most p network nodes as service center locations so that the maximal disutility perceived by the worst situated user is minimal. Solution of the problem can be used as the first step of the lexicographic min-max algorithm. To obtain the solution by most effective way, we have suggested a bisection method [7], which uses the above mentioned sequence $d^0 < d^1 < \ldots < d^u$ of all possible $u+1$ distance values between possible center locations and the users' locations. We define a zero-one constant a_{ij}^s for each triple $[i, j, s]$, where $i \in I$, $j \in J$, $s \in [0 \ldots u]$. The constant a_{ij}^s is equal to 1, if the distance d_{ij} between the user location j and the possible center location i is less than or equal to d^s, otherwise a_{ij}^s is equal to 0. Additionally to the zero-one location variables $y_i \in \{0, 1\}$ for $i \in I$, we introduce the variables x_j, to indicate, whether the user's disutility at location $j \in J$ following from the nearest located center is greater than d^s. In this case, the variable takes the value of 1, and it takes the value of 0 otherwise. The corresponding model can be formulated as follows.

$$Minimize \sum_{j \in J} x_j \tag{6}$$

$$Subject\ to: \quad x_j + \sum_{i \in I} a_{ij}^s y_i \geq 1 \quad for\ j \in J \tag{7}$$

$$\sum_{i \in I} y_i \leq p \tag{8}$$

$$y_i \in \{0, 1\} \quad for\ i \in I \tag{9}$$

$$x_j \geq 0 \quad for\ j \in J \tag{10}$$

In this model, the objective function (6) represents the number of user locations, which perceive disutility greater than d^s. The constraints (7) ensure that the variables x_j are allowed to take the value of 0, if there is at least one center located in radius d^s from the user location j and constraint (8) limits the number of located service centers by p. Note that even if no in-

tegrality constraint is imposed on x_j, the resulting solution of (6) – (10) will contain no variable x_j, with value different from 0 or 1.

The solution of the problem (6) – (10) indicates, whether there exists a solution of the p-center problem with objective function value less than d^s, what can be easily used by bisection process for determination of the lowest subscript v+1, where the associated value of (6) is zero. Let the subscript value $c=w-v+1$ correspond to the first nonzero component $B_c(\mathbf{y})$ of the distribution vector $[B_0(\mathbf{y}), B_1(\mathbf{y}) \ldots B_w(\mathbf{y})]$. The following lexicographical minimization of the distribution vector processes step-by-step the components $B_c(\mathbf{y}), B_{c+1}(\mathbf{y}) \ldots B_w(\mathbf{y})$. In accordance with [5], [10], we suggested an iterative process of lexicographic minimization based on solving the problem (11) – (17) for the components $B_t(\mathbf{y})$, where $t=c \ldots w$. Additionally to the zero-one location variables $y_i \in \{0, 1\}$ for $i \in I$, we introduce the variables x_{js}, to indicate, whether the distance from the user's location $j \in J$ to the nearest located center is greater than d^s. In this case, the variable takes the value of 1, and it takes the value of 0 otherwise. In the associated model, the symbol e_s denotes the difference $d^{s+1} - d^s$ and value $\underline{B}^*_k$ corresponds to the objective function value of (11) obtained in the steps preceding the step t.

$$\textit{Minimize} \quad \sum_{j \in J} b_j \sum_{s=v+c-t}^{v} e_s x_{js} \tag{11}$$

$$\textit{Subject to:} \quad x_{js} + \sum_{i \in I} a_{ij}^s y_i \geq 1 \quad \textit{for } j \in J, \quad s = 0, 1, \ldots, v \tag{12}$$

$$\sum_{i \in I} y_i \leq p \tag{13}$$

$$\sum_{j \in J} b_j \sum_{s=v+c-k}^{v} e_s x_{js} \leq \underline{B}^*_k \quad \textit{for } k = c, \ldots t-1 \tag{14}$$

$$\sum_{i \in I} a_{ij}^{v+1} y_i \geq 1 \quad \textit{for } j \in J \tag{15}$$

$$x_{js} \geq 0 \quad \textit{for } j \in J, \quad s = 0, 1, \ldots, v \tag{16}$$

$$y_i \in \{0, 1\} \quad \textit{for } i \in I \tag{17}$$

In this model, the constraints (12) ensure that the variables x_{js} are allowed to take the value 0, if there is at least one service center located in the distance d^s from the users' location j. The constraint (13) puts the limit p on the number of located service centers. The constraints (14) and (15) prevent the components $B_0(\mathbf{y}), B_1(\mathbf{y}) \ldots B_{t-1}(\mathbf{y})$ from worsening. It holds that $B_t(\mathbf{y})$ takes the value of zero for $t=0 \ldots c-1$ due to (15). Note that even if no integrality constraint is imposed on x_{js}, the resulting solution of (11) – (17) will contain no variable x_{js}, with value different from 0 or 1.

In the previously described first phase, the process of lexicographical minimization can be prematurely stopped at some level $r < w$ just after computation $B_r(\mathbf{y})$. The second phase of the suggested algorithm consists of the objective function (2) minimization subject to additional condition that the first r+1 components of the associated distribution vector cannot be worsened. Making use of the radial formulation and the above-introduced constants and variables, the second phase can be performed by solving the following problem.

$$\begin{aligned} &\textit{Minimize} \quad \sum_{j \in J} b_j \sum_{s=0}^{v} e_s x_{js} \\ &\textit{Subject to} \quad (12) - (17) \quad \textit{for } t = r+1. \end{aligned} \tag{18}$$

4 NUMERICAL EXPERIMENTS

To explore the price and loss of fairness depending on level r of the semi-fair deployment algorithm, we performed a series of numerical experiments. The used instance was derived from the real emergency health care system, which was originally implemented in Slovak Republic. The system covers demands of all communities - towns and villages spread over the region by given number of ambulance vehicles. In the benchmark, the set of communities represents both the set J of users' locations and also the set I of possible center locations. The cardinalities of these sets are equal to the number 2916. The number p of deployed centers was derived from the original design and it equals to 273. The service center deployment problem was solved first as the min-sum problem using radial approach [6]. The maximal distance $d^{w+1} = d^{24}$ from a user to the nearest located center was found by the analysis of the associated solution $\boldsymbol{y}^s$. The following usage of the bisection radial min-max algorithm (see the starting procedure of the first phase of suggested method) led to the result $d^{v+1} = d^{13}$, which means that the semi-fair deployment starts with the level $c=w-v+1= 12$. Then the two-phase method was performed for r=12, 13 … 23. The resulting solutions $\boldsymbol{y}^r = \boldsymbol{y}^{12}$ ... $\boldsymbol{y}^{23}$ were analyzed from the point of price of fairness and loss of fairness. The price of fairness was computed according to (4), where $\boldsymbol{y}^f$ was substituted by $\boldsymbol{y}^r$. The loss of fairness (LOF) was expressed using (19), where $\boldsymbol{y}^l$ lexicographical min-max optimal solution corresponds to $\boldsymbol{y}^{23}$.

$$LOF = \frac{E(\boldsymbol{B}(\boldsymbol{y}^r)) - E(\boldsymbol{B}(\boldsymbol{y}^l))}{E(\boldsymbol{B}(\boldsymbol{y}^l))} \tag{19}$$

Obtained issues are presented in Table 1 and some of them are depicted in Figure 1.

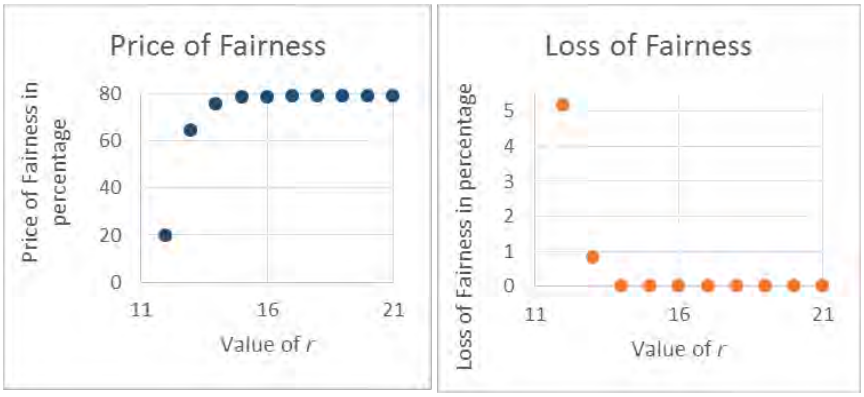


Figure 1: Trends of the price of fairness and loss of fairness depending on the level r

Table 1: Price of fairness (POF) and the loss of fairness (LOF) of the semi-fair designs for various levels r and computational time in minutes for the individual designs

r	0	12	13	14	15	16	17	18	19	20
POF	0.00	19.58	64.29	75.54	78.30	78.52	78.78	78.90	78.90	78.90
LOF		5.18	0.83	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Time [min]	0.29	0.19	3.18	8.35	8.64	16.99	16.54	35.32	68.40	58.58

To solve the problems described in the previous sections, the optimization software FICO Xpress 7.7 (64-bit, release 2014) was used and the experiments were run on a PC equipped with the Intel® Core™ i7 2630 QM processor with the parameters: 2.0 GHz and 8 GB RAM.

5 CONCLUSIONS

We suggested a semi-fair design of public service system using the radial approach, where various levels of the fairness were considered. The lexicographic min-max deployment of service centers was taken as the fair optimal public service system design. The semi-fair design consists of replacing a portion of less preferred levels of the min-max optimization by performing one min-sum optimization process. We studied loss and gain of the average user and the most exposed users. On one side, we evaluated the price of fairness to express the loss of the average user, when the semi-fair approach is applied and on the other side; we evaluated the loss of fairness for the most exposed users. We have found, that at most four most prior levels of lexicographical minimization play role. This finding seems to be very important from the point of practical service system designing as the associated computational time of the lexicographic minimization dramatically grows with increasing number of considered levels. The future research will be focused on a possible relaxation of the biggest distance representing the first level of the lexicographic minimization with the intention to mitigate the differences between the system optimal solution and the min-max solution.

Acknowledgment: This work was supported by the research grants VEGA 1/0518/15 "Resilient rescue systems with uncertain accessibility of service" and VEGA 1/0339/13 „Advanced microscopic modelling and complex data sources for designing spatially large public service systems" and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks". This work was also supported by the project University Science Park of the University of Žilina (ITMS: 26220220184) supported by the Research & Development Operational Program funded by the European Regional Development Fund. We would like to thank to „Centre of excellence for systems and services of intelligent transport" (ITMS 26220120050) for built up the infrastructure, which was used.

References

[1] Avella, P., Sassano, A. and Vasil'ev, I. (2007). Computational study of large scale p-median problems. Mathematical Programming, 109: 89-114.

[2] Bertsimas, D., Farias, V. F., Trichakis, N. (2011). The Price of Fairness. Oper. Res. 59: 17–31.

[3] Buzna, Ľ., Kohání, M. and Janáček, J. (2014). An approximation algorithm for the facility location problem with lexicographic minimax objective. Journal of Applied Mathematics. 2014: 1–12.

[4] García, S., Labbé, M. and Marín, A. (2011). Solving large p-median problems with a radius formulation. INFORMS Journal on Computing, 23(4): 546-556.

[5] Janáček, J. and Kvet, M. (2014). Lexicographic optimal public service system design. In Conference Olomouc Proceedings of 32nd International Conference Mathematical Methods in Economics, September 10-12 2014 (pp 366-371). Olomouc: Palacký University.

[6] Janáček, J. and Kvet, M. (2014). Relevant Network Distances for Approximate Approach to the p-median Problem. In Selected Papers of the International Conference of the German operations research society (GOR), September 4-7 2012 (pp 123-128). Hannover: Springer.

[7] Kvet, M. and Janáček, J. (2015). Min-max optimal public service system design. Croatian Operational Research Review, 6(1): 17-27.

[8] Marsh, M. and Schilling, D. (1994). Equity measurement in facility location analysis. European Journal of Operational Research, 74: 1–17.

[9] Nash, J. (1950). The bargaining problem. Econometrica, 18 (2): 155–162.

[10] Ogryczak, W. and Slivinski, T. (2006). On Direct Methods for Lexicographic Min-Max Optimization. In: Gavrilova, M. (Eds.): ICCSA 2006, LNCS 3982, (pp. 802-811). Berlin: Heidelberg: Springer.

FUZZY LOGIC BASED QUALITY ASSESSMENT IN KINEMATIC GNSS-POSITIONING

Polona Pavlovčič Prešeren, Bojan Stopar, Miran Kuhar and Oskar Sterle
University of Ljubljana, Slovenia
{polona.pavlovcic, bojan.stopar, miran.kuhar, oskar.sterle}@fgg.uni-lj.si

Abstract: This paper presents a fuzzy logic application in kinematic GNSS (Global Navigation Satellite System) positioning, which is often used method in geodetic surveying. The aim of this paper is to examine the quality of dual frequency carrier-phase observation results, obtained under various measurement conditions. The latter are described by the satellite constellation (given in factors of dilution of precision *DOP*) and by the occupation duration on a point, where the observations are performed. In our experiment the quality factors were obtained using the commercial GNSS post- processing software. Further we have used them as input data in our fuzzy logic software with the aim to get final ratings of the GNSS-positioning quality achievement. We have shown that fuzzy logic approach could be an useful tool in decision making whether GNSS-processing results (coordinates and their accuracy) from kinematic positioning upon specific measurement conditions will be successful or not.

Key-words: fuzzy logic, kinematic GNSS-positioning, geodetic surveying, dual frequency carrier-phase observations, dilution of precision *DOP*, occupation duration

1 INTRODUCTION

Global navigation satellite systems (GNSS), that in present time include mostly GPS and GLONASS, but also growing Galileo and Compass, has revolutionized 3D geodetic positioning. In Slovenia GNSS positioning has gained more important role since 2008, but also previously has been often used. 1st January 2008 is the establishment date of the new national horizontal reference system D96/TM. Since then all technical issues regarding land cadastral surveying have to be performed in the new D96/TM coordinate system. GNSS allows us real-time positioning in global coordinates system of a homogenous accuracy. This is the reason that GNSS has become very often used technique in geodetic surveying. Among GNSS methods kinematic method that allows us real-time coordinate determination is of particular importance.

Real time kinematic (RTK) as one of the GNSS methods uses carrier-phase observations and provides a centimeter-level accuracy of coordinates in real time. The method is fast, but depends on optimal measurement conditions, that are referred to good satellite geometry and no physical obstructions to reception of the satellite signal in the vicinity of point's location. It should be emphasized, that RTK has its benefits in real-time positioning; however, it does not achieve the best possible positioning accuracy. Static GNSS method outperforms RTK in accuracy achievement, but it has its drawbacks in the need of longer occupation duration and in the need of post-processing.

In kinematic point positioning there are still several topics to discuss, which should be answered by empirical tests. These answers could be used at the field-work as valuable information. For example, it should be clarified how the occupation duration affects the accuracy. Some authors already showed the benefit of averaging point positions acquired several times over 1-2 minute long occupation duration with the intermediate interval of 10-30 minutes [1]. It is well known fact that most of the errors in kinematic positioning could be solved using redundant measurements after the elapse of a sufficient time. There is also not clarified, how much the obstructions affect the achieved positional accuracy. Interesting results have been introduced from investigation of the RTK achievable accuracy and repeatability under different satellite constellations and measurement conditions [2]. In

several situations we have different information at disposal, but further we have to make the decision on the GNSS positioning quality achievement, estimated from available information. Traditionally all quality models are based on probability distribution of position errors. In contrast, fuzzy logic in its wider sense is derived from a fuzzy set theory, which deals with classes of objects with non-sharp boundaries and permits assessment of a specific element membership by membership functions [3].

There are several attempts of introducing fuzzy logic into GNSS positioning; using C/A-code based GPS-measurements (static positioning) in combination with the signal-to-noise ratio (SNR) in order to show the quality of position determination [4]. Correlation between cut-off angle and the occupation duration in static GNSS-processing in different vicinities from the base stations have been introduced [5]. In current paper we have focused on the fuzzy logic use in kinematic and real-time kinematic method with different occupation durations and different measurement conditions.

1.1 Kinematic and real-time kinematic GNSS positioning

RTK method originates from the mid-1990s. The method itself has a lot of in common with its predecessor DGNSS (differential GNSS), which was the first geodetic method used for real-time positioning. Both methods are relative and in both cases we obtain 3D-coordinates in a global coordinate system while measuring. On the other hand methods differ considerably. While RTK is based on better quality carrier-phase measurements, DGNSS uses less accurate code observations. The final accuracy of the coordinates, acquired from DGNSS, is from 0.5 to 1 m, but in case of RTK the horizontal accuracy is about 1-2 cm and vertical is twice as that [6]. Relative principle and carrier-phase usage enables cancellation of main errors that affect stand-alone positioning.

Figure 1 presents the principle of relative positioning: the unknown point coordinates (left) are computed relatively to the known coordinates of the GNSS base-station. Real-time observation processing is based on a communication link between two GNSS receivers, where transmission takes place in accordance with the industrial standard RTCM SC 104.

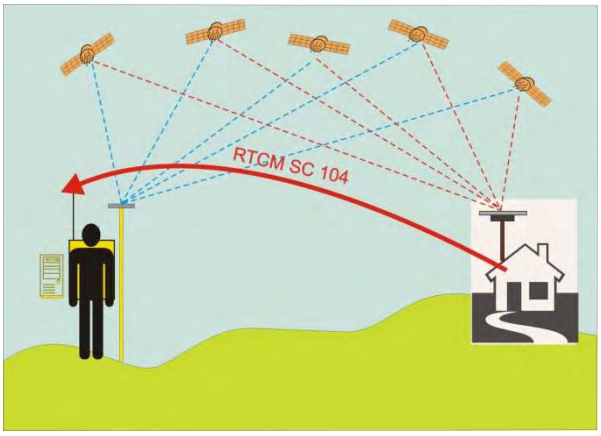


Figure 1: Real time kinematic GNSS positioning

In RTK the quality of coordinates can be checked while measuring. The most important in the RTK-performance is the integer ambiguity resolution, which is specific only for the methods, which use carrier-phase observations. The quality of the coordinates depends on the occupation duration and on the number of visible GNSS-satellites at the specific location.

Usually, *DOP* factors are used to describe the quality of satellite configuration, which considerably affects the accuracy of final coordinates.

1.2 GNSS measurement conditions

GNSS is based on a trilateration or resection method, where the unknown point coordinates are determined on the basis of measured distances from points with known coordinates, i.e. satellites. To solve the resection problem in a space at least 3 distances from the satellites are needed. But we all know in GNSS positioning four measured distances are needed, three of them for calculating 3D position and the fourth for time synchronization of a receiver's clock with more accurate satellites' clocks. One of the major factors that affect the point coordinates' accuracy is the geometry of satellites in view. When the satellites are located relative close to each other, the overlapping area of position uncertainty is larger and vice versa. Dilution of precision (*DOP*) is the ratio between coordinates and measurement accuracy. As *DOP* values increase, both the horizontal and vertical precision of point position decrease [7].

The satellite-receiver geometry can be described by the assumption of having only four satellites in view. There four unit vectors forming a tetrahedron (Figure 2). The volume of tetrahedron and *DOP* factors are functionally dependent. Larger tetrahedron's volume means smaller *DOP* and smaller *DOP* indicates good measurement conditions.

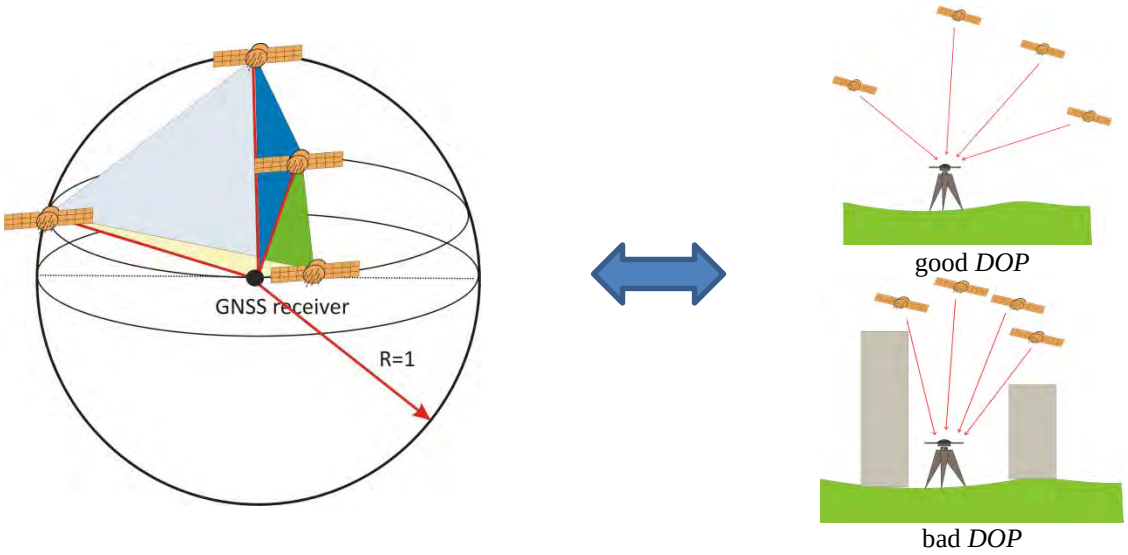


Figure 2: Determination of *DOP* on the unit sphere from geometry of four best satellites (left) and presentation of good and bad DOPs (right).

The overall quality assessment of static or kinematic measurements under different conditions (cut-off angles and occupation durations) could be described by:

- *DOP* values (maximal value of *PDOP* shouldn't exceed 6; but more information on *DOP* values is described in [8]),
- covariance matrix Σ of the estimated receiver's coordinates (x_i, y_i, z_i) given by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{x_i}^2 & \sigma_{x_i y_i} & \sigma_{x_i z_i} \\ \sigma_{y_i x_i} & \sigma_{y_i}^2 & \sigma_{y_i z_i} \\ \sigma_{z_i x_i} & \sigma_{z_i y_i} & \sigma_{z_i}^2 \end{bmatrix} \quad (1)$$

In our particular case both factors we computed using commercial GNSS carrier-phase processing software.

In GNSS positioning it is well known fact: lower the *DOP* and longer occupation durations yields to better position determination. Since the measurements cannot always be performed in ideal conditions, we have we have to find the compromise also for situations between with the compromise we want to find the ratings for the quality assessment by different measurement conditions.

2 FUZZY LOGIC IMPLEMENTATION

In our study we estimated the influence of static point occupation duration in kinematic surveying as well as different cut-of elevation angles on final position accuracy. Since we wanted to avoid extreme cases, known as »good« or »bad« positioning, determined only from one specific parameter (only from *DOP* for example), we determined various states between those two. We used fuzzy logic, conceived by Lotfi Zadeh in 1960s, because it incorporates a rule-based approach to control the problem [9]. Output linguistic variables are described by the terms (in case of three terms: *good medium, bad*) or by pre-defined numbers for ratings. It is based on an empirical solution and relies on the experiences rather than on technical understanding of the system [10].

Fuzzy logic algorithm consists of four steps: first, defining inputs (in our case *DOP* and occupation durations) and outputs (accuracy). Second we define membership functions, and then creating the rules and finally in fourth step by simulating results of fuzzy logic system to get the ratings under different conditions. In our particular case the rating were described by weights. Information from experiences in GNSS quality positioning, described by rules, was put into fuzzy logic system. We have followed the rules:

- if *PDOP* is high (above 6), the solution is not good (1st membership function)→ *low weight*
- if *PDOP* is medium and accuracy of coordinates is small (2nd membership function)→ *weight is medium*
- if accuracy of coordinates is small (as far we have oriented ourselves on kinematic positioning, we have used the 5 cm accuracy for 3D point positioning as a good result), we have good solution (2nd membership function)→ *weight is high*

Numerical values of all above rules were –1, 0, +1 and weight of the specific rule *i* was computed by equation [11]:

$$p_i = \frac{min(parameter)}{value_i} \tag{2}$$

Where $min(parameter)$ corresponded to the minimal parameter of the specific membership function (in our case Z-shaped membership function was used). Computed values of weights (Equation 2) were in the range of [0, 1]. For this particular case we have defined the weight value 0.25 for bad solution, 0.50 for medium and 0.75 for good solution. If the value exceeded 0.75, the point positioning would be of a better quality as expected, if the value was below 0.25, the solution would be even worse as expected.

3 NUMERICAL SIMULATIONS

Numerical simulations using different values of *DOP* and position accuracies were performed using fuzzy logic scripts, implemented in MATLAB R2014a [3]. After performing all previously needed computations on point position quality description (*DO*P

factors, accuracies of coordinates), software gives us several ratings to identify quality of point coordinates under specific conditions (occupation durations and factors *DOP*). The larger the value of computed weight, the better the quality of point coordinates and vice versa. An interesting parameter, which affects the quality of GNSS coordinates, is the elevation cut-off angle. Signals from satellites with low elevation angle are noisier, but 3D positioning accuracy with satellites' signals above certain elevation angle affects the point coordinates quality, especially the height. Occupation duration on site prior performing measurements is the other factor in kinematic positioning, which affects the quality of computed coordinates.

In Table 1 we show the ratings (described in weights) for post-processed kinematic GNSS observations from the baseline length of approximately 4 km. Observations were performed on 20th March 2015 and were post-processed using precise ephemerides.

Table 1: Fuzzy logic results for different occupation durations and cut-off angles (kinematic GNSS positioning)

	Elevation angle 5°	**Elevation angle 10°**	**Elevation angle 15°**	**Elevation angle 20°**	**Elevation angle 25°**	**Elevation angle 35°**
Occupation duration	Computed weight	Computed weight	Computed weight	Computed weight	Computed weight	Computed weight
5 min	0.80	0.80	0.80	0.80	0.72	0.50
2.5 min	0.80	0.80	0.72	0.72	0.72	0.50
1 min	0.72	0.72	0.72	0.65	0.65	0.45
: 30 s	0.65	0.65	0.65	0.50	0.45	0.45
15 s	0.30	0.30	0.30	0.25	**0.22**	**0.22**
5 s	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

According to the results from Table 1 we can see there are small differences between 5-minute and 2.5-minute occupation durations, but there are significant differences of 2.5 minutes to just a few seconds' occupation durations. In case of a few-second occupation duration coordinates are determined only by code-solution with the accuracy of several meters, so fuzzy logic solution was marked as extremely bad (0.00). Other values, presented in Table 1, are according to our expectation – it is commonly known that longer occupation duration leads to well-determined point coordinates. But we should know that for some other locations or for different occupation durations the results of coordinate determination could be different, since GNSS performance follows dynamic satellite constellation changing during the day (in case of GPS constellation repeats after 12 hours). That means that in some particular cases, where more satellites could be available, coordinate determination could be faster, in some cases slower, but in the cases with less than five satellites, which is minimal number for the kinematic algorithms, carrier-phase solution could even not be obtained.

4 CONCLUSIONS

Our study showed fuzzy logic can be used as an interesting additional tool for positioning quality description, especially in cases, when one of the parameters (cut-off angle that affects *DOP* or occupation duration) could not be provided well. In future several topics should be taken into consideration in order to get answers on questions:

- what is the quality of positioning achievement on the same locations during different times of the day, namely from different GNSS constellations;
- what is the quality of positioning performance on the same locations with different measurement conditions (different time of year);
- what is the positioning performance at different locations, where real obstructions of GNSS satellite signals exist (in our case there we no obstructions, but we have simulated them using different cut-off elevation angles).

Acknowledgement
The authors would like to thank to the referee, who immediately reviewed the earlier version of this paper and provided valuable suggestions and comments.

References

[1] Janssen, V., Haasdyk, J., McElroy, S. and Kinsyside, D. (2011). CORSnet-NSW: Improving positioning infrastructure for New South Wales, to appear in Proceedings of Surveying & Spatial Sciences Institute Biennal International Conference (SSSC2011), Wellington, New Zealand.

[2] Pirti, A., Yucel, M. A. and Gumus K. (2013). Testing real time kinematic GNSS (GPS and GPS/GLONASS) methods in obstructed and unobstructed sites. Geodetski vestnik 57(3): 498-512.

[3] Mathworks. (2014). Fuzzy logic Toolbox: User's Guide (r2014). Retrieved from: www.mathworks.com/help/pdf_doc/gads/gads_tb.pdf

[4] Crocetto, N., Ponte, S., Pugliano, G. and Savino, L., 2007. Fuzzy logic based methodologies for mobile mapping: enhancing positioning accuracy of GPS/GNSS measurements. ISPRS proceedings XXVVI/5-C55, Padua Italy.

[5] Kostov, G., 2012. Study on the overall quality of the planned fast static GNSS measurements, if certain values of the parameters are applied in the system, using fuzzy logic. FIG Working week 2012, Rome Italy. Retrieved from: https://www.fig.net/resources/proceedings/fig_proceedings/fig2012/papers/ts03d/TS03D_kostov_5678.pdf

[6] Gakstatter, E., 2014. Centimeter-level RTK accuracy more and more available – for less and less. GPS World, 2014. Retrieved from: http://gpsworld.com/centimeter-level-rtk-accuracy-more-and-more-available-for-less-and-less/

[7] Hofmann-Wellenhof, B., Lichtenegger, H. and Collins, J., 2002. GPS Theory and practice. Springer-Verlag Wien.

[8] Langley, R. B., 1999. Dilution of precision. GPS World, 1999. Retrieved from: http://www.nrem.iastate.edu/class/assets/nrem446_546/week3/Dilution_of_Precision.pdf

[9] Hajek, P., 2010. Fuzzy logic. The Stanford Encyclopedia of Philosophy *(Fall 2010 Edition)*, E.N. Zalta. Retrieved from: http://plato.stanford.edu/archives/fall2010/entries/logic-fuzzy

[10] Rojas, R. (1996). Fuzzy logic. In Rojas, R. Neural networks – a systematic introduction (pp. 289-310). Springer-Verlag, Berlin, New-York.

[11] Kostov, G., 2005. Assessment of the Quality of Geodetic Networks Using fuzzy Logic (pp. 2-7). Vienna University of Technology.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 8:

Environment and Human Resources

ANALYSIS OF STUDENTS' FLOW IN HIGHER EDUCATION STUDY PROGRAMMES USING DISCRETE HOMOGENEOUS MARKOV CHAINS

Alenka Brezavšček and Alenka Baggia
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
E-mail: alenka.brezavscek@fov.uni-mb.si; alenka.baggia@fov.uni-mb.si

Abstract: The paper examines an application of Markov chains to the flow of students at the undergraduate study programme. The student flow has been modelled by a reducible discrete Markov chain with five transient and two absorbing states. The probabilities of absorption (graduating and withdrawal) were obtained. Furthermore, the fundamental matrix was calculated to determine the expected length of students' stay before finishing the study. The model was tested on the first degree professional study programme at the University of Maribor, Faculty of Organizational Sciences.

Keywords: student flow, higher education, discrete Markov chain, probability of absorption, expected time until absorption.

1 INTRODUCTION

Stochastic processes provide a well-investigated subset of problems of interest for operations research. An important family of stochastic processes are Markov chains, defined as a random sequence in which the dependency of the successive events goes back only one unit in time. In other words, the future probability behaviour of the process depends only on the present state of the process and is not influenced by its past history. This is called the Markovian property. Despite a very simple structure, Markov chains are extremely useful in a wide variety of practical probability problems [5]. The application of Markov chains can be found in various branches of natural sciences, engineering, and medical sciences (see e.g. [3]). Students progressing toward completing their undergraduate degrees possess all the pertinent stochastic characteristics, and can therefore be modelled as a Markov chain. The application of Markov chain to the student flows provides a means for projecting the number of students graduating and withdrawing by age, by gender, and by study programme. As such, the model also provides estimates of the average time a student stays in the system, the probability of completion as well as the average time to complete the study [1, 2]. In this paper we present the model which can be used for analysing the student flow in Slovenian higher education. The model will be applied to the first degree professional study programmes at the University of Maribor, Faculty of Organizational Sciences.

2 DISCRETE-TIME MARKOV CHAINS

A discrete-time Markov chain is a sequence of random variables $\{X_n, n = 0,1,...\}$ with a discrete state space *S*. The state space *S* is a set of all possible values of X_n in times n=0,1,... The set *S* is assumed to be finite or countably infinite.

The probability of transition between the two states of a Markov chain in a given time interval is called the transition probability. The *one-step transition probabilities* are of particular importance. We denote the one-step transition probability between the states *i* and *j* by the symbol p_{ij}. If the state space is finite and equal to $S = \{1, 2, \ldots, N\}$ then all of the one-step transition probabilities can be represented by a *probability transition matrix* **P**:

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & N \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1,N} \\ p_{21} & p_{22} & \cdots & p_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,N} \end{bmatrix} \end{array} \tag{1}$$

A Markov chain is said to be *homogeneous* if transition probabilities p_{ij} in **P** are constant over time, i.e. independent on the time parameter *n*.

The states of a Markov chain fall into distinct types according to their limiting behaviour. If the ultimate return to the initial state *j* is a certain event, then the state *j* is called *recurrent*. If the ultimate return to the initial state *j* has probability less than 1, the state *j* is called *transient*. The states of an arbitrary Markov chain can thus be divided into two sets (one of them may be empty): one set composed of all the recurrent states, and the other set composed of all the transient states. The recurrent states can be decomposed uniquely into *closed sets*. A closed set is a set of recurrent states from which there is no escape. Once a closed set is entered it is never vacated. A state *j* is *absorbing* if {*j*} is a closed set. A Markov chain consisting of both sets of states, recurrent and transient, is *reducible*. The set of the transient states is usually denoted by *T*, while the closed sets are denoted by C_i, i=1...*l*. If the set *T* is empty (i.e. all the states are recurrent), the Markov chain is *irreducible*.

When a finite reducible Markov chain contains *l* closed sets, the probability transition matrix **P** can be written in the following form:

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R}_1 & \cdots & \mathbf{R}_l \\ \mathbf{0} & \mathbf{P}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_l \end{bmatrix} \tag{2}$$

where the matrix **Q** contains one-step transition probabilities between the transient states in the set *T*, the matrices $\mathbf{P}_i$, i=1, 2, …, *l*, contain one-step transition probabilities within the *i*-th closed set C_i, and the matrices $\mathbf{R}_i$ i=1, 2, …, *l*, contain one-step transition probabilities from the transient states into the closed set C_i. All the elements in the matrices denoted by **0** are zeroes.

Once the matrices **Q**, $\mathbf{P}_i$, and $\mathbf{R}_i$, i=1, 2, …, *l*, are known, one can determine the probabilities that the Markov chain leaves the transient states and is *absorbed* into one of the closed sets of the recurrent states. Besides, *the expected time until absorption* can be calculated. In order to calculate these characteristics of a Markov chain we need the *fundamental matrix* **N** which can be obtained as follows:

$$\mathbf{N} = \left(\mathbf{I} - \mathbf{Q}\right)^{-1} \tag{3}$$

where **I** denotes the identity matrix. The elements of **N** represent the expected number of visits to a given transient state.

The probabilities of absorption from a given transient state to the closed set C_i, $i = 1, 2, \ldots, l$, can be obtained from the matrix $\mathbf{f}_{\{C_i\}}$ which is calculated from the equation [4]:

$$\mathbf{f}_{\{C_i\}} = \mathbf{NR} \tag{4}$$

where **R** represents the matrix composed from all the matrices $\mathbf{R}_i$ $i = 1, 2, \ldots, l$.

The expected time until absorption can be obtained from the column vector $\boldsymbol{\mu}$ calculated from the equation [4]:

$$\boldsymbol{\mu} = \mathbf{N1} \tag{5}$$

where **1** is the column identity vector.

3 THE MODEL

A typical first degree study programme at Slovenian universities lasts three years. After finishing the third year, a student can enrol into the so-called candidate year. During this year the student needs to write a thesis but is not obliged to attend the lectures. Therefore, to model the student flow we will use a finite Markov chain with the following states:

1 – the student is enrolled into first year of the study programme
2 – the student is enrolled into second year of the study programme
3 – the student is enrolled into third year of the study programme
C – the student is enrolled into the candidate year
I – the student is currently inactive
G – the student has graduated and successfully finished the study programme
W – the student has withdrawn from the study programme

The probability transition matrix for this Markov chain is based on the following assumptions:

- The student who is currently enrolled into the first or second year of the study programme can next year either progress to a higher level or repeat a year, staying at the same level.
- The student who is currently enrolled into the third year of the study programme can next year be either enrolled into the candidate year, or can graduate and finish the study.
- Irrespective of the level of the study, every year some students can become inactive.
- The student who is inactive for more than one year is classified as having withdrawn from the study programme.
- The student who has withdrawn will never finish this study programme. We have not noted whether or not he/she has been transferred to another study programme.
- The student who has graduated and successfully finished the study will never apply for the same study programme again. We have not noted whether or not he/she has applied for another study programme or continued the education at postgraduate level.

The probability transition matrix describing the progression of students from the first study year towards graduation is:

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \\ G \\ W \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & C & I & G & W \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & 0 & 0 & p_{1I} & 0 & p_{1W} \\ 0 & p_{22} & p_{23} & 0 & p_{2I} & 0 & p_{2W} \\ 0 & 0 & p_{33} & p_{3C} & p_{3I} & p_{3G} & p_{3W} \\ 0 & 0 & 0 & p_{CC} & p_{CI} & p_{CG} & p_{CW} \\ p_{I1} & p_{I2} & p_{I3} & p_{IC} & 0 & p_{IG} & p_{IW} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \tag{6}$$

It is evident that Markov chain (6) is reducible. It consists of two closed sets of absorbing states $C_1 = \{G\}$ and $C_2 = \{W\}$, and of five transient states $T = \{1, 2, 3, C, I\}$. The matrices **Q**, $\mathbf{R}_1$ and $\mathbf{R}_2$ can be obtained from (6) as follows:

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & C & I \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & 0 & 0 & p_{1I} \\ 0 & p_{22} & p_{23} & 0 & p_{2I} \\ 0 & 0 & p_{33} & p_{3C} & p_{3I} \\ 0 & 0 & 0 & p_{CC} & p_{CI} \\ p_{I1} & p_{I2} & p_{I3} & p_{IC} & 0 \end{bmatrix} \end{array} \quad \mathbf{R}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{c} G \end{array} \\ \begin{bmatrix} 0 \\ 0 \\ p_{3G} \\ p_{CG} \\ p_{IG} \end{bmatrix} \end{array} \quad \mathbf{R}_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{c} W \end{array} \\ \begin{bmatrix} p_{1W} \\ p_{2W} \\ p_{3W} \\ p_{CW} \\ p_{IW} \end{bmatrix} \end{array} \tag{6}$$

Using the matrix **Q** from (6) we can calculate the fundamental matrix **N** according to (3):

$$\mathbf{N} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & C & I \end{array} \\ \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{1C} & n_{1I} \\ n_{21} & n_{22} & n_{23} & n_{2C} & n_{2I} \\ n_{31} & n_{32} & n_{33} & n_{3C} & n_{3I} \\ n_{C1} & n_{C2} & n_{C3} & n_{CC} & n_{CI} \\ n_{I1} & n_{I2} & n_{I3} & n_{IC} & n_{II} \end{bmatrix} \end{array} \tag{7}$$

The elements n_{ij} in **N** represent the expected number of years the student spends to complete the *j*-th level of the study when he is currently at the *i*-th study level.

The probability of absorption is calculated according to (4). The result is the matrix $\mathbf{f}_{\{C_1,C_2\}}$:

$$\mathbf{f}_{\{C_1,C_2\}} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{cc} G & W \end{array} \\ \begin{bmatrix} p_{1G} & p_{1W} \\ p_{2G} & p_{2W} \\ p_{3G} & p_{3W} \\ p_{CG} & p_{CW} \\ p_{IG} & p_{IW} \end{bmatrix} \end{array} \tag{8}$$

The values p_{iG} in the first column of (8) represent the fraction of students, currently at the *i*-th study level, who will actually graduate and successfully finish the study. However, the values in the second column of (8), p_{iW}, represent the fraction of students who will withdraw from the study programme and will never finish it.

The expected time until absorption is calculated according to (5). The result is the column vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \\ \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_C \\ \mu_I \end{bmatrix} \end{array} \tag{9}$$

The particular value μ_i in (9) can be interpreted as the expected duration of the study (till graduation or withdrawal) from the *i*-th level of the study programme.

4 NUMERICAL EXAMPLE

To test the model, data were collected from the students' intake records at the University of Maribor, Faculty of Organizational Sciences. In our analysis only the first degree professional study programmes were included. The cumulative data of student progression during six academic years from 2008/09 to 2014/15 are listed in Table 1.

Table 1: Student progression through years from 2008/09 to 2014/15

2008/09 → 2009/10								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2008/09**
1	19	60	0	0	93	0	44	**216**
2	0	20	78	0	16	0	0	**114**
3	0	0	0	0	44	21	0	**65**
C	-	-	-	-	-	-	-	**0**
I	0	19	6	0	0	0	121	**146**
new	213							
Σ 2009/10	**232**	**99**	**84**	**0**	**153**	**21**	**165**	

2009/10 → 2010/11								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2009/10**
1	27	88	0	0	109	0	8	**232**
2	0	14	64	0	21	0	0	**99**
3	0	0	0	38	8	38	0	**84**
C	0	0	0	0	0	0	0	**0**
I	0	19	4	0	0	0	130	**153**
new	210							
Σ 2010/11	**237**	**121**	**68**	**38**	**138**	**38**	**138**	

2010/11 → 2011/12								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2010/11**
1	17	90	0	0	130	0	0	**237**
2	0	40	63	0	18	0	0	**121**
3	0	0	0	59	0	9	0	**68**
C	0	0	0	0	0	38	0	**38**
I	0	0	0	0	0	3	135	**138**
new	222							
Σ 2011/12	**239**	**130**	**63**	**59**	**148**	**50**	**135**	

2011/12 → 2012/13								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2011/12**
1	12	75	0	0	152	0	0	**239**
2	0	25	63	0	42	0	0	**130**
3	0	0	0	63	0	0	0	**63**
C	0	0	0	0	2	57	0	**59**
I	0	0	0	0	0	0	148	**148**
new	212							
Σ 2012/13	**224**	**100**	**63**	**63**	**196**	**57**	**148**	

2012/13 → 2013/14								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2012/13**
1	10	59	0	0	155	0	0	**224**
2	0	18	40	0	42	0	0	**100**
3	0	0	0	43	0	20	0	**63**
C	0	0	0	5	11	27	20	**63**
I	0	5	0	0	0	0	191	**196**
new	166							
Σ 2013/14	**176**	**82**	**40**	**48**	**208**	**47**	**211**	

2013/14 → 2014/15								
	1	**2**	**3**	**C**	**I**	**G**	**W**	**Σ 2013/14**
1	4	40	0	0	132	0	0	**176**
2	0	16	28	0	38	0	0	**82**
3	0	0	0	35	0	5	0	**40**
C	0	0	0	0	17	31	0	**48**
I	2	2	20	0	0	0	184	**208**
new	167							
Σ 2014/15	**173**	**58**	**48**	**35**	**187**	**36**	**184**	

Data in Table 1 were used to estimate the transition probabilities p_{ij} in (6). The values p_{ij} were calculated as the average transition probabilities considering all six academic years. The probability transition matrix is:

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \\ G \\ W \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & C & I & G & W \end{array} \\ \begin{bmatrix} 0.066 & 0.307 & 0 & 0 & 0.588 & 0 & 0.040 \\ 0 & 0.202 & 0.513 & 0 & 0.285 & 0 & 0 \\ 0 & 0 & 0 & 0.646 & 0.129 & 0.225 & 0 \\ 0 & 0 & 0 & 0.026 & 0.188 & 0.680 & 0.106 \\ 0.001 & 0.062 & 0.023 & 0 & 0 & 0.003 & 0.910 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Considering **P**, the matrices (6) are then equal to:

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & C & I \end{array} \\ \begin{bmatrix} 0.066 & 0.307 & 0 & 0 & 0.588 \\ 0 & 0.202 & 0.513 & 0 & 0.285 \\ 0 & 0 & 0 & 0.646 & 0.129 \\ 0 & 0 & 0 & 0.026 & 0.188 \\ 0.001 & 0.062 & 0.023 & 0 & 0 \end{bmatrix} \end{array} \quad \mathbf{R}_1 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{c} G \end{array} \\ \begin{bmatrix} 0 \\ 0 \\ 0.225 \\ 0.680 \\ 0.003 \end{bmatrix} \end{array} \quad \mathbf{R}_2 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{c} W \end{array} \\ \begin{bmatrix} 0.040 \\ 0 \\ 0 \\ 0.106 \\ 0.910 \end{bmatrix} \end{array}$$

Using **Q** the fundamental matrix (7) is calculated:

$$\mathbf{N} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & C & I \end{array} \\ \begin{bmatrix} 1.072 & 0.477 & 0.264 & 0.175 & 0.883 \\ 0.001 & 1.295 & 0.677 & 0.449 & 0.541 \\ 0 & 0.021 & 1.017 & 0.674 & 0.264 \\ 0 & 0.016 & 0.013 & 1.035 & 0.201 \\ 0.001 & 0.081 & 0.066 & 0.044 & 1.040 \end{bmatrix} \end{array}$$

Finally, the probability of absorption (8) and the expected time until absorption (9) are calculated:

$$\mathbf{f}_{\{C_1,C_2\}} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \begin{array}{cc} G & W \end{array} \\ \begin{bmatrix} 0.18 & 0.82 \\ 0.46 & 0.54 \\ 0.69 & 0.31 \\ 0.71 & 0.29 \\ 0.05 & 0.95 \end{bmatrix} \end{array} \qquad \boldsymbol{\mu} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ C \\ I \end{array} \begin{array}{c} \\ \begin{bmatrix} 2.82 \\ 2.96 \\ 1.98 \\ 1.26 \\ 1.23 \end{bmatrix} \end{array}$$

5 RESULTS AND CONCLUSIONS

The values in the column vector $\boldsymbol{\mu}$ represent the expected time (in years) needed to finish the study (graduation or withdrawal) from a particular level of the study programme. We can see that the student, who is currently enrolled in the first year, needs on average 2.82 years to finish the study. Unfortunately, the results in the matrix $\mathbf{f}_{\{C_1,C_2\}}$ show that only 18% of these students will actually graduate, while the withdrawal probability is very high (82%). This is probably because most of the first year students are rather confused due to change of educational environment and their inability to understand the tenets of academic work. We can see that the probability of graduation increases and the probability of withdrawal decreases as the students progress to higher levels. This may be the result of the fact that they understand the system better as they pass from one level to another. Results in $\mathbf{f}_{\{C_1,C_2\}}$ and $\boldsymbol{\mu}$ also show that that the majority of inactive students (95%) will never graduate. On average they leave the programme after 1.23 years.

In our opinion such analysis can be very useful in education planning. Application of the presented model can be used by policy makers at government agencies to check a particular educational policy of the institution. Having estimated the future minimum enrolment the school management will be able to adjust the policy when necessary [2].

References

[1] Adam, R. Y. (2015). An Application of Markov Modeling to the Student Flow in Higher Education in Sudan. International Journal of Science and Research (IJSR), 4(2): 49–54.

[2] Adeleke, R. A., Oguntuase, K. A and Ogunsakin, R. E. (2014). Application of Markov Chain to the Assessment of Students' Admission and Academic Performance in Ekiti State University. International Journal of Scientific & Technology Research, 3(7): 349–357.

[3] Beichelt, F. (2006). Stochastic Processes in Science, Engineering and Finance. Boca Raton: Chapman & Hall/CRC.

[4] Hudoklin-Božič, A. (2003). Stohastični procesi, 6th Edition, Kranj: Založba Moderna organizacija.

[5] Tijms, H. C. (2003). A First Course in Stochastic Models. Chichester: John Wiley & Sons.

APPROXIMATION ALGORITHMS FOR THE TEACHERS ASSIGNMENT PROBLEM

Katarína Cechlárová[1], Pavlos Eirinakis[2], Tamás Fleiner[3], Dimitrios Magos[4], Ioannis Mourtos[2] and Eva Oceľáková[1]

[1]Institute of Mathematics, Faculty of Science, P.J. Šafárik University, Košice, Slovakia
[2]Department of Management Science & Technology, Athens University of Economics and Business, 76 Patission Ave., 104 34 Athens, Greece
[3]Budapest University of Technology and Economics and MTA-ELTE Egerváry Research Group, Budapest, Hungary
[4]Department of Informatics, Technological Educational Institute of Athens, Ag. Spyridonos Str., 12210 Egaleo, Greece

Abstract: We study the assignment of trainee teachers to schools for a practical placement. The starting point is the situation characteristic for Slovak and Czech education system where each pre-service teacher specializes in two subjects. It is known that when each school has a certain upper limit for the number of assignees whose specialization involves a given subject, the problem of assigning the maximum number of trainee teachers is NP-hard. In this paper we propose several approximation algorithms for this problem.

Keywords: assignment problem; approximation.

1 INTRODUCTION

The traditional study of teachers-to-be in Slovakia and Czech republic involves specialization of each student in two subjects, e.g. Mathematics and Physics, Chemistry and Biology, Slovak language and English etc. Each curriculum of teachers' study contains also several practical placements in a real school. During such placements students teach pupils themselves, always under supervision of an experienced and qualified teacher approved by the university for taking this responsibility. Students might try to find suitable schools and supervising teachers by themselves, but to ensure the quality of such a placement, each faculty usually provides a list of such schools and teachers, and students are assigned to them by the faculty staff.

Finding an acceptable placement of students is not easy and it usually takes several days and many iterations. The aim is to find a place for each student. However, even if the number of approved supervising teachers is sufficient for the current number of students to be placed, this is not always possible, as the structure of available places might not be suitable, not all schools provide supervisors for all subjects, or they may not have enough classes to accept several students for a particular subject, or some students cannot be placed to some schools because of, say, time-consuming commuting.

The classical problems of combinatorial optimization like the maximum cardinality bipartite matching problem, assignment problem, or flow problem have been successfully applied to a range of variants of manpower allocation problems, see e.g. applications reviewed in [1], Chapter 12. Requirements brought about by concrete practical applications can also lead to some NP-complete cases, like the one described in [8]. A problem motivated by scheduling teaching practices was modelled as the 3-dimensional assignment problem in [6]. Since in our case, each future teacher is connected with 2 subjects, this setting also resembles the hospital/residents problem with couples [10]. Still, our problem is different as the structure of subjects is more specific.

In our previous publications [2, 3] and [4] we showed that under the requirement to teach both subjects at one school during the same placement, then the problem to find a matching for the maximum possible number of students is polynomial only if there are altogether only two specialization subjects, or there are three subjects but each school can accept at most one student for each subject (irrespectively of her other specialization); all other cases are NP-hard.

Although an integer linear program for the teachers assignment problem in [3] was successfully applied to real data, from the theoretical point of view, further research for NP-hard problems is necessary. In this paper we explore the possibilities of polynomial approximation algorithms in this context. In Section 4 we propose some other possible research directions.

2 Definitions and notation

An instance J of the Teachers Assignment problem, TAP for short, involves a set A of n applicants (students, trainee teachers), a set S of m schools and a set $P = \{1, 2, \ldots, k\}$ of k subjects, like Mathematics, Physics, Informatics or Biology etc.

Each applicant $a \in A$ is characterized by a pair of different subjects $\mathbf{p}(a) = \{p_1(a), p_2(a)\} \subseteq P$. The set of applicants whose specialization involves a subject $p \in P$ will be denoted by A_p and the set of those whose specialization is exactly $\{p, r\}$ by $A_{p,r}$. We also suppose that each applicant a provides a list $S(a)$ of *acceptable* schools, i.e. schools to which she (we shall refer to the applicants as females) is willing to go.

Each school $s \in S$ has a certain capacity for each subject, the vector of capacities of school s will be $\mathbf{c}(s) = (c_1(s), \ldots, c_{|P|}(s)) \in \mathbb{N}^{|P|}$. An entry of $\mathbf{c}(s)$ is a *partial capacity* of school s. Here, $c_p(s)$ is the maximum number of applicants from A_p that school s is able to accept.

An *assignment* $\mathcal{M}$ is a subset of $A \times S$ such that each applicant $a \in A$ is a member of at most one pair in $\mathcal{M}$. We shall write $\mathcal{M}(a) = s$ if $(a, s) \in \mathcal{M}$ and say that applicant a is *assigned* (to school s); if there is no such school, applicant a is *unassigned*. The set of applicants assigned to a school s will be denoted by $\mathcal{M}(s) = \{a \in A; (a, s) \in \mathcal{M}\}$. If $\mathcal{M}$ is any assignment, we denote by $\mathcal{M}_p$ and by$\mathcal{M}_{p,r}$ its restrictions to applicants from the sets A_p and $A_{p,r}$, respectively. More precisely,

$$\mathcal{M}_p(s) = \{a \in A; (a, s) \in \mathcal{M} \ \& \ p \in \mathbf{p}(a)\}$$

and

$$\mathcal{M}_{p,r}(s) = \{a \in A; (a, s) \in \mathcal{M} \ \& \ \{p, r\} = \mathbf{p}(a)\}.$$

An assignment $\mathcal{M}$ is a *matching* if $\mathcal{M}(a) \in S(a)$ for each $a \in A$ and $|\mathcal{M}_p(s)| \leq c_p(s)$ for each school s and each subject p.

A matching is *maximal* if no school can accept an unassigned applicant without violating at least one of its partial capacities. A maximal matching can be found quite easily, for example by the following procedure: take each applicant in turn and assign her to any acceptable school that has still enough capacity for both her subjects. A matching is *maximum* if it assigns the maximum number of applicants.

Example 1. Clearly, not each maximal matching is maximum. In the TAP instance with the set of subjects $P = \{1, 2, 3\}$, given in Figure 1, the maximum matching is $\mathcal{M}_1 = \{(a_1, s_1), (a_2, s_1), (a_3, s_2)\}$ of size 3. However, matching $\mathcal{M}_2 = \{(a_3, s_1)\}$ is maximal and its size is 1. This shows that the size of a maximal matching can be only one third of a maximum matching. Theorem 2.1 (proved below) says that it cannot be less.

school	capacities for			applicant	type	acceptable
	1	2	3			schools
s_1	2	1	1	a_1	$\{1, 2\}$	s_1
s_2	0	1	1	a_2	$\{1, 3\}$	s_1
				a_3	$\{2, 3\}$	s_1, s_2

Figure 1: Instance J for Example 1.

Theorem 2.1 *If $\mathcal{M}$ is any maximal matching and $\mathcal{M}^*$ is a maximum matching then $|\mathcal{M}| \geq |\mathcal{M}^*|/3$.*

Proof. Suppose that we have a maximum matching $\mathcal{M}^*$ in an arbitrary instance J of TAP and we want to insert into J the pairs matched by a maximal matching $\mathcal{M}$ one by one. If $(a, s) \in \mathcal{M}$ belongs to $\mathcal{M}^*$ too, we do nothing. But, if $(a, s) \notin \mathcal{M}^*$, then the pairs matched by $\mathcal{M}^*$ may prevent (a, s) from being inserted. To create enough place for this pair, we might need to drop from $\mathcal{M}^*$ at most one pair for each subject of a's specialization, plus the pair (a, s') if $(a, s') \in \mathcal{M}^*$ and $s \neq s'$. Hence, $|\mathcal{M}| \geq \frac{1}{3}|\mathcal{M}^*|$.

3 Approximation algorithms

MAX-TAP denotes the problem to find a matching with maximum cardinality for a given instance J of TAP. The special case of MAX-TAP where the number of subjects is fixed to k will be denoted by MAX-k-TAP.

Trivially, MAX-2-TAP and MAX-3-TAP with all partial capacities bounded by 1 are polynomially solvable. In the former case all the applicants are equivalent and each school s can accept at most $\min\{c_1(s), c_2(s)\}$ applicants. In the latter case, each school can accept at most one applicant. So both problems reduce to the simple bipartite maximum cardinality (b-)matching problem (for polynomial algorithms see e.g. [11]). By contrast, MAX-3-TAP is NP-complete even if all partial capacities are bounded by 2 and MAX-4-TAP is NP-complete even when all partial capacities are at most 1 (see [2, 3]) .

In this section we study approximation algorithms for MAX-TAP. Theorem 2.1 gives a trivial 3-approximation algorithm and Example 1 shows that this approximation bound is tight. Below we propose two approximation algorithms with better approximation guarantees. Both are based on finding a maximum cardinality matching $\mathcal{M}^p$ in an instance of MAX-TAP for a subset of applicants A_p where subject p is fixed.

First we describe how to find matching $\mathcal{M}^p$ by employing network flow methods. For an instance J of TAP and a fixed subject $p \in P$ we create a network $N(J_p)$. Its vertices are: a vertex v_a for each applicant $a \in A_p$, a vertex $u_{s,r}$ for each school and for each subject r such that $c_r(s) > 0$, plus two vertices σ and τ (source and sink). The arcs of $N(J_p)$ and their capacities are given in Figure 2.

(σ, v_a)	capacity 1	for all $a \in A_p$
$(v_a, u_{s,r})$	capacity 1	for all subjects $r \neq p$, for all $a \in A_{p,r}$ and $s \in S(a)$
$(u_{s,r}, u_{s,p})$	capacity $c_r(s)$	for all $s \in S$ and all subjects $r \neq p$
$(u_{s,p}, \tau)$	capacity $c_p(s)$	for all $s \in S$

Figure 2: Arcs and their capacities for the flow network $N(J_p)$.

As all capacities are integral, by [5] there exists an integer maximum flow f that can be used to define a feasible matching as follows: for an applicant $a \in A_{p,r}$ we set $\mathcal{M}(a) = s$ if $f(v_a, u_{s,r}) = 1$. It is easy to see that $\mathcal{M}$ is indeed a matching: each applicant is assigned to an acceptable school and no partial capacity is exceeded.

The number of vertices in $N(J_p)$ is bounded by $n + mp + 2$, the upper bound for the number of its arcs is $n + nm + 2mp + 2$. The push-relabel algorithm with dynamic trees uses $O(NM\ log(N^2/M))$ operations for a network with N vertices and M arcs [7]. This means that a maximum cardinality matching $\mathcal{M}^p$ can also be found in time that is polynomial in the number of applicants and schools.

3.1 Algorithm Greedy1

Let us consider now the algorithm depicted in Figure 3.

---

begin fix the order of subjects $1, 2, \ldots, k$;
for $p := 1$ to k **do**
  begin find a maximum cardinality matching $\mathcal{M}^p$ for A_p;
    reduce the set of applicants and partial capacities of schools accordingly
  end
end

Figure 3: Algorithm Greedy1

Theorem 3.1 *The approximation guarantee of algorithm* Greedy1 *is* 2.

Proof. We prove by induction on the number k of subjects. For $k = 2$, Greedy1 clearly finds a maximum size matching, so the theorem holds. Assume now that the assertion holds for at most $k-1$ subjects and consider an instance J with k subjects. Let $\mathcal{M}^*$ be any maximum matching of J and matching $\mathcal{M}^G = \cup_{p=1}^{k} \mathcal{M}^p$ be the output of some realization of Greedy1 in subject order $1, 2, \ldots, k$. Observe that

$$|\mathcal{M}_1^*| \leq |\mathcal{M}_1^G|, \tag{1}$$

where MM_1^* and $\mathcal{M}_1^G$ are the restrictions of matchings $\mathcal{M}^*$ and $\mathcal{M}^G$ to A_1.

Take the assignment $\mathcal{M}^* \setminus \mathcal{M}_1^*$ and add individual matched pairs of $\mathcal{M}_1^G$ one by one in an arbitrary order. To be able to add the next pair of $\mathcal{M}_1^G$, say (a, s) for some $a \in A_{1,r}$, then we might need to displace at most one pair of $\mathcal{M}_{r,p}^*$ for some $r \neq 1$ to create free places (there is clearly no obstruction in subject 1, thanks to inequality (1)). So if $\mathcal{M}'$ is the set of displaced matched pairs outside $\mathcal{M}_1^*$ then $|\mathcal{M}'| \leq |\mathcal{M}_1^G|$. After the process we get a matching $\mathcal{M} = \mathcal{M}^* \cup \mathcal{M}_1^G \setminus (\mathcal{M}_1^* \cup \mathcal{M}')$ such that

$$|\mathcal{M}_1^G| = |\mathcal{M}_1| \geq \frac{1}{2}(|\mathcal{M}_1^*| + |\mathcal{M}'|) \ . \tag{2}$$

As $\mathcal{M}_1 = \mathcal{M}_1^G$, the matching $\mathcal{M} \setminus \mathcal{M}_1$ is a matching of the instance that has one less subject and that we got after the first phase of Greedy. Hence by the induction hypothesis we have that

$$|\mathcal{M}^G| - |\mathcal{M}_1^G| = |\mathcal{M}^G \setminus \mathcal{M}_1^G| \geq \frac{1}{2}|(\mathcal{M} \setminus \mathcal{M}_1)| = \frac{1}{2}|(\mathcal{M}^* \setminus (\mathcal{M}_1^* \cup \mathcal{M}'))| = \frac{1}{2}(|\mathcal{M}^*| - (|\mathcal{M}_1^*| + |\mathcal{M}'|)) \ . \tag{3}$$

Combination of inequalities (2) and (3) implies

$$|\mathcal{M}^G| = |\mathcal{M}_1^G| + |\mathcal{M}^G \setminus \mathcal{M}_1^G| \geq \frac{1}{2}(|\mathcal{M}_1^*| + |\mathcal{M}'|) + \frac{1}{2}(|\mathcal{M}^*| - (|\mathcal{M}_1^*)| + |\mathcal{M}\ '|)) = \frac{1}{2}|\mathcal{M}^*|,$$

hence the assertion follows.

Example 2. Take the TAP instance with four subjects, six applicants $a_1, a_2, a_3, b_1, b_2, b_3$, three schools s_1, s_2, s_3 with all partial capacities equal 1 and the characteristics of applicants given in Figure 4. Here, the greatest size of a matching, namely 6, is achieved by assigning the pair of applicants a_i, b_i to school s_i for $i = 1, 2, 3$. Suppose that Greedy1 works with the order of subjects $1, 2, 3, 4$. There are 6 different maximum matchings for A_1. However, if the first phase of Greedy1 produces e.g. $\mathcal{M}_1^G = \{(a_1, s_2), (a_2, s_3), (a_3, s_1)\}$ then no further applicant can be matched and Greedy1 outputs a matching whose size is half of the maximum. So the bound of Theorem 3.1 is also tight, already for four subjects.

applicant	type	acceptable schools	applicant	type	acceptable schools
b_1	$\{3,4\}$	s_1	a_1	$\{1,2\}$	s_1, s_2, s_3
b_2	$\{2,4\}$	s_2	a_2	$\{1,3\}$	s_1, s_2, s_3
b_3	$\{2,3\}$	s_3	a_3	$\{1,4\}$	s_1, s_2, s_3

Figure 4: Instance J for Example 2.

3.2 Algorithm Greedy2

Our second algorithm, called Greedy2, is given in Figure 5.

---

begin for $p := 1$ to k find a maximum cardinality matching $\mathcal{M}^p$ of applicants A_p;
keep $\mathcal{M}^j$ whose size is maximum;
add applicants from $A \setminus A_j$ arbitrarily to get a maximal matching
end

Figure 5: Algorithm Greedy2

Theorem 3.2 Greedy2 *is a* $\frac{k}{2}$*-approximation algorithm.*

Proof. Let $\mathcal{M}^*$ be a maximum matching in an instance of MAX-TAP. Then

$$|\mathcal{M}_1^*| + |\mathcal{M}_2^*| + \cdots + |\mathcal{M}_k^*| = 2|\mathcal{M}^*|$$

as each matched pair (a, s) is counted twice in the left-hand side, namely in $|\mathcal{M}_p^*|$ and $|\mathcal{M}_r^*|$ if $a \in A_{p,r}$. By Pigeonhole principle, at least one term on the left, say $|\mathcal{M}_1^*|$, has the size at least $|\mathcal{M}^*|/k$. As $|\mathcal{M}_1^*|$ is not greater than $\max\{|\mathcal{M}^p|, p = 1, 2, \ldots, k\}$, Algorithm Greedy2 outputs a matching of size at least $|\mathcal{M}^*|/k$.

Example 3. In the TAP instance of Figure 6, the maximum matching $\mathcal{M} = \{(a_1, s_2), (a_2, s_2), (a_3, s_1)\}$ is of size 3. For each subject p, the cardinality of maximum matching $\mathcal{M}^p$ is 2, but if Greedy2 chooses $p = 1$ and matching $\mathcal{M}_1 = \{(a_1, s_1), (a_2, s_1)\}$ then the matching output by the algorithm will be of size 2. This is finally a tight example for Greedy2.

school	capacities for			applicant	type	acceptable schools
	1	2	3			
s_1	2	1	1	a_1	$\{1,2\}$	s_1, s_2
s_2	2	1	1	a_2	$\{1,3\}$	s_1, s_2
				a_3	$\{2,3\}$	s_1

Figure 6: Instance J for Example 3.

Notice that Greedy2 outperforms Greedy1 only for three subjects, giving the guarantee $\frac{2}{3}$.

4 Conclusion and open questions

We showed that the MAX-TAP problem, in spite of being intractable, allows relatively easy approximation algorithms. Notice, however, that the tight examples for all the proposed algorithms are such that the lower bound is achieved if in the first step an "incorrect" matching

is chosen. Since there is always a choice leading to an optimal solution, a question is whether these algorithms could be refined to obtain a better approximation bound. In general, the obtained bounds leave a lot of space for improvement, even more so, as so far no lower bound of approximation has been obtained. Or, is it possible that MAX-TAP could be APX-complete? We are inclined to believe in this possibility because the NP-completeness provided in [2, 3] uses as the starting point 3-DIMENSIONAL MATCHING, that is APX-complete thanks to the result of Kann [9]. On the other hand, a result in [4] implies that the problem of *minimizing the number of unassigned students* does not admit a polynomial approximation algorithm with guarantee $n^{1-\varepsilon}$ for any $\varepsilon > 0$, if P$\neq$ NP.

Finally, it could be worthwhile to study the MAX-TAP problem from the parameterized complexity angle. One possible parameter is the maximum partial capacity of schools.

5 Acknowledgements

This work was supported by VEGA grants 1/0344/14 and 1/0142/15, University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF, (Cechlárová, Oceľáková), by OTKA K108383 and the ELTE-MTA Egerváry Research Group (Fleiner). The project is implemented under the Act "Bilateral S&T cooperation Greece - Slovak Republic 2011 - 2012". It is also co-financed by the European Union (European Regional Development Fund -ERDF) and the Greek Ministry of Education (GSRT) and is a part of the Operational Programme "Competitiveness & Entrepreneurship" (OPCE I). (Eirinakis, Magos, Mourtos). The authors also thank for the support of the COST Action IC1205.

References

[1] R. K. Ahuja, T. L. Magnanti, J. B. Orlin, *Network flows: theory, algorithms, and applications*, Prentice Hall, 1993.

[2] K. Cechlárová, T. Fleiner and E. Potpinková, *Practical placement of trainee teachers to schools*, in Proc. SOR 13, eds. L. Zadnik Stirn, J. Žerovnik, J. Povh, S. Drobne, A. Lisec (2013), 129-134.

[3] K. Cechlárová, T. Fleiner, D. Manlove, I. Mc Bride and E. Potpinková, *Modelling practical placement of trainee teachers to schools*, CEJOR, DOI 10.1007/s10100-014-0356-5, 2014.

[4] K. Cechlárová, T. Fleiner, D. Manlove and I. Mc Bride, *Stable matchings of teachers to schools*, http://arxiv.org/ abs/1501.05547.

[5] G. B. Danzig, D. R. Fulkerson, *On the max-flow min-cut theorem of networks*, in Linear Inequalities and Related Systems, Eds. H. W. Kuhn, A. W. Tucker, Princeton University Press (1956), 215-221.

[6] A. M. Frieze, J. Yadegar, *An algorithm for solving 3-dimensional Assignment problem with application to scheduling a teaching practice*, J. of OR society 32(11), 989-995, 1981

[7] A.V. Goldberg, R.R. Tarjan, *A new approach to the maximum flow problem*, Journal of the ACM 35 (1988), 921-940.

[8] A. Hefner, P. Kleinschmidt, *A constrained matching problem*, Annals of OR 57: 135–145 (1995).

[9] V. Kann, *Maximum bounded 3-dimensional matching is MAX SNP-complete*, Inf. Process. Lett. 37 (1991) 2735.

[10] E. Ronn, *NP-complete stable matching problems*, Journal of Algorithms 11(2), 285-304, 1990.

[11] A. Schrijver: *Combinatorial Optimization: Polyhedra and Efficiency*, Springer, Algorithms and Combinatorics 24 (2003).

REGIONAL UNEMPLOYMENT IN CROATIA: EVIDENCE FROM DYNAMIC PANEL MODEL

Nataša Erjavec
Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
nerjavec@efzg.hr

Saša Jakšić
Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
sjaksic@efzg.hr

Abstract: The paper analyses the impact of regional economic activity, regional wages and education on regional unemployment rate in Croatian counties for the period from 2004 to 2012. The results of the empirical analysis point to the conclusion that the regional labor market adjustment mechanisms in Croatia are not functioning. It is especially important to note the lack of wage flexibility that makes it difficult for the labor market to adapt to changes in the macroeconomic surrounding.

Keywords: Croatia, regional unemployment, wages, dynamic panel model.

1 INTRODUCTION

More than 20 years since the beginning of the transition process from centrally planned on the way to market oriented economies, the issues of high unemployment rates on the country level and differences among regional level unemployment rates are still present in most of the Central and Southeast European (CSEE) countries. Croatia is not an exception, its unemployment rate being 17 percent and regional unemployment rates ranging from 37 to 10.8 percent. Apart from the size of the figures, of prominent importance is the persistence of the regional unemployment rate disparity. One of the key factors of its persistence is the lack or inexistence of adjustment mechanisms.

Benchmark model for studying regional unemployment dynamics was established in [2]. However, as the model presented in [2] is based on US economy and its regional labor markets adjustment mechanisms, other authors ([5], [6] and [9]) investigated regional labor market dynamics in Europe. Their results indicate different adjustment mechanisms in Europe and US regional labor markets: while migrations are the key adjustment mechanism in US, adjustments in labor force participation are the main channel of adjustment in Europe [5]. To get a better insight into potential adjustment mechanisms, a detailed overview on empirical research and variables employed in regional unemployment studies is provided in [6] and [9].

Literature on Croatian regional unemployment is scarce. Among previous research, main contributions were made by [3], [4] and [8]. [3] found evidence of diverging regional unemployment dynamics which have eventually led to increasing persistence among Croatian regions. Using Labor Force Survey (LFS) data, [4] found that both individual and regional characteristics played an important role in determination of employment and earnings. Furthermore, a large part of the differences in regional labour market performance is attributed to the differences in human capital endowment. [8] supplements these findings by analysing the influences of regional characteristics on unemployment rates and finds that expectations of unemployed have a pronounced influence on regional unemployment rates. The author also addressed the issue of geographical immobility and the failure of migration as an adjustment mechanism on the regional labor market in Croatia. Therefore, previous research of Croatian regional unemployment indicates the importance of education in

explaining differences in regional unemployment and inefficiency of adjustment mechanisms that could reduce large differences in the regional unemployment rates.

This paper analyses the influence of economic activity, regional wages and education on regional unemployment rate in Croatian counties for the period from 2004 to 2012. Apart from the variables used in the final model specification, there are other variables that were obvious candidates as a proxy for labor market adjustment mechanisms: vacancies, migration, population size and employment. However, all these variables turned out to be statistically insignificant. In the empirical analysis few alternative model specifications were applied. In order to capture unobserved heterogeneity across Croatian counties, static panel data models (fixed and random effects models) were applied. Additionally, the dynamic panel model was estimated as a robustness check.

The obtained results and the fact that most of the employed variables are not statistically significant indicate that the adjustment mechanisms prevailing in the established market economies are still not functioning in Croatia. Among the adjustment mechanisms it is important to stress out the limited wage flexibility which is the main contributor to the low flexibility of the labor market to occurrence of the macroeconomic shocks.

The remainder of the paper is organised as follows: the second section gives details on the data set employed in this paper. Third section describes the applied methodology and reports the main results of the empirical analysis and the forth section concludes and gives some suggestions for further research.

2 DATA

The empirical analysis is performed for a cross-section of Croatian counties ($\boldsymbol{N = 21}$) for the period from 2004 to 2012 ($\boldsymbol{T = 9}$). In accordance with theoretical and empirical literature on regional unemployment ([6] and [9]), individual and regional characteristics in unemployment (unemployment rate for Croatian counties: variable *unr*) were explained by following commonly used explanatory variables: regional real GDP per capita (deflated by harmonised index of consumer prices (HICP), 2005=100, variable *gdppc*) in thousand HRK represents the degree of the economic activity in the County; data on regional wages (average monthly net earnings deflated by HIPC, 2005=100) (variable *wages*) in HRK; counties'population size (variable *pop*); vacancies registered by counties (variable *vac*); counties' net migration (variable *mig*); employment (variable empl); and data on share of counties' employed population with university degree (*degree*) as a proxy for education. All data employed in the paper were obtained from the Croatian Bureau of Statistics, except of vacancies which were obtained from Croatian Employment Service.

Prior to the empirical analysis, all variables were pretested for stationarity. The battery of panel unit root tests was performed. Namely, Levin, Lin and Chu test; Im, Pesaran and Shin test; ADF-Fisher test and Pesaran's CADF test. Tests rejected the null of a unit root process for all variables. Thus, the analysis was performed by treating all variables as being stationary.

3 METHODOLOGY AND EMPIRICAL RESULTS

Several alternative models were estimated. Using a pooled ordinary least squares (OLS) estimator resulted in heteroskedasticity and serial correlation in the error term. The null hypotheses of constant variance and no autocorrelation were strongly rejected. Thereafter, the static version of the panel model using fixed (FE) and random effects (RE) estimators was estimated. The proposed static panel model can be specified as follows:

$$y_{i,t} = \alpha + \beta X_{i,t} + \nu_i + \varepsilon_{i,t}, i = 1,2,\dots,N, t = 1,2\dots,T. \quad (1)$$

i denotes individual county and *t* denotes time. Dependent variable $y_{i,t}$ is defined as average county unemployment rate, and *X* is a set of regressors. As variables population (*pop*), vacancies (*vac*), migration (*mig*) and employment (*empl*) were not statistically significant in both models (fixed and random effects) they were excluded as regressors. Therefore, the models were estimated using variables overall income per capita (*gdppc*), earnings (*wages*) and share of population with university degree (*degree*) as explanatory variables for the regional unemployment rate. Panel model allows controlling for other (observed and unobserved) variables that differ from county to county (such as share of prevailing activities and involvement in the unofficial sector) but do not change over time. Furthermore, this also enables controlling for variables that vary through time, but not across states.

When tested for the significance of individual effects, both *F*-test for fixed effects and Breusch-Pagan Lagrangian multiplier test for random effects indicated the existence of significant differences across counties. Within group fixed effect estimates and generalised least squares (GLS) random effects estimates with robust standard errors are presented in Table 1.

Table 1: Estimation results (dependent variable: regional unemployment rate).

Variable	***OLS***	***FE***	***RE***	***Dynamic panel***
Lagged dependent variable				0.4651
				(0.0505)**
gdppc	-0.4000	-0.1597	-0.1925	-0.2988
	(0.0285)**	(0.0577)*	(0.0461)*	(0.0471)*
wages	0.0003	-0.0101	-0.0088	-0.0017
	(0.0022)	(0.0020)**	(0.0019)**	(0.0015)
degree	0.9294	0.8013	0.7669	0.4975
	(0.1699)**	(0.1559)**	(0.1445)**	(0.1294)**
Model diagnostics				
N	189	189	189	147
F or Wald test	166.82**	139.29**	94.67**	258.52**
R^2	0.6391	0.9486		
SSE or $\hat{\sigma}_e$	4.3394	1.6623	1.6623	
$\hat{\sigma}_v$		4.9465	4.1081	

Note: A constant is also included in the model specification but is not reported. Robust standard errors are in parenthesis;
** denotes statistical significance at 5%,*
*** denotes statistical significance at 1%.*

Even though both fixed and random effects proved to be significant, and both model specifications give similar estimates, formal test was also performed in order to decide which model is appropriate.

Analysed data failed to meet the asymptotic assumptions of the Hausman test. Hence a robust version of Hausman's specification test was applied with the null hypothesis that a difference in estimates is not systematic. As the obtained chi-squared test statistic of 1.83 (*p*-value=0.6089) was small enough not to reject the null hypothesis the conclusion was that a random effect model is a preferred model specification.

Additionally, to account for possible dynamics, lagged dependent variable was added in the model as an additional variable in the dynamic panel model. The defined dynamic panel model was also a robustness check for the results obtained by static panel models (fixed and random effects models). The analysed dynamic model specification can be written as follows:

$$y_{i,t} = \alpha + \gamma y_{i,t-1} + \beta X_{i,t} + \nu_i + \varepsilon_{i,t}, \; \boldsymbol{i = 1, 2, \ldots, N}, \; \boldsymbol{t = 1, 2 \ldots, T}. \quad (2)$$

Counties are denoted by subscript *i*, while *t* stands for years. Model (2) is estimated using Arellano–Bond one-step estimator with robust standard errors. [1] derived a consistent generalized method of moments (GMM) estimator in order to solve possible endogeneity problems arising from the correlation between regressors and the error term making standard estimators inconsistent.

As the robust standard errors were employed, Sargan test cannot be performed. It is well known that the asymptotic distribution for Sargan test statistic is not known. Sargan test has an asymptotic chi-squared distribution only for a homoskedastic error term and overrejects the null hypothesis in the presence of heteroskedasticity [1]. The Arellano–Bond tests for first- and second-order autocorrelation in the first-differenced errors reject the null in both cases. The corresponding *p*-value for testing the first-order autocorrelation was 0.0431 and for the second-order autocorrelation *p*- value was 0.0009.

Considering the parameters of the estimated models, the fact that in the initial models most of the variables (population, vacancies, migration and employment) were not statistically significant is very indicative on the lack of the adjustment mechanisms in Croatian regional labor markets. Therefore, mechanisms prevailing in established market economies such as migration do not act as a factor that would reduce the differences in regional unemployment rates. The results are in line with findings of [4] from the research performed almost ten years ago which suggests high persistence of regional unemployment disparities in Croatia.

Statistically significant variables have the expected sign, i.e. higher economic activity in the county leads to a decrease in unemployment. A positive sign for the education variable implies that higher the level of education of existing employees, it is harder for unemployed to enter the workforce. This also coincides with the variable vacancies being statistically insignificant, which could imply low level of job creation and that qualification and skills of unemployment do not meet the employer's needs.

Turning to the estimated parameters of the dynamic model, it is important to note that the lagged dependent variable is statistically significant which suggests that the specified dynamic panel model is an appropriate model for describing regional unemployment dynamics. The main difference between the dynamic panel model and static panel models is that the wages are not statistically significant in the dynamic model specification. Although, it could be argued that the obtained result is specification sensitive, actually all models yield similar result as the impact of wages on unemployment rate in static panel models was statistically significant but the intensity of the impact is almost negligible.

Obtained results and the fact that wages are not statistically significant, are in line with European Commission's country specific recommendations published in May 2015. In its guidelines the European Commission highlights limited wage adjustment and generally low flexibility of the system which makes it difficult to adapt to changes in the macroeconomic surrounding.

Possible factors that could be behind the failure of the adjustment mechanisms are numerous. For instance, [4] indicates that the majority of the Croatian housing stock is owner occupied (which makes it relatively costly to move) and that migrants tend to seek a job within the immediate commuting area. All these arguments could present obstacles for the malfunctioning of the migrations as an adjustment channel. The share of employees in the public sector, numerous bodies of local government units and publicly owned companies, collective agreements, negotiating power of syndicates could be the main culprits for the lack of wage flexibility.

There are many other probable factors behind the failure of the adjustment mechanisms and the investigation of the importance of each individual factor is a large potential for further research.

4 CONCLUSION

The process of transition from centrally planned to market oriented economies is very difficult and along the way countries are faced with many challenges. Among the challenges, regional unemployment is one of the most persistent issues to be dealt with. The findings of this paper indicate that the regional unemployment issue is a very serious issue in Croatia and the absence of adjustment mechanisms prevailing in established market economies are the main obstacle to reducing the regional unemployment disparities.

One of the main shortcomings of this paper (and all papers referred in the text) is its reliance on yearly data for the description of regional unemployment dynamics. In that manner, some features of the regional labor markets that could potentially be important for the description of regional unemployment dynamics cannot be modelled. For instance, seasonal characteristics of certain occupations prevailing in some of the Croatian regions could be an important factor for the description of the dynamics. Taking into consideration higher frequency data (monthly, quarterly) could yield different results and hence lead to different conclusions on relevance of adjustment mechanisms.

References

[1] Arellano, M., and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies, Vol 58: 277–297.

[2] Blanchard, O.J. and Katz, L.F. (1992). Regional Evolutions. Brookings Papers on Economic Activity, Vol 1: 1–61.

[3] Botrić, V. (2004). Regional Aspects of Unemployment in Croatia. Croatian Economic Survey, Vol 6(1): 77-95.

[4] Botrić, V. (2007). Regional Labor Market Differences in Croatia: Evidence from the 2000-2005 Labor Force Survey Data. Privredna kretanja i ekonomska politika, Vol 17(113): 26-51.

[5] Decressin J. and Fatás A. (1995). Regional labor market dynamics in Europe. European Economic Review, Vol 39: 1627–1655.

[6] Elhorst, J.P. (2003). The Mystery of Regional Unemployment Differentials: Theoretical and Empirical Explanations. Journal of Economic Surveys, Vol 17(5): 709-748.

[7] European Commission (2015). Country-specific Recommendations. http://ec.europa.eu/europe2020/pdf/csr2015/csr2015_croatia_en.pdf [Accessed 9/6/2015].

[8] Luo, X. (2007). Regional Disparities In Labor Market Performance In Croatia: The Role Of Individual And Regional Structural Characteristics. The World Bank Policy Research. Working Paper Series No. 4148. Washington, DC: The World Bank.

[9] Zeilstra, A.S., and Elhorst, J.P. (2014). Integrated Analysis of Regional and National Unemployment Differentials in the European Union. Regional Studies, Vol 48(10): 1739-1755.

EMERGENCY SYSTEM DESIGN WITH TEMPORARILY FAILING CENTERS

Jaroslav Janáček
University of Žilina, Faculty of Management and Informatics,
Univerzitná 1, 010 26 Žilina, Slovak Republic,
jaroslav.janacek@fri.uniza.sk

Marek Kvet
University of Žilina, Faculty of Security Engineering
1. mája 32, 010 26 Žilina, Slovak Republic,
marek.kvet@fbi.uniza.sk

Abstract: This paper deals with the emergency service system design with temporarily failing centers. We assume that the users may get the service from the second or third nearest located center, if the nearest one is occupied. Attempts at solving large instances of the problem described by a location-allocation model often fail due to lack of memory or computational time. We suggest an approximate approach based on a radial formulation for both min-max and min-sum optimization. We present a two-phase approach to emergency service system design with temporarily failing centers, where the first phase is based on so-called exposing constraints to minimize the disutility of the worst situated users and then the total discomfort of an average user is minimized.

Keywords: emergency service system, failing centers, exposing constraints, radial formulation

1 INTRODUCTION

The design of emergency service system [3], [8] includes determination of center locations, from which the associated service is distributed to all users of the system. The objective in the standard formulation of the design problem is to minimize some sort of disutility as the social costs, which are proportional to the distances between served objects and the nearest service centers. In addition, another simplification often used by service system designers constitutes in the assumption that a user is serviced only from the nearest located service center. This simplification can be used, if the structure of a public administration system is designed, but the assumption does not hold, when an emergency service system is designed due to random occurrence of the demand for service and limited capacity of the service centers. This paper is focused on such methods of the emergency service system design, where the generalized disutility is considered instead of simple distance. It follows the idea of random occurrence of the demand on service and limited capacity of the service centers in real emergency rescue systems [9]. At the time of the current demand for service, the nearest service center may be occupied by some other user, for whom this service center is also the nearest one. When such situation occurs, the current demand is usually served from the second nearest center or from the third nearest center, if the second one is also occupied. Thus we assume that the service is generally provided from more located service centers and the individual contributions from relevant centers may be weighted by reduction coefficients depending on a center order. A location-allocation model of this problem with generalized disutility was formulated in [11]. The previously performed research proved that the location-allocation formulation is considerably outperformed by the radial formulation, when a commercial IP-solver is used to design the optimal service center deployment subject to simple disutility is considered. The radial formulation was successfully extended for the above-mentioned generalized disutility in min-sum location problems [5], [9]. In the previous research, the average user´s disutility was minimized. This min-sum approach may cause that the disutility perceived by the worst situated user is inadmissibly high. This unfairness may

be mitigated by the initial phase of the min-max process [10]. In this paper, we focus on two phase emergency system design method, where the generalized disutility is considered. The both phases are based on the radial formulation of the problems, which can considerably accelerate the associated solving process [1], [2], [5], [9] and outperform the location-allocation formulations. The first phase of the suggested method searches for a design of min-max optimal public service system. The result is processed by the second phase, where the average user´s disutility is minimized subject to the condition, that the minimal disutility of the worst situated user cannot be worsened. The remainder of the paper is organized as follows. The min-sum radial formulation of an emergency system design subject to generalized disutility is concisely described in Section 2. Section 3 is devoted to the min-max iterative approach applied on the radial formulation extended to the generalized disutility. The composed algorithm of the semi-fair emergency system design is suggested in Section 4 and associated numerical experiments are performed in Section 5. The results and findings are summarized in Section 6.

2 RADIAL FORMULATION OF MIN-SUM PROBLEM WITH GENERALIZED DISUTILITY

The necessity of solving large instances of the *p*-median problem has originally led to the radial formulation [1], [2], [4] and [6], which enables effective emergency system design for the case of simple disutility proportional to the distance from a user to the nearest service center. This approach was generalized [5], [9] to be applicable on the cases, when the user's disutility is influenced by distances from the user to *r* nearest located centers. To formulate the public service system design problem with the optimal generalized disutility, we use denotation of the set of users' locations by symbol *J* as above and a set of possible service center locations by symbol *I*. At most *p* locations from *I* must be chosen so that the sum of users' disutilities is minimal. In this paper, the generalized disutility for any user is modeled by a sum of weighted disutility contributions from the *r* nearest centers. The weight coefficients q_k for $k = 1 \ldots r$ are positive real values, which meet the inequalities $q_1 \geq q_2 \geq \ldots \geq q_r$. The *k*-th weight can be proportional to the probability of the case, that the *k*-1 nearest located centers are occupied and the *k*-th nearest center is available [11].

The network distance of a possible location *i* from user location *j* is denoted as d_{ij}. The decisions which determine the designed system can be modeled by location variables y_i for $i \in I$. The variable $y_i \in \{0,1\}$ models the decision on service center location at the place *i*. The variable takes the value of 1 if a facility is located at *i* and it takes the value of 0 otherwise.

The approximate radial formulation starts from partitioning of the range $[d^0, d^m]$ of all possible distance values $d^0 < d^1 < \ldots < d^m$ from a user to a possible center location into $v+1$ zones. The zones are separated by dividing points $D_1, D_2 \ldots D_v$ chosen from the sequence $d^0 < d^1 < \ldots < d^m$, where $0 = d^0 = D_0 < D_1$ and $D_v < D_{v+1} = d^m$. The zone *s* corresponds with the interval $(D_s, D_{s+1}]$. The length of the *s-th* interval is denoted by e_s for $s = 0 \ldots v$. To describe the problem, auxiliary zero-one variables x^k_{js} for $s = 0 \ldots v$ and $k = 1 \ldots r$ are introduced. The variable x^k_{js} takes the value of 1 if the *k*-th smallest distance from the user at $j \in J$ to the located center is greater than D_s and it takes the value of 0 otherwise. Then the expression $e_0 x^k_{j0} + e_1 x^k_{j1} + e_2 x^k_{j2} + e_3 x^k_{j3} + \ldots + e_v x^k_{jv}$ constitutes an upper approximation of the *k*-th smallest disutility contribution d^k_{j*} for the user at *j*. If the disutility d^k_{j*} belongs to the interval $(D_s, D_{s+1}]$, then the value of D_{s+1} is the upper estimation of d^k_{j*} with a maximal possible deviation e_s. To complete the associated radial model, we introduce a zero-one constant a_{ij}^s for each triple *i, j, s* where $i \in I$, $j \in J$ and $s = 0 \ldots v$. The constant a_{ij}^s is equal to 1 if and only if the distance d_{ij} of a user at the location *j* from the possible center location *i* is less than or equal to D_s, otherwise a_{ij}^s is equal to 0. Then the model can be formulated as follows:

$$Minimize \sum_{j\in J} b_j \sum_{s=0}^{v} e_s \sum_{k=1}^{r} x_{js}^k \quad (1)$$

$$Subject\ to \quad \sum_{k=1}^{r} x_{js}^k + \sum_{i\in I} a_{ij}^s y_i \geq r \quad for\ j \in J,\ s = 0..v \quad (2)$$

$$\sum_{i\in I} y_i \leq p \quad (3)$$

$$x_{js}^k \in \{0, 1\} \quad for\ j \in J,\ s = 0,..v,\ k = 1..r \quad (4)$$

$$y_i \in \{0,1\} \quad for\ i \in I \quad (5)$$

The constraints (2) ensure that the sum of variables x^k_{js} over k expresses the number of the service centers in the radius D_s from the user location j, which remains to the number r. The constraint (3) puts a limit p on the number of located facilities. Even though the constraints do not ensure the above declared meaning of the individual variables x^k_{js}, the objective function (1) gives the upper bound of the sum of disutility values.

3 APPROXIMATE RADIAL APPROACH TO MIN-MAX PROBLEM

Under the above-mentioned preconditions, we can describe the min-max optimal emergency service system design problem adding one nonnegative auxiliary variable h to model an upper bound of the disutility perceived by the worst situated user.

$$Minimize\ h \quad (6)$$

$$Subject\ to \quad \sum_{k=1}^{r} q_k \sum_{s=0}^{m-1} e_s x_{jsk} \leq h \ \ for\ j \in J \quad (7)$$

and (2) - (5)

The zero-one coefficients a_{ij}^s are defined here for $i \in I$, $j \in J$ and $s \in [0 \ldots m\text{-}1]$. The coefficients are derived from the disutility contribution values, which range only over non-negative integers of all possible disutility values $d^0 < d^1 < \ldots < d^m$.

The link-up constraints (7) ensure that each perceived disutility is less than or equal to the upper bound h. As the solvability of the min-max problem formulated above is questionable considering computational time, we focus on the approximate approach based on bisection method applied on so-called exposing structures introduced in [7].

The triple $[u, S, G]$ is denoted complete exposing structure, if its components satisfy the following rules. The first component u is a positive integer less than or equal to r. The second component S is an u-tuple $[S(1), \ldots, S(u)]$ of nonnegative increasing integers where $0 \leq S(1) < S(2) < \ldots < S(u) \leq m$. The third component is u-tuple $[G(1), \ldots, G(u)]$ of positive increasing integers where $1 \leq G(1) < G(2) < \ldots < G(u) = r$. Using the above introduced location variables $y_i \in \{0,1\}$ for $i \in I$, the following set of constraints can be formulated for the exposing structure $[u, S, G]$.

$$\sum_{i\in I} a_{ij}^{S(w)} y_i \geq G(w) \quad for\ j \in J,\ w \in [1..u] \quad (8)$$

If a feasible solution $\boldsymbol{y}$ of the constraints (3), (5), (8) structure exists for a complete $[u, S, G]$, then each user location j must lie at least in the radius $d^{S(1)}$ from $G(1)$ located service centers and in the radius $d^{S(2)}$ from $G(2)$-$G(1)$ additional service centers and so on to the radius $d^{S(u)}$ from the $G(u)$-$G(u\text{ -}1)$ service centers. It means that the worst situated user perceives the generalized disutility less than or equal to the value of (9).

$$H_{[u,S,G]} = d^{S(1)} \sum_{k=1}^{G(1)} q_k + \sum_{w=2}^{u} d^{S(w)} \sum_{k=G(w-1)+1}^{G(w)} q_k \quad (9)$$

4 COMPOSED ALGORITHM OF EMERGENCY SYSTEM DESIGN

The suggested approximate algorithm of the emergency system design problem consists of two phases. The first phase is based on the partial search over set of non-dominated complete valid exposing structures, where inspected structures are obtained by the lexicographic maximal completion of "seed" structure [1, [s^*], [k]] for $k = 1 \ldots r$ according to the following algorithm.

Step 0. Initialize the best found complete valid exposing structure [bu, bS, bG] with value of bH by the structure [1, [s^*], [r]], where the minimal subscript s^* is found by simple bisection process over range [0...m].

Step 1. Repeat the following steps for k=1 … r.

Step 2. Initialize the starting incomplete structure [u, S, G] by valid incomplete exposing structure [1, [s^*], [k]], where the subscript s^* is found, by simple bisection process over the range [sm...sM]. The limit sm is set at zero for k=1 and it equals to s^* at the next steps for k-1. The limit sM is specified using the value bH.

Step 3. Apply the procedure Complete on the structure [u, S, G] and if a valid complete exposing structure is found and $H_{[u,\ S,\ G]} < bH$ holds, then update the exposing structure [bu, bS, bG] with value of bH by the newly found structure.

The second phase starts with the best found exposing structure [bu, bS, bG], where the component $bS(bu)$ yields maximal index $\underline{m}$ of the element from the sequence $d^0 < d^1 < \ldots < d^m$, to take into consideration in the second phase.

The succeeding approximate radial formulation starts from the reduced sequence of distance values $d^0 < d^1 < \ldots < d^{\underline{m}}$. Dividing points D_1, D_2, …, D_v are chosen from the sequence $d^0 < d^1 < \ldots < d^{\underline{m}}$ so that $0 = d^0 = D_0 < D_1$ and $D_v < D_{v+1} = d^{\underline{m}}$, where the set of the dividing points must contain the elements $d^{S(w)}$ for $w \in 1 \ldots bu$. The components of the list bS must be replaced by corresponding subscripts of the dividing points. The zero-one variables x^k_{js} for $s = 0 \ldots v$ and k=1 … r are introduced as in the Section 2. Then the model of the second phase problem can be formulated as follows:

$$Minimize \sum_{j \in J} b_j \sum_{s=0}^{v} e_s \sum_{k=1}^{r} x^k_{js} \tag{10}$$

$$Subject\ to \quad \sum_{k=1}^{r} x^k_{js} + \sum_{i \in I} a^s_{ij} y_i \geq r \quad for\ j \in J,\ s = 0..v \tag{11}$$

$$\sum_{i \in I} y_i \leq p \tag{12}$$

$$\sum_{i \in I} a^{bS(w)}_{ij} y_i \geq bG(w) \quad for\ j \in J,\ w \in [1..bu] \tag{13}$$

$$x^k_{js} \in \{0, 1\} \quad for\ j \in J,\ s = 0,..v,\ k = 1..r \tag{14}$$

$$y_i \in \{0,1\} \quad for\ i \in I \tag{15}$$

The constraints (11) and (12) have the same meaning as constraints (2) and (3). The constraints (13) ensure that the worst situated user perceives the generalized disutility less than or equal to the value of bH obtained in the first phase.

5 NUMERICAL EXPERIMENTS

To compare the min-sum, min-max and composed approaches to the emergency system design with the temporarily failing centers, we performed the series of numerical experiments. The used benchmarks were derived from the real emergency health care system, which was

originally implemented in seven regions of Slovak Republic, i.e. Banská Bystrica, Košice, Nitra, Prešov, Trenčín, Trnava and Žilina. These sub-systems cover demands of all communities - towns and villages spread over the particular regions by given number of ambulance vehicles. In the benchmarks, the set of communities represents both the set J of users' locations and also the set I of possible center locations. The cardinalities of these sets vary from 250 to 650 according to the considered region. The number p of located centers was derived from the original design and it varies from 30 to 80. Various scenarios were studied as concerned the number r of the considered nearest centers and the associated weights $q_1 \ldots q_r$. The condensed results of the experiments are presented in Table 1, where each row corresponds to one of the tested scenario. The following parameters of the obtained service system designs were observed, when particular approaches were applied. The parameter *sgap* gives the deviation of the min-sum objective function value of the solution from the optimal solution of the min-sum problem. The deviation is given in percentage, where the base equals to the objective of the min-sum optimal solution. The parameter *mgap* gives the deviation of the maximal disutility perceived by users in the suggested solution from the best known solution of the problem, where the maximal disutility is minimized. The parameter *mgap* is given also in percentage and the base is min-max objective value of min-max problem solution. Together with *mgap*, there is presented parameter *afl*, which gives percentage of population, which perceive disutility greater than the maximal disutility resulting from the min-max solution. This parameter is computed only for solutions obtained by min-sum approach. All entries of the Table 1 are average values obtained by solving the above-described seven instances.

Table 1: Average results of numerical experiments for individual scenarios in percentage

Scenario						Min-sum			Min-max		Composed	
r	*q1*	*q2*	*q3*	*q4*	*q5*	*sgap*	*afl*	*mgap*	*sgap*	*mgap*	*sgap*	*mgap*
3	1	0.2	0.1	-	-	0,9	1,7	65,9	55,0	1,9	31,6	1,2
3	1	0.2	0.05	-	-	0,9	2,0	78,8	62,1	1,3	34,6	0,3
3	1	0.5	0.2	-	-	1,0	1,4	50,4	51,6	0,8	31,5	1,7
3	1	0.8	0.5	-	-	1,2	1,0	41,4	49,9	0,0	30,0	2,8
5	1	0.5	0.2	0.1	0.05	2,3	1,1	47,3	49,5	0,6	28,0	1,5
5	1	0.8	0.5	0.2	0.1	2,8	1,2	42,9	46,5	0,3	27,4	2,2

To solve the described benchmarks, the optimization software FICO Xpress 7.7 (64-bit, release 2014) was used and the experiments were run on a PC equipped with the Intel® Core™ i7 2630 QM processor with the parameters: 2.0 GHz and 8 GB RAM.

The computational time are not reported here, but we note that the computational time for any instance solved by radial min-sum approach does not exceed 18 seconds whereas min-max solution was completed at most in five minutes due to involved iterative process.

6 CONCLUSIONS

We suggested a composed approach to the emergency system design, where failing centers are considered. The suggested approach has been compared to the min-sum and min-max approaches, which were also based on effective radial formulation of the underlying location

problem. Concerning the min-sum optimal solution, we can notice that the attempt at min-max fairness costs approximately from 50 to 60 percent losses (*sgap*) of the original min-sum objective value, what is also called the price of fairness. The price of fairness can be reduced by the suggested algorithm by 20 percent, if the composed approach is used. On the other side, the losses of price of fairness can be compensated by reducing the disutility (*mgap*) perceived by the worst situated users to the value from 40 to 80 percent, when the best found min-max solution is taken as hundred percent. It can be concluded that the suggested composed approach is able to considerably reduce the unfairness perceived by one percent of users, but it pays for it by 30 percent losses of system objective value. The future research will be focused on a more detailed research of the used exposing structures to find whether big losses of the system objective are inevitable.

Acknowledgment: This work was supported by the research grants VEGA 1/0518/15 "Resilient rescue systems with uncertain accessibility of service" and VEGA 1/0339/13 „Advanced microscopic modelling and complex data sources for designing spatially large public service systems" and APVV-0760-11 "Designing of Fair Service Systems on Transportation Networks". This work was also supported by the project University Science Park of the University of Žilina (ITMS: 26220220184) supported by the Research & Development Operational Program funded by the European Regional Development Fund. We would like to thank to „Centre of excellence for systems and services of intelligent transport" **(**ITMS 26220120050) for built up the infrastructure, which was used.

References

[1] Avella, P., Sassano, A. and Vasil'ev, I. (2007). Computational study of large scale p-median problems. Mathematical Programming, 109: 89-114.

[2] García, S., Labbé, M. and Marín, A. (2011). Solving large p-median problems with a radius formulation. INFORMS Journal on Computing, 23(4): 546-556.

[3] Ingolfsson, A., Budge, S. and Erkut, E. (2008). Optimal ambulance location with random delays and travel times. Health Care management science, 11(3): 262-274.

[4] Janáček, J. (2008). Approximate Covering Models of Location Problems. In Lecture Notes in Management Science. In Proceedings of the 1st International Conference on Applied Operational Research ICAOR ´08, Vol. 1, Sept. 2008 (pp. 53-61). Yerevan.

[5] Janáček, J. (2014). Radial Approach to the Emergency Public Service System Design with Generalized System Utility. Intern. J. of Applied mathematics and Informatics, 8: 7-14.

[6] Janáček, J. and Kvet, M. (2014). Relevant Network Distances for Approximate Approach to the p-median Problem. In Selected Papers of the International Conference of the German operations research society (GOR), September 4-7 2012 (pp 123-128). Hannover: Springer.

[7] Janáček, J. and Kvet, M. (2015). Min-Max Optimization of Emergency Service System by Exposing Constraints. Communications –Scientific Letters of the University of Žilina, 17(2): 15-22

[8] Jánošíková, L. (2007). Emergency Medical Service Planning. Communications –Scientific Letters of the University of Žilina, 9(2): 64-68.

[9] Kvet, M. (2014). Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: Proc. of the 10th Intern. IEEE Conference Digital Technologies 2014 (pp. 198-208). Žilina.

[10] Ogryczak, W. and Slivinski, T. (2006). On Direct Methods for Lexicographic Min-Max Optimization. In: Gavrilova, M. (Eds.): ICCSA 2006, LNCS 3982, (pp. 802-811). Berlin: Heidelberg: Springer.

[11] Snyder, L. V. and Daskin, M. S. (2005). Reliability Models for Facility Location; The Expected Failure Cost Case. Transport Science, 39(3): 400-416.

REESTIMATING THE PURCHASING POWER PARITY SYSTEM FOR EMERGING ASIAN CONTRIES

Alenka Kavkler
Faculty of Economics and Business, University of Maribor, Department of Quantitative Economic Analysis
Razlagova 14, 2000 Maribor, Slovenia
alenka.kavkler@uni-mb.si

Jani Bekő
Faculty of Economics and Business, University of Maribor, Department of Political Economy
Razlagova 14, 2000 Maribor, Slovenia
jani.beko@uni-mb.si

Abstract: This paper investigates the validity of Purchasing Power Parity (PPP) for several emerging Asian economies in the period from January 1995 till December 2013 with respect to US dollar. Because of the documented weaknesses of linear specifications in examining this exchange rate concept, we employ a nonlinear unit root test based on the exponential smooth transition autoregressive model. The results of unit root tests for the US dollar-based real exchange rate series indicate that PPP is not valid for the majority of economies in the sample.

Keywords: purchasing power parity, nonlinear unit root test, smooth transition autoregression

1 INTRODUCTION

The theory of Purchasing Power Parity (PPP) states that exchange rate movements are primarily determined by changes in domestic and foreign prices. This simple but powerful exchange rate theory with only slight modifications since its inception remains one of the basic ingredients for a wide range of contemporary international macroeconomic models. The popularity of PPP has been additionally boosted by a huge body of studies which examine the long-run empirical relationship between exchange rate movements and shifts in relative prices. In order to provide robust empirical support for the PPP theory, researchers frequently experiment with new estimation techniques, with different data sets and by testing the theory on large and various country samples. As emphasized in [9], if the speed of convergence toward the PPP level is increasing with the growing deviation from the equilibrium exchange rate parity, then the otherwise widely used linear specification of PPP testing is inappropriate. Paper [8] suggests that the adjustment of exchange rates can be nonlinear because of transaction costs in international trade and significant trade barriers. Important nonlinearities can also be detected in cases of heterogeneous interactions of traders in the foreign exchange market concerning the expected exchange rate adjustment. The author quotes an additional factor of exchange rate nonlinearity stemming from the coordination effect of intervention by monetary authorities on the main traders in situations of fundamental exchange rate misalignments.

While the testing of nonlinear properties of real exchange rates for developed market economies is reported by a number of studies ([6], [7] and [10]), the application of a nonlinear framework by PPP tests for emerging countries is still poorly documented ([2], [4]) with mixed results. Examining the validity of PPP in emerging countries is particularly important, since it is widely recognized that these countries during the process of introducing market mechanisms were exposed to a range of institutional changes as well as to numerous real shocks which very likely affected their real exchange rates. In this paper, therefore, we try to fill this empirical gap by employing, in addition to the conventional augmented Dickey-Fuller (ADF) test, the Kapetanios et al. [3] (KSS hereafter) unit root test to detect the stationarity characteristics of real exchange rates in several emerging Asian economies.

The paper is divided into following four sections. After the introduction, section 2 presents the procedure for the KSS test. The characteristics of data and the empirical results are summarized in section 3. The main implication of the study is given in the concluding part of the article.

2 PRESENTATION AND METHODOLOGY

Kapetanios *et al.* [3] developed a test for the null hypothesis of unit root against the alternative hypothesis of a nonlinear stationary smooth transition autoregressive (STAR) model. The authors attempted to distinguish between the nonstationary linear processes and the stationary nonlinear ones. The motivation for the development of the new test lies in the persistent failure of the standard ADF test to reject the null of a unit root. Consequently, two alternative frameworks for unit root testing were proposed in recent years. The first approach utilizes panel tests and their higher power in comparison to standard unit root tests. The second approach incorporates stationary models other than the simple AR or ARMA under the alternative hypothesis, including nonlinear transition dynamics. The authors of the test extended the last framework by analyzing a particular kind of nonlinear dynamics, namely exponential smooth transition autoregressive (ESTAR) models.

The smooth transition autoregressive (STAR) model of order 1 is given by the equation

$$y_t = \beta y_{t-1} + \beta^* y_{t-1} G(\gamma, c; y_{t-d}) + \varepsilon_t,\ t = 1,2,\ldots,T,\ \ d \geq 1, \tag{1}$$

where β and β^* are unknown parameters and ε_t is a sequence of independent identically distributed errors. Initially, y_t is assumed to be a zero-mean process, but the framework can easily be extended to include more general processes with non-zero mean and time trend. G represents a continuous transition function bounded between 0 and 1. The slope parameter γ is an indicator of the speed of transition between 0 and 1, whereas the threshold parameter c points to where the transition takes place. y_{t-d} is the transition variable and stands for the variable y lagged d times. The most popular functional forms are the Logistic Smooth Transition Autoregressive (LSTAR) form with logistic transition function and ESTAR with exponential transition function. The LSTAR transition function is monotonously increasing, while ESTAR is U-shaped around c and thus enables reswitching. The ESTAR functional form can be defined as

$$G(\gamma, c; y_{t-d}) = 1 - \exp\left(-\gamma (y_{t-d} - c)^2\right). \tag{2}$$

Kapetanios *et al.* [3] applied the ESTAR transition function with c equal to zero. By substituting G in equation (1) with the ESTAR transition function from equation (2), we obtain the ESTAR model

$$y_t = \beta y_{t-1} + \beta^* y_{t-1}\left[1 - \exp\left(-\gamma \cdot y_{t-d}^2\right)\right] + \varepsilon_t. \tag{3}$$

The null hypothesis of unit root implies $\beta = 1$ and $\gamma = 0$, since $G(0; y_{t-d}) = 0$. Model (3) postulates the nonstationary linear First-Order Autoregressive (AR(1)) model

$$y_t = \beta y_{t-1} + \varepsilon_t \tag{4}$$

under the null hypothesis and a stationary model (with $\gamma > 0$)

$$y_t = \left[\beta + \beta^* G(\gamma; y_{t-d})\right] y_{t-1} + \varepsilon_t,\ \ 0 < G(\gamma; y_{t-d}) < 1 \tag{5}$$

under the alternative. When y_{t-d} is close to zero, model (5) resembles a unit root process, since $G(\gamma;0)=0$. For large values of y_{t-d}, on the other hand, we obtain an approximation of the linear AR(1) with the root equal to $\beta+\beta^*$. We assume that $-1<\beta+\beta^*<1$ (i.e. $-2<\beta^*<0$), as this condition implies stable roots and a stationary AR(1) model.

The null hypothesis $H_0:\gamma=0$ needs to be tested against the alternative $H_1:\gamma>0$. β^* is not identified under the null, and testing such a hypothesis is not feasible. To overcome this problem, the authors of the test used the Taylor series approximation, as interpreted in [1]. In the first step, the authors assumed *d* = 1 (which can be done without loss of generality) and respecified the ESTAR model (3) as

$$\Delta y_t=\beta^* y_{t-1}\left[1-\exp\left(-\gamma\cdot y_{t-d}^2\right)\right]+\varepsilon_t. \tag{6}$$

After replacing the right-hand side expression with its first-order Taylor approximation, one obtains the following auxiliary regression:

$$\Delta y_t=\delta y_{d-1}^3+error. \tag{7}$$

Using the t-statistic approach, the Nonlinear Augmented Dickey-Fuller (NLADF) statistic is defined as

$$NLADF=\frac{\hat{\delta}}{\sigma_{\hat{\delta}}}, \tag{8}$$

where $\hat{\delta}$ denotes the Ordinary Least Squares (OLS) estimate from auxiliary regression (7) and $\sigma_{\hat{\delta}}$ its standard error.

In a more general framework, when the errors of model (6) are serially correlated, the equation is augmented with lagged differences of the process y_t:

$$\Delta y_t=\sum_{j=1}^{p}\beta_j\Delta y_{t-j}+\beta^* y_{t-1}\left[1-\exp\left(-\gamma\cdot y_{t-d}^2\right)\right]+\varepsilon_t, \tag{9}$$

as first proposed by Dickey and Fuller in the derivation of the ADF test. The number of lags (*p*) is defined as the minimal number that removes residual autocorrelation. Auxiliary regression augmented with *p* lagged differences can be given as

$$\Delta y_t=\sum_{j=1}^{p}\beta_j\Delta y_{t-j}+\delta y_{d-1}^3+error. \tag{10}$$

The NLADF test statistic is calculated from equation (8), as before. Paper [3] derives the limiting nonstandard distribution of the NLADF statistic that involves Brownian motion.

3 DATA AND EMPIRICAL ESTIMATES

The investigated sample consisted of the following emerging Asian economies: Bangladesh, Indonesia, Pakistan, Philippines, South Korea and Vietnam.

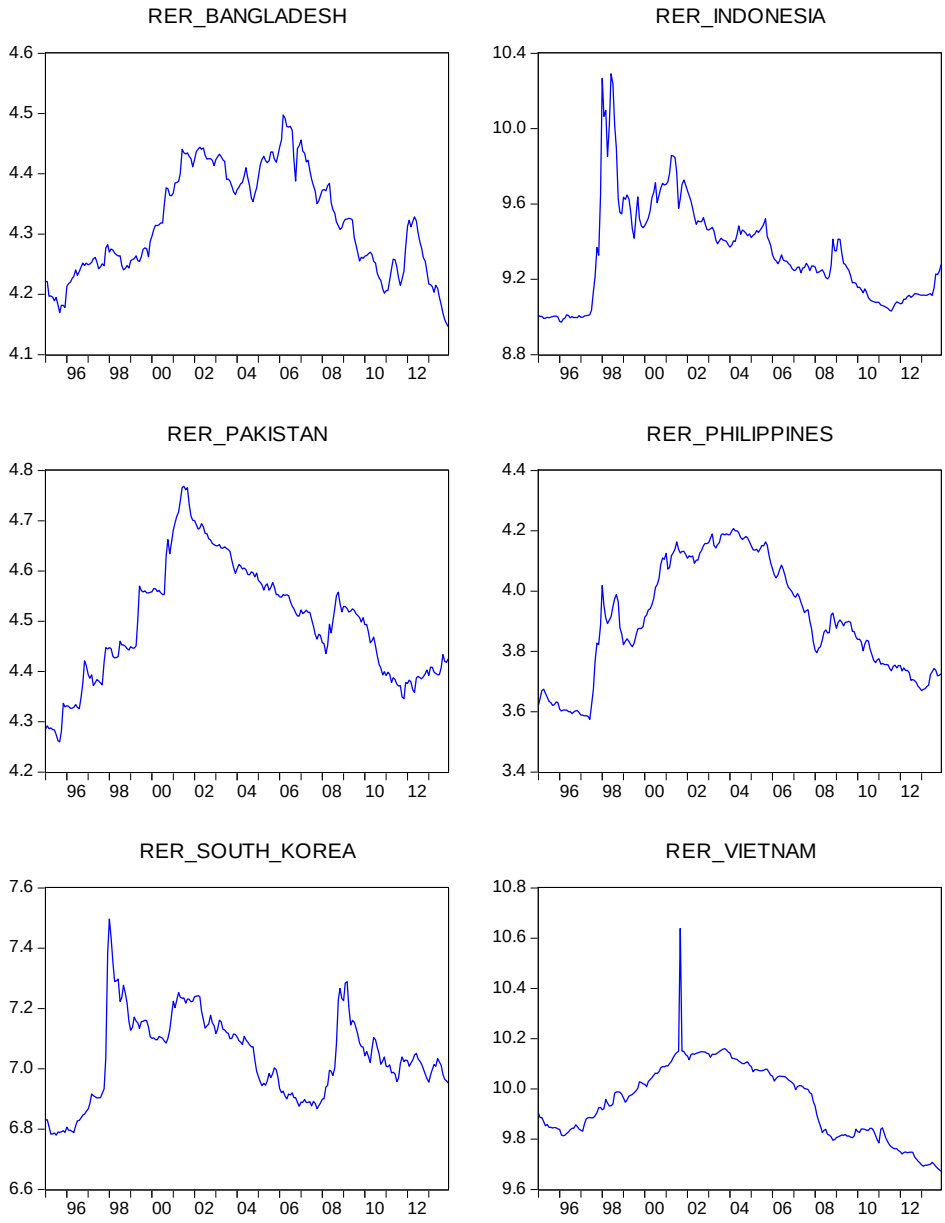


Figure 1: The real exchange rates of emerging Asian economies

The monthly averages of nominal exchange rates and consumer price indices were obtained from the IMF International Financial Statistics and from the Eurostat, with the exception of Pakistan, where the data for the exchange rate from May 1999 till July 1999 were obtained from the State Bank of Pakistan. The US dollar-based real exchange rates covered the period from January 1995 to December 2013. For all countries in the sample, the consumer price indices refered to year 2010. The dynamics of real exchange rates of individual countries is depicted in Figure 1.

The results of the KSS test and the standard ADF test for models with constant and for models with constant and time trend are given in Table 1.

Table 1: Results of unit root test for selected countries

Country	Trend p-value	Intercept			Intercept and time trend		
		No. of lags	KSS	ADF	No. of lags	KSS	ADF
Bangladesh	0,4739	1	-1,9721	-1,0836	1	-1,8162	-0,9555
Indonesia	0,0000	0	-4,0351***	-2,1571	0	-3,9307***	-2,4467
Pakistan	0,8435	1	-1,7356	-1,7894	1	-1,7505	-1,8328
Philippines	0,8634	1	-1,2292	-1,4105	1	-1,2300	-1,3945
South Korea	0,7944	2	-2,8699*	-2,5187	2	-2,8922	-2,5045
Vietnam	0,0000	0	-12,7362***	-2,0654	0	-13,2976***	-2,5795

Notes: The number of lags in the auxiliary regression is defined as the minimal number that removes residual autocorrelation. The 1%, 5% and 10% asymptotic critical values for ADF with intercept are -3.46, -2.88 and -2.57, respectively. The 1%, 5% and 10% asymptotic critical values for ADF with intercept and trend are -4.01, -3.43 and -3.14, respectively. The 1%, 5% and 10% asymptotic critical values for KSS with intercept are -3.48, -2.93 and -2.66, respectively. The 1%, 5% and 10% asymptotic critical values for KSS with intercept and trend are -3.93, -3.40 and -3.13, respectively. The critical values for KSS are taken from [3].
***, ** and * statistically significant at the 1%, 5% and 10% levels, respectively.

Results from the standard ADF unit root test with and without the trend suggest that we cannot reject the null hypothesis of nonstationarity of real exchange rate for any of the countries. By employing the KSS test in models without the trend element, the nonstationarity hypothesis of real exchange rates is rejected in the case of Indonesia, South Korea and Vietnam. When the trend element is examined within the KSS framework, the stationarity of bilateral real exchange rates among the members of examined group holds again for just two countries: Indonesia and Vietnam. As the time trend is significant for Indonesia and Vietnam, but not significant for South Korea, the results suggest that in the linear framework the null hypothesis of unit root holds for all countries, while the nonlinear KSS test rejects nonstationarity for Indonesia and Vietnam.

4 CONCLUSION

This paper evaluated the PPP proposition by scrutinizing the stationarity properties of US dollar-based real exchange rates for the set of emerging Asian economies. The testing results clearly indicate that even after taking into account the nonlinear reversion of real exchange rates of examined countries, we are able to confirm the validity of PPP only in the case of two countries. The PPP concept is therefore not appropriate for determining the equilibrium exchange rates of the majority of countries in our sample. Lack of PPP evidence probably also reflects insufficient coordination of monetary and exchange rate policies among the selected countries. Further empirical examination is needed to solve the PPP puzzle for emerging Asian economies.

References

[1] Granger, C. W. J. and Teräsvirta, T. (1993) *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[2] Janjua, Saeed Anwar, and Eatzaz Ahmad (2006), *Tests of Purchasing Power Parity for South Asian Countries*. Pakistan Economic and Social Review, Vol. 44, 235-243.

[3] Kapetanios, G., Shin, Y. and Snell, A. (2003) Testing for a unit root in the nonlinear STAR framework, *Journal of Econometrics*, **112**, 359–79.

[4] [4] Kim, Hyung-Gun, and Sang Young Jei (2013), *Empirical Tests of Purchasing Power Parity using a Time-Varying Parameter Model: Japan and Korea Cases*, Applied Economic Letters, Vol. 20, 525-529.

[5] [5] Lee, CH. and Chou, PI. (2013) The behavior of real exchange rate: nonlinearity and breaks, *International Review of Economics and Finance*, **27**, 125–33.

[6] Lothian, J. R. and Taylor, M. P. (2008) Real exchange rates over the past two centuries: how important is the Harrod-Balassa-Samuelson effect?, *The Economic Journal*, **118**, 1742–63.

[7] Ramajo, J. and Ferré, M. (2010) Purchasing power parity revisited: evidence from old and new tests for an organisation for economic co-operation and development panel, *Applied Economics*, **42**, 2243–60.

[8] Taylor, M. P. (2006) Real exchange rates and purchasing power parity: mean-reversion in economic thought, *Applied Financial Economics*, **16**, 1–17.

[9] Taylor, A. M. and Taylor, M. P. (2004) The purchasing power parity debate, *Journal of Economic Perspectives*, **18**, 135–58.

[10] Zhou, S. and Kutan, A. M. (2011) Is the evidence for PPP reliable? A sustainability examination of the stationarity of real exchange rates, *Journal of Banking & Finance*, **35**, 2479–90.

INCORPORATION OF SYSTEM BASED EMERGY INDICATORS IN OPTIMISATION OF DAIRY PRODUCTION IN SLOVENIA

Tina Kocjančič, Jaka Žgajnar and Luka Juvančič
University of Ljubljana, Biotechnical faculty, department for Animal Science
Groblje 3, 1230 Domžale, Slovenia
tina.jaklic@bf.uni-lj.si; jaka.zgajnar@bf.uni-lj.si; luka.juvancic@bf.uni-lj.si

Abstract: Emergy analysis is an environmental accounting approach that supports the idea that the natural capital invested in the production of goods and services determines their real value. In this paper emergy based performance indicators are incorporated into standard socioeconomic optimisation model of milk production sector in Slovenia. The multiple-criteria optimisation model based on goal programming (GP) investigates the alternatives to the sector's reorganisation. Scenario analysis suggests that incorporation of emergy criterion into the optimisation model results in a balanced production structure and more favourable economic and biophysical performance of the sector, which confirms complementarity of emergy and economic evaluation approaches.

Keywords: system approach, bio-economic model, emergy, goal programming, agriculture, milk production.

1 INTRODUCTION

Agriculture is a complex system in which the economic principles of production are directly entwined with its ecological characteristics ([1], [2], [3]). Rising demand for food, increasing resource scarcity, high market volatility and growing environmental pressures, challenge the agricultural sector not only to increase productivity, but to do so in a more sustainable manner ([4], [5]). Various objectives of the common agricultural policy and an increased demand for interdisciplinary research approaches had an important role in the development of bio-economic models. These are in general known as (mathematical) models that link different disciplines in order to answer multi-dimensional questions about the organisation of agricultural production systems ([6]). Bio-economic models as decision support tools have to be based on a comprehensive economic evaluation that considers the laws of the natural environment in which the agriculture operates ([7]). However the integration of biophysical and economic components in technical and conceptual sense still remains the most significant challenge in this field ([6], [8],[9]).

Emergy analysis ([10], [11], [12]) is an environmental accounting approach that adopts a biophysical understanding of value. It adheres to the notion that the work of the geo-biosphere is a driving force of all global processes and supports the idea that the natural capital invested in the production of goods and services determines their real value. Emergy (measured in solar emergy Joules) is a cumulative measure which defines the amount of available energy required to deliver a given product or service or to support a specific flow. As such it may be used to quantify environmental costs/value of market and non-market resources and therefore to define the contribution of ecological processes in the production of any good or service. Contrary to a standard economic evaluation, that is anthropocentric in its nature, emergy analysis provides an eco-centric perspective on agricultural activity. In order to contribute to a more complete perspective on agriculture and to improve the quality of decision-making processes, this paper intents to present an innovative attempt to incorporate emergy analysis into a standard optimisation (decision-making) model, which is illustrated through the case of the milk production sector in Slovenia.

2 METHODOLOGICAL APPROACH

2.1 The system investigated

The system investigated is the milk production sector in Slovenia. This selection considered that milk production is the predominating and most important agricultural activity in the country. In the last decade the sector witnessed intensive economically driven structural changes that lead to concentration and specialisation in cattle breeding. This resulted in a substantial decrease in the number of dairy farms, a doubling of the average herd size, and an increase in the milk yield and milk quality ([13]). These changes are economically driven and are likely to continue.

2.2 The development of a modelling tool at the national level

2.2.1 Preliminary analysis at a farm level

The mathematical model of milk production sector was developed in two stages. Firstly, in a preliminary analysis Slovenian milk production sector was categorised into nine farm types. These reflect the diversity within organisation of dairy farming in Slovenia and include small, subsistence oriented farms (F1), semi-subsistence farms (F2), organic farms with a varying degree of production intensity and market presence (F3 and 4), and various types of conventional production systems, significantly differing in herd size, breeds, agricultural land area, proportion of arable land, and the amount of compound feed (F5 to 9). Basic farm characteristics (Table 1) that derived from the Agricultural Census 2010 performed by Statistical office of Slovenia and from the Central Cattle Breeding database from Agricultural institute of Slovenia served as a starting point to formulate technological, economic and environmental parameters of each farm type and to quantify key human-controlled and environmental outputs and input flows to the milk production systems.

Table 1: Basic farm type characteristics

	F1	F2	F3	F4	F5	F6	F7	F8	F9
farm type	subsi-stance	half-subsistance	extensive organic	intensive organic	conven-tional	smaller intensive	highly intensive	larger intensive	agricult. enterprise
*Breed**	S, BS	S, BS	S, BS	S, BS	S, BS	HF, S, BS	HF	HF	HF
Dairy cows	2	8	4	26	20	46	51	105	654
Milk yield (kg/cow)	3,600	4,500	3,000	4,500	5,500	7,400	9,300	7,500	7,000
UAA (ha)	4	9	9	44	17	37	37	90	762
crop field	11%	19%	8%	13%	37%	56%	59%	53%	58%
terrain	steep/hilly	steep, hilly, flat	steep/ hilly	hilly/flat	hilly/flat	hilly/flat	flat	flat	flat

* S-Simmental, BS- Brown Swiss (BS) HF- Holstein–Friesian breed

Several socioeconomic and emergy analysis performance indicators were calculated. These provided an insight into the differences between the farm types' in their profitability, productivity and farmer's income independence and environmental impact of the production (socioeconomic indicators), as well as biophysical efficiency, system's interaction with local environment and system sustainability (emergy indicators). A more in depth description of the methodology, selected indicators and the results of the preliminary analysis can be found in Jaklič et al. [14].

Secondly, based on re-aggregation of the farms' characteristics, a reference model that represents the sector for the year 2010 was specified. The sector in 2010, i.e. its structure, performance and other characteristics, served as a baseline reference to the model solutions obtained from the optimisation model developed.

2.2.2 Definition of the model

The integrated multiple-criteria optimisation modelling tool developed is supported by (weighted) goal programming (GP) and aims to look for an optimal structure of the sector, that will utmost satisfy several often conflicting objectives. The number of dairy farms within each farm type denotes a key model variable and the original model solution that determines values of all other characteristics of the sector.

The sector's boundaries are defined by agricultural land intended for milk production in 2010 and remain fixed through the entire modelling process. Furthermore, another constrain built in the model intends to control transformations of a farm from one type into another if significant differences in technological parameters exist between them (e.g. comparability of farming conditions).

The model is applied to investigate three scenarios of agricultural production planning. The scenarios differ in their objective set that aims to reflect two opposing agro-political orientations (scenario 1 and 2) and their logical combination (scenario 3). The first scenario (SC1) reflects protectionist attitude to agricultural production planning and is focused on achieving a) high productivity, b) employment and c) total income in the sector. The second scenario (SC2) considers an orientation that emphasises more ecologically and socially sustainable agriculture. The objectives incorporated in the model that reflect such attitude are a) maximum total income in the sector, b) lowest possible environmental impact (GHG emission) and c) a minimal pressure of the production process on local environment. The latter is measured by the emergy indicator "Environmental Loading Ratio", which is the ratio of non-renewable to renewable emergy use in a system. A compromise approach that supports protectionist and eco-social objectives in production planning is presented in scenario 3 (SC3). This optimisation is performed based on weighted GP, which signifies that (unlike in the other two scenarios) different weights are assigned to individual objectives. As the objectives related to environmental protection and long term sustainability are only gradually being incorporated into nowadays still prevailing production and income related policy orientation, they have been assigned lower significance.

All of the objective targets applied in the scenarios reflect the extreme values that were preliminary determined by single-criterion optimisation modelling. The solutions of scenario analysis were evaluated and compared according to socioeconomic and emergy indicators of sector's performance (Table2).

Table 2: Socioeconomic and emergy indicators of sector's performance

Socioeconomic indicators		Emergy indicators	
Income	Total income in the sector is the aggregated income of dairy farms	Emergy use	Unit Emergy Value (UEV) indicates biophysical efficiency of a system in emergy use and renewability of a system
Production	Total production in the sector is the aggregated production of dairy farms		Emergy Density (ED) is emergy per hectare and denotes emergy use intensity
Employment	Number of employed persons (1 person equals 2000 working hours)	Interaction with local environment	Emergy Yield Ratio (EYR) measures the ability of a system to exploit free local resources
Public payments	Total amount of public payments (PP)		Environmental Loading Ratio (ELR) indicates pressure of the system on local environment
Income stability	Share of PP in total income	System sustainability	Fraction of renewable emergy in total emergy use (%R)
	Hourly wage is income received per hour of labour (PP incl.)		Emergy Sustainability Index (ESI) is a ratio between ability to exploit free local resources and pressure of a system to local environment
	Income sufficiency is a share of work that is fully paid with income earned (PP excl.)		
Environmental impact	Greenhouse gas emissions (GHG) GHG per unit of production	Emergy exchange	Emergy exchange ratio(EER) indicates the relative trade advantage in emergy exchange (producer vs. purchaser)

3 RESULTS AND DISCUSSION

Figure 1 shows farm type representation in the total milk production according to the results of multiple-criteria model scenario analysis. The results show distinctive differences in the structure when pursuing protectionist or eco-socially oriented targets. Compared to the production in 2010 the reorganisation that is based on pursuing protectionist goals results in a significant increase in highly intensive production (F7; 25%). Such structure is a result of the intensification of the sector that would occur by reorganisation of small subsistence (F1) and organic farms (F3) into half-subsistence farms (F2), a reorganisation of larger organic (F4) and averaged sized conventional farms (F5) into larger and more intensive farm types (F6) and with a concentration of largest farms (F6, 8 and 9) into highly intensive production systems (F7). This model solution fully attains only the production target, whereas income and employment are not reached entirely (87% and 89% respectively).

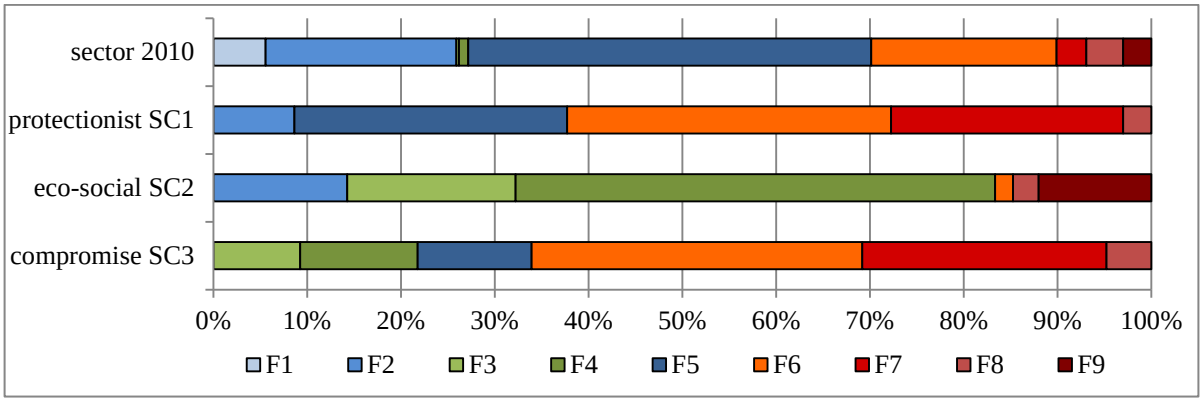


Figure 1: Farm type representation in the milk production structure

Contrary to the protectionist scenario which clearly favours conventional and intensive production, eco-social scenario leads to reorganisation of the sector in which the less intensive organic farms (F3 and F4; 83%) contribute 69% to the total milk production. In addition, compared to the year 2010 agricultural enterprises (F9) significantly increase their contribution to the total production (12%). Such structural differences result mainly from the reorganisation of less intensive conventional production (F1, F2) to extensive organic production (F3), as well as from the reorganisation of conventional production at middle-sized farms (F5, F6) into intensive organic production (F4). In this model solution both environmental and biophysical objectives are reached, whereas total income in the sector is 21% lower that targeted.

Finally, the solution of the compromise scenario (SC3) results in a relatively well-balanced production structure. The structure mainly comprises of (highly) intensive production at F6 and F7 (35% and 26% respectively) but also includes more than 20% of organically produced milk. This solution assumes a restructuring of the smaller into larger farms as well as a reorganisation of the production technology from conventional to organic. The model solution fully attains solely the income targeted. Larger deviations from the target values are noted for the objectives that relate to environmental impact and pressure on local environment (+81% and +66%). This is mainly attributed to the formulation of the model that assumes lesser importance of these two objectives in comparison to the others.

Figure 2 shows the deviations in the presented model solutions from the reference performance of the sector in 2010. Indicators that characterise performance of the sector were normalised and aggregated into logical groups (Table 2). The results show that protectionist

and eco-social scenarios provide almost diametrical solutions. The protectionist attitudes in agricultural production planning result in higher income of the sector, improved productivity and as a result also high biophysical emergy use efficiency. However, such reorganisation of the sector that is largely based on intensive production (Figure 1) has harmful effects on the local and global environment and represents a step-back in system sustainability. This is due to a high dependence on non-renewable resources.

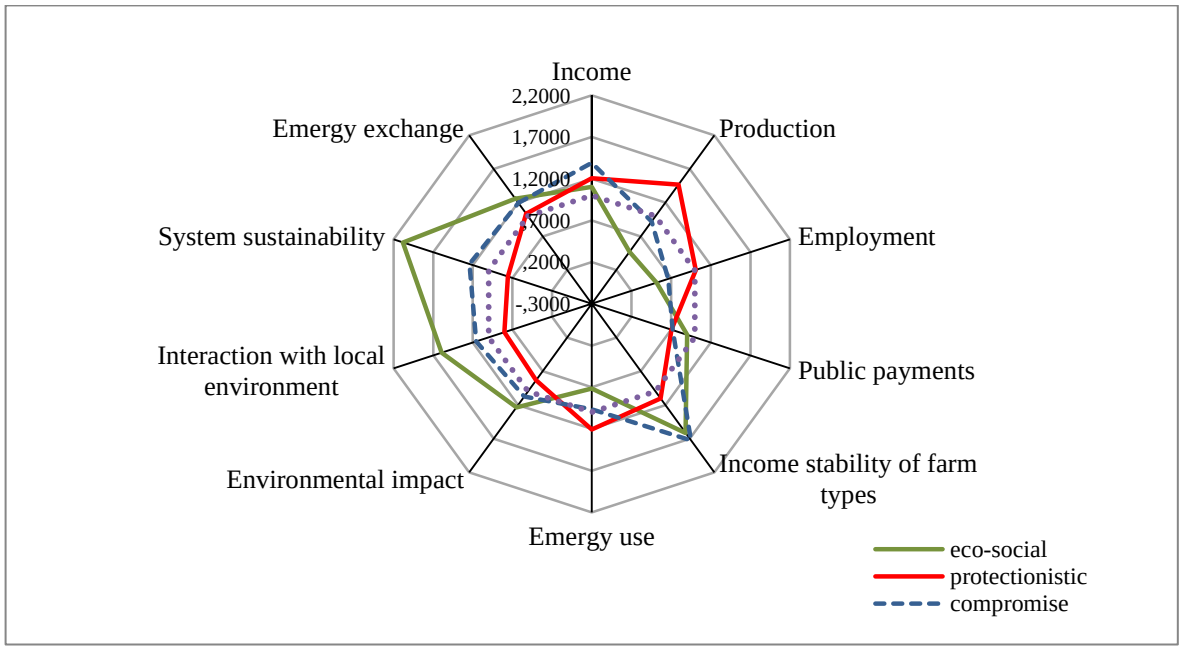


Figure 2: Relative variations in selected characteristics of milk production sector between the baseline (2010) and multiple-criteria solutions

Conversely, ecologically and socially sustainable oriented agricultural production planning, which also includes emergy targets, results in a highly sustainable production that originates from the system's low pressure on local environment and its high ability to exploit free local resources. A significant disadvantage of such a solution that favours organic production is a considerably low productivity of the sector and a substantial decline in employment.

A solution from the model that simultaneously considers the targets of both of the extreme approaches shows to offer a compromise between their conflicting requirements. Such an integrated solution emphasises the advantages and reduces the weaknesses of the previous (extreme) solutions. It produces a sector that achieves significantly better results than the baseline scenario of the sector in 2010, albeit at the expense of lower employment, slightly higher public expenses and a less pronounced decline in productivity.

4 CONCLUSION

In this paper we present an innovative attempt to incorporate emergy analysis into standard economic evaluation models that support agricultural production planning. By incorporating both, an economic (anthropocentric) and emergy based (eco-centric) indicators, the multiple-criteria model aims to provide more comprehensive evaluation of the sector's performance and of various alternatives to its reorganisation.

We confirm that emergy and economic evaluation are not mutually exclusive, instead their joint application provides an informative insight into agricultural activity. The results do not negate a link between intensification and overall improvement of the sector. Moreover, a

solution that is characterised by a wide and diverse range of agricultural holdings and a balanced production structure leads to the conclusion that the integration of system-based sustainability indicators into standard economic optimization models can improve socio-economic performance and biophysical functioning of the sector.

A key limitation of the proposed approach is that the model is simplified in such manner that does not allow reallocation of resources between agricultural sectors. We believe that the shortcomings of this simplification can be properly avoided by extending the modelling tool to include other potentially competitive sectors. The development in this direction would improve the applicative value of the modelling tool.

References

[1] Smith C. A. S., McRae T. 2000. Understanding and Assessing the Environmental Sustainability of Agriculture. V: Environmental Sustainability of Canadian Agriculture: Report of the Agri Environmental Indicator Project. A Summary. McRae, T., Smith, C.A.S., Gregorich L.J. (eds.). Ottava, Agriculture and Agri-Food Canada: 2

[2] OECD. 2000. Environmental Indicators for Agriculture. Methods and Results. Executive Summary 2001. OECD Publishing: 53 pp. http://www.oecd.org/greengrowth/sustainable-agriculture/1916629.pdf

[3] van Zanten B. T., Verburg P. H., Espinosa M., Gomez-y-Paloma S., Galimberti G., Kantelhardt J., Kapfer M., Lefebvre M., Manrique R., Piorr A., Raggi M., Schaller L., Targetti S., Zasada I., Viaggi D. 2014. European agricultural landscapes, common agricultural policy and ecosystem services: a review. Agronomy for Sustainable Development, 34, 2: 309-325

[4] Godfray H., Charles J., Beddington J. R., Crute I. R., Haddad L., Lawrence D., Muir J. F., Pretty J., Robinson S., Thomas S. M., Toulmin C. 2010. Food security: the challenge of feeding 9 Billion people. Science, 327: 812-818

[5] OECD/FAO. 2012. OECD-FAO Agricultural Outlook 2012-2021. OECD Publishing and FAO: 286 pp. http://dx.doi.org/10.1787/agr_outlook-2012-en

[6] Flichman G., Allen T. 2012. Bio-economic modeling: State-of-the-art and key priorities. CGIAR, Research Program on Policies, Institutions and Markets. http://www.pim.cgiar.org/files/2013/12/FlichmanThomas_BioEconModeling_Final.pdf

[7] Daily G. C., Söderqvist T., Aniyar S., Arrow K., Dasgupta P., Ehrlich P. R., Folke C., Jansson A., Jansson B., Kautsky N., Levin S., Lubchenco J., Maler K. G., Simpson D., Starrett D., Tilman D., Walker B. 2000. The value of nature and the nature of value. Science, 289: 395-396

[8] Gasparatos A., El-Haram M., Horner M. 2009. The argument against a reductionist approach for measuring sustainable development performance and the need for methodological pluralism. Accounting Forum, 33, 3: 245-256.

[9] Gasparatos A., Scolobig A. 2012. Choosing the most appropriate sustainability assessment tool. Ecological Economics, 80: 1-7.

[10] Odum H. T. 1983. Systems Ecology: An Introduction. New York, John Wiley: 644 pp.

[11] Odum H. T. 1988. Self-organization, transformity, and information. Science, 242: 1132-1139

[12] Odum H. T. 1996. Environmental accounting. Energy and environmental decision making. New York, John Wiley & Sons,: 370 pp.

[13] KIS. 2011. Poročilo o stanju kmetijstva, živilstva in gozdarstva v letu 2010. Pregled po kmetijskih trgih. Ljubljana, Kmetijski inšitut Slovenije. http://arhiv.kis.si/pls/kis/!kis.web?m=36&j=SI

[14] Jaklič T., Juvančič L., Kavčič S., Debeljak M. 2014. Complementarity of socio-economic and emergy evaluation of agricultural production systems: The case of Slovenian dairy sector. Ecological Economics, 107: 469-481.

HAS THE LONG-TERM RECESSION CHANGED MANAGERS' ASSESSMENTS AND EXPECTATIONS IN THE CROATIA'S RETAIL TRADE SECTOR?

Nataša Kurnoga and Mirjana Čižmešija
University of Zagreb
Faculty of Economics & Business
Trg J. F. Kennedyja 6
10000 Zagreb, Croatia
{nkurnoga, mcizmesija}@efzg.hr

Abstract: The aim of the paper was to explore whether long-term recession influenced the changes of managers' assessments and expectations in Croatia's retail trade sector. The analysis was divided in two main parts: analysis of association between variables in Croatia's business survey in the first quarter of 2015, and comparison of the obtained results with the same results for the first quarter of 2009. It was concluded that the Croatia's managers, during six-year recession have changed their assessments and expectations, which influenced the changes in the character of relationship of the variables in Croatia's business survey for retail trade sector.

Keywords: Business survey, RTCI – Retail Trade Confidence Indicator, measures of association, Pearson *Chi*-square test, contingency coefficient

1 INTRODUCTION

Business survey is a qualitative survey based on managers' estimates and expectations about their past and present business situation concerning four sectors of national economies: manufacturing industry, construction, retail trade and services sector. Surveys basis are questionnaires in which managers express their opinions. Qualitative data as managers' opinions are translated into the quantitative composite indicators: Industrial Confidence Indicator (ICI), Construction Confidence Indicator (BCI), Retail Trade Confidence Indicator (RTCI) and Services Confidence Indicator (SCI) [8].

Croatia has been conducting surveys in accordance with the harmonised EU methodology continually since 1995. Initially, Croatia's survey was conducted on a quarterly basis. Croatia's questionnaires are harmonised with the same in EU, but Croatia's questionnaire contain additional question about the liquidity of the company: "Liquidity of your firm is: (1) good, (2) temporary difficulties, (3) bad" [3].

Previous researches have shown the existence of association between variables in Business survey, i.e. relationship between manager's assessments and expectations exists [1,4,7]. The aim of this research was to empirically explore whether long-term recession in Croatia influenced the changes of managers' attitudes in the retail trade sector related to assessments and expectations about their business position. The study is based on a comparison of the results of the same kind of research [4] conducted for the first quarter of 2009 and for the first quarter of 2015, i.e. at the beginning and at the end of recession period in Croatia. Differences are, among other things, a product of managers' pessimism during six-year recession. It resulted from managers' more pessimistic views on expected production, expected business situation, current level of overall order books, on export order books, on the volume of stocks etc. Primarily negative views on managers' assessment and expectations are in accordance with the general economic situation in Croatia: unemployment, decreased export, increased public debt, etc.

As in the previously mentioned paper, research examines the changes in relationships among the variables of business survey for retail trade sector.

2 METHODOLOGY

Business survey is a qualitative survey based on questionnaires in which managers express their opinions and expectations about their business situation. The questions in the Croatia's business survey questionnaire for the retail trade sector [3,8] are presented in Table 1.

Table 1: Questions in Croatia's business survey questionnaire for the retail trade sector

Questions
Present business position, Q1
Turnover over the past quarter, Q2
Turnover over the same quarter in one year ago, Q3
Stock of finished products, Q4
Prices over the past quarter, Q5
Expected turnover in the next quarter, Q6
Expected business position in the next 2 quarters, Q7
Ownership, Q8
Present employment position, Q9
Liquidity, Q10

Almost all questions, except ownership (Q8), present employment situation (Q9) and liquidity (Q10) have a similar response pattern where the answers are given according to a three-option scale e.g. higher, unchanged and less or good, satisfactory and bad. To the question about the company's liquidity the offered answers are good, temporary difficulties and bad. Questions that are related to ownership and employment are not of the interest in this research.

The purpose of this research was to analyse the association between the selected variables in business survey for retail trade sector. The empirical analysis was divided (detailed) in four parts. In the first part the analysis of association between the variable liquidity and variables which are components of RTCI, was conducted. RTCI is a composite indicator calculated as a simple average of seasonally adjusted balances of three variables: present business position, stock of finished products and expected business position. The second part of the research was the analysis of association between variable liquidity and the remaining variables. The third part was the analysis of association between selected variable (e.g. present and expected situation). In the fourth part, the comparison of research results for the first quarter of 2009 and the first quarter of 2015 (at the beginning and at the end of the recession) was conducted.

For these purposes, the nonparametric measures of association [5,6] were used, as well as tests of their significance. The Pearson Chi-square test is the most commonly used in researches [1,2,4,7] for testing the presence of relationship between two categorical variables but it does not evaluate the strength of the relationship. Based on the Chi-square statistic it is possible to calculate different coefficients of association which measure the intensity of the relationship between two variables. The nonparametric measures of association which evaluate the existence and strength of the relationship between two categorical variables are Phi coefficient, contingency coefficient and Cramer's V. In this paper Chi-square test statistics, corresponding p-values and contingency coefficients are given.

The hypotheses of the Pearson Chi-square test are:

$$\begin{aligned} H_0 &: \text{the variables are independent} \\ H_1 &: \text{the variables are not independent,} \end{aligned} \tag{1}$$

where the null hypothesis (H_0) test states that there is no relationship between two variables and the alternative hypothesis (H_1) contradicts the null hypothesis. Formally, the hypotheses are:

$$H_0 : p(x_i, y_j) = p(x_i)p(y_j) \forall \mathrm{i} = 1,2,...,\mathrm{r}, \forall \mathrm{j} = 1,2,...,\mathrm{c},$$
$$H_1 : \exists i, j \quad p(x_i, y_j) \neq p(x_i)p(y_j), \tag{2}$$

where $p(x_i, y_j), p(x_i), p(y_j)$ are joint probability density function, marginal column and marginal row probabilities, respectively.

The survey was conducted for the first quarter of 2015. The sample consisted of 50 companies in the retail trade sector and it is representative for the entire Croatia. The results obtained on that sample were compared to the results of research conducted for the first quarter of 2009 [4]. In that study the sample was composed of 102 companies in the retail trade sector. It is evident that the sample in this study is smaller than in the previous research. It is a consequence of a long-term recession period where the population was reduced. The response rate is also lower which is in accordance with managers' pessimism due to a long recession that lasts for six years.

The data source is *Centre for Business Research* (the periodical *Privredni vjesnik* – a business magazine in Croatia).

3 EMPIRICAL RESULTS

As mentioned above, the empirical analysis was divided (detailed) in four parts. The first part is related to the analysis of association between the variable liquidity and the variables which are components of RTCI: present business position, stock of finished products and expected business position. The results are shown in Table 2.

Table 2: Measures of association – liquidity and components of RTCI

Variables	***Chi-square***	***p-value***	***Contingency coefficient***
Present business position, Q1	9.972395	0.00159	0.4077782
Stock of finished products, Q4	3.684533	0.05492	0.2619791
Expected business position, Q7	3.684533	0.05492	0.2619791

The presented results indicate that there are associations between variable liquidity and components of RTCI. All *p*-values are smaller than theoretical significance level ($\alpha = 0.1$). Consequently, the null hypotheses of no association are rejected. The results could indicate that there is an association between variable liquidity and present business position, liquidity and stock of finished products and liquidity and expected business position. The values of contingency coefficients for variables stock of finished products and expected business position show low association (<0.3), while the value for variable present business position point out the moderate association between variables.

The second part concerns the analysis of relationship between the variable liquidity and the remaining variables (Table 3).

Table 3: Measures of association –liquidity and remaining variables

Variables	*Chi-square*	*p-value*	*Contingency coefficient*
Turnover over the past quarter, Q2	5.546668	0.01852	0.3160000
Turnover over the same quarter in one year ago, Q3	4.919276	0.02656	0.2992872
Prices over the past quarter, Q5	0.000167	0.98968	0.0018297
Expected turnover in the next quarter, Q6	0.047954	0.82666	0.0309542

The results in Table 3 show that *p*-values for two variables related to the past turnovers are smaller than theoretical significance level. Hence, the null hypotheses of no association are rejected. The results may indicate that there is an association between variables liquidity and turnover over the past quarter and also between variables liquidity and turnover over the same quarter in one year ago. The values of contingency coefficients show moderate association for variable turnover over the past quarter and law association for variable turnover over the same quarter in one year ago. *p*-values for variables prices over the past quarter and expected turnover in the next quarter are greater than any theoretical significance level. Therefore, the null hypotheses of no association cannot be rejected.

The third part includes the analysis of association between the variables relating to current and future business position and turnover. The results are given in Table 4.

Table 4: Measures of association – present and expected situation

Variables	*Chi-square*	*p-value*	*Contingency coefficient*
Present business position, Q1 ***Expected business position in the next 2 quarters, Q7***	9.183674	0.00244	0.3939193
Turnover over the past quarter, Q2 ***Expected turnover in the next quarter, Q6***	6.088259	0.01361	0.3294660
Turnover over the same quarter in one year ago, Q3 ***Expected turnover in the next quarter, Q6***	13.25021	0.00027	0.4576995

The results in Table 4 show that all *p*-values are smaller than theoretical significance level. Hence, the null hypotheses of no association are rejected, i.e. it may indicate that variables present and expected business position, and variables past turnover and expected turnover are dependent. The values of contingency coefficients show moderate association for all relationships.

The fourth part concerns the comparison of research results for the first quarter of 2009 and the first quarter of 2015, i.e. previous research refers to the beginning of the recession and current research to the end of the recession (Table 5 and Table 6).

Table 5: Without changes in the character of relationship, p-values

Variables	*p-value, 2009*	*p-value, 2015*
Liquidity and present business position	0.000	0.002
Liquidity and prices over the past quarter	0.808	0.990
Present and expected business position	0.000	0.002
Turnover over the past quarter and expected turnover in the next quarter	0.000	0.000
Turnover over the same quarter in one year ago and expected turnover in the next quarter	0.002	0.000

It can be seen that character of relationship between variables liquidity and present business situation remained unchanged regardless of the stage of the economic cycle (at the beginning

and at the end of recession period). The same conclusion was passed for managers' assessments and expectations on the business position and on the turnover. The assessments on liquidity and assessments on prices obviously were not affected by the recession (in accordance with the survey results; it is not managers' opinion).

Table 6: The changes in the character of relationship due to long-term recession, p-values

Variables	***p-value, 2009***	***p-value, 2015***
Liquidity and stock of finished products	0.208	0.055
Liquidity and expected business position	0.393	0.055
Liquidity and turnover over the past quarter	0.915	0.019
Liquidity and turnover over the same quarter in one year ago	0.574	0.027
Liquidity and expected turnover in the next quarter	0.026	0.827

After six years of the recession, just before the end of recession period, stocks and liquidity are dependent variables. The situation in the beginning of the recession in Croatia in 2009 was different. Apparently, the managers with the deepening of crisis and recession have become more pessimistic. The same conclusion can be made for the managers' assessments on liquidity and on expected business position, on liquidity and on turnover over the past quarter and also on liquidity and on turnover over the same quarter in one year ago.

4. CONCLUSION

Questionnaires in Croatia's business survey include additional question about the company's liquidity. Consequently, much of the research relates to the variable liquidity and to its relationship with the remaining variables of business survey in retail trade sector.

Assessments on liquidity and present business position, assessments and expectations of the business situation, expected turnover and turnover over the past quarter and expected turnover and turnover over the same quarter in one year ago were 2009 dependent variables, as well as in 2015. The change did not occur even in the character of the relationship between variables liquidity and prices.

It was also concluded that among some variables there was a change in the character of the relationship: liquidity and stock of finished products, liquidity and expected business position, liquidity and turnover over the past quarter, liquidity and turnover over the same quarter in one year ago. The reverse situation is in the character of the relationship of liquidity and expected turnover in the next quarter.

Based on the research results it was concluded that the Croatia's managers, during the recession have changed the criteria in the evaluations and expectations, which influenced the character of relationship among the variables in Croatia's business survey for retail trade sector. This is a result of long-term recession and increased manager's pessimism.

Acknowledgement
This paper has been fully supported by the Croatian Science Foundation under the project No. 3858.

References

[1] Bahovec, V., Čižmešija, M. (2005). Test of no Association between variables – components of Consumer Confidence Indicator and stratification variables for Croatian Consumer survey. Proceedings of the 8th International Symposium on Operationa Research SOR '05, Zadnik Stirn, L., Drobne, S. (Eds.): 405-410.

[2] Bodrožić, I., Jurun, E., Pivac, S. (2007). Chi-Square versus proportions testing – Case study on tradition in Croatian brand. Proceedings of the 9th International Symposium on Operational Research SOR '07, Zadnik Stirn, N., Drobne, S. (Eds.): 415-420.

[3] Čižmešija, M. (2008). Konjunkturni testovi Europske unije i Hrvatske. Privredni vjesnik, Zagreb.

[4] Čižmešija, M. (2011), Neovisnost varijabli konjunkturnog testa trgovine u Hrvatskoj u recesijskom razdoblju. Zbornik Ekonomskog fakulteta u Zagrebu, Vol. 1, No. 1: 17-32.

[5] Gibbons, J.D. (1993), Nonparametric Measures of Association (Quantitative Applications in the Social Sciences, Sage Publications L.t.d., London.

[6] Hitchcock, D. B. (2009), Yates and Contingency Tables: 75 Years Later, Journal Électronique d'Histoire des Probabilités et de la Statistique, Volume 5, Issue 2: 1-14.

[7] Kurnoga, N., Erjavec, N., Bahovec, V. (2015) The Analysis of Association between the Variables in Croatian Business Survey for Services Sector. The Business & Management Review, Datta, P. R. (Eds.): 263-268.

[8] The joint harmonised EU programme of business and consumer surveys, User guide (updated 21/03/2014). European Economy. European Commission, Directorate-General for Economic and financial affairs, http://ec.europa.eu/economy_finance/indicators/business_consumer_surveys/userguide_en.pdf [Accessed 06/06/2015]

DISCRETE-EVENT SIMULATION MODEL OF CUSTOMER SUPPORT SERVICE IN TELECOMMUNICATIONS

Andrea Majić
Poslovanje 2 Ltd.
Vijeća Europe 95, 32000 Vukovar, Croatia
andrea.majic@gauss.hr

Adela Has and Marijana Zekić-Sušac
University of Josip Juraj Strossmayer in Osijek, Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
adela.has@efos.hr; marijana@efos.hr

Abstract: The objective of this paper is to create a simulation model for customer service in the telecommunication sector with the aim of identifying problems and reducing waiting time and queues. The model is based on discrete-event simulation and uses triangular and normal random distribution functions in modeling business processes. The data for the model is obtained by interviewing employees of a large Croatian telecommunication company. The activity cycle diagram was built showing the flow of business processes, and life cycles for all types of entities that exist in the model. After the simulation model is created and ran in a number of iterations, the simulation results are used to determine expected waiting time, size of queues in two departments, the cost, and the utilization of resources. A what-if analysis is conducted to propose possible improvements of the business processes. The model can serve as a support in making operational decisions in customer relationship management.

Keywords: discrete-event simulation, modeling, queues, business processes

1 INTRODUCTION

Simulations are able to explore the impact of various decisions in a real system without any consequences that might appear if the same decision were actually implemented. Simulation modeling can be applied for various problems, such as the effect reduction or increase in the number of employees, profitability of opening a new plant, reduction of queues in a specific business process at a facility etc. In this paper, the focus is on the research of a customer support service in a large telecommunication company in order to simulate their work processes by discrete-event simulations. It is shown how the simulation modelling enables to improve business processes by detecting issues, and suggesting improvements by reorganizing resources and redesigning process flow in customer support service. The proposed solution can contribute in optimizing business processes and thus creating more satisfied customers.

2 PREVIOUS RESEARCH

Discrete-event simulation can be applied to different areas, including services, manufacturing, and other areas. Zec [11] uses discrete-event simulation to determine the number of ships in the passenger lines that satisfy the needs and decrease the cost. Kaurić et al. [5] created a simulation model of an operation desk office for bus ticket sales in a transportation company, using the GPSS simulation tool. A number of papers investigate its usage in telecommunications, focused on improving the performance of call centers. Ding et al. [4] investigated the true demand in call centers with redials and reconnects, and proposed a model to estimate the number of fresh calls, and the redial and reconnect probabilities. Some researchers [6] applied forecasting techniques to estimate incoming calls for a mobile telecommunication company and developed a simulation model to enhance its

performance. Ma et al. [7] proposed a new performance assessment methodology for human-in-the-loop call center systems by focusing on individual or team performance rather than queues. The simulation model developed by Pichitlamken et al. [8] enables what-if analysis, as well as continuous-time Markov chain (CTMC) queueing models. Brown et al. [2] developed several statistical techniques for analysis of the basic components in a call center simulation, such as testing if a point process is a Poisson process, or estimating the mean function in a nonparametric regression.
Previous research illustrates the efficiency of discrete-event simulation modeling in improving various aspects of the customer service performance. However, there is no general model that could serve all, and each system deserves special attention.

3 METHODOLOGY

According to [1], simulation is the imitation of the operation of a real-world process or system over time, while the behavior of a system over time is studied by developing a simulation model. Two groups of methods are used for simulation modeling of dynamic systems; system dynamics and discrete-event simulation [3]. In this paper the discrete-event simulation was used which has shown success in analysis of waiting queues. Banks et al. [1] define discrete-event simulation as modeling of systems in which the state variable changes only at a discrete set of points in time. The system behavior is described with the discontinuous (discrete) method, in the form of a sequence of events and activities in the system. The basic elements of discrete simulations are entity, event, activity and process. The entities are the main objects of the system, and can be permanent or temporary. The event represents a change in the system state, which occurs at some point of time. The activity represents an interaction of entities that takes some time, and causes a change in the state of entities. The process can be described as set of logically related, consecutive events through which temporary entities pass by. The simulation clock measures the elapsed time of the simulation [3]. Since the duration of the processes and the time of arrival of entities is difficult to determine in advance and the same process can often have different duration, the random variables can be used with various probability distribution functions (pdf). In this model three distribution functions were used: exponential, normal (Gaussian), and triangular distribution. The exponential distribution is used to determine the time of arrivals of entities, according to the following pdf [1]:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0\,, & x < 0 \end{cases} \quad (1)$$

where λ is the mean number of occurences per time unit. In our research $\lambda = 1$. The normal (Gaussian) distribution requires the mean and the standard deviation as the input parameters. The triangular distribution requests three input parameters: two endpoints (minimum and maximum), and the mode value [1]. In our experiments, the endpoints and the mode (most likely) values for each of the processes are determined by interviewing the operators.

The conceptual model for discrete-event simulation provides great opportunities in presentation and formalization a real system by displaying the system at various levels [10]. Graphical methods that are frequently used to display the conceptual model are activity cycle diagrams (ACD diagrams) and Petri nets. Activity cycle diagram can be defined as sequence of actions which go through various types of entities [3]. Entities may be active or idle. The activity duration is determined in advance by a time distribution function, while the time of waiting in queue is not easy to determine because it is a consequence of the dynamics of arrivals requirements and dynamics of serving these requests.

4 CONCEPTUAL MODEL AND DANA

The main problem in the observed customer support service was the size of the queue, which is often around 100 users in the queue. The estimated distribution of inquires is 65% in TS, and 35% in IA department. The company is interested to reduce the queues by not increasing the labor cost. In order to describe the problem, the ADC diagram of the conceptual model is created. ACD diagram is particularly appropriate to illustrate the conceptual model for problems with a strong queuing structure. In the customer service support there are three types of permanent and one temporary type of entities. Permanent entities are: the Interactive Voice Response system (IVR), operators, and providers of technical support and account info, while the user is a temporary entity that passes through the system. The ACD for the described model is presented in Figure 1.

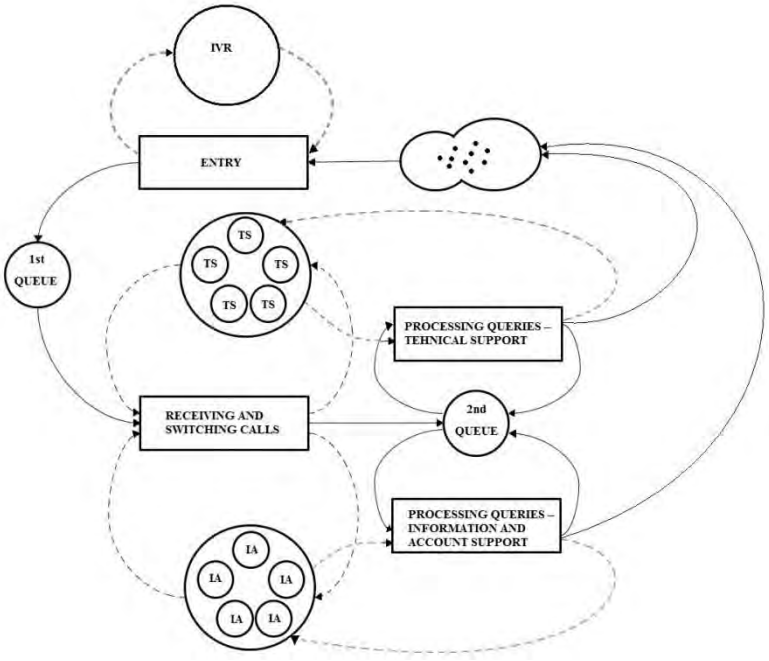


Figure 1: ACD of the conceptual model of customer service in telecommunications

It can be seen in Figure 1 that a user can be in the following queues: (1) after pressing a button on the IVR machine while waiting for the operator to answers the call, and (2) after switching while waiting for either the information and account operator or the technical support operator receives the call. For clarity reasons, only five operators in each department are graphically shown at the diagram. Upon the completion of the processing, the call is terminated, i.e. the user exits the system.

Data were collected by interviewing a telecommunication company's employees and screening the process in the customer support service department. The customer service in the Technical support (TS) department currently employs about 160 operators, while the Information and Account (IA) department employs about 120 operators. The average number of employees working at the same time is 45 in TS and 25 in IA. The salary of operators differs on workdays, Sundays and holidays. Working hours of agents are flexible, they can select a number of working hours from 6:00 am until 1:00 am every day, depending on request. Therefore, the variables needed to compute the cost in the whole process are:

- price of work p_w, p_s, p_h ,
- working hours $h_{wi}, h_{si}, h_{hi},\ i = 1,2, \dots, n$

where p_w is the price per hour of work on workdays, , p_s is the price per hour of work on Sundays, p_h is the price per hour of work on holidays, h_{wi} is the number of working hours of an operator i on workdays, h_{si} on Sundays, h_{hi} on holidays, while n is the total number of operators. The total cost c is computed by

$$c = \sum_{i=1}^{n} h_{wi}\, p_w + h_{si} p_s + h_{hi} p_h \quad (2)$$

The processing time t_p $(p = 1, 2, \dots, m)$ is the time needed to complete a customer request in a process p, and is estimated on the basis of a normal or a triangular distribution functions (a suitable functions is selected for each process). The number of customers entering the process is determined by the exponential distribution. The simulation aims to find the cost c which enables the satisfactory average size of queues, waiting time, and the utilization of employees.

The real process flow in the customer service support could be described as follows. The process begins with a customer call, where the customer waits for the first free operator to answer the call. After establishing the connection a customer briefly explains the problem or seeks for a certain information while at the same time the operator opens the application with customer information. Then the process of authentication follows, which involves identity verification. The minimum duration of this process is 30 seconds, but there are users who do not have ready information needed, and the process can sometimes take up to 5 minutes. According to estimates, 90% of users typically passes authorization, while 10% of them do not. After successful authentication, the customer requests from the operator some information about a specific query. The conversation with a customer in the IA department usually takes considerably less time than the conversation in the TS department. The main reason is that in the TS department the operators usually have to repair some failure or interfere. It has been estimated that the request about the services or account usually takes minimum 30 seconds, maximum 6 minutes, and most likely 4 minutes. The estimated duration of a conversation in the TS is minimum 2 minutes, maximum 10 minutes, and most likely 5 minutes. If the problem cannot be resolved by the operator in the TS department, the operators forwards the request to technicians. It is estimated that 70% of faults is resolved by the operators in the customer service, while 30% of cases is forwarded to other departments. Estimated number of calls in 1 hour reach up to 500 calls, and users wait for a free operator from one minute to 40 minutes (if previously not give up and terminate the call), and most likely around 10 minutes.

5 RESULTS

On the basis of the ACD diagram, the simulation is built in Arena Simulation software tool. Arena is a discrete-event simulation tool which uses SIMAN language to create the simulation program. It also enables what-if analyses to test the effect of various ideas, policies and strategies before they are implemented in the real system [9]. The basic time unit was the minute. Users enter the system randomly on average 1 entity every 2 minutes. The model is simulated in duration of 30 days, 19 working hours per day, and ran in 5 replications with distribution functions (normal and triangular) for determining the duration of the processes. The process flow diagram shown in Figure 2 uses rectangular shapes to present processes, while the decisions are represented in rhombus.

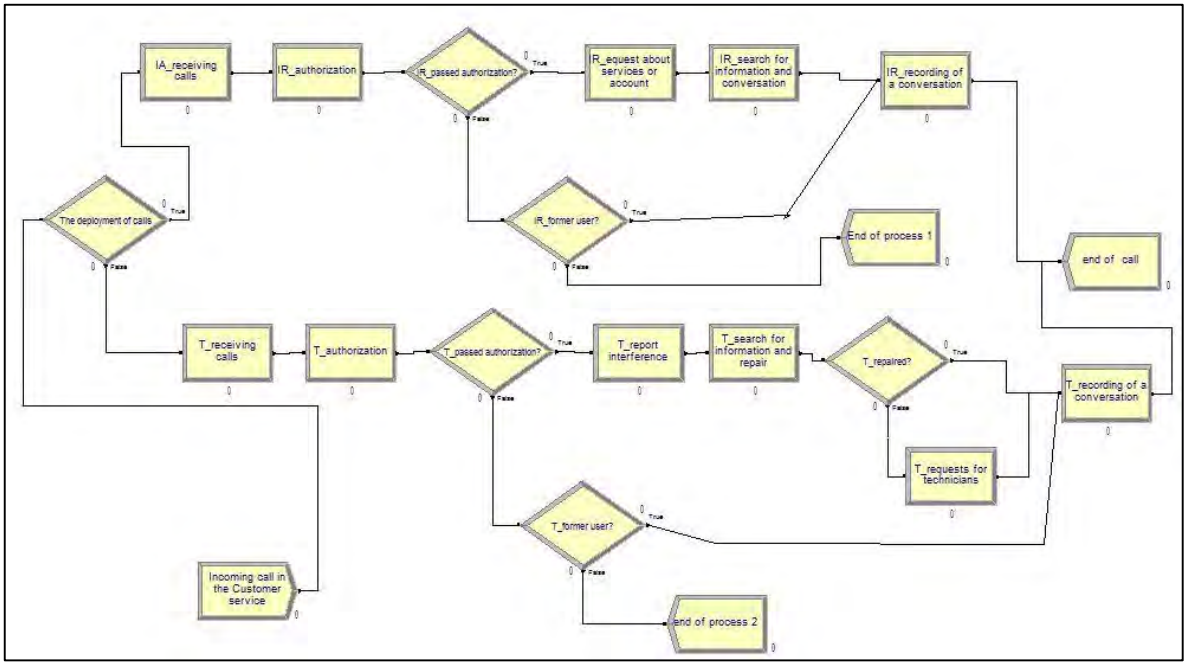


Figure 2: Simulation model of the customer support service

The results presented in Table 1 show that the problem of large queues and long average waiting time is present in scenario 1 which simulates the real situation, especially in the TS department. The average size of the queue in the TS department is 170, while in the IA department is 85. The resource utilization also shows that operators in the TS department have a higher percentage of effectiveness.

Table 1: The results of the simulation process in two different scenarios

Output variable	Value in Scenario 1	Value in Scenario 2
Ave. time a user spends in the process	10.43 minutes	10.49 minutes
Min. time a user spends in the process	1.03 minutes	1.11 minutes
Max. time a user spends in the process	19.67 minutes	19.17 minutes
Ave. waiting time for user	32.15 minutes	4.48 minutes
Max. waiting time for user	287.03 minutes	71.55 minutes
Ave. size of the queue	255 (IA department 85, Technical department 170)	17 in total (IA department 1, Technical department 16)
Total monthly cost of the system	1.509.553 Euro	1.509.620 Euro
Users processed in the system per month	289 561	292 553
Resource utilization: IA operator Technical operator	 73.29% of time 81.66% of time	 50.2% of time 71.38% of time

In order to perform a what-if analysis, the scenario 2 was created by increasing the number of operators in the TS department for 5, and decreasing the number of operators in the IA department for 5. The results of scenario 2 show that the although the average time a user spends in the process is approximately the same as in scenario 1, there are significant improvements in the waiting time and in the queue size. The average waiting time for user is now only 4.48 minutes comparing to 32.15 minutes in scenario 1, while the maximum waiting time is also four times shorter in scenario 2. Concerning the average size of the queue, the scenario 2 has only 17 users in queue in average, which is 15 times lower than in

scenario 2. Such improvements are not negligible, showing the benefit of simulation modelling.

6 DISCUSSION AND CONCLUSION

The paper presents a procedure to use simulation modeling in order to analyze the real situation and to suggest some improvements in business processes of a customer support service in telecommunications. The created model showed that the discrete-event simulation with uncertainty distribution functions is able to identify problems, and to offer improvements. In order to provide a realistic model, the ADC diagram of the conceptual model is created, and the simulation model is built with entities, resources, processes, and decisions. The processing time in each process is estimated by using uncertainty distribution functions: normal and triangular. Based on the simulation results the problem of long queues in the technical department is observed in the real-case scenario. In order to conduct a what-if analysis, the second scenario is created in which the number of operators is re-scheduled between the departments. The analysis showed that significant improvements could be made in decreasing the waiting time and the size of the queue. Some processes that take longer time than desired are also identified, such as the process of information seeking in the system, which could be solved by centralizing the database and applications. This work could be useful to researchers in the field of business simulations, and managers in customer service support, who could use this methodology to improve the efficiency of their business processes.

References

[1] Banks, J., Carson J.S., Nelson, B.L., Nicol, D.M. (2010). Discrete-Event System Simulation, Fifth edition, Prentice Hall, Upper Saddle River.

[2] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L. (2005). Statistical analysis of a telephone call center, A queueing-science perspective, Journal of the American Statistical Association Vol. 100(469), pp. 36-50.

[3] Čerić V., Varga M. (2004). Informacijska tehnologija u poslovanju (Information Technology in Business), Ekonomski fakultet u Zagrebu, Element.

[4] Ding, S., Koole, G., van der Mei, R.D. (2015). On the estimation of the true demand in call centers with redials and reconnects, European Journal of Operational Research, Vol. 246(1), pp. 250-262

[5] Kaurić S., Pogarčić I. (2013). Simulacija prodaje pretplatničkih karata javnoga prijevoza (Simulation of prepaid ticket sale for public transportation), Zbornik Veleučilišta u Rijeci. Vol. 1(1), pp. 307-318.

[6] Kim, Y.B., Lee, H., C, H-G. (2005). A Discrete Event Simulation Study for Incoming Call Centers of a Telecommunication Service Company, Artificial Intelligence and Simulation, Lecture Notes in Computer Science Vol. 3397, pp 390-399.

[7] Ma, J., Kim, N., Rothrock, L. (2011). Performance assessment in an interactive call center workforce simulation, Simulation Modelling Practice and Theory, Vol.19(1), pp. 227-238.

[8] Pichitlamken, J., Deslauriers, A., L'Ecuyer, P., Avramidis, A.N. (2003). Modelling and simulation of a telephone call center, Simulation Conference 2003. Proceedings of the 2003 Winter, 7-10 December 2003, Vol. 2(1), pp. 1805-1812.

[9] Rockwell Automation (2015). Arena Simulation Software, http://www.arenasimulation.com/arena_Home.aspx [Accessed: 16. June 2015.]

[10] Severović K., Žajdela N., Cvetković Šoštarić B. (2009). Konceptualni model kao alat za upravljanje kvalitetom bankarskih usluga (Conceptual model as a tool for quality management in bank services). Ekonomski vjesnik, 2(1) pp. 147-160.

[11] Zec D. (2003). Optimalna veličina ro-ro putničkog broda u obalnoj plovidbi (The optimal size of the ro-ro passenger ship in coastal navigation). Pomorski zbornik, Vol.40 No.1, pp. 35-50.

COMPUTER CRIMES IN MODERN SOCIETY

Slavko Šimundić
University of Split, Faculty of Law
Domovinskog rata 8, 21000 SPLIT, CROATIA
slavko.simundic@pravst.hr

Danijel Barbarić
University of Split, Faculty of Law
Domovinskog rata 8, 21000 SPLIT, CROATIA
danijel.barbaric@pravst.hr

Abstract: Today people are faced with several forms of computer crimes. In addition to the most common forms of computer crimes (unauthorised access to computer system, computer espionage, computer sabotage, computer fraud, computer forgery, software piracy), many other forms of computer crimes also exist. This paper will analyze several forms of computer crimes, while data mining will aid with the analysis of these crimes.

Keywords: computer, crimes, forms, data mining

1 INTRODUCTION

Computer Crimes are widespread worldwide. Everyone can become victim of cyber attacks. Today we have different types of computer crimes. It is known that cyber crime can come from many sources [10]. Some people with vast knowledge can hack everything (email accounts, smartphones, etc.). However, it is really dangerous and serious if someone hacks the nuclear program of a country (for example Stuxnet), governments data, medical devices, airplane information systems, power networks etc. There are many types of computer crimes. The most common forms of computer crimes are unauthorised access to computer system, computer espionage, computer sabotage, computer fraud, computer forgery, software piracy, botnet, denial of service, identity theft, cheating, keystroke logging, hoax, spam, misrepresentation, social engineering, backdoor, meeting hijackings, computer piracy, pirating by end users, overusing a client server, internet piracy, installing illegal software to newly composed computers, forgive software, pirating audio-visual content, etc. It is imperative that analysis is completed for each type of computer crime and while research for new crimes is conducted for ones yet to be discovered.

2 DATA MINING AND MOST COMMON FORMS OF COMPUTER CRIMES

Data mining can be used for one of the model crime detection problems, which can help in detecting the crimes and speeding up the process of solving crimes [7]. Data mining, as one of the operational research methods, can be helpful with analyzing the most common forms of computer crimes. Sindhu and Meshram (2012) claim that: “Cyber Crime Data mining is the extraction of Computer crime related to determine crime patterns”[8]. Data mining as one aspect of operational research, can be connected with computer crimes and help identify more information about this field. Different forms of computer crimes and their statistics data should be analyzed worldwide with data mining. Also, while analyzing the most common forms of computer crimes, perpetrators (or attackers) and theirs behaviours must also be closely examined with the assistance of network forensics to collect and analyze log and status of information [9]. Web crime mining has its process which includes several modules: pre-processing web texts, candidate features, feature reconstruction, data mining

[6]. Several data mining techniques which include association analysis, classification and clustering analysis are often used to identify patterns in data [6].

Nowadays computers and information and communication technology are part of our everyday life. Privacy should be protected from different forms of computer crimes. The usual crimes being committed in cyberspace are business espionage, Internet fraud, cyber terrorism and more are on the rise [8]. According to Sindhu and Meshram (2012), statistics data of emerging cyber crimes are as follows: "Data theft (33%), Email abuse (22%), Unauthorized Access (19%), Data alteration (15%), Virus attacks (5%), DoS attacks (3%), Others (3%)" [8].

Most common forms of computer crimes are: unauthorised access to computer systems (hacking), computer espionage, computer sabotage, computer fraud, computer forgery, software piracy, harmful and illegal contents and many other forms of misuse [12, 4].

2.1 Unauthorised access to computer systems (hacking)

Unauthorised access to computer systems are acts, with the aim of avoiding checks to the access of systems which enable the perpetrator to, under the guise of being authorised, use the services and resources of the systems. Checking access is implemented in two ways: by identifying the person accessing and by establishing who has authorised access, that is, establishing all that the person is allowed to do when gaining access to the system. The consequences of hacking are breach of confidentiality, access to data and programmes, and to services.

Perpetrators are persons who have been allowed access to the system where they use the terminals of other employees or they are persons who access the system through modems. They are divided into internal and external perpetrators. Internal perpetrators gain access, in general, through the carelessness of other employees while external perpetrators use various methods to gain access to the password to enter the system. Perpetrators ensure that the acts they commit through the computer systems of others or through the use of cloned mobiles remain unknown in order to erase their tracks and prevent others from discovering them.

2.2 Computer espionage

Computer espionage includes manipulation where the aim is unauthorised access to secret data and information stored in systems or in transmission through telecommunication channels. The sale of such procured information causes illegal gain.

Perpetrators are persons who in this way want to prevent competition and gain information which they can use. They are persons who are highly educated in information and telecommunication technology. Sometimes they work alone, while some work for others. When they happen to come across information, they offer it on the market. External entities attempt to remain unnoticed when gaining access to the system to avoid existing measures of physical protection. Professionals who have means at their disposal, take advantage of all the weaknesses of the system in order to reach their goal.

2.3 Computer sabotage

Computer sabotage consists of activities which aim to disenable the normal functioning of the system or preventing its use, that is, the use of its resources. This includes erasing, changing, damaging data, damaging programmes with the aim of disenabling its further use and function. Computer sabotage is particularly dangerous when hacking into business systems, military and all other infrastructures.

Perpetrators are persons with technical knowledge who use sophisticated methods and means. The most common perpetrators are hackers, terrorists, criminals motivated by revenge, political gain and other beliefs. This is a physical activity leading to further disenabled or interfered computer use.

2.4 Computer fraud

Computer fraud is related to the types of manipulations with data for illegal gain. This occurs by entering using, storing, and exchanging data within a certain informational system or within exchanging information on the internet. It is carried out easily, and uncovered with difficulty because the perpetrators use the personal data of others.

Perpetrators are mainly employees of legal entities with authorised access to computer systems. Internal perpetrators in the absence of their negligent colleagues use their terminals. Often former employees are involved, motivated by revenge (usually due to losing their job), act in this way.

2.5 Computer forgery

There are two types of manipulation. In the first, the computer is used for forging the existing digital form of documents of other people, and in the other, the computer is used in order to create such documents and commit forgery. Authorised and unauthorised access to the system where such documents can be found enables changing the existing data or copying a new document. Scanners are used for transmission onto the computer, programmes the contents and form of which are changed and printers which print these documents out on paper.

Perpetrators have a wider knowledge and use of sophisticated legal and illegal equipment which today is no longer a deciding factor because technology is becoming cheaper every day. With relatively little investment, significant results can be achieved.

2.6 Software piracy

Software piracy consists of the unauthorised reproduction and use of protected programmes. It is one of the most common forms of computer crime. The digital form enables fast and simple reproduction and the exchange of programmes to which the lack of legal regulation has contributed.

There are many perpetrators because for this form of computer crime particular knowledge or equipment is not necessary. It must be emphasised here that the perpetrators of software piracy can be divided into two categories: those who do this for amusement and those who engage in piracy for illegal gain. In other words, the second category contains perpetrators who consciously multiply and distribute protected programmes without the permission of the bearer of copy right and profit from this illegal act. Limited technical knowledge is necessary and the advancements of the modem and the Internet have helped.

3 OTHER FORMS OF COMPUTER CRIMES

Together with development of information technology and development in computer power, more sophisticated methods of computer crimes now exist [11]. Year after year, different forms of computer crimes are discovered.

Since computer crimes happen through networks, it can be concluded that “data mining forms are useful to extract unexpected network patterns, eg. “it can be useful to introduce a research of processing tasks of anomaly and data mining based botnet detection systems.[2].”

Data mining can also be used in other forms of computer crimes like: botnet, denial of service, identity theft, cheating, keystroke logging, hoax, spam, mispresentation, social engineering, backdoor, meeting hijackings, computer piracy etc. [12].

3.1 Botnet

A botnet attack consists of sending unwanted e-mails from one personal computer and from one Internet access while the owner of a personal computer remains oblivious. Numerous "infected computers" appear in practice where a vast number of e-mails to many e-mail addresses is sent. The infected computers are called "Zombie" computers and hackers, in this way, hide behind other individuals and find victims worldwide. A botnet is also known as a "Web of Infected Computers."

3.2 Denial of service

Denial of service is a kind of attack, more familiarly know by the term DoS attack. It is a kind of an attack by which an attempt is made to overload the network equipment and web providers. When this happens, computers and providers are no longer able to achieve legitimate transmission. DDoS attack (Distributed Denial of Service) is a kind of attack where the sources that overload the network equipment and web providers are located in several places on the Internet. Mostly attacked computers are used to attack other networks and computers which have access to the Internet.

3.3 Identity theft

Identity theft (Phishing) can manifest in either two ways. The first way is stealing important information from data bases, economic subjects or stealing from institutions which store such information. In the other way, the attacker steals important information directly from the user or gets it fraudulently. By using identity theft the hackers have the goal of gaining access to passwords of the victims' bank accounts. They also use this technology to uncover some other important data.

3.4 Cheating

Cheating (Spoofing) is in no way a harmless form of computer attack. It consists of creating a false version of something and this is usually a specific location on the Internet and e-mail addresses. The user registers with his valid information and then this information finds its way to the hacker who misuses it to access the true location.

3.5 Keystroke logging

Keystroke logging manifests itself so that the infected computer records all the keys pressed on the keyboard by the user. In this way, the hackers can uncover a range of important pieces of information and misuse them. By keystroke logging the hackers can steal e-mails and go through the Recycle Bin and illegally find certain data or information.

3.6 Hoax

Hoax is an e-mail containing false content sent with the aim of intimidating or falsely informing the receiver. A hoax is sent in order for it to be forwarded to as many e-mail addresses as possible. One of the main indicators of a hoax is the following sentence: "Send this message to as many addresses as possible." Hoax creators use professional terminology and try to establish credibility by referring to known companies so it is difficult to establish the credibility of the e-mail. The receivers indeed do forward them thinking they are helping others by doing so. Most common forms of a hoax can be found in warnings about harmful programmes, chain mail, false requests for help, intimidating and threatening messages, false petitions, compromising and harmless messages [12, 13].

3.7 Spam

Spam messages are unwanted messages of an advertising nature which lure the receivers into buying goods or services on the Internet. It is not problematic if these kinds of messages appear periodically, but the real problem is when these messages start to appear in electronic mail boxes on a daily basis. In recent years, spamming has become a serious problem [3].

3.8 Mispresentation

The perpetrator uses the identity of another person with the aim of accessing important data. The access to important data can be achieved by using a code or a password which enables access to a computer system, that is, to a protected area in it. Important data can be obtained by using some of the afore described illegal acts.

3.9 Social engineering

Social engineering consists of perpetrators manipulating other persons who can be used for unauthorized actions in the informational systems or a personal computer. In other words, obtaining data is achieved through an unauthorised access to an informational system or personal computer which can be manipulated while the owner of the two is unaware of the actions taking place. It is about information to which the attacker cannot gain access to in a legal way. The perpetrators can be all of those people who have access to that same computer system while the illegally obtained data is used to accomplish their own goals.

3.10 Backdoor

This kind of attack on computers consists of bridging security systems. Authors of programme support use the "backdoor" for a more quality processing of programme procedure and for skipping some of its parts. The perpetrator can change some important programme details so it is necessary to supervise the informational system more often. After completion, the back door has to be removed from the final version of the programme support, but this is sometimes unintentionally forgotten.

3.11 Meeting hijackings

This is quite an unusual term for a type of attack which happens in moments of distraction of persons who legally use an informational system or a personal computer. Under the term meeting hijacking a strong connection at a meeting layer is included, that is, at a layer of

network protocol among equal users or between clients and providers. In this way, the attacker overtakes the meeting with the aim of gaining unauthorised access to data and services or control over planting forged data. While opening a meeting authorisation is mandatory by inserting certain information. The perpetrators uncover this information and the attack can be carried out during the meeting.

3.12 Computer piracy

Under the term computer piracy, software pirating and pirating audio-visual content is included. In the context of considering criminal acts which can be committed with the help of computers and other modern technologies, it is about criminal acts which are characterised as specific criminal acts from the field of computer crimes.

It must be pointed out that software is one of the most significant and most valuable factors which can be found inside modern informational technologies. Pirating starts from software because it not only enables copying of existing computer programmes, but also duplicating illegal computer programmes from the Internet. By buying software, the buyer is not its owner. By buying software a right of its usage is actually bought with the obligation of respecting the limits set by its owner (most often the software developer). Detailed rules connected to its usage are explained in the supporting documents and to which the buyer must abide.

There are more kinds of software pirating today and these are: pirating by employees of a firm, overusing the client server, internet piracy, installing an illegal software to newly composed computers, forging software and pirating audio-visual content.

4 CONCLUSION

Sudden technological advancement over the last three decades has brought about a mass of new criminal acts and challenges. Information and communication technology ease our lives but at the same time if we become victim of cyber crime we can have some problems in our lives. Today we have various data mining based approaches and skills which are used to prevent Cyber attacks and crimes [5]. Data mining methods can be used to confine cyber crimes but unfortunately they cannot completely eliminate them [1].

Everybody who frequently uses the computer should work on their own computer literacy and security. Adequate protection for email and similar services is the password which consists of various letters, numbers and other punctuation. The password should be changed more often than not. It would be useful to have several email addresses and use various passwords and use those passwords that are difficult to remember or that we cannot remember, but rather we have them saved in a document which we keep on a USB stick where nothing else is saved and which we keep in a safe place. Of course, it is difficult to expect that someone would use something like this or say a fingerprint or the eye's iris or some third type of identification. The great majority of people use precisely easily memorable passwords and do not change them often. That is why it is a good idea to be on guard, to not give anyone your password and to pay a great deal of attention when we use other ways of accessing the Internet apart from the one we have at home. When we use a computer for our work and deal with sensitive problem areas (e.g. finances) it is good to use a computer that does not have the Internet at all and to never on that computer link the memory cards (e g. USB stick) of others.

Danger is also present during Internet shopping, thus much care should be taken. It is recommended that the protection of the web page being used for shopping is checked carefully and that a special bank card is used which does not contain many funds and on

which regular payments are not received. If the web pages are not protected, Internet shopping should be avoided.

References

[1] Aboubekeur, Hamdi-Cherif; Kara-Mohamed, Chafia (alias Hamdi-Cherif) (2015); Data mining technology for countering cybercrime - A Compulatational Intelligence Perspective, Proceedings of the 9th International Conference on Computer Engineering and Applications (CEA '2015), Dubai, United Arab Emirates, pp. 522-527

[2] Alparslan, Erdem; Karahoca, Adem; Karahoca, Dilek; (2012), BotNet Detection: Enhancing Alaysis by Using Data Mining Techniques, in: Karahoca, Adem (ed.), Advances in Data Mining Knowledge Discovery and Applications, InTech, Chapters published, pp. 349-366

[3] Bača, Miroslav; Ćosić, Jasmin (2013) Prevencija računalnog kriminaliteta, Policija i sigurnost, Vol. 22 No.1, pp. 146-158

[4] Dragičević, Dražen (1999) Kompjutorski kriminalitet i informacijski sustavi, Informator, Zagreb

[5] Sayyada Sara Banu, et. all. (2015), Data Mining based Soft Computing Skills towards Prevention of Cyber Crimes of the Web, International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3. Issue. 3, pp. 1226-1231

[6] Sharma, Anshu; Sharma, Shilpa (2012); An Intelligent Analysis of Web Crime Data Using Data Mining, International Journal of Engineering and Innovative Technology (IJEIT) Vol. 2, Issue 3, pp. 203-206

[7] Shyam Varan, Nath, (2006), Crime Pattern Detection Using Data Mining, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, December, 2006, pp. 41-44

[8] Sindhu, K.K.; Meshram, B.B., (2012) A Digital Forensic Tool for Cyber Crime Data mining, IRACST – Engineering Science and Technology: An International Journal (ESTIJ), Vol. 2., No.1., pp.117-124

[9] Sindhu, K.K.; Meshram, B.B., (2012) Digital Forensics and Cyber Crime Datamining, Journal of Information Security, 3, pp.196-201

[10] Stepheson, Peter (2000) Investigating Computer-Related Crime, Boca Raton, Fl; CRC Press, United States of America

[11] Šimundić, Slavko (2007) Pravna informatika, Sveučilišta u Splitu, Pravni fakultet Split

[12] Šimundić, Slavko; Franjić, Siniša. (2009) Računalni kriminalitet, Sveučilišta u Splitu, Pravni fakultet Split, Split

[13] Šimundić, Slavko; Franjić, Siniša; Vdovjak, Krešimir, (2012) HOAX, Zbornik radova Pravnog fakulteta u Splitu, Vol. 49. No.3, pp. 459-480

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Session 9:

OR Perspectives

WEATHER RISK HEDGING - AN OPERATIONS PERSPECTIVE

Frank Chen and Houmin Yan
Department of Management Sciences
College of Business
The City University of Hong Kong, Hong Kong
youhchen@cityu.edu.hk and cbhyan@cityu.edu.hk

Abstract

Deviations from expected weather conditions can challenge many company's revenues, costs or profits, especially those in in energy, retailing, agriculture, transportation, construction, travel, food processors, and real estate. Weather risk here is referred to as potential financial losses caused by unusual yet non-catastrophic weather conditions, and its management is a type of risk management done by organizations to address those losses. In this talk, we will provide a literature survey on weather risk management, with a focus on hedging and operations issues.

Advance in meteorological intelligence and e-commerce technology, combined with big data, can forecast weather-dependent product demand, not just in aggregate, but also down to the city. The weather risk management market enables businesses to actively manage the financial impact of weather through risk transfer instruments based on a defined weather element, such as temperature, rain, snow, wind, etc. Today, growing numbers of companies are using this market for weather-related business risks.

By making a payment (a "premium") to a separate party that will assume the financial weather risk for them, a company essentially is buying a type of insurance - the party assuming the risk will pay the buyer a pre-set amount of money that will correspond to the loss or cost increase caused by the disruptive weather.

We will survey the literature and practices regarding how companies with weather sensitive demand or yields have used weather risk hedging in their operations decisions, and how their operational and hedging decisions should be made jointly and optimally.

RACING WITH RESET UTILIZING FULL INFORMATION

Jawad Elomari
ORTEC B.V.
Houtsingel 5, 2719 EA, The Netherlands
jawad.elomari@ortec.com

Abstract: Racing-with-reset was originally created for *F-Race* in situations where there is a fixed budget constraint, and early termination is of no value. Upon reset, however, *F-Race* only makes use of part of the information gathered so far, as it relies on statistical tests that require equal sample sizes, for instance the *F*-test. This work replaces the statistical tests with ones that allow for unequal sample sizes, and hence can utilize full information. The proposed method will be shown to achieve a lower probability of incorrect selection, compared to *F-Race* and equal allocation, over a range of scenarios.

Keywords: racing, racing with reset, multiple comparisons, selection algorithms, parameter tuning

1 INTRODUCTION

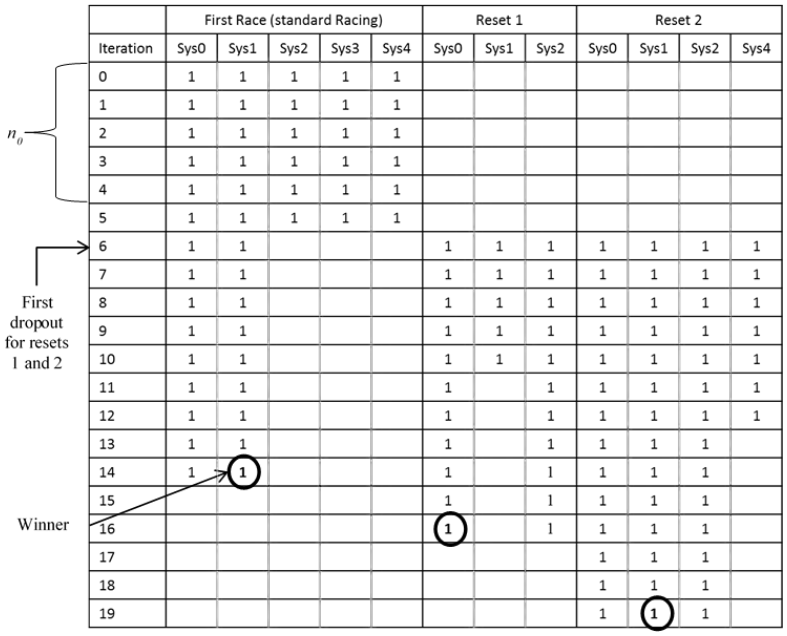
Racing algorithms were created for the model selection problem in Artificial Intelligence, but they can be applied to the more general ranking and selection problem, where one is to select the best among a finite set S of k stochastic "systems", e.g. stochastic algorithms, simulated scenarios, medical treatments, etc. Assuming no prior knowledge about the systems, Racing begins with sampling each system n_0 times to estimate its characteristics, like the mean and variance. Then, it iteratively applies an Analysis of Variance (*ANOVA*) test, followed by pair-wise comparisons with the current best (say lowest mean) to identify the system(s), possibly none, that perform significantly worse than the current best. In such a case, inferior systems are discarded from the race, and the rest are sampled once more. If the *ANOVA* is insignificant, all systems are sampled once again. This continues until one system remains, a winner, or until the sampling budget N is consumed [1].

In situations where there is a fixed budget constraint, and it is of no value to terminate early, Racing is not applicable. A work-around was previously proposed in [2] where by the algorithm "resets" if a winner is identified and N is not consumed. Racing-with-reset begins with a relatively high initial significance level α_0 (say 0.3 or 0.2), and runs until a single system remains. If at this point N is not consumed, the algorithm rolls back to the iteration where the first drop out occurred, lowers the significance level α by a factor γ, and then applies the tests again to *all* systems. With the new smaller α Racing will either discard all previously discarded systems at that iteration (less likely), or pick some up. The reset continues until N is consumed, each time the previous α is discounted by γ. See Figure 1.

Since racing-with-reset was created for *F-Race* [3], it had a drawback; the statistical tests in *F-Race* rely on a two-way *ANOVA* model and therefore require equal sample sizes. This limits the amount of information that can be used at the reset iteration and afterwards. For example, in Figure 1, reset 1 occurs at iteration 5, if a two-way test is applied, only samples 0-5 for all systems can go enter the *ANOVA* test with the new $\alpha \leftarrow \alpha \cdot \gamma$. A one-way test, however, can utilize all 15 samples of systems 0 and 1, plus all 6 samples of the remaining systems. The ability to use full information, in addition to the reduction in α, is expected to improve the algorithm's chances in making a correct selection.

This work introduces a new racing-with-reset algorithm for a fixed N, suitable for independent or low correlated systems. It relies on the Kruskal-Wallis *one*-way *ANOVA* by ranks [4], followed by the Bonferroni-Dunn multiple comparisons test [5], and shall be called *KW-RaceR*, where "*R*" stands for reset. Empirical comparison against *F-Race* and Equal Allocation (*EA*) will show that the proposed method achieves a lower probability of *incorrect*

selection (*PICS*) using the same N for a wide range of scenarios. The test-bed consists of probability distributions with pre-determined means, variances, and covariances that simulate the performance of stochastic systems.



	First Race (standard Racing)					Reset 1			Reset 2			
Iteration	Sys0	Sys1	Sys2	Sys3	Sys4	Sys0	Sys1	Sys2	Sys0	Sys1	Sys2	Sys4
0	1	1	1	1	1							
1	1	1	1	1	1							
2	1	1	1	1	1							
3	1	1	1	1	1							
4	1	1	1	1	1							
5	1	1	1	1	1							
6	1	1				1	1	1	1	1	1	1
7	1	1				1	1	1	1	1	1	1
8	1	1				1	1	1	1	1	1	1
9	1	1				1	1	1	1	1	1	1
10	1	1				1	1	1	1	1	1	1
11	1	1				1		1	1	1	1	1
12	1	1				1		1	1	1	1	1
13	1	1				1		1	1	1	1	
14	1	(1)				1		1	1	1	1	
15						1		1	1	1	1	
16						(1)		1	1	1	1	
17									1	1	1	
18									1	1	1	
19									1	(1)	1	

Figure 1: An example of a racing-with-reset algorithm.

This paper is organized as follows, Section 2 offers a brief literature review. Computational details are in Section 3. Experimental setup is in Section 4, followed by the results and analysis in Section 5. Finally, Section 6 presents future work and concludes.

2 RELEVANT WORK

F-Race was initially created to select the best parameter setting (a system) under which a stochastic algorithm performs best. It was later improved to handle a large number of systems in [6], and has been since applied in many fields, for instance: neural network training [7], portfolio selection [8], and bioinformatics [9]. *F-Race* lacks a "search" component; that is, if a better system exists outside S, it cannot find it. [10] proposed a remedy by inserting a new competitor into the race in each iteration, while [3] introduced an iterative version called *I\F-Race* that adds a meta-level estimation of distribution algorithm that iteratively provides the lower-level *F-Race* with new candidates to select from.

In situations where the budget is fixed and there is no advantage of terminating early, *F-Race* was extended to *F-RaceR* that is able to consume any given budget, and adapt the α parameter, see [2]. They argued that *F-RaceR*'s performance is robust to the newly introduced parameter γ, and showed that *F-RaceR* outperforms *F-Race* in terms of *PICS*. Finally, a more comprehensive sensitivity analysis of the parameters of *F-Race* and *I\F-Race* can be found in [11], who observed, among other things, the importance of properly setting the α parameter.

3 COMPUTATIONAL ASPECTS OF THE KW-TEST

Starting with the *ANOVA* test, the null hypothesis is: for k *independent* systems ($k > 2$), all systems represent medians θ of the same population, while the alternative hypothesis is: at least two systems represent medians of two different populations.

$$h_0: \theta_1 = \theta_2 = \cdots = \theta_k$$
$$h_1: Not\ h_0. \tag{1}$$

If h_0 is rejected for some α, there is at least one system that is significantly different than the rest. In that case, follow-up tests should be conducted to identify which. Each observed sample i from system j (y_{ij}) is replaced with a rank relative to all other samples across all systems. A rank of 1 is assigned to the lowest y_{ij} and a rank of $\tau = \sum_{j=1}^{k} n_j$ to the highest y_{ij}, where n_j is to number of samples observed for system *j* so far. Finally, the ranks are placed back into their original positions, and rank averages per system are calculated.

The test statistic approximately follows a Chi-square distribution, with $k-1$ degrees of freedom. If there are no, or few, ties, the test statistic is calculated as

$$H = \frac{12}{\tau(\tau+1)} \sum_{j=1}^{k} \left[\frac{\sum_{i=1}^{n} r(y_{ij})^2}{n_j} \right] - 3(\tau+1). \tag{2}$$

A correction factor may be used in case of excessive ties, it is calculated as

$$C = 1 - \frac{\sum_{i=1}^{\epsilon} (t_i^3 - t_i)}{\tau^3 - \tau}, \tag{3}$$

where ϵ is the number of sets of ties, and t_i is the number of tied scores in the i^{th} set. The corrected test statistic H_c is calculated by dividing (2) by (3).

If the *ANOVA* test is significant, pair-wise comparisons with the current best are required. The one used here is the Bonferroni-Dunn test. It relies on a minimum difference (CD_{KW}) in ranks averages, required for systems i and j to be significantly different at a specified α.

$$CD_{KW} = z_{adj} \sqrt{\frac{\tau(\tau+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, \tag{4}$$

where z_{adj} is obtained from normal distribution tables, and is adjusted such that the overall error rate does not increase above a user specified value. To do so, the Bonferroni correction is applied. Any absolute difference between systems' average ranks $\left|\bar{R}_i - \bar{R}_j\right|$ that is greater than or equal to CD_{KW}, indicates that they come from two populations with unequal medians. This is used as a basis to discard inferior systems from the race.

4 EXPERIMENTAL SETUP

The proposed *KW-RaceR* is compared to *F-RaceR*, and *EA* on the basis of the *PICS* vs. sampling effort curve, that is estimated experimentally over $r = 10\,000$ independent replications. As N is fixed here, it is more informative to look at the area under the curve (*AUC*), rather than the final *PICS* achieved, especially that all methods will eventually reach a zero *PICS* if N is large enough. The fixed experimental settings are: $k = 10, N = 2000, n_0 = 10, \alpha_0 = 0.1, \gamma = 0.5$, with pair-wise correlation level ρ: 0.0, 0.3, 0.6, 0.9.

The performances of stochastic systems is simulated by drawing data from *normal* distributions with the following characteristics: monotonically increasing means with constant variance (*MIM-CV36*) $\sim \mathcal{N}(i, 6^2) \forall i = 0, \ldots, k-1$. Monotonically increasing means with exponentially increasing variance (*MIM-expIV*) $\sim \mathcal{N}(i, (i+1)^4) \forall i = 0, \ldots, k-1$. Monotonically increasing means with exponentially decreasing variance (*MIM-expDV*)

$\sim\mathcal{N}(i,(k-i)^3)\forall i=0,\ldots,k-1$. Means and variances are drawn at random from a uniform distribution (*RAND*) $\sim\mathcal{N}\big(U(0,k),U(10{,}50)\big)$, with exact values used shown in Table 1. Finally, let the best be the system with the *lowest* mean.

Table 1: The exact means and variances drawn from $\sim\mathcal{N}\big(U(0,k),U(10{,}50)\big)$.

Set	System	0	1	2	3	4	5	6	7	8	9
rand1	Mean	0.10	0.98	1.32	3.27	6.21	6.49	8.03	8.34	9.10	9.78
	Variance	35.93	44.34	42.10	24.42	43.39	28.12	44.49	34.35	24.31	39.72
rand2	Mean	0.23	0.50	1.09	1.65	5.51	5.87	7.50	8.31	8.38	9.85
	Variance	39.70	39.40	39.43	29.34	46.50	34.92	35.69	37.39	45.04	40.77

5 RESULTS AND ANALYSIS

Starting with the *AUC* performance measure, it is clear from Table 2 that *KW_RaceR* (labelled III) achieves the best (lowest) values for all cases and correlation levels tested. *EA* performs worse as expected, and *F_RaceR* is somewhere in between. Note that the *AUC* measurement is only an approximation, as the *PICS* curves are not very smooth with 10 000 replications.

Table 2: AUC values for EA (I), F_RaceR (II), and KW_RaceR (III) for all tested case and correlation levels.

	MIMCV36			MIMexpDV			MIMexpIV			rand1			rand2		
Corr.	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
0.0	281	171	**105**	993	988	**910**	1070	1023	**932**	841	676	**601**	429	239	**170**
0.3	203	128	**77**	924	922	**844**	894	877	**815**	770	592	**523**	328	182	**123**
0.6	111	86	**46**	890	851	**747**	887	742	**681**	652	467	**413**	195	126	**71**
0.9	14	9	**8**	830	785	**478**	673	594	**524**	448	387	**170**	32	12	**14**

Before analysing the results further, the following is noted: *F-RaceR* relies on a two-way *ANOVA* model that attributes the variation observed in the samples to two factors: the system, and the instance (one row of observations across all systems). The instance effect is simulated by introducing correlation; hence, *F-RaceR* should be more appropriate for the high correlation case, compared to *KW-RaceR*.

When *F-RaceR* is used with 0.0 correlation, it is expected to consume more samples to try and (mistakenly) assess an instance effect. Thus, a single *F* race should be longer than a *KW* race on average, leading to fewer resets. Both these conjectures are supported by figures 2 and 3 for the independent case, where the $\boldsymbol{\alpha}$-curve for *F-RaceR* is higher, and the survival plots show that *F-RaceR* indeed samples inferior systems more often, at the expense of the more crucial top ones. All of this leads to a slower convergence of the *PICS* curve. Note that the discontinuities in the *PICS* curves correspond to zero values that do not appear on the log scale.

At high correlation *F-RaceR* identifies a winner earlier since the problem is generally easier, compared to 0.0 correlation, and *F-RaceR* quickly assesses the instance effect and utilizes it in discarding inferior systems. Yet, this leads to more resets and causes $\boldsymbol{\alpha}$ to drop to low values early on. Eventually, *F-RaceR* becomes similar to *EA* because it does not make use of all the samples previously gathered, and its performance suffers. *KW-RaceR*, on the other hand, will never become similar to *EA*, regardless of how many times it resets, as it uses full information upon reset. Comparable behaviour was observed for the other cases.

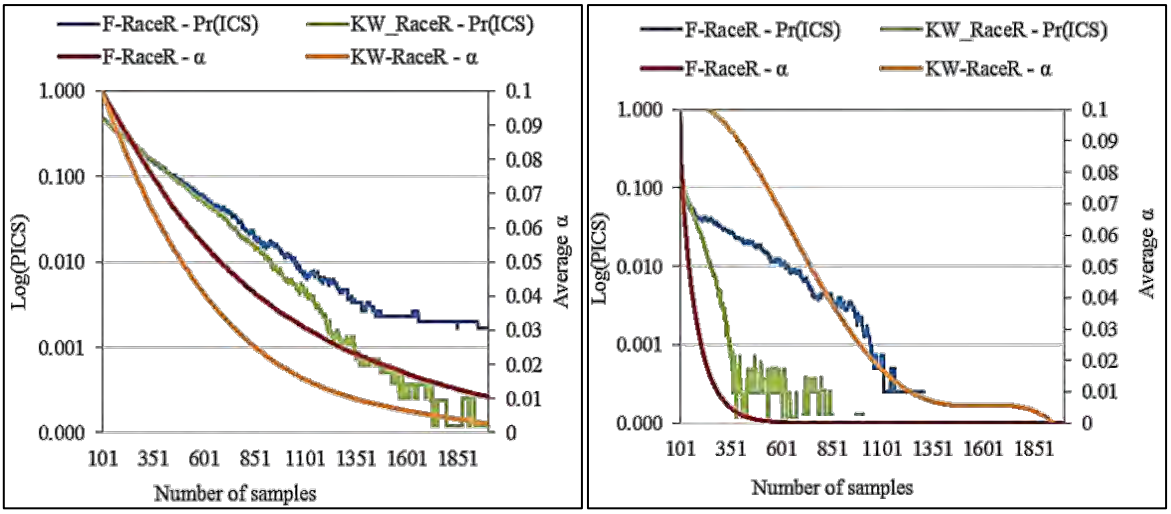


Figure 2: PICS curves for KW-RaceR and F-RaceR at 0.0 (left) and 0.9 (right) correlations for the MIM-CV case. The figure also displays the average nominal α value obtained due to resetting.

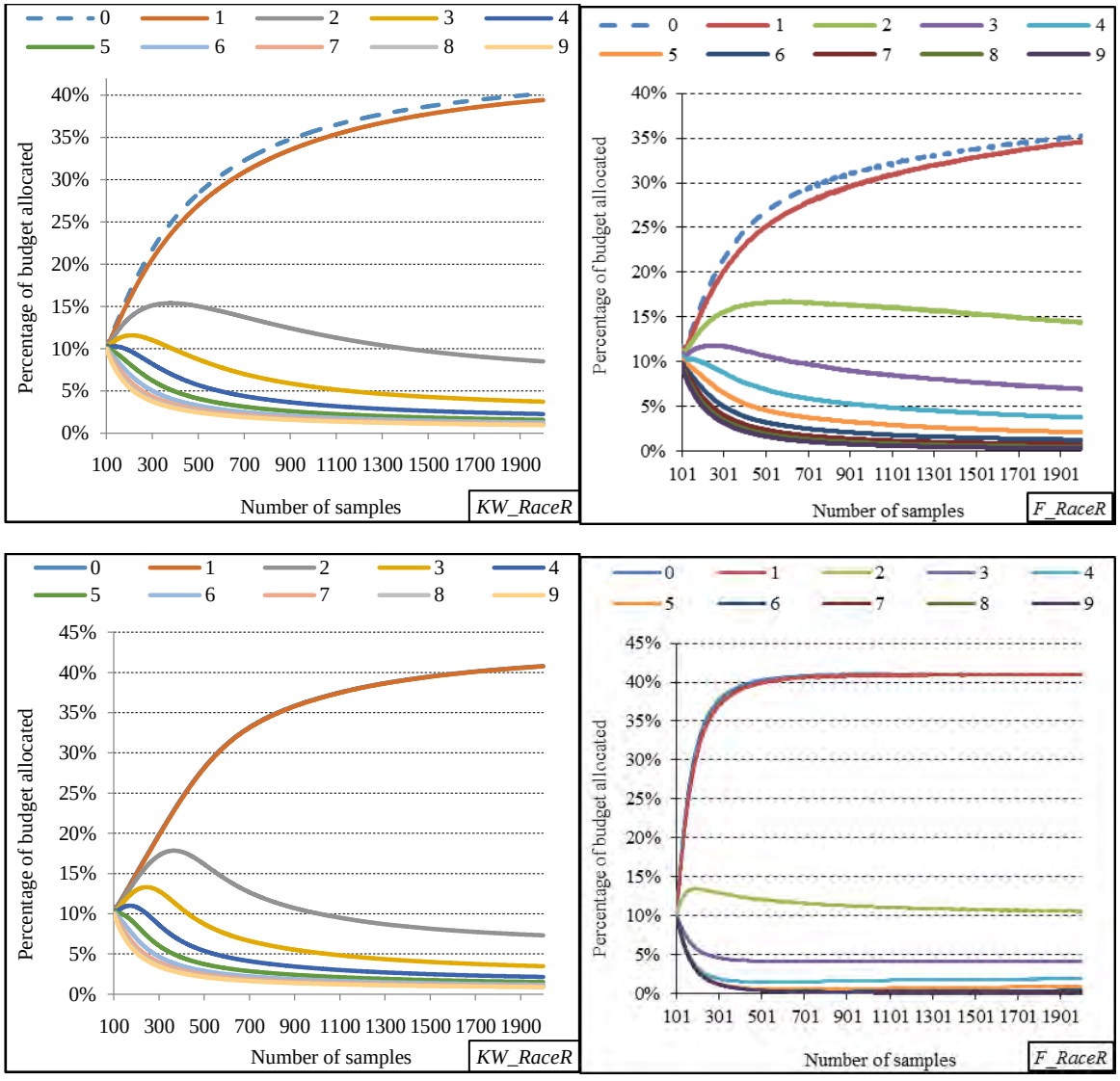


Figure 3: Survival plots for for the MIM-CV case under 0.0 (above) and 0.9 (below) correlation levels.

5 CONCLUSION AND FUTURE WORK

A new racing-with-reset algorithm was presented here for the fixed sampling budget case. It relies on the Kruskal-Wallis one-way *ANOVA* by ranks test, followed by the Bonferroni-Dunn multiple comparisons test. The advantage of *KW-RaceR* is its ability to utilize full information upon reset, compared to *F-RaceR* that uses partial information due to the equal sample size requirement of its statistical tests. Empirical evaluation between these two methods and *EA*, revealed that *KW-RaceR* achieves lower *PICS* values by the end of the run, and on average, as indicated by the *AUC* measure.

Future work entails testing on non-stationary distributions, non-normal data, and creating normal-model based *ANOVA* (parametric) Racing algorithms. It is interesting to test, and study, the following two hypotheses: first, can parametric Racing outperform non-parametric Racing with normal data. Second, can non-parametric Racing outperform parametric Racing with non-normal data. The author is currently investigating these questions.

References

[1] Maron, O. and A. Moore. (1994). Hoeffding races: accelerating model selection search for classification and function approximation. Advances in neural information processing systems. 6(1): pp. 59-66.

[2] Branke, J. and J. Elomari. (2013). Racing with a Fixed Budget and a Self-Adaptive Significance Level. In G. Nicosia and P. Pardalos, (Eds.). Learning and Intelligent Optimization (pp. 272-280). Berline: Springer.

[3] Birattari, M., Z. Yuan, P. Balaprakash, and T. Stützle. (2010). F-Race and Iterated F-Race: An overview. In T. Bartz-Beielstein, M. Chiarandini, L. Paquete, and M. Preuss, (Eds.). Experimental Methods for the Analysis of Optimization Algorithms (pp. 311-336). Berlin: Springer.

[4] Kruskal, W.H. and W.A. Wallis. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association. 47(260): pp. 583-621.

[5] Dunn, O.J. (1961). Multiple comparisons among means. Journal of the American Statistical Association. 56(293): pp. 52-64.

[6] Balaprakash, P., M. Birattari, and T. Stützle. (2007). Improvement Strategies for the F-Race Algorithm: Sampling Design and Iterative Refinement. In T. Bartz-Beielstein, M. Blesa Aguilera, C. Blum, B. Naujoks, A. Roli, G. Rudolph, and M. Sampels, (Eds.). Hybrid Metaheuristics (pp. 108-122). Berline: Springer.

[7] Blum, C. and K. Socha. (2005). Training feed-forward neural networks with ant colony optimization: an application to pattern classification. In Hybrid Intelligent Systems. Rio de Janeiro 6-9 Nov. Rio de Janeiro: IEEE.

[8] Gaspero, L., G. Tollo, A. Roli, and A. Schaerf. (2007). Hybrid Local Search For Constrained Financial Portfolio Selection Problems. In P. Hentenryck and L. Wolsey, (Eds.). Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (pp. 44-58). Berlin: Springer.

[9] Di Gaspero, L. and A. Roli. (2008). Stochastic local search for large-scale instances of the haplotype inference problem by pure parsimony. Journal of Algorithms. 63(3): pp. 55-69.

[10] Chiarandini, M., M. Birattari, K. Socha, and O. Rossi-Doria. (2006). An effective hybrid algorithm for university course timetabling. Journal of Scheduling. 9(5): pp. 403-432.

[11] Pérez Cáceres, L., M. López-Ibáñez, and T. Stützle. (2014). An Analysis of Parameters of I-Race. In C. Blum and G. Ochoa, (Eds.). Evolutionary Computation in Combinatorial Optimisation (pp. 37-48). Berlin: Springer.

OPTIMAL AND NEAR-OPTIMAL STRATEGIES IN SOLVING BI-CRITERIA DISCRETE STOCHASTIC DYNAMIC PROGRAMMING PROBLEM

Maciej Nowak and Tadeusz Trzaskalik
University of Economics in Katowice, Department of Operations Research
ul. 1 Maja 50, 40-287 Katowice, Poland
{maciej.nowak, tadeusz.trzaskalik}@ue.katowice.pl

Abstract: In the paper a multiperiod bi-criteria discrete decision making process under risk is considered. We apply a discrete, stochastic dynamic programming approach based on Bellman's principle of optimality. The decision maker defines a priority criterion and aspiration levels for both criteria. The final solution is identified interactively. A modification of the aspiration levels during the procedure makes it possible to find a satisfactory solution. The method is illustrated by a numerical example in which expected profit and probability of success are considered as criteria. The notes on applicability of the procedure are included in conclusions.

Keywords: multiobjective dynamic programming, decision making under risk, quasi-hierarchical approach, interactive methods.

1 INTRODUCTION

Many decision making problems are dynamic, which means that not a single decision, but a series of interdependent decisions must be made at different times in order to achieve an overall goal. As consequences of these decisions become apparent in the near or remote future, their effects are unknown. If it is possible to estimate probability distributions of future outcomes, we deal with a dynamic stochastic decision making problem.

In this paper, we analyse a bi-criteria dynamic stochastic decision making problem, which can be described as follows:

1. The decision process consists of T periods. At each period, a decision must be made. Any decision made at period t determines the characteristics of the problem at period $t + 1$.
2. Risk is taken into account: the consequences of the decision made at period t are uncertain.
3. Two conflicting criteria are considered.
4. The aim is to identify a strategy under which satisfactory values of both criteria are obtained.

Our method is based on Bellman's principle of optimality [1]. We use this approach to identify optimal solutions with respect to each criterion separately. As these solutions are usually different, we propose to use a quasi-hierarchical approach. We assume that the decision maker is able to define a hierarchy of criteria and to determine the extent to which the optimal value of a higher-priority criterion can be made worse in order to improve the value of lower-priority criterion. To solve the problem we must identify near-optimal solutions for both criteria. The algorithm that we use is based on the observation made previously in [3]. In order to identify the strategy which is worse than the optimal strategy only, we must consider all the strategies that differ from the optimal strategy in one of the feasible states and to select best of them. This rule can be also applied to find the next strategies in the ranking.

In dynamic programming it is not possible to identify solutions satisfying all requirements directly formulated by the decision maker. Instead, we look for solutions satisfying each requirement separately. Starting from the strategy that is optimal with respect to the more important criteria, we analyse the value of the second criterion and if it is satisfactory, we end

the procedure. Otherwise, we look for the strategy which is next in the ranking with respect to the more important criteria. The procedure is finished when a strategy satisfactory with respect to both criteria is identified.

In stochastic dynamic decision making problems the expected value is usually used to evaluate feasible solutions. However, decision makers are usually also interested in measures. In the example presented below, we consider two objectives: maximization of expected profit and maximization of probability of success, which is defined as reaching one of predefined final states.

2 SINGLE-CRITERION STOCHASTIC DYNAMIC PROGRAMMING

We will use the following notation [5, 6]:
T – number of stages of the decision process under consideration,
y_t – state of the process at the beginning of stage t ($t = 1,\ldots,T$),
$\mathbf{Y}_t$ – finite set of process states at stage t,
x_t – feasible decision at stage t,
$\mathbf{X}_t(y_t)$ – finite set of decisions feasible at stage t, when the process is in state $y_t \in \mathbf{Y}_t$ at the beginning of this stage,
$F_t(y_{t+1} \mid y_t, x_t)$ – value of stage criterion at stage t for the transition from state y_t to state y_{t+1}, when the decision taken was $x_t \in \mathbf{X}_t(y_t)$,
$P_t(y_{t+1} \mid y_t, x_t)$ – probability of the transition at stage t from state y_t to state y_{t+1}, when the decision taken was $x_t \in \mathbf{X}_t(y_t)$.
$P(y_1)$ – probability of distribution in the set of initial stages $y_1 \in \mathbf{Y}_1$.
$\{x\}$ – strategy – a function assigning to each state $y_t \in \mathbf{Y}_t$ exactly one decision $x_t \in \mathbf{X}_t(y_t)$,
$\{\mathbf{X}\}$ – the set of all strategies of the process under consideration,
$\{x_{\overline{t,T}}\}$ – shortened strategy, encompassing stages from t to T.

The following equation holds:

$$\forall_{t\in\overline{1,T}} \forall_{y_t\in\mathbf{Y}_t} \forall_{x_t\in\mathbf{X}_t(y_y)} \sum_{y_{t+1}\in\mathbf{Y}_{t+1}} P_t(y_{t+1} \mid y_t, x_t) = 1 \tag{1}$$

Let us assume that we have selected a certain strategy $\{\bar{x}\} \in \{\mathbf{X}\}$. The expected value of the process realisation for the shortened strategy $\{\bar{x}_{\overline{T,T}}\}$ is calculated as follows:

$$G_T(y_T, \{\bar{x}_{\overline{T,T}}\}) = \sum_{y_{T+1}\in\mathbf{Y}_{T+1}}' F_T(y_{T+1} \mid y_T, \bar{x}_T) P_T(y_{T+1} \mid y_T, \bar{x}_T) \tag{2}$$

The expected value of the realisation for the shortened strategy $\{\bar{x}_{\overline{t,T}}\}$ when at the beginning of this stage the process was in state $y_t \in \mathbf{Y}_t$ is found from the formula:

$$G_t(y_t, \{\bar{x}_{\overline{t,T}}\}) = \sum_{y_{t+1}\in\mathbf{Y}_{t+1}} (F_t(y_{t+1} \mid y_t, \bar{x}_t) + G_{t+1}(y_{t+1}, \{\bar{x}_{\overline{t+1,T}}\})) P_t(y_{t+1} \mid y_t, \bar{x}_t) \tag{3}$$

The expected value of the process realisation for the fixed strategy $\{\bar{x}\}$ is calculated from the formula:

$$G\{\bar{x}\} = \sum_{y_1\in\mathbf{Y}_1}' G_1(y_1, \{\bar{x}\}) P_1(y_1) \tag{4}$$

Using Bellman's optimality principle [1], we determine the optimal strategy.

Algorithm 1

1. For each state $y_T \in \mathbf{Y}_T$ we calculate the optimal values:

$$G_T^*(y_T) = \max_{x_T \in \mathbf{X}_t(y_t)} \sum_{y_{T+1} \in \mathbf{Y}_{T+1}} F_T(y_{T+1} \mid y_T, x_T) P_T(y_{T+1} \mid y_T, x_T) \tag{5}$$

   and find the decision $x_t^*(y_t)$, for which this maximum is attained. This decision forms a part of the optimal strategy being constructed.
2. For stage t, $t \in \overline{T-1,1}$ and each state $y_t \in \mathbf{Y}_t$ we calculate the optimal values

$$G_t^*(y) = \max_{x_t \in \mathbf{X}_t(y_t)} \sum_{y_{t+1} \in \mathbf{Y}_{t+1}} (F_t(y_{t+1} \mid y_t, x_t) + G_{t+1}^*(y_{t+1})) P_t(y_{t+1} \mid y_t, x_t) \tag{6}$$

   and find the decision $x_t^*(y_t)$, for which this maximum is attained. This decision forms a part of the optimal strategy being constructed.
3. The optimal expected value of the process realisation is calculated from the formula:

$$G\{x^*\} = \sum_{y_1 \in \mathbf{Y}_1} G_1^*(y_1, \{x^*\}) P_1(y_1) \tag{7}$$

3 DETERMINATION OF NEAR-OPTIMAL STRATEGIES

Let Z be a threshold given by the decision maker and

$$\alpha = G\{x^*\} - Z \tag{8}$$

The strategy $\{x^m\}$ is called near-optimal if the expected value of its realisation differs from the expected value of the realisation of the optimal strategy $\{x^*\}$ by at most the given value α, that is:

$$G\{x^*\} - G\{x^m\} \le \alpha \tag{9}$$

By **LS**(Z) we denote the set of strategies for which the objective function is at least Z:

$$\mathbf{LS}(Z) = \{\{x\} \in \{\mathbf{X}\} : G\{x\} \ge Z\} \tag{10}$$

The procedure for identifying **LS**(Z) was proposed in [4]. It can be described as follows.

Algorithm 2

The algorithm modifies the strategies by changing a decision in one state only. For each new strategy we check if it generates a solution different from the ones determined previously. If so, we calculate the expected value of the given criterion and check if it satisfies the condition formulated by the decision maker. If this is not the case, such a strategy does not have to be further analysed, since its further modification cannot lead to an improvement of the value of the criterion. The procedure ends when all the possibilities are exhausted.

4 A BI-CRITERIA APPROACH

In this section we present an application of the quasi-hierarchical approach in a bi-criteria case.

Algorithm 3

1. Determine the optimal solutions of the problem with respect to criterion 1 and criterion 2, applying Algorithm 1.
2. Present the optimal values of both criteria to the decision maker.

3. Ask the decision maker to determine the aspiration thresholds Z_1 and Z_2, that is the values which should be attained by each criterion in the final solution.
4. Determine the sets $\mathbf{LS}_1(Z_1)$ and $\mathbf{LS}_2(Z_2)$ of strategies satisfying the requirements determined by the decision maker, applying Algorithm 2.
5. Determine the set $\mathbf{LS}(Z_1, Z_2)$ which is the intersection of the sets $\mathbf{LS}_1(Z_1)$ and $\mathbf{LS}_2(Z_2)$:

$$\mathbf{LS}(Z_1, Z_2) = \mathbf{LS}_1(Z_1) \cap \mathbf{LS}_2(Z_2) \tag{11}$$

6. If $\mathbf{LS}(Z_1, Z_2) = \varnothing$, go to 8.
7. Ask the DM to analyse solutions from $\mathbf{LS}(Z_1, Z_2)$ and choose the best one. Go to 9.
8. Ask the DM to choose the new values Z_1 and Z_2. Go to 4.
9. End of the procedure.

5 NUMERICAL EXAMPLE

We consider a two-stage decision process. The sets of states for the consecutive stages are:

$$\mathbf{Y}_1 = \{1, 2, 3\} \qquad \mathbf{Y}_2 = \{4,5, 6\} \qquad \mathbf{Y}_3 = \{7, 8, 9\}$$

The sets of feasible decisions are:

$$\mathbf{X}_1(1) = \{A, B\} \qquad \mathbf{X}_1(2) = \{C, D\} \qquad \mathbf{X}_1(3) = \{E, F\}$$
$$\mathbf{X}_2(4) = \{G, H\} \qquad \mathbf{X}_2(5) = \{I, J\} \qquad \mathbf{X}_2(6) = \{K, L\}$$

The possible stage realisations of the process, probabilities of their occurrence, as well as the values of the stage criteria functions, are presented in Table 1.

Table 1: Stage realisations of the process

Stage	$(y_{t+1} \mid y_t, x_t)$	P(·)	$F^1(\,)$	$F^2(\cdot)$	Stage	$(y_{t+1} \mid y_t, x_t)$	P(·)	$F^1(\,)$	$F^2(\cdot)$
1	(4| 1,A)	0.4	1	20	2	(7| 4,G)	0.5	0	18
1	(5| 1,A)	0.6	1	17	2	(8| 4,G)	0.5	0	12
1	(4| 1,B)	0.7	1	15	2	(7| 4,H)	0.3	0	15
1	(6| 1,B)	0.3	1	13	2	(9| 4,H)	0.7	1	9
1	(4| 2,C)	0.6	1	11	2	(7| 5,I)	0.2	0	11
1	(5| 2,C)	0.4	1	14	2	(8| 5,I)	0.8	0	14
1	(5| 2,D)	0.2	1	20	2	(8| 5,J)	0.9	0	20
1	(6| 2,D)	0.8	1	12	2	(9| 5,J)	0.1	1	11
1	(5| 3,E)	0.9	1	16	2	(8| 6,K)	0.2	0	12
1	(6| 3,E)	0.1	1	13	2	(9| 6,K)	0.8	1	14
1	(5| 3,F)	0.5	1	15	2	(7| 6,L)	0.9	0	15
1	(6| 3,F)	0.5	1	15	2	(9| 6,L)	0.1	1	10

The probability distribution in the set of initial states is:

$$P(1) = 0.1, \qquad P(2) = 0.3 \qquad P(3) = 0.6$$

Thanks to the small size of this illustrative problem, the existing strategies can be written down and numbered from 1 to 64. This numbering is presented in Table 2.

Table 2: The set of strategies

No	Decisions	No	Decisions	No	Decisions	No	Decisions
1	(A,C,E,G.I,K)	17	(A,D,E,G.I,K)	33	(B,C,E,G.I,K)	49	(B,D,E,G.I,K)
2	(A,C,E,G,I,L)	18	(A,D,E,G.I,L)	34	(B,C,E,G.I,K)	50	(B,D,E,G.I,L)
3	(A,C,E,G.J,K)	19	(A,D,E,G.J,K)	35	(B,C,E,G.J,K)	51	(B,D,E,G.J,K)
4	(A,C,E,G.J,L)	20	(A,D,E,G.J,L)	36	(B,C,E,G.J,L)	52	(B,D,E,G.J,L)
5	(A,C,E,H.I,K)	21	(A,D,E,H.I,K)	37	(B,C,E,H.I,K)	53	(B,D,E,H.I,K)
6	(A,C,E,H.I,L)	22	(A,D,E,H.I,L)	38	(B,C,E,H.I,L)	54	(B,D,E,H.I,L)
7	(A,C,E,H.J,K)	23	(A,D,E,H.J,K)	39	(B,C,E,H.J,K)	55	(B,D,E,H.J,K)
8	(A,C,E,H.J,L)	24	(A,D,E,H.J,L)	40	(B,C,E,H.J,L)	56	(B,D,E,H.J,L)
9	(A,C,F,G.I,K)	25	(A,D,F,G.I,K)	41	(B,C,F,G.I,K)	57	(B,D,F,G.I,K)
10	(A,C,F,G.I,L)	26	(A,D,F,G.I,L)	42	(B,C,F,G.I,L)	58	(B,D,F,G.I,L)
11	(A,C,F,G.J,K)	27	(A,D,F,G.J,K)	43	(B,C,F,G.J,K)	59	(B,D,F,G.J,K)
12	(A,C,F,G.J,L)	28	(A,D,F,G.J,L)	44	(B,C,F,G.J,L)	60	(B,D,F,G.J,L)
13	(A,C,F,H.I,K)	29	(A,D,F,H.I,K)	45	(B,C,F,H.I,K)	61	(B,D,F,H.I,K)
14	(A,C,F,H.I,L)	30	(A,D,F,H.I,L)	46	(B,C,F,H.I,L)	62	(B,D,F,H.I,L)
15	(A,C,F,H.J,K)	31	(A,D,F,H.J,K)	47	(B,C,F,H.J,K)	63	(B,D,F,H.J,K)
16	(A,C,F,H.J,L)	32	(A,D,F,H.J,L)	48	(B,C,F,H.J,L)	64	(B,D,F,H.J,L)

We consider a bi-criteria process. The aim is to maximize the probability of completing the process in state 9, pointed by the DM (criterion 1). The second criterion is to maximize the expected profit (criterion 2). We will illustrate algorithm 3.

Step 1.

We determine the optimal solutions of the problem with respect to criterion 1 and criterion 2, applying Algorithm 1. W obtain: $G^{1*}\{x^8\} = 0.772$, $G^{2*}\{x^3\} = 28.316$.

Step 2.

We present the optimal values of both criteria to the decision maker.

Step 3.

We ask the decision maker to determine the aspiration thresholds. Let us assume, that the DM accepted a 2% decrease of the value of the more important criterion 1 and a 2% decrease of the value of the criterion 2. The threshold value are as follows: $Z_1 = 0.756$, $Z_2 = 27.750$.

Step 4.

We determine the sets $\mathbf{LS}_1(Z_1)$ and $\mathbf{LS}_2(Z_2)$, by applying Algorithm 2. W obtain:

$\mathbf{LS}_1(0.756) = \{\{x^8\}, \{x^{24}\}, \{x^{40}\}, \{x^{56}\}\}$

$\mathbf{LS}_2(27{,}750) = \{\{x^3\}, \{x^4\}, \{x^{19}\}, \{x^{23}\}, \{x^{35}\}, \{x^{51}\}, \{x^{36}\}, \{x^{11}\}, \{x^{20}\}, \{x^{27}\}\}$

Step 5.

We determine the set $\mathbf{LS}(0.756; 27.750) = \mathbf{LS}_1(0.756) \cap \mathbf{LS}_2(27.750) = \varnothing$

Step 6.

As $\mathbf{LS}(0.756; 27.750) = \varnothing$, we go to 8.

Step 8.

We ask the DM to choose new values Z_1 and Z_2. Let us assume that the DM accepted a 3% decrease of the value of the less important criterion 2. The threshold value for the criterion 1 is now equal to 27.467.

Step 4.

We determine the set $LS_2(27.467)$ by applying Algorithm 2. We obtain:

$\mathbf{LS}_2(27.467) = \mathbf{LS}_2(27.750) \cup \{\{x^{24}\}, \{x^{31}\}, \{x^{43}\}, \{x^{55}\}, \{x^{59}\}, \{x^{52}\}\}$

Step 5.

We determine the set $\mathbf{LS}(0.756; 27.467) = \mathbf{LS}_1(0.756) \cap \mathbf{LS}_2(27.467) = \varnothing$

Step 6.

As $\mathbf{LS}(0.756; 27.467) = \varnothing$, we go to 8.

Step 8.

We ask the DM to choose new values Z_1 and Z_2. Let us assume that the DM accepted a 5% decrease of the value of the criterion 1, and a 3% decrease of the value for the criterion 2. The threshold values are now: $Z_1 = 0.733$, $Z_2 = 27.467$.

Step 4.

We determine the set $\mathbf{LS}_1(0.733)$ by applying Algorithm 2. We obtain:

$$\mathbf{LS}_1(0.733) = \mathbf{LS}_1(0.756) \cup \{\{x^7\}, \{x^{16}\}, \{x^{32}\}, \{x^{48}\}, \{x^{39}\}, \{x^{20}\}, \{x^{64}\}\}$$

Step 5.

We determine the set $\mathbf{LS}(0.756; 27.467) = \mathbf{LS}_1(0.756) \cap \mathbf{LS}_2(27.467) = \{\{x^{20}\}\}$

Step 6.

As $\mathbf{LS}(0.756; 27.467) \neq \emptyset$, we ask the DM to analyse the obtained solution. We have $G^{1*}\{x^{20}\} = 0.738$ and $G^{2*}\{x^{20}\} = 27.818$, so $\{x^{20}\} \in \mathbf{LS}(0.738; 27.818)$. The values of the criteria for the final solution differ from the optimal values by 4.4% and 1.76%, respectively.

6 CONCLUSIONS

Numerous decision problems are dynamic by nature. In real-world situations the decision process can rarely be formulated in terms of a single choice. Often, a series of interdependent decisions must be made at different times in order to achieve an overall goal. Moreover, as the future is unknown, the results of these decisions are usually unknown. Stochastic discrete muliobjective dynamic programming is an efficient tool to model and to evaluate such processes.

In the paper a new procedure for a bi-criteria problem is presented. It is based on Bellman's principal of optimality and on the quasi-hierarchical approach. The final solution is found by an interactive procedure.

Our procedure can be applied in various fields. One example of the application of stochastic dynamic programming is presented in [2]. The problem consists in identifying the best strategy for tender preparation. A company considers entering a new market. It can operate as a general contractor or cooperate with a local company. The final decision must specify whether the company should make efforts to find a partner, and what role it should eventually play. In such a case the probability of success is a very important issue which has to be considered together with the profit maximization criterion.

Acknowledgement

This research was supported by National Science Centre, decision no DEC-2013/11/B/HS4/01471.

References

[1] Bellman, R. (1957). Dynamic Programming. Princeton University Press.

[2] Nowak, B., Nowak, M. and Trzaskalik, T. (2012). Decision aiding in engineering project planning using multiobjective dynamic programming and simulation. Journal of Global Business Administration, 4(1): 1-17.

[3] Nowak, M. (2015). Quasi-hierarchical approach in multiobjective decision tree. University of Economics in Katowice Research Papers (in press, in Polish).

[4] Nowak, M. and Trzaskalik, T. Quasi-hierarchical approach to discrete multiobjective stochastic dynamic programming. (submitted to ISME 2015 Symposium).

[5] Trzaskalik T (1998) Multiobjective Analysis in Dynamic Environmnent. The Karol Adamiecki University of Economics in Katowice Press, Katowice, 1998.

[6] Trzaskalik T. (1990). Multiobjective Discrete Dynamic Programming. Theory and Applications in Economics. The Karol Adamiecki University of Economics in Katowice Press, Katowice, 1990 (in Polish)

ON THE OPTIMIZATION APPROACH TO POLYMATRIX GAMES

Andrei Orlov
Matrosov Institute for System Dynamics & Control Theory SB RAS,
Lermontov str., 134, Irkutsk, 664033, Russia

Abstract: The problem of numerical finding of a Nash equilibrium in a polymatrix game is considered. Such a game turns out to be equivalent to the solving a nonconvex optimization problem with a bilinear structure in the objective function. Special methods of local and global search for the 3-player game are proposed and investigated.

Keywords: polymatrix games, Nash equilibrium, nonconvex optimization problems, local search, global search, computational simulation

1 INTRODUCTION

It is well known that the problem of numerical finding of equilibrium points in the Game Theory [4, 8] is one of the pivotal issues for the contemporary mathematical optimization theory and methods [9]. The simplest problem of the Game Theory is the classical matrix game that can be reduced to two dual linear programming (LP) problems [4, 8]. So, one can say that the matrix games are Convex Structures, and there are no fundamental difficulties for their solution. The first extension of a matrix game is the non-antagonistic bimatrix game. Such a game already represents Nonconvex (bilinear) Structure [4, 7, 8]. For seeking equilibria in bimatrix games, Lemke-Howson type methods are very popular [3]. On the other hand, the nonconvex optimization approach to numerical finding of Nash equilibrium in bimatrix games, elaborated in our group, turns out to be effective for large-scale problems [7, 14, 16].

In this work, we extend our approach to polymatrix games [1, 4, 12]. For example, a 3-player polymatrix game can be completely described by six matrices, therefore, we will call it the hexamatrix game. New methods of local and global search in hexamatrix games are proposed and its convergence is investigated. These methods are based on the equivalence theorem for the game and a special mathematical optimization problem with a bilinear structure in the objective function [12]. This special problem is solved by the Global Search Theory for nonconvex problems with (d.c.) functions of A.D. Alexandrov [10, 11, 13].

2 PROBLEM FORMULATION AND THE REDUCTION THEOREM

Consider the following hexamatrix game with mixed strategies:

$$\left.\begin{array}{c} F_1(x,y,z) \triangleq \langle x, A_1y + A_2z\rangle \uparrow \max\limits_{x},\ x \in S_m,\quad F_2(x,y,z) \triangleq \langle y, B_1x + B_2z\rangle \uparrow \max\limits_{y},\ y \in S_n, \\ F_3(x,y,z) \triangleq \langle z, C_1x + C_2y\rangle \uparrow \max\limits_{z},\ z \in S_l, \end{array}\right\}$$

where $S_p = \{u = (u_1, \ldots, u_p)^T \in \mathbb{R}^p \mid u_i \geq 0,\ \sum\limits_{i=1}^{p} u_i = 1\}$, $p = m, n, l$, and the symbol "$\triangleq$" means "equals by definition".

Definition 2.1 The triple $(x^*, y^*, z^*) \in S_m \times S_n \times S_l$ satisfying the inequalities

$$\left.\begin{array}{ll} v_1^* = v_1(x^*, y^*, z^*) \triangleq F_1(x^*, y^*, z^*) \geq F_1(x, y^*, z^*) & \forall x \in S_m, \\ v_2^* = v_2(x^*, y^*, z^*) \triangleq F_2(x^*, y^*, z^*) \geq F_2(x^*, y, z^*) & \forall y \in S_n, \\ v_3^* = v_3(x^*, y^*, z^*) \triangleq F_3(x^*, y^*, z^*) \geq F_3(x^*, y^*, z) & \forall z \in S_l, \end{array}\right\}$$

will be henceforth called a *Nash equilibrium point* in the game $\Gamma_3 = \Gamma(A, B, C)$ ($A = (A_1, A_2)$, $B = (B_1, B_2)$, $C = (C_1, C_2)$). Herewith, the strategies x^*, y^*, and z^* will be called the *equilibrium strategies*. The numbers v_1^*, v_2^*, and v_3^* will be called the *payoffs of players* 1, 2, and 3, respectively, at the equilibrium point (x^*, y^*, z^*). Denote the *set of all Nash equilibrium points* of the game $\Gamma_3 = \Gamma(A, B, C)$ by $NE = NE(\Gamma_3) = NE(\Gamma(A, B, C))$. □

It is well known that in the case of the game $\Gamma_3 = \Gamma(A, B, C)$ due to Nash's Theorem [12], there exists a Nash equilibrium point in mixed strategies.

Further consider the following optimization problem ($\sigma \triangleq (x, y, z, \alpha, \beta, \gamma)$):

$$\left.\begin{array}{c} \Phi(\sigma) \triangleq \langle x, A_1 y + A_2 z\rangle + \langle y, B_1 x + B_2 z\rangle + \langle z, C_1 x + C_2 y\rangle - \alpha - \beta - \gamma \uparrow \max\limits_{\sigma}, \\ \sigma \in D \triangleq \{(x, y, z, \alpha, \beta, \gamma) \in \mathbb{R}^{m+n+l+3} \mid x \in S_m, \;\; y \in S_n, \;\; z \in S_l, \\ A_1 y + A_2 z \le \alpha e_m, \quad B_1 x + B_2 z \le \beta e_n, \quad C_1 x + C_2 y \le \gamma e_l\}, \end{array}\right\} \qquad (\mathcal{P})$$

where $e_p = (1, 1, ..., 1) \in \mathbb{R}^p, p = m, n, l$.

Theorem 2.2 [12] *A point (x^*, y^*, z^*) is a Nash equilibrium point in the hexamatrix game $\Gamma(A, B, C) = \Gamma_3$ if and only if it is a part of a global solution $\sigma_* \triangleq (x^*, y^*, z^*, \alpha_*, \beta_*, \gamma_*) \in \mathbb{R}^{m+n+l+3}$ of Problem $(\mathcal{P})$. At the same time, the numbers α_*, β_*, and γ_* are the payoffs of the first, the second, and the third players, respectively, in the game Γ_3: $\alpha_* = v_1(x^*, y^*, z^*)$, $\beta_* = v_2(x^*, y^*, z^*)$, $\gamma_* = v_3(x^*, y^*, z^*)$. In addition, an optimal value $\mathcal{V}(\mathcal{P})$ of Problem $(\mathcal{P})$ is equal to zero:*

$$\mathcal{V}(\mathcal{P}) = \Phi(x^*, y^*, z^*, \alpha_*, \beta_*, \gamma_*) = 0. \qquad (1)$$

Corollary 2.3 *Suppose, $(x^*, y^*, z^*) \in NE(\Gamma(A, B, C))$ with the payoffs α_*, β_*, and γ_*. Then $\alpha_* = \max\limits_i (A_1 y^* + A_2 z^*)_i$, $\beta_* = \max\limits_j (B_1 x^* + B_2 z^*)_j$, $\gamma_* = \max\limits_t (C_1 x^* + C_2 y^*)_t$.*

Thus, one can conclude that the seeking for a Nash equilibrium can be carried out by solving Problem $(\mathcal{P})$. However, we have to note that at present a numerical solution of the nonconvex Problem $(\mathcal{P})$ seems to be rather difficult [2, 5], because the classical methods of convex optimization (see e.g. [5]) are not able to provide with a global solution to nonconvex problems, and they are not capable of escaping a local optimum.

In this work, in order to solve Problem $(\mathcal{P})$, we will use an approach based on Global Search Theory [10, 11, 13]. According to this theory the Global Search consists of two principal stages: 1) a local search, which takes into account the structure of the problem under scrutiny; 2) the procedures based on Global Optimality Conditions (GOC) [10, 11], which allow to improve the point provided by the local search method, in other words, to escape a local pit.

This methodology turned out to be rather effective and promising in numerical solution of several actual nonconvex problems of Operations Research [6, 7, 14, 15, 16].

3 LOCAL SEARCH

To implement a local search in Problem $(\mathcal{P})$, let us apply the ideas, first, of splitting variables in several groups, and, after that, of consecutive solving of specially constructed LP problems with respect to the groups of variables. These ideas have previously demonstrated its efficiency in bimatrix games [7, 14, 16], bilinear programming problems [6, 14], and bilevel problems [15].

In order to do it, consider the following LP problems:

$$\left.\begin{array}{c} f_1(x, \beta) \triangleq \langle x, (A_1 + B_1^T) v + (A_2 + C_1^T) w\rangle - \beta \uparrow \max\limits_{(x,\beta)}, \\ (x, \beta) \in X(v, w, \bar\gamma) \triangleq \{(x, \beta) \mid x \in S_m, \\ B_1 x - \beta e_n \le -B_2 w, \; C_1 x \le \bar\gamma e_l - C_2 v\}; \end{array}\right\} \qquad (\mathcal{LP}_x(v, w, \bar\gamma))$$

$$\left.\begin{array}{c} f_2(y,\gamma) \triangleq \langle y, (B_1 + A_1^T)u + (B_2 + C_2^T)w\rangle - \gamma \uparrow \max\limits_{(y,\gamma)}, \\ (y,\gamma) \in Y(u,w,\bar{\alpha}) \triangleq \{(y,\gamma) \mid y \in S_n, \\ A_1 y \leq \bar{\alpha} e_m - A_2 w,\ C_2 y - \gamma e_l \leq -C_1 u\}; \end{array}\right\} \qquad (\mathcal{LP}_y(u,w,\bar{\alpha}))$$

$$\left.\begin{array}{c} f_3(z,\alpha) \triangleq \langle z, (C_1 + A_2^T)u + (C_2 + B_2^T)v\rangle - \alpha \uparrow \max\limits_{(z,\alpha)}, \\ (z,\alpha) \in Z(u,v,\bar{\beta}) \triangleq \{(z,\alpha) \mid z \in S_l, \\ A_2 z - \alpha e_m \leq -A_1 v,\ B_2 z \leq \bar{\beta} e_n - B_1 u\}. \end{array}\right\} \qquad (\mathcal{LP}_z(u,v,\bar{\beta}))$$

Here $(u, v, w, \bar{\alpha}, \bar{\beta}, \bar{\gamma}) \in D$ is a feasible point in Problem $(\mathcal{P})$.

For the sake of simplicity of denotations, let us introduce the following functions:

$$\varphi(v,w,\bar{\gamma}) \triangleq \sup\limits_{(x,\beta)} \{f_1(x,\beta) \mid (x,\beta) \in X(v,w,\bar{\gamma})\}, \tag{2}$$

$$\psi(u,w,\bar{\alpha}) \triangleq \sup\limits_{(y,\gamma)} \{f_2(y,\gamma) \mid (y,\gamma) \in Y(u,w,\bar{\alpha})\}, \tag{3}$$

$$\chi(u,v,\bar{\beta}) \triangleq \sup\limits_{(z,\alpha)} \{f_3(z,\alpha) \mid (z,\alpha) \in Z(u,v,\bar{\beta})\}. \tag{4}$$

Let $(x^0, y^0, z^0, \alpha_0, \beta_0, \gamma_0) \in D$ be a starting point. For example, one can use the barycenters of standard simplexes as follows:

$$\begin{array}{c} x_i^0 = \dfrac{1}{m},\ i = 1, ..., m; \quad y_j^0 = \dfrac{1}{n},\ j = 1, ..., n; \quad z_t^0 = \dfrac{1}{l},\ t = 1, ..., l; \\ \alpha_0 = \max\limits_i (A_1 y^0 + A_2 z^0)_i; \beta_0 = \max\limits_j (B_1 x^0 + B_2 z^0)_j; \gamma_0 = \max\limits_t (C_1 x^0 + C_2 y^0)_t. \end{array} \tag{5}$$

Denote: $\Phi_s \triangleq \Phi(\sigma_s)$. Now let us describe the local search algorithm.

YZ_γ-procedure

Step 0. Set $s := 1$, $y^s := y^0$, $z^s := z^0$, $\gamma_s := \gamma_0$.

Step 1. Using an LP technique, find a $\rho_s/3$-solution (x^{s+1}, β_{s+1}) to the Problem $(\mathcal{LP}_x(y^s, z^s, \gamma_s))$, so that the following inequality holds:

$$\begin{array}{c} f_1(x^{s+1}, \beta_{s+1}) + \rho_s/3 = \\ = \langle x^{s+1}, (A_1 + B_1^T)y^s + (A_2 + C_1^T)z^s\rangle - \beta_{s+1} + \rho_s/3 \geq \varphi(y^s, z^s, \gamma_s). \end{array} \tag{6}$$

Step 2. Find a $\rho_s/3$-solution (y^{s+1}, γ_{s+1}) to Problem $(\mathcal{LP}_y(x^{s+1}, z^s, \alpha_s))$, so that the following inequality takes place:

$$\begin{array}{c} f_2(y^{s+1}, \gamma_{s+1}) + \rho_s/3 = \langle y^{s+1}, (B_1 + A_1^T)x^{s+1} + (B_2 + C_2^T)z^s\rangle - \\ -\gamma_{s+1} + \rho_s/3 \geq \psi(x^{s+1}, z^s, \alpha_s). \end{array} \tag{7}$$

Step 3. Find a $\rho_s/3$-solution (z^{s+1}, α_{s+1}) to Problem $(\mathcal{LP}_z(x^{s+1}, y^{s+1}, \beta_{s+1}))$, so that the following inequality holds:

$$\begin{array}{c} f_3(z^{s+1}, \alpha_{s+1}) + \rho_s/3 = \langle z^{s+1}, (C_1 + A_2^T)x^{s+1} + (C_2 + B_2^T)y^{s+1}\rangle - \\ -\alpha_{s+1} + \rho_s/3 \geq \chi(x^{s+1}, y^{s+1}, \beta_{s+1}). \end{array} \tag{8}$$

Step 4. If the following inequality takes place:

$$\Phi_{s+1} - \Phi_s \leq \tau, \tag{9}$$

where τ is a given accuracy, then stop computing, else set $s := s + 1$, and loop to Step 1. $\square$

Note that the all components of the point $(x^0, y^0, z^0, \alpha_0, \beta_0, \gamma_0)$ are not required to start the YZ_γ-procedure, only the part (y^0, z^0, γ_0) is sufficient.

The main new result concerning the convergence of the local search method is the following.

Theorem 3.1 *Suppose,* $\rho_s > 0,\ s = 0, 1, 2, \ldots,\ \sum\limits_{s=0}^{\infty} \rho_s < +\infty$. *Then the sequence* $\sigma^s \triangleq (x^s, y^s, z^s, \alpha_s, \beta_s, \gamma_s)$, *produced by the* YZ_γ*-procedure, converges to the point* $\hat{\sigma} \triangleq (\hat{x}, \hat{y}, \hat{z}, \hat{\alpha}, \hat{\beta}, \hat{\gamma})$, *which is satisfying the following inequalities:*

$$\Phi(\hat{\sigma}) \geq \Phi(x, \hat{y}, \hat{z}, \hat{\alpha}, \beta, \hat{\gamma}) \quad \forall (x, \beta) \in X(\hat{y}, \hat{z}, \hat{\gamma}), \tag{10}$$

$$\Phi(\hat{\sigma}) \geq \Phi(\hat{x}, y, \hat{z}, \hat{\alpha}, \hat{\beta}, \gamma) \quad \forall (y, \gamma) \in Y(\hat{x}, \hat{z}, \hat{\alpha}), \tag{11}$$

$$\Phi(\hat{\sigma}) \geq \Phi(\hat{x}, \hat{y}, z, \alpha, \hat{\beta}, \hat{\gamma}) \quad \forall (z, \alpha) \in Z(\hat{x}, \hat{y}, \hat{\beta}). \tag{12}$$

Definition 3.2 Any 6-tuple $\hat{\sigma}$, satisfying the inequalities (10), (11), and (12), will henceforth be called a *critical point* for Problem $(\mathcal{P})$. If the inequalities (10), (11), and (12) are satisfied with certain accuracy at some point, then this point is called an *approximately critical point.*

4 GLOBAL SEARCH

In this section we develop a Global Search Method (GSM) for Problem $(\mathcal{P})$, which is based on GOC for nonconvex problems with d.c. objective functions [10, 13]. Indeed, the objective function of Problem $(\mathcal{P})$ can be represented as a difference of two convex functions, for example, as follows:

$$\Phi(x, y, z, \alpha, \beta, \gamma) = f(x, y, z) - g(x, y, z, \alpha, \beta, \gamma), \tag{13}$$

where

$$\begin{aligned} f(x, y, z) = \frac{1}{4}\Big(&\|x + A_1 y\|^2 + \|x + A_2 z\|^2 + \|B_1 x + y\|^2 + \|y + B_2 z\|^2 + \\ &+ \|C_1 x + z\|^2 + \|C_2 y + z\|^2\Big), \quad g(\sigma) = \frac{1}{4}\Big(\|x - A_1 y\|^2 + \|x - A_2 z\|^2 + \\ &+ \|B_1 x - y\|^2 + \|y - B_2 z\|^2 + \|C_1 x - z\|^2 + \|C_2 y - z\|^2\Big) + \alpha + \beta + \gamma. \end{aligned} \tag{14}$$

It is easy to see that these functions are convex on (x, y, z) and σ, respectively.

Now, let us consider the GOC for Problem $(\mathcal{P})$, which are represented in a contrapositive form. These conditions are based on GOC for nonconvex problems with d.c. objective functions [10, 11, 13, 14].

Define $\alpha(y, z) \triangleq \max\limits_{1 \leq i \leq m} (A_1 y + A_2 z)_i$, $\beta(x, z) \triangleq \max\limits_{1 \leq j \leq n} (B_1 x + B_2 z)_j, \gamma(x, y) \triangleq \max\limits_{1 \leq t \leq l} (C_1 x + C_2 y)_t$, so that, according to Corollary 2.3, $\alpha_* = \alpha(y^*, z^*), \beta_* = \beta(x^*, z^*), \gamma_* = \gamma(x^*, y^*)$.

Theorem 4.1 [10, 11, 13, 14] *If a feasible 6-tuple* $\sigma_* = (x^*, y^*, z^*, \alpha_*, \beta_*, \gamma_*)$ *is not a global solution to Problem* $(\mathcal{P})$, *then there exists a triple* $(u, v, w) \in \mathbb{R}^{m+n+l}$, *a vector* $(\bar{x}, \bar{y}, \bar{z}) \in S_m \times S_n \times S_l$, *and a scalar* ξ, *such that*

$$\begin{gathered} f(u, v, w) - \xi = \zeta \stackrel{\triangle}{=} \Phi(\sigma_*) < 0, \\ g(u, v, w, \alpha(v, w), \beta(u, w), \gamma(u, v)) \leq \xi \leq \sup(g, D), \end{gathered} \tag{15}$$

and the following inequality takes place:

$$g(\bar{x}, \bar{y}, \bar{z}, \bar{\alpha}, \bar{\beta}, \bar{\gamma}) - \xi < \langle \nabla_{xyz} f(u, v, w), (\bar{x}, \bar{y}, \bar{z}) - (u, v, w) \rangle, \tag{16}$$

where $\bar{\alpha} = \alpha(\bar{y}, \bar{z}),\ \bar{\beta} = \beta(\bar{x}, \bar{z}),\ \bar{\gamma} = \gamma(\bar{x}, \bar{y})$. □

These GOC possess a so-called algorithmic (constructive) property [10, 13, 14]. For Problem $(\mathcal{P})$, it means that if one was successful to find the 4-tuple $(\bar{u}, \bar{v}, \bar{w}, \bar{\xi})$ from (15), and the point $\bar{\sigma} \triangleq (\bar{x}, \bar{y}, \bar{z}, \bar{\alpha}, \bar{\beta}, \bar{\gamma})$, $(\bar{x}, \bar{y}, \bar{z}) \in S_m \times S_n \times S_l$, $\bar{\alpha} = \alpha(\bar{y}, \bar{z})$, $\bar{\beta} = \beta(\bar{x}, \bar{z})$, $\bar{\gamma} = \gamma(\bar{x}, \bar{y})$, such that the inequality (16) holds:

$$g(\bar{\sigma}) - \bar{\xi} < \langle \nabla_{xyz} f(\bar{u}, \bar{v}, \bar{w}), (\bar{x}, \bar{y}, \bar{z}) - (\bar{u}, \bar{v}, \bar{w}) \rangle,$$

then, due to the convexity of the function $f(\cdot)$, and the equality in (15) one obtains $\Phi(\bar{\sigma}) > \Phi(\sigma_*)$. It means that the 6-tuple $\bar{\sigma} = (\bar{x}, \bar{y}, \bar{z}, \bar{\alpha}, \bar{\beta}, \bar{\gamma})$ is better than the 6-tuple $\sigma_* = (x^*, y^*, z^*, \alpha_*, \beta_*, \gamma_*)$. In order to see more details on GOC, we refer the readers to the papers and books [7, 10, 13, 14].

Now, let us pass to the description of the GSM based on Theorem 4.1 taking into account the d.c. decomposition (13)–(14). Let there be given an approximate critical point $\sigma_k \triangleq (x^k, y^k, z^k, \alpha_k, \beta_k, \gamma_k)$ in Problem $(\mathcal{P})$ obtained by the local search procedure. Denote $\zeta_k \triangleq \Phi(\sigma_k)$. Then, one has to carry out the following chain of operations.

1) Choose the number $\xi \in [\xi_-, \xi_+]$, where $\xi_- \stackrel{\triangle}{=} \inf(g, D)$, $\xi_+ \stackrel{\triangle}{=} \sup(g, D)$ and construct an approximation $\mathcal{A}_k = \{(u^1, v^1, w^1), \ldots, (u^{N_k}, v^{N_k}, w^{N_k}) \mid f(u^p, v^p, w^p) = \zeta_k + \xi, \ \ p = 1, ..., N_k\}$ of the level surface $U(\zeta_k) = \{(x, y, z) \mid f(x, y, z) = \zeta_k + \xi\}$ of the convex function $f(x, y, z)$.

2) For each point (u^p, v^p, w^p) of $\mathcal{A}_k$ verify the inequality

$$g(u^p, v^p, w^p, \alpha(v^p, w^p), \beta(u^p, w^p), \gamma(u^p, v^p)) \leq \xi, \ \ p = 1, \ldots, N_k, \tag{17}$$

according to GOC (see the inequality in (15)).

If the inequality (17) is fulfilled, then the approximation point (u^p, v^p, w^p) will be used at the following steps, otherwise, this point has to be excluded from the consideration.

3) Using the point (u^p, v^p, w^p) of the approximation $\mathcal{A}_k$, selected at the stage 2), find an approximate solution $(\bar{x}^p, \bar{y}^p, \bar{z}^p, \bar{\alpha}_p, \bar{\beta}_p, \bar{\gamma}_p)$ of the following linearized problem [10, 13, 14]:

$$g(\sigma) - \langle \nabla f(u^p, v^p, w^p), (x, y, z) \rangle \downarrow \min_{\sigma}, \quad \sigma \in D. \tag{$\mathcal{PL}_p$}$$

Note that the Problem $(\mathcal{PL}_p)$ is convex quadratic, and, consequently, it can be solved by one of the contemporary convex optimization methods [5].

4) Proceeding from the points $(\bar{x}^p, \bar{y}^p, \bar{z}^p, \bar{\alpha}_p, \bar{\beta}_p, \bar{\gamma}_p) \in D$, produce critical points $(\hat{x}^p, \hat{y}^p, \hat{z}^p, \hat{\alpha}_p, \hat{\beta}_p, \hat{\gamma}_p)$, $p \in \{1, \ldots, N_k\}$, by means of the local search method.

5) Further, for each $p \in \{1, \ldots, N_k\}$ solve the following level problem:

$$\langle \nabla f(u, v, w), (\hat{x}^p, \hat{y}^p, \hat{z}^p) - (u, v, w) \rangle \uparrow \max_{(u,v,w)}, \quad f(u, v, w) = \zeta_k + \xi. \tag{$\mathcal{U}_p$}$$

Let (u_0^p, v_0^p, w_0^p) be an approximate solution to $(\mathcal{U}_p)$.

6) At the final stage, for each $p \in \{1, \ldots, N_k\}$ verify the following inequality (cf. (16)):

$$g(\hat{x}^p, \hat{y}^p, \hat{z}^p, \hat{\alpha}_p, \hat{\beta}_p, \hat{\gamma}_p) - \xi < \langle \nabla f(u_0^p, v_0^p, w_0^p), (\hat{x}^p, \hat{y}^p, \hat{z}^p) - (u_0^p, v_0^p, w_0^p) \rangle.$$

If for some p the latter inequality is fulfilled then the point $(\hat{x}^p, \hat{y}^p, \hat{z}^p, \hat{\alpha}_p, \hat{\beta}_p, \hat{\gamma}_p)$ turns out to be better than σ_k.

Note that the crucial moment of the above GSM consists in constructing an approximation of the level surface of the convex function $f(\cdot)$, which generates the basic nonconvexity in the problem under consideration. In particular, in Problem $(\mathcal{P})$ the approximation $\mathcal{A}_k = \mathcal{A}(\zeta_k)$ has been constructed with the help of special sets of directions [6, 7, 14, 15, 16]:

$$Dir1 = \{(e^i, e^j, e^t), \ \ i = 1, ..., m, \ j = 1, ..., n, \ t = 1, ..., l\}, \tag{18}$$

$$Dir2 = \{(e^i + x^k, e^j + y^k, e^t + z^k), \ \ i = 1, ..., m, \ j = 1, ..., n, \ t = 1, ..., l\}, \tag{19}$$

where $e^i \in I\!R^m, e^j \in I\!R^n, e^t \in I\!R^l$ are the basic Euclidean vectors, (x^k, y^k, z^k) is a part of a current critical point of Problem $(\mathcal{P})$.

As a result, taking into account the features of Problem $(\mathcal{P})$ and basing on the stages of the global search 1)-6), we have constructed and implemented the Global Search Algorithm in the hexamatrix games. The efficiency of methods developed for hexamatrix games is demonstrated by the results of computational solving of the test problems. Our future work will be direct to the elaborating of local and global search methods for polymatrix games with more than 3 players.

5 Acknowledgements

This work is carried out under financial support of Russian Science Foundation (project no. 15-11-20015).

References

[1] Audet, C., Belhaiza, S. and Hansen, P. (2006). Enumeration of all the extreme equilibria in game theory: bimatrix and polymatrix games. J. Optim. Theory Appl., 129(3): 349–372.

[2] Horst, R. and Tuy, H. (1993). Global Optimization. Deterministic Approaches. Berlin: Springer-Verlag.

[3] Lemke, C.E., Howson, J.J. (1964). Equilibrium points of bimatrix games. J. Soc. Ind. Appl. Math., 12: 413–423.

[4] Mazalov, V. (2014). Mathematical Game Theory and Applications. New York: John Wiley & Sons.

[5] Nocedal, J. and Wright, S.J. (2000). Numerical Optimization. New York-Berlin-Heidelberg: Springer-Verlag.

[6] Orlov, A.V. (2008). Numerical solution of bilinear programming problems. Comput. Math. Math. Phys., 48(2): 225–241.

[7] Orlov, A.V. and Strekalovsky, A.S. (2005). Numerical search for equilibria in bimatrix games. Comput. Math. Math. Phys., 45(6): 947–960.

[8] Owen, G. (1995). Game Theory. San Diego: Academic Press.

[9] Pang, J.-S. (2010). Three modeling paradigms in mathematical programming. Math. program. Ser.B., 125(2): 297–323.

[10] Strekalovsky, A.S. (2003). Elements of Nonconvex Optimization (in Russian). Novosibirsk: Nauka.

[11] Strekalovsky, A.S. (2014). On Solving Optimization Problems with Hidden Nonconvex Structures. In Rassias, T.M., Floudas, C.A., Butenko, S. (Eds.). Optimization in Science and Engineering. (pp. 465–502). New-York: Springer.

[12] Strekalovsky, A.S. and Enkhbat, R. (2014). Polymatrix games and optimization problems. Autom. Remote Control, 75(4): 632–645.

[13] Strekalovsky, A.S. and Orlov, A.V. (2007a). A new approach to nonconvex optimization. Numer. Methods Prog., 8: 160–176.
http://num-meth.srcc.msu.ru/english/index.html. [Accessed 24 June 2015]

[14] Strekalovsky, A.S. and Orlov, A.V. (2007b). Bimatrix Games and Bilinear Programming (in Russian). Moscow: FizMatLit.

[15] Strekalovsky, A.S., Orlov, A.V. and Malyshev, A.V. (2010). On computational search for optimistic solutions in bilevel problems. J. Glob. Optim., 48(1): 159–172.

[16] Vasilyev, I.L., Klimentova, K.B. and Orlov, A.V. (2007). A parallel search of equilibrium points in bimatrix games (in Russian). Numer. Methods Prog., 8: 233–243.
http://num-meth.srcc.msu.ru/english/index.html. [Accessed 24 June 2015]

USING DATA MINING METHODS IN WEB ANALYTICS

Zoran Selinger
Wolfgang Digital
Palmerston House, Denzille Lane, Dublin 2, Ireland
zselinger@gmail.com

Marijana Zekić-Sušac[1]
University of Josip Juraj Strossmayer in Osijek, Faculty of Economics
Gajev trg 7, 31000 Osijek, Croatia
marijana@efos.hr

Abstract: The paper deals with the problem of extracting knowledge about customers from the data collected by web analytics tools. The analytics of a web shop was observed in order to discover patterns of user behaviour by data mining methods. Artificial neural networks, decision trees, and association rules were used in order to model revenue from online purchases in relation to devices, channels, and the number of sessions used by customers when accessing a web site. Although the predictive models of user behaviour created by neural networks and decision trees did not generate high accuracy, they identified a similar set of variables as important for understanding user behaviour. Also, the method of association rules extracted some very confident interesting patterns that could influence a company's marketing decisions.

Keywords: data mining, decision trees, neural networks, association rules, web analytics

1 INTRODUCTION

In contemporary business systems, the quality of customer relations became one of the most important characteristics of a company's business success. In new marketing campaigns, it is important to identify specific segments of consumers and target each segment with a personalized message that will generate a higher response rate, higher profit, and customer satisfaction. Although the target marketing could be based on some standard statistical techniques that analyze customers' frequent habits, data mining methods could go a step forward and provide customer segmentation, as well prediction models of their behaviour. The reason for using data mining methods is to reveal hidden knowledge among a large amount of data collected by companies in each segment of their businesses [8]. Due to their success, the methods of data mining became very accepted in today's businesses [5]. Web mining, i.e. data mining on web user data is the process of acquiring useful information from the database of user history on a web site. The data are usually collected by specific web analytics tools such as Google Analytics [6]. Web analytics tools capture data on user behaviour, sources of visits, conversions of visitors into customers, etc. Those tools are capable of producing reports, target warnings, and results of various analyses. Besides providing a huge amount of data, web analytics is valuable for its capability to combine data sets and find relationships among them [8]. In this paper, the focus is on identifying the relationship among the revenue a user produces on a web site and its behaviour.

2 OVERVIEW OF PREVIOUS RESEARCH

The need for intelligent methods usage in web mining is emphasized by many authors [2] as an important area in creating the new generation of intelligent web that will be able to adopt to user needs. Intelligence on the web site provided by the user can be divided into three categories: (1) direct information/intelligence provided by the user, such as reviews,

[1] Corresponding author

comments, recommendations, ratings, etc., (2) indirect information within blogs, unstructured text, contribution to online communities and wikis, and (3) intelligence provided by data mining techniques, where the patterns of user behaviour are discovered and used for recommendations, predictive analysis, market segmentation etc. [2] In recent years, a number of papers deals with data mining and web mining. Carmona et al. [6] used data mining techniques on web analytics data in order to discover rules that could be used by web designers to improve a web site structure and design. Arbelaitz et al. [4] showed that data mining methods could generate profiles of users with similar navigation habits. Similar research was conducted by Romero et. al. in [11] who discovered four interesting navigation patterns that assisted in understanding the interaction of user with the site content. Tao et al. [12] used data mining methods to generate recommendations for budget allocation among various commercials in a toy store. Zhang et al. [15] also used data mining methods that produced a high product adjustment rate of 50%. Hung et al. [7] used clustering to categorize elderly customers into three categories with different patterns of appliances usage. However, the methodological tool tested so far is mostly based on statistical descriptive analysis, or on single or competitive usage of data mining methods. For these reasons, this paper aims to suggest a methodological procedure for discovering patterns of web user behaviour.

3 METHODOLOGY

Data mining represents a methodological basis of business intelligence, by including techniques that are able to identify hidden relationships among the variables, such as statistical, mathematical methods, expert systems, machine learning methods, and others [10]. Data mining reveals patterns and rules from the data, and represent analytical processes which results in creating a new knowledge [5].

3.1 Data preparation

The dataset used for this research comes from the web analytics of a US company from California which produces organic herbs and spices. The company uses the Google Analytics tool to collect data from their web site. The sample consisted of 34044 cases of web site visitors, covering the time period from January 1 to December 31, 2013. In this research a random sample of 1726 cases was used for modelling purposes, from which some cases were omitted in order to obtain an equal distribution of output categories. Web user activity is captured through sessions. In analytics tools, a session is a group of interactions of a user that take place on a website within a time frame. Sessions could represent multiple screen or page views, events, social interactions, and e-commerce transactions. A session expires after 30 minutes of a user inactivity, or by changing a channel from which a user comes from. Due to the fact that users with more than seven sessions were very rare, the first seven sessions of each visitor were used in the modelling purposes. The input variables are: the channel used by a visitor to access the web site in each session (c_1,...,c_7), the device used to access the web site (d), while the revenue (value of the purchase) was the output variable y. So, the model can be formulated as: $y = f(c_1,...,c_7,d)$. Eight channels were observed: (1) D - Direct source – no specific source, or the visitor entered URL address directly in a browser, or had it in a bookmark, (2) O - „Organic“ search – a visitor used a search engine result to access the web site, (3) P – Paid search – a visitor used a paid add on an advertising platform such as Google AdWords or Bing Ads, (4) R – Referral – a visitor came by using a link on another web site, (5) E – E-mail – a visitor accessed the web site by using a link in an e-mail message, (6) S – Social network – a visitor accessed the web through a social network, (7) X – Display – a visitor accessed through a commercial aid announced by a

display network (such as AdSense, BuySellAds, AdRoll, and others), and Null – Unavailable – no session (used when a visitor sequence is over or the data is missing). The output variable was the Revenue value (value made by a purchase) in categories: (1) Low: 0-114.15 USD, (2) Medium: 114.16 – 570.75 USD, and (3) High: more than 570.75 USD. Besides developing a model that captures all visitors' behaviour, the separate models for personal computer users, and for mobile devices users were created. For the purpose of neural network (NN) and decision tree (DT) modelling, each dataset is divided into the train, test, and validation subsamples (see Table 1).

Table 1: Subsampling for the purpose of modelling

Sample	*Dataset 1 – All visitors*			*Dataset 2 – PC visitors*			*Dataset 3 – Mobile visitors*		
	Revenue 1	*Revenue 2*	*Revenue 3*	*Revenue 1*	*Revenue 2*	*Revenue 3*	*Revenue 1*	*Revenue 2*	*Revenue 3*
Train (70%)	205	205	205	174	174	174	30	30	30
Test (10%)	29	29	29	25	25	25	4	4	4
Validation (20%)	58	58	58	50	50	50	9	9	9
Total (100%)	292	292	292	249	249	249	43	43	43

Train subsample is used for NN learning phase, the test subsample is used for optimizing NN parameters (number of learning iterations, number of hidden units), while the final validation set is used to test the model on the hold-out sample. DTs were estimated on the 80% of the sample, and tested on the same validation sample (20%) as NNs.

3.2 Data mining methods used

Three data mining methods were used: neural networks (NNs), decision trees (DT), and association rules. The NNs are mathematical models that try to approximate a function between inputs and the output by using historical samples, in order to be applied on new data in prediction, classification or association [9]. The most frequently used NN is the multilayer perceptron (MLP), a network that is able to use various algorithms to minimize the objective function. NNs are structured in layers. The input layer of a NN consists of n input units $x_i \in R$, $i=1,2,...,n$, and randomly determined initial weights w_i usually from the interval [-1,1]. The hidden layer produces an output y_c computed by:

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where f is the activation function selected by the user, which can be logistic, tangent hyperbolic, exponential, linear, step or other [9]. The computed output is compared to the actual output y_a, and the local error ε is computed. The error is then used to adjust the weights of the input vector according to a learning rule, usually the Delta rule [9]. The above process is repeated in a number of iterations (epochs), where the gradient descent or other algorithm is used to minimize the error. In this research, the output layer of all NN models consisted of three categories of revenue generated by a visitor. The number of hidden units varied from 1 to 20. The NN structure and training time was optimized by a split-sample procedure. Besides the MLP NN, we also tested the Radial Basis Function Network (RBFN) which is based on clustering and computes the distance between each input vector and a

centre. It uses Euclidean distance measure (see [13]). In order to produce rules that could be used for classifying web customers, we used DTs that are described in more details in [14]. In a DT, a parent node is splitted into two child nodes according to the selected split point. A binary tree is created, and the process is repeated using cross-validation to select the right-sized tree (we used 10-fold crossvalidation). The evaluation function used for splitting is the Gini index defined in [3]. Additionaly, the method of association rules is used to extract the rules that have a high probability. A rule can be represented as $X \rightarrow Y, X \cap Y = \emptyset$ where *X* and *Y* are itemsets in a transaction database, and the rule is interpreted as *If X then Y*. Most common measures used to select the important rules are support *s* and confidence *c*. Support s is computed by [1]:

$$s(X \rightarrow \mathrm{Y}) = \frac{number\ \ of\ transactions\ containing\ X\ and\ Y}{number\ \ of\ transactions\ containing\ X \cup Y} \tag{2}$$

Confidence s is computed by:

$$c(X \rightarrow \mathrm{Y}) = \frac{number\ \ of\ transactions\ containing\ X\ and\ Y}{number\ \ of\ transactions\ containing\ X} \tag{3}$$

4 RESULTS

Separate models for each of the three datasets were developed by NN to find relationship between customer behaviour and purchase they made. Twenty NN architectures were trained, tested, and validated in each dataset, but here we bring only the most successful ones for each model.

Table 2: Best neural network results for all three models

Model	***Method***	***NN structu-re***	***Activati-on function***	***Classifi-cation rate for Revenue 1***	***Classifi-cation rate for Revenue 2***	***Classifi-cation rate for Revenue 3***	***Total classifi-cation rate***
1 – All visitors	MLP NN	51-45-3	Logistic	67.80%	30.36%	70.00%	56.57%
2 – PC Visitors	MLP NN	49-18-3	Logistic	42.00%	54.35%	39.62%	40.94%
3 – Mobile Visitors	MLP NN	40-45-3	Logistic	27.27%	50.00%	100.00%	52.00%
1 – All visitors	CART	-	-	50.00%	12.07%	70.69%	44.25%
2 – PC Visitors	CART	-	-	8.00%	84.00%	22.00%	38.00%
3 – Mobile Visitors	CART	-	-	44.00%	0.00%	22.00%	22.00%

It can be seen from Table 2 that MLP network produced the highest total classification rate of 56.57% in Model 1 – All visitors. The accuracy was lower (52.00%) in Model 3 - Mobile visitors, while the lowest rate was obtained in Model 2 – PC visitors (40.94%). Overall, the total accuracy of any of the three NN models is not satisfactory, revealing that the selection of input variables is not sufficient to explain the output. However, the classification rate of category 3 of the *Revenue* is very high (100% in Model 3, 70.00% in Model 2). It reveals that behaviour of customers that buy more expensive products can be predicted with a very high accuracy by NN. The sensitivity analysis obtained on the validation subsample in each of the three models showed that the most important variable for Model 1 is s_5 (*Channel in session 5*), followed by c_4 (*Channel in session 4*), then c_2 (*Channel in session 2*), and device type. The variable c_2 was also shown as the most important in Model 2, while in Model 3, the variable c_4 was the most important for predicting the output, followed by c_5 and c_1. The

accuracy of all three DT models was below 50%. The DT Model 1 produced the total classification rate of 44.25%, while DT Model 2 and Model 3 have total classification rate below 40%. Similar to NN Model 1 and NN Model 2, the DT Model 1 recognizes visitors in category 3 more accurate than in other two categories (70.69%). However, the DT Model 2 is more accurate in recognizing category 2 (84%). The sensitivity analysis performed on DT models showed that *Channel in session 4* is extracted as a very important variable, followed by *Channel in session 5* and *Channel in session 2*. Due to the fact that the tree split was made according to D channel or Null channel, it can be implied that users who make a lower number of steps before realizing a transaction, mostly use direct access, and yield lower revenue.

In order to extract more detailed insight into customers' behaviour, the association rules were conducted on separate datasets for PC visitors and mobile visitors. Two important rules were generated for PC visitors. A high *s* (63%) and *c* (95%) is obtained for the first rule „IF *Direct channel* THEN *Null*" meaning that if a visitor in one of the sessions access the web site by a direct channel, than the next step will be the null channel, i.e. the visit will end by a purchase or the source is unknown. The second rule „IF *Paid search channel* THEN *Null*" shows that the paid search channel also frequently ends the transaction (with confidence of 96%). When association rules are used in Model 3 – mobile visitors, only the rule „IF *Direct channel* THEN *Null*" was extracted as important with the support s of 73% and confidence s of 95%. The results show that visitors from mobile devices have similar behaviour as PC visitors, except they more frequently use the direct channel.

4.1 Evaluation and usage of revealed knowledge

The above results emphasize the importance of direct channel and paid search channel as the most frequent ones and also as the best predictors of the revenue category. The NN models show slightly higher total classification rate than the DT models while the classification rate for customers in the highest revenue category is very high (ranging from 70% to 100% in some NN models). This implies that customers that spend more money on the web share some common recognizable behaviour. Also, separate models should be developed for each category of revenue, and additional variables should be included in modelling medium and low categories. Improvements could be also directed to variable reduction procedures by removing the variables that were found less important in this research. In all models the step 4 was emphasized as very important to predict the revenue, therefore the marketing manager should focus on this step and pay more attention to visitors who have performed only four steps before the transaction and come from a direct channel or from a paid search channel.

5 CONCLUSION

The paper aimed to extract useful knowledge about customers from the web analytics by data mining methods, such as neural networks, decision trees, and association rules. The methods were used to model the revenue on the basis of devices and channels the visitors use, and a number of sessions they perform until they make the transaction on a web site. The models had a high accuracy for the category of high revenue, while the prediction of categories of low and medium revenue was not accurate enough. All three tested methods emphasized the importance of direct channel and paid search channel as the most important for realizing a transaction, while the four-session activity describes the behaviour of majority of visitors before making a purchase. Having such knowledge on user behaviour, the web sites could adjust its personalization strategy, improve the system performance, and contribute to development of intelligence on the web. The improvements could be focused

on developing separate models for different categories of revenue. The limitations of this study are in a single dataset observed with relatively small selection of variables. The initial models given in this paper could be enriched by additional variables, and tested on the whole available data instead on a random sample, therefore enabling big data analysis. Such model could be implemented in an online business intelligence system to assist customer relationship managers in designing their strategies on budget allocation.

References

[1] Abdullah, Z., Herawan, T., Ahmad, N., Mat Deris, M. (2011). Mining significant association rules from educational data using critical relative support approach, Procedia - Social and Behavioral Sciences, 28, 97-101.

[2] Alag, S. (2009). Collective Intelligence in Action. Greenwich, CT: Manning Publications.

[3] Apte C, Weiss S. (1997). Data Mining with Decision Trees and Decision Rules. Future Generation Computer Systems; 13(2): 197-210.

[4] Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Perez, J.M. i Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. Expert Systems with Applications. 40(18), pp. 7478-7491.

[5] Berry, M.J.A i Linoff, G.S. (2004). Data mining techniques for marketing, sales, and customer relationship management. Indianapolis: Wiley Publishing Inc.

[6] Carmona, C.J., Ramirez-Gallego, S., Torres, F., Bernal, E., del Jesus, M.J. i Garcia S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications. 39(12), pp. 11243-11249.

[7] Hung, Y., Chen, K.B., Yang, C. i Deng, G. (2013). Web usage mining for analysing elder self-care behavior patterns. Expert Systems with Applications. 40(2), pp. 775-783.

[8] Kursan, I. i Mihić, M. (2010). Business Intelligence: The Role of the Internet in Marketing Research and Business Decision-Making. Management: Journal of Contemporary Management Issues. 15(1), pp. 69-86.

[9] Masters T. (1995). Advanced Algorithms for Neural Networks. New York: John Wiley & Sons.

[10] Panian, Ž i Klepac, G. (2003). Poslovna inteligencija (Business Intelligence). Zagreb: Masmedia.

[11] Romero, C., Ventura, S., Zafra, A. i de Bra, P. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. Computers & Education. 53(3), pp. 828-840

[12] Tao, Y., Hong, T., Lin, W. i Chiu, W. (2009). A practical extension of web usage mining with intentional browsing data toward usage. Expert Systems with Applications. 36(2), pp. 3937-3945

[13] Tonković, Z., Zekić-Sušac, M. i Somolanji, M. (2009). Predicting natural gas consumption by neural networks. Technical Gazette. 16(3), pp. 51-61.

[14] Witten, I.H., Frank, E. (2000) Data Mining, Practical Machine Learning Tools and Techniques With Java Implementations, San Francisco, CA: Morgan Kauffman Publishers.

[15] Zhang, X., Edwards, J. i Harding, J. (2007). Personalised online sales using web usage data mining. Computers in Industry. 58(8-9), pp. 772-782.

NONCONVEX OPTIMIZATION: MODERN VIEW

Alexander S. Strekalovskiy
Matrosov Institute for System Dynamics and Control Theory of SB RAS,
Lermontov St., 134, Irkutsk, 664033, Russia

Abstract: Here we address two optimization problems: the search for Nash equilibria in polymatrix games and the quadratic-linear bilevel programming problem. It can be shown that each of the problems possesses a hidden nonconvexity and, as a consequence, a rather large number of local solutions which are different from global ones. Further, the principal elements of Global Search Theory are sketched. Finally, we present the results of computational solutions.

Keywords: nonconvex optimization, polymatrix games, Nash equilibria, bilevel programming.

1 INTRODUCTION: EXAMPLES OF APPLIED PROBLEMS

1.1 Search for an Equilibrium

As example of equilibrium problems, let us consider the polymatrix games [14] which reflect the conflict of N parties (players), each one having a finite number of strategies. After having introduced the mixed strategies, we obtain

$$\left.\begin{array}{c} F_i(x) = \sum\limits_{j\neq i}\langle x_i, A_{ij}x_j\rangle = \langle x_i, \sum\limits_{j\neq i} A_{ij}x_j\rangle \uparrow \max\limits_{x_i}, \quad x_i \in S_i, \\ S_i = \{x_i = (x_{i1},\ldots,x_{in_i})^T \in \mathbb{R}^{n_i}_+ \mid \sum\limits_{j=1}^{n_i} x_{ij} = 1\}, \quad i = 1,\ldots,N, \end{array}\right\} \tag{1}$$

where A_{ij} are matrices $(n_i \times n_j)$, $i, j = 1,\ldots,N$.

Next we will use the following denotation: $(x \parallel z_i) \triangleq (x_1,\ldots,x_{i-1},z_i,x_{i+1},\ldots,x_N)$.

Some economics, engineering and ecological problems can be represented in the form of polymatrix games, in which the Nash equilibrium is the common concept and can be represented as follows: find an equilibrium situation $x^* = (x_1^*,\ldots,x_N^*) \in S \triangleq S_1 \times S_2 \times \ldots \times S_N$:

$$F_i(x^* \parallel x_i) \leq F_i(x^*) \quad \forall x_i \in S_i, \quad i = 1,\ldots,N. \tag{2}$$

In formulae (1) and (2), from the first sight, any nonconvexity is not yet visible.

However, it turns out that the search for the Nash equilibrium can be reduced [14] to solving the following nonconvex (in general) problem of mathematical programming:

$$\left.\begin{array}{c} \Phi_N(\sigma) \triangleq \sum\limits_{i=1}^{N}[\alpha_i - F_i(x)] = \sum\limits_{i=1}^{N}(\alpha_i - \langle x_i, \sum\limits_{j\neq i} A_{ij}x_j\rangle) \downarrow \min\limits_{\sigma}, \\ \sigma = (x,\alpha) \in \mathbb{R}^n \times \mathbb{R}^N, \quad x = (x_1,\ldots,x_N) \in S, \quad \alpha = (\alpha_1,\ldots,\alpha_N) \in \mathbb{R}^N, \\ \sum\limits_{j\neq i} A_{ij}x_j \leq \alpha_i e_{n_i}, \quad i = 1,\ldots,N. \end{array}\right\} \tag{3}$$

where $e_{n_i} = (1,1,\ldots,1)^\top \in \mathbb{R}^{n_i}$. Note, that the numbers α_i^* in a global solution (x^*,α^*) to Problem (3) are the optimal profits of the players in the game (1)–(2), while the vector $(x_1^*,...,x_N^*)$ turns out to be just a Nash equilibrium point in the game (1)–(2).

On account of the formulation (3) it becomes clear that a way (method) of finding a Nash point strongly depends on the properties of the matrices A_{ij}.

So, the conclusion is obvious here and consists in the fact, that the initial statement (1)–(2) of a polymatrix game is deceptive in the sense that it has, in general, a hidden (implicit) nonconvexity.

1.2 Hierarchical Optimization Problems

Hierarchical problems are encountered in practice because of impossibility of accumulation of the total available information at the upper level in the process of investigation of structurally complex control systems (social, economic, ecological-economic ones etc.) and, as a consequence, possess some hidden nonconvexity generated by just hierarchical structures.

For example, the financial systems in the economic power countries are usually constructed as bilevel systems. As to mathematical aspects of the statement, problems of bilevel programming represent extremum problems, that side by side with standard constraints which are expressed in terms of equalities and inequalities, include the constraints described with the aid of optimization subproblem representing the lower level of the bilevel problem.

To begin with, let us consider the linear bilevel problem

$$(\mathcal{LBP}): \quad \left\{ \begin{array}{c} F(x,y) \triangleq \langle c,x\rangle + \langle d,y\rangle \downarrow \min\limits_{x,y}, \quad x \in X = \{\, x \in \mathbb{R}^m \mid Ax \le b \,\}, \\ y \in Y_*(x) \triangleq \operatorname{Arg}\min\limits_{y}\{\langle d_1,y\rangle \mid y \in Y(x)\}, \quad Y(x) = \{\, y \in \mathbb{R}^n \mid A_1x + B_1y \le b_1 \,\}, \end{array} \right.$$

where $c \in \mathbb{R}^m$, $d, d_1 \in \mathbb{R}^n$, $b \in \mathbb{R}^p$, $b_1 \in \mathbb{R}^q$, and A, A_1, B_1 are matrices of corresponding dimensions. Even in this very simple case it is easy to construct an example showing the nonconvexity of the problem $(\mathcal{LBP})$.

Example 1.1. [1] Consider the problem

$$F(x,y) = x + 3y \downarrow \min_{x,y}, \quad x,y \in \mathbb{R}, \quad 1 \le x \le 6, \quad y \in Y_*(x) = Sol(\mathcal{P}_L), \qquad (\mathcal{LBP}_1)$$

$$(\mathcal{P}_L): \qquad f(y) = -y \downarrow \min_{y}, \quad x + y \le 8, \quad x + 4y \ge 8, \quad x + 2y \le 13.$$

Regardless the convexity of the set

$$Z = \{\, (x,y) \in \mathbb{R}^2 \mid 1 \le x \le 6, \;\; x + y \le 8, \;\; x + 4y \ge 8, \;\; x + 2y \le 13 \,\},$$

it is easy to see even geometrically that the set $Z_* = \{\, (x,y) \in Z \mid y \in Y_*(x) \,\}$ is nonconvex that implies the nonconvexity in the problem $(\mathcal{LBP}_1)$.

2 OPTIMIZATION PROBLEMS WITH D.C. FUNCTIONS

The targets of our presentation can be bounded by consideration of the class $DC(\mathbb{R}^n)$ of the functions $f(\cdot)$ which can be represented as the difference of two convex functions (d.c. functions). Nowadays, this class is viewed by the specialists [2–5, 9, 12, 19] to be rather wise for consideration. Furthermore, the $DC(\mathbb{R}^n)$ possesses several remarkable properties.

Besides, the number of problems with d.c. functions is so large that the majority of the specialists, who have a long-time experience of solving problems of d.c. programming are sure [2–5, 19] that almost all nonconvex optimization problems are really d.c. problems.

Hence, the following optimization problem can be viewed as rather general:

$$\left. \begin{array}{l} f_0(x) = g_0(x) - h_0(x) \downarrow \min\limits_{x}, \;\; x \in S, \\ f_i(x) = g_i(x) - h_i(x) \le 0, \;\; i = 1,\dots,m; \\ f_j(x) = g_j(x) - h_j(x) = 0, \;\; j = 1,\dots,N. \end{array} \right\} \qquad (4)$$

Here g_i, g_j, h_i, h_j are convex functions and S is convex set from $\mathbb{R}^n$.

Apparently, almost all the specialists in optimization areas could estimate Problem (4) as very difficult and unsolvable by the existing approaches and methods even for the case of small dimension (say, $n = 100, \dots, 1000$.)

Therefore, let us start with rather simple (from the first glance) nonconvex optimization problems.

1. *D.C. minimization*

$$(\mathcal{P}): \qquad f(x) = g(x) - h(x) \downarrow \min, \quad x \in D, \tag{5}$$

where $g(\cdot)$, $h(\cdot)$ are convex functions, and D is a convex set, $D \subset \mathbb{R}^n$.

2. *D.C. constraint problem*

$$(\mathcal{DCC}): \qquad \left.\begin{array}{l} f_0(x) \downarrow \min\limits_x, \ x \in S, \\ F(x) = g(x) - h(x) \le 0, \end{array}\right\} \tag{6}$$

where $g(\cdot)$ and $h(\cdot)$ are as above, $S \subset \mathbb{R}^n$, $f_0(\cdot)$ is a continuous function.

Note, that any quadratic optimization problem with arbitrary matrices occurs in the classification (5)–(6) or takes the form (4).

3 GLOBAL SEARCH APPROACH

Since in our approach the general global search procedure includes two principal parts: a) local search; b) procedures of escape a critical point provided by a local search method, we are going, first, to consider special local search methods.

3.1 Local Search

The ideas of the most of local search methods (LSM) are rather simple and can be reduces to solution of the convex (partially) linearized problems/

For example, as regards the problem of d.c. minimization $(\mathcal{P})$–(5), the basic element, the "corner stone" of the Global and Local Search Methods is the solving the following (linearized at a current iteration point $x^s \in D$) convex problem

$$(\mathcal{PL}_s): \qquad \Phi_s(x) \triangleq g(x) - \langle h'(x^s), x\rangle \downarrow \min_x, \quad x \in D, \tag{7}$$

where $h'(x^s) = h'_s \in \partial h(x^s)$, $s = 1, 2, \ldots$ is a subgradient of the convex function $h(\cdot)$ at the point x^s [4]. Furthermore, the LSM itself for $(\mathcal{P})$–(5) may consists in the consecutive solving (likewise in the method of "direct iterations") Problems $(\mathcal{PL}_s)$–(7). More precisely, given $x^s \in D$, we can find $x^{s+1} \in D$ as an approximate solution to $(\mathcal{PL}_s)$ by means of some suitable convex optimization method (for example, BFGS), or one of packages of applied software (Xpress-MP, IBM CPLEX etc).

So, we produce the sequence $\{x^s\}$ according to the inequality:

$$\Phi_s(x^{s+1}) \triangleq g(x^{s+1}) - \langle h'(x^s), x^{s+1}\rangle \ - \delta_s \le \inf_x\{g(x) - \langle h'(x^s), x\rangle \mid \ x \in D\} \tag{8}$$

where the sequence $\{\delta_s\}$ fulfils the condition $\sum\limits_{s=0}^{\infty} \delta_s < +\infty$, $\delta_s > 0$, $s = 1, 2, \ldots$. It was rather surprising that, in this case, the process converges in the sense [10, 11].

Note, that very frequently for small dimensions ($n \le 7, 8, 10$) cases LSM (8) provides for a global solution to $(\mathcal{P})$–(5).

3.2 Global Optimality Conditions

The second step in the global search methodology can be viewed as the most important one and even crucial, because the question is how to escape a critical point (provided by a LSM and that is not a global solution).

Such a procedure is substantiated by the theoretical basis produced with the help of so-called Global Optimality Condition (GOC) which for the case of d.c. minimization problem $(\mathcal{P})$–(5) takes the following form.

Theorem 3.1 *If z is a global solution to $(\mathcal{P})$, $z \in Sol(\mathcal{P})$, $\zeta \triangleq f(z)$, then*

$$(\mathcal{E}): \qquad \begin{cases} \forall (y,\beta) \in \mathbb{R}^n \times \mathbb{R}: \quad h(y) = \beta - \zeta, \\ g(x) - \beta \geq \langle h'(y), x - y \rangle \ \ \forall x \in D. \end{cases} \tag{9}$$

So, when selecting the "perturbation parameters" (y, β) satisfying (9) and solving the linearized problem (sf. (7))

$$\Phi_y(x) \triangleq g(x) - \langle h'(y), x \rangle \downarrow \min_x, \quad x \in D, \tag{10}$$

(where $y \in \mathbb{R}^n$ is not obligatory feasible!) we obtain a family of starting points $x(y,\beta)$ for a further (assume) local search.

Moreover, at each level $\zeta_k = f(z^k)$ it is not necessary to investigate all the pair (y,β) satisfying (9), $\zeta_k = \beta - h(y)$, but it is sufficient to discover the violation of the variational inequality (9) only for one pair $(\hat{y}, \hat{\beta})$.

After that, one proceeds to the next iteration of the global search: $z^{k+1} := \hat{x}$, $\zeta_{k+1} := f(z^{k+1})$, and starts the procedure from the very beginning. So, the idea of the global search method becomes considerably more clear.

3.3 Global Search Methods

In order to deal with nonconvex optimization problems and, in addition, on account of rather large computational experience [12, 13, 15–18] we propose two rules on which a search for a global solution to d.c. optimization problems is based.

These rules may be explained as follows.

1. Never apply CGM or BFGS if you are not convinced that your problem is convex.

2. Try to separate the data of your problem into two parts — convex and anticonvex.

For example, dealing with a quadratic function of the kind $q(x) = \frac{1}{2}\langle x, Qx \rangle$, where the matrix Q is indefinite, you have to separate the matrix $Q(n \times n)$ into a difference $Q = Q_1 - Q_2$ of two symmetric positive definite matrices $Q_i = Q_i^T > 0$, $i = 1, 2$.

Let us return now to the construction of a GSM (strategy) based on GOC presented in Theorem 3.1 and specialized only for Problem $(\mathcal{P})$–(5).

The basic stages of such a GSM (strategy) can be described as follows.

I. Find a critical point z by means of the special LSM ((8), for example).

II. Choose a number $\beta \in [\beta_-, \beta_+]$, where $\beta_- = \inf(g, D)$, $\beta_+ = \sup(g, D)$ can be approximated by rather rough estimates.

III. Construct an approximation

$$\mathcal{A}(\beta) = \{y^1, \ldots, y^N \mid h(y^i) = \beta - \zeta, \ \ i = 1, \ldots, N = N(\beta)\}$$

of the level surface of the function $h(\cdot)$.

IV. Starting at any point y^i of the approximation $\mathcal{A}(\beta)$ find a feasible point u^i by means of the special local search algorithm (8).

V. Verify the VI (9) from GOC

$$g(u^i) - \beta \geq \langle h'(w^i), u^i - w^i \rangle \quad \forall i = 1, \ldots, N, \tag{11}$$

where w^i may be found as the projection of the point u^i onto the convex set

$$\mathcal{L}(h, \beta - \zeta) = \{\, x \in \mathbb{R}^n \mid h(x) \leq \beta - \zeta \,\}.$$

VI. If $\exists j \in \{1, \ldots, N\}$ such that (11) is violated, then set $x^{k+1} := u^j$ and return to Stage I. Otherwise change β and return to Stage III.

Table 1: Computational results.

m =n	F_0	F_k	st	LP	Loc	$Time$
200	40.3811	0	2	972	65	00:43.66
250	51.1617	0	4	4843	321	5:36.30
300	60.1209	0	2	28	2	00:44.17
400	75.0987	0	6	16168	978	59:05.49
500	75.3494	0	2	158	20	05:57.88
600	84.1025	0	2	82	7	07:33.92
700	89.0439	0	2	54	6	11:50.34
800	99.8335	0	2	48	3	15:28.70
900	100.0419	0	3	136	11	29:31.38
1000	106.8368	0	3	178	18	45:04.34

4 COMPUTATIONAL RESULTS FOR BIMATRIX (BM) GAMES

The numerical experiments were conducted with software programs implementing the global search algorithms above. The computational simulations have been separated into several stages, and the first results of these experiments have been published in [8].

Further, the analysis of the results allowed us to conclude about some shortcomings of the software programs developed.

As a consequence, for solving the BM games of rather high dimension (up to 1000×1000) we decided to apply ILOG CPLEX 9.1 especially oriented to LP problems. In addition, in order to create the worst conditions for the global search software the entries of matrices A and B have randomly been generated from the interval $[-n, n]$, where $n = m$.

The software programs of global search were run on Pentium 4, CPU 3GHz with 512Mb of RAM and have been implemented by post-graduated students without a long computational experience.

Nevertheless, the results of computational solving of BM games ($m = n$) can be viewed as rather promising from the point of view of analysis of numeric results of Table 1.

In Table 1, $m = n$ is the number of pure strategies of players 1 and 2, F_0 stands for the value of the goal function at the starting points, F_k is the corresponding value at the best obtained point, st is the number of iterations of global search algorithms (or, what is the same, the number of critical (stationary) points passed by GS algorithms), LP and Loc represent the number of linearized problems solved and the number of local search algorithm's applications respectively. It can readily be seen, that in the cases of $m = n = 250$ and 400 it happened to randomly generate very difficult problems. Nevertheless, all test-problems have successfully been solved that certifies on the computational effectiveness of the software program created on the basis of the global search algorithm and the Global Search Theory.

Now, we are preparing to attack the bimatrix game of dimension $m = n = 10^4$ and the similar three-person-game of dimension $m = n = l = 5$, and 10.

5 Acknowledgements

The work is supported by the Russian Science Foundation, project No. 15-11-20015.

References

[1] Dempe, S. (2002). Foundations of Bilevel Programming. Dordrecht, Netherlands: Kluwer Academic Publishers.

[2] Floudas, C.A. and Pardalos, P.M. (eds.) (2004). Frontiers in Global Optimization. New York: Kluwer Academic Publishers.

[3] Hiriart-Urruty, J.-B. (1985). Generalized Differentiability, Duality and optimization for Problems dealing with Difference of Convex Functions. Lecture Notes in Economics and Mathem. Systems. Berlin, Springer-Verlag, 256: 37–69.

[4] Hiriart-Urruty, J.B. and Lemaréchal, C. (1993). Convex Analysis and Minimization Algorithms. Berlin: Springer Verlag.

[5] Horst R. and Tuy H. (1993). Global Optimization. Deterministic Approaches. Berlin: Springer-Verlag.

[6] Mangasarian, O.L. (1964). Equilibrium points in bimatrix games. Journal of the Society for Industrial and Applied Mathematics, 12: 778–780.

[7] Mills, H. (1960). Equilibrium points in finite games. Journal of the Society for Industrial and Applied Mathematics, 8(4): 397–402.

[8] Orlov, A.V. and Strekalovsky A.S. (2005). Numerical search for equilibria in bimatrix games. Comput. Math. and Math. Physics, 45(6): 947–960.

[9] Pardalos, P.M. and Resende, M.G.C. (eds.) (2002). Handbook of Applied Optimization. Oxford Univesity Press, Inc.

[10] Strekalovsky, A.S. (2015). On local search in d.c. optimization problems. Applied Mathematics and Computation, 255: 73–83.

[11] Strekalovsky, A.S. (2014). On Solving Optimization Problems with Hidden Nonconvex Structures. In Rassias, T.M., Floudas, C.A., Butenko, S. (Eds.). Optimization in Science and Engineering (pp. 465–502). New York: Springer.

[12] Strekalovsky, A.S. (2003). Elements of Nonconvex Optimization. Novosibirsk: Nauka Publ. (In Russian).

[13] Strekalovsky, A.S. (1987) On problem of global extremum. The Proceedings of the USSR Academy of Sciences, 292(5): 1062–1066.

[14] Strekalovsky, A.S. and Enkhbat, R. (2014). Polymatrix games and optimization problems. Autom. Remote Control, 75(4): 632–645.

[15] Strekalovsky, A.S. and Orlov, A.V. (2007). Bimatrix Games and Bilinear Programming. Moscow: PHYSMATHLIT (in Russian).

[16] Strekalovsky, A.S., Orlov, A.V., and Malyshev, A.V. (2010). Local search in a quadratic-linear bilevel programming problem. Numerical Analysis and applications, 3(1): 59–70.

[17] Strekalovsky, A.S., Orlov, A.V., and Malyshev, A.V. (2010). Numerical solution of a class of bilevel programming problems. Numerical Analysis and applications, 3(2): 165–173.

[18] Strekalovsky, A.S., Orlov, A.V., and Malyshev, A.V. (2010) On computational search for optimistic solutions in bilevel problems. Journal of Global Optimization, 48: 159–172.

[19] Tuy, H. (1995). D.c. Optimization: Theory, Methods and Algorithms. In: Horst, R. and Pardalos, P.M. (eds.) Handbook of Global optimization (pp. 149–216). Kluwer Academic Publisher.

The 13th International Symposium on
Operational Research in Slovenia

SOR '15

Bled, SLOVENIA
September 23 - 25, 2015

Appendix

Authors' addresses

Addresses of SOR'15 Authors

(The 13th International Symposium on OR in Slovenia, Bled, SLOVENIA, September 23 – 25, 2015)

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
1.	Darko	Aleksovski	Jožef Štefan Institute, Knowledge Technologies Department	Jamova cesta 39	1000	Ljubljana	Slovenia	darko.aleksovski@ijs.si
2.	Arianna	Alfieri	Politecnico di Torino	C.so Duca degli Abruzzi 24	10129	Torino	Italy	arianna.alfieri@polito.it
3.	Ayca	Altay	Istanbul Technical University, Industrial Engineering Department	ITU Isletme Fakultesi, Macka	34357	Istanbul	Turkey	altaya@gmail.com
4.	Josip	Arnerić	University of Zagreb, Faculty of Economics and Business Zagreb	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	jarneric@efzg.hr
5.	Nuno	Azevedo	Middle East Technical University, Üniversiteler Mahallesi; Institute of Applied Mathematics	Dumlupınar Bulvarı 1	06800	Çankaya Ankara	Turkey	
6.	Anna	Azzi	Department of Management and Engineering, University of Padua	Stradella San Nicola 3	36100		Italy	annaazzi.uni@ gmail.com
7.	Zoran	Babić	University of Split, Faculty of Economics	Cvite Fiskovića 5	21000	Split	Croatia	babic@efst.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
8.	Alenka	Baggia	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	alenka.baggia@ fov.uni-mb.si
9.	Vlasta	Bahovec	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	vbahovec@efzg.hr
10.	Danijel	Barbarić	University of Split, Faculty of Law	Domovinskog rata 8	21000	Split	Croatia	danijel.barbaric@ pravst.hr
11.	Nejc	Bat	Arctur d.o.o.	Industrijska cesta 5	5000	Nova Gorica	Slovenia	nejc.bat@arctur.si
12.	Daria	Battini	Department of Management and Engineering, University of Padua	Stradella San Nicola 3	36100	Padua	Italy	daria.battini@unipd.it
13.	Jani	Bekő	Faculty of Economics and Business, University of Maribor, Department of Quantitative Economic Analysis	Razlagova 14	2000	Maribor	Slovenia	alenka.kavkler@ uni-mb.si
14.	Aharon	Ben-Tal	William Davidson Faculty of Industrial Engineering and Management Technion - Israel Institute of Technology	Technion city	32000	Haifa	Israel	abental@ ie.technion.ac.il
15.	Ján	Bendík	University of Žilina, Faculty Management Science and Informatics, Department of Mathematical Method and Operations Research	Univerzitná 8215/1	01026	Žilina	Slovak Republic	Jan.Bendik@fri.uniza.sk
16.	Tomaž	Berlec	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	tomaz.berlec@fs.uni-lj.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
17.	Ana	Bilandžić	Faculty of Economics in Osijek, J.J. Strossmayer University of Osijek	Trg Ljudevita Gaja 7	31000	Osijek	Croatia	anag@efos.hr
18.	David	Bogataj	MEDIFAS-CERRISK	Mednarodni prehod 6 Vrtojba	5290	Šempeter pri Gorici	Slovenia	dbogataj@actuary.si
19.	Marija	Bogataj	MEDIFAS-CERRISK	Mednarodni prehod 6 Vrtojba	5290	Šempeter pri Gorici	Slovenia	marija.bogataj@ guest.arnes.si
20.	Marko	Bohanec	Jožef Stefan Institute, Department of Knowledge Technologies	Jamova cesta 39	1000	Ljubljana	Slovenia	marko.bohanec@ijs.si
21.	Marko	Bohanec	Salvirt ltd.	Dunajska cesta 136	1000	Ljubljana	Slovenia	Marko.Bohanec@ salvirt.com
22.	Drago	Bokal	University of Maribor; Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	drago.bokal@um.si
23.	Đula	Borozan	J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek		31000	Osijek	Croatia	borozan@efos.hr
24.	Luka	Borozan	J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek		31000	Osijek	Croatia	lborozan@mathos.hr
25.	Andrej	Bregar	Informatika d.d.	Vetrinjska ulica 2	2000	Maribor	Slovenia	andrej.bregar@ informatika.si
26.	Alenka	Brezavšček	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	alenka.brezavscek@ fov.uni-mb.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
27.	Sergio	Cabello	Faculty of Mathematics and Physics, University of Ljubljana	Jadranska 19	1111	Ljubljana	Slovenia	sergio.cabello@ fmf.uni-lj.si
28.	Katarína	Cechlárová	Institute of Mathematics, Faculty of Science, P. J. Šafárik University	Jesenná 5	040 01	Košice	Slovak Republic	katarina.cechlarova@ upjs.sk
29.	Frank	Chen	Department of Management Sciences, College of Business	The City University of Hong Kong			Hong Kong	youhchen@cityu.edu.hk
30.	Michaela	Chocholatá	University of Economics in Bratislava, Department of Operations Research and Econometrics	Dolnozemská cesta 1	852 35	Bratislava,	Slovak Republic	chocholatam@ yahoo.com
31.	Stefano	Cozzini	Centro DEMOCRITOS, Istituto Officina dei Materiali CNR-IOM; c/o Scuola Internazionale Superiore di Studi Avanzati (SISSA)	Via Bonomea 265	34136	Trieste	Italy	stefano.cozzini@sissa.it
32.	Dajana	Cvrlje	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	dcvrlje@efzg.hr
33.	Peter	Czimmermann	University of Žilina			Žilina	Slovak Republic	peter.czimmermann@ fri.uniza.sk
34.	Anita	Čeh Časni	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedy 6	10000	Zagreb	Croatia	aceh@efzg.hr
35.	Draženka	Čizmić	University of Zagreb, Faculty of Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	dcizmic@efzg.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
36.	Mirjana	Čižmešija	University of Zagreb, Faculty of Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	mcizmesija@efzg.hr
37.	Marko	Debeljak	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	marko.debeljak@ijs.si
38.	Petra	Drevenšek	Faculty of computer and information science, University of Ljubljana		1000	Ljubljana	Slovenia	
39.	Damjana	Drobne	University of Ljubljana, Biotechnical faculty, Dept. of Biology	Večna pot 111	1000	Ljubljana	Slovenia	damjana.drobne@ bf.uni-lj.si
40.	Samo	Drobne	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	samo.drobne@ fgg.uni-lj.si
41.	Ksenija	Dumičić	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedy 6	10000	Zagreb	Croatia	kdumicic@efzg.hr
42.	Mehmet Bulent	Durmusoglu	Istanbul Technical University/Industrial Engineering Department		34367	Macka, Istanbul	Turkey	durmusoglum@itu.edu.tr
43.	Sašo	Džeroski	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	saso.dzeroski@ijs.si
44.	Pavlos	Eirinakis	Department of Management Science & Technology, Athens University of Economics and Business	76 Patission Ave.	104 34	Athens	Greece	
45.	Jawad	Elomari	ORTEC B.V.	Houtsingel 5	2719	EA	The Netherlands	jawad.elomari@ortec.com

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
46.	Nataša	Erjavec	Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	nerjavec@efzg.hr
47.	Liljana	Ferbar Tratar	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1000	Ljubljana	Slovenia	liljana.ferbar.tratar@ef.uni-lj.si
48.	Bogdan	Filipič	Jožef Stefan Institute, Department of Intelligent Systems	Jamova cesta 39	1000	Ljubljana	Slovenia	bogdan.lipic@ijs.si
49.	Tamás	Fleiner	Budapest University of Technology and Economics and MTA-ELTE Egerváry Research Group	Magyar tudósok körútja 2	1117	Budapest	Hungary	fleiner@cs.bme.hu
50.	Maria Victoria	de-la-Fuente Aragon	Technical University of Cartagena, Research Group Industrial Engineering and Management	S/ Dr. Fleming, s/n	30202	Cartagena	Spain	marivi.fuente@upct.es
51.	Andrea	Furková	University of Economics in Bratislava, Department of Operations Research and Econometrics	Dolnozemská cesta 1	852 35	Bratislava,	Slovak Republic	andrea.furkova@euba.sk
52.	Ioannis	Ganas	Technological Educational Institute of Epirus,Department of Accounting and Finance	Psathaki	481 00	Preveza	Greece	ganas@teiep.gr
53.	Helena	Gaspars-Wieloch	Poznan University of Economics, Department of Operations Research	Al. Niepodleglosci 10	61-875	Poznan	Poland	helena.gaspars@ue.poznan.pl

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
54.	Tanja	Gologranc	Faculty of Natural Sciences and Mathematics, University of Maribor	Koroška cesta 160	2000	Maribor	Slovenia	tanja.gologranc@um.si
55.	Ekaterina	Grakova	VŠB - Technical University of Ostrava, IT4Innovations	17. listopadu 15/2172	708 00	Ostrava	Czech Republic	ekaterina.grakova.st@vsb.cz
56.	Petra	Grošelj	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	petra.groselj@bf.uni-lj.si
57.	Robert W.	Grubbström	Linköping Institute of Technology	83 Linköping	SE-58183	Linköping	Sweden	robert@grubbstrom.com
58.	Gregory	Gurevich	Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering	Bialik Sts. 56, P.O. Box 950	84100	Beer Sheva	Israel	gregoryg@sce.ac.il
59.	Nebojša	Gvozdenović	University of Novi Sad, Faculty of Economics	Segedinski put 9-11	24000	Subotica	Serbia	nebojsa.gvozdenovic@gmail.com
60.	Yossi	Hadad	Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering	Bialik Sts. 56, P.O. Box 950	84100	Beer Sheva	Israel	yossi@sce.ac.il
61.	Jana	Hančlová	VŠB-Technical University of Ostrava, Faculty of Economics, Department of European Integration	Sokolská třda 33	701 21	Ostrava	Czech Republic	jana.hanclova@vsb.cz
62.	Adela	Has	University of Josip Juraj Strossmayer in Osijek, Faculty of Economics	Gajev trg 7	31000	Osijek	Croatia	adela.has@efos.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
63.	Michal	Haviar	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	Michal.Haviar@ st.fri.uniza.sk
64.	Niko	Herakovič	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 4	1000	Ljubljana	Slovenia	niko.herakovic@ fs.uni-lj.si
65.	Milan	Hladík	Charles University, Faculty of Mathematics and Physics, Department of Applied Mathematics	Malostranské nám. 25	11800	Prague	Czech Republic	milan.hladik@matfyz.cz
66.	Eloy	Hontoria	Technical University of Cartagena, Business Management Department		30202	Cartagena	Spain	eloy.hontoria@upct.es
67.	Domen	Hudoklin	University of Ljubljana		1000	Ljubljana	Slovenia	domen.hudoklin@ fe.uni-lj.si
68.	Tomi	Ilijaš	Arctur, d.o.o.	Industrijska cesta 15	5000	Nova Gorica	Slovenia	tomi.ilijas@arctur.si
69.	Saša	Jakšić	Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	sjaksic@efzg.hr
70.	Marko	Jakšič	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1000	Ljubljana	Slovenia	marko.jaksic@ ef.uni-lj.si
71.	Marta	Janackova	University of Zilina, Department of Applied Mathematics	Univerzitna 1	01026	Zilina	Slovak Republic	marta.janackova@ fstroj.uniza.sk
72.	Jaroslav	Janáček	University of Žilina, Faculty of Management and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	jaroslav.janacek@ fri.uniza.sk

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
73.	Ľudmila	Jánošíková	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	Ludmila.Janosikova@ fri.uniza.sk
74.	Slawomir	Jarek	University of Economics in Katowice, Department of Operations Research	ul. 1 maja 50	40-287	Katowice	Poland	slawomir.jarek@ ue.katowice.pl
75.	Marina	Jeger	Faculty of Economics in Osijek, J.J. Strossmayer University of Osijek	Trg Ljudevita Gaja 7	31000	Osijek	Croatia	marina@efos.hr
76.	Janja	Jerebic	Faculty of Natural Sciences and Mathematics, University of Maribor	Koroška cesta 160	2000	Maribor	Slovenia	
77.	Eva	Jordan	Kekon d.o.o	Grajski trg 15	8360	Žužemberk	Slovenia	eva.e.jordan@ gmail.com
78.	Elza	Jurun	University of Split, Faculty of Economics Split, Department of Quantitative Methods	Cvite Fiskovića 5	21000	Split	Croatia	elza@efst.hr
79.	Luka	Juvančič	University of Ljubljana, Biotechnical faculty, Department for Animal Science	Groblje 3	1230	Domžale	Slovenia	luka.juvancic@ bf.uni-lj.si
80.	David	Kaljun	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	david.kaljun@fs.uni-lj.si
81.	Tadej	Kanduč	Faculty of Information Studies, Laboratory of Decision Support Systems	Ulica talcev 3	8000	Novo mesto	Slovenia	tadej.kanduc@ fis.unm.si
82.	Pinar	Karagoz	METU Computer Engineering Department			Ankara	Turkey	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
83.	Alenka	Kavkler	Faculty of Economics and Business, University of Maribor, Department of Quantitative Economic Analysis	Razlagova 14	2000	Maribor	Slovenia	alenka.kavkler@ uni-mb.si
84.	Baruch	Keren	Department of Industrial Engineering and Management, SCE - Shamoon College of Engineering	Bialik Sts. 56, P.O. Box 950	84100	Beer Sheva	Israel	baruchke@sce.ac.il
85.	Mirjana	Kljajić Borštnar	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	mirjana.kljajic@ fov.uni-mb.si
86.	Marko	Kobal	Arctur d.o.o.	Industrijska cesta 5	5000	Nova Gorica	Slovenia	
87.	Tina	Kocjančič	University of Ljubljana, Biotechnical faculty, Department for Animal Science	Groblje 3	1230	Domžale	Slovenia	tina.jaklic@bf.uni-lj.si
88.	Michal	Kohání	University of Zilina, Faculty of Management Science and Informatics	Univerzitna 8215/1	01026	Zilina	Slovak Republic	michal.kohani@ fri.uniza.sk
89.	Branko	Kontić	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	branko.kontic@ijs.si
90.	Peter	Korošec	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	
91.	Dujam	Kovač	University of Split, Faculty of Economics	Cvita Fiskovića 5	21000	Split	Croatia	dkovac01@ student.efst.hr
92.	Renata	Kožul Blaževski	University of Split, University Department of Professional Studies	Kopilica 5	21000	Split	Croatia	rkozulb@oss.unist.hr
93.	Jaka	Kranjc	Faculty of Information Studies		8000	Novo mesto	Slovenia	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
94.	Mira	Krpan	University of Zagreb, Faculty of Economics and Business	Trg J. F. Kennedy 6	10000	Zagreb	Croatia	mkrpan@efzg.hr
95.	Miran	Kuhar	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	miran.kuhar@ fgg.uni-lj.si
96.	Nataša	Kurnoga	University of Zagreb, Faculty Economics and Business,	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	nkurnoga@efzg.hr
97.	Janez	Kušar	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	janez.kusar@fs.uni-lj.si
98.	Vladimir	Kuzmanovski	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	vladimir.kuzmanovski@ ijs.si
99.	Marek	Kvet	University of Žilina, Faculty of Security Engineering	1. mája 32	010 26	Žilina	Slovak Republic	marek.kvet@fbi.uniza.sk
100.	Mitja	Lakner	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	mitja.lakner@ fgg.uni-lj.si
101.	Ivana	Lolić	University of Zagreb, Faculty of Economics and Business Zagreb	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	ilolic@efzg.hr
102.	Borut	Lužar	Faculty of Information Studies	Ulica talcev 3	8000	Novo mesto	Slovenia	
103.	Dimitrios	Magos	Department of Informatics, Technological Educational Institute of Athens	Ag. Spyridonos Str.	12210	Egaleo	Greece	
104.	Andrea	Majić	Poslovanje 2 Ltd.	Vijeća Europe 95	32000	Vukovar	Croatia	andrea.majic@gauss.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
105.	Luka	Mali	Faculty of Electrical Engineering, University of Ljubljana		1000	Ljubljana	Slovenia	
106.	Branka	Marasović	University of Split, Faculty of Economics	Cvita Fiskovića 5	21000	Split	Croatia	branka.marasovic@efst.hr
107.	Ivan	Marijanović	University of Split, Faculty of Economics Split	Cvite Fiskovića 5	21000	Split	Croatia	marijanovic2015@efst.hr
108.	Jonathan	Marks-Perreau	ARVALIS, Institut du végétal		CS 30200	Estrées Mons	France	j.marksperreau@arvalisinstitutduvegetal.fr
109.	Jan	Martinovič	VŠB - Technical University of Ostrava, IT4Innovations	17. listopadu 15/2172	708 00	Ostrava	Czech Republic	jan.martinovic@vsb.cz
110.	Lukáš	Melecký	VŠB-Technical University of Ostrava, Faculty of Economics, Department of European Integration	Sokolská třda 33	701 21	Ostrava	Czech Republic	lukas.melecky@vsb.cz
111.	Matej	Mihelčić	Laboratory for Information Systems, Division of Electronics, Ruđer Bošković Institute	Bijenicka cesta 54	10000	Zagreb	Croatia	
112.	Miha	Mlakar	Jožef Stefan Institute, Department of Intelligent Systems	Jamova cesta 39	1000	Ljubljana	Slovenia	miha.mlakar@ijs.si
113.	Ioannis	Mourtos	Department of Management Science & Technology, Athens University of Economics and Business	76 Patission Ave.	104 34	Athens	Greece	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
114.	Nolberto	Munier	Valencia Polytechnic University, INGENIO - Avenue des Tarongers S/N	Avda. Peris i Valero 157	46006	Valencia	Spain	nolmunier@yahoo.com
115.	Athanassios	Nikolakopoul-os	National Technical University of Athens, Department of Chemical Engineering	Heroon Polytechneiou 9	15780	Athens	Greece	nikolako@mail.ntua.gr
116.	Maciej	Nowak	University of Economics in Katowice, Department of Operations Research	ul.1 Maja 50	40-287	Katowice	Poland	maciej.nowak@ ue.katowice.pl
117.	Eva	Oceľáková	Institute of Mathematics, Faculty of Science, P.J. Šafárik University			Košice	Slovak Republic	
118.	Andrei	Orlov	Matrosov Institute for System Dynamics & Control Theory SB RAS	Lermontov str., 134	664033	Irkutsk	Russian Federation	
119.	Irena	Palić	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedy 6	10000	Zagreb	Croatia	ipalic@efzg.hr
120.	Gregor	Papa	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	
121.	Erika	Pastore	Politecnico di Torino	C.so Duca degli Abruzzi 24	10129	Torino	Italy	erica.pastore@polito.it
122.	Polona	Pavlovčič Prešeren	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	polona.pavlovcic@ fgg.uni-lj.si
123.	Martin	Pavlovič	University of Maribor, Faculty of agriculture and life science	Pivola 11	2311	Hoče	Slovenia	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
124.	Viljem	Pavlovič	University of Maribor, Faculty of agriculture and life science	Pivola 11	2311	Hoče	Slovenia	
125.	Karmen	Pažek	University of Maribor, Faculty of agriculture and life science	Pivola 11	2311	Hoče	Slovenia	
126.	Martin	Pečar	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	m.pecar@ijs.si
127.	Mirjana	Pejić Bach	University of Zagreb, Faculty Economics and Business, Department of Informatics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	mpejic@efzg.hr
128.	Tunjo	Perić	University of Zagreb, Faculty Economics and Business, Department of Informatics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	tperic@efzg.hr
129.	Alessandro	Persona	Department of Management and Engineering, University of Padua	Stradella San Nicola 3	36100		Italy	alessandro.persona@ unipd.it
130.	Štefan	Peško	University of Žilina			Žilina	Slovak Republic	stefan.pesko@ fri.uniza.sk
131.	Georgi	Petrov Dimitrov	University of Library Studies and Information Technologies	bul.Tsarigradsko shose 119	1784	Sofia	Bulgaria	geo.p.dimitrov@ gmail.com
132.	Diogo	Pinheiro	Middle East Technical University, Üniversiteler Mahallesi; Institute of Applied Mathematics	Dumlupınar Bulvarı 1	06800	Cankaya Ankara	Turkey	
133.	Snježana	Pivac	University of Split, Faculty of Economics	Cvita Fiskovića 5	21000	Split	Croatia	snjezana.pivac@efst.hr
134.	Tea	Poklepović	University of Split, Faculty of Economics	Cvite Fiskovića 5	21000	Split	Croatia	tea.poklepovic@efst.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
135.	Latif	Pourkarimi	Department of Mathematics, Razi University			Kermanshah	Iran	lp_karimi@yahoo.com
136.	Janez	Povh	Faculty of Information Studies in Novo mesto	Ulica talcev 3	8000	Novo mesto	Slovenia	janez.povh@ fis.unm.si
137.	Andreja	Pucihar	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	andreja.pucihar@ fov.uni-mb.si
138.	Uroš	Rajkovič	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55 a	4000	Kranj	Slovenia	uros.rajkovic@ fov.uni-mb.si
139.	Vladislav	Rajkovič	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55 a	4000	Kranj	Slovenia	vladislav.rajkovic@ fov.uni-mb.si
140.	Lukáš	Rapant	VŠB - Technical University of Ostrava, IT4Innovations	17. listopadu 15/2172	708 00	Ostrava	Czech Republic	lukas.rapantg@vsb.cz
141.	Nada	Ratković	University of Split, Faculty of Economics Split, Department of Quantitative Methods	Cvite Fiskovića 5	21000	Split	Croatia	nada.ratkovic@efst.hr
142.	Benoit	Real	ARVALIS, Institut du végétal		CS 30200	Estrées Mons	France	b.real@ arvalisinstitutduvegetal. fr
143.	Maja	Repnik	University of Maribor, Faculty of Energy Technology	Hočevarjev trg 1	8270	Krško	Slovenia	maja.repnik@um.si
144.	Tomáš	Režnar	VŠB - Technical University of Ostrava, IT4Innovations	17. listopadu 15/2172	708 00	Ostrava	Czech Republic	tomas.reznar@vsb.cz
145.	Marko	Robnik-Šikonja	University of Ljubljana, Faculty of Computer and Information Science	Večna pot 113	1000	Ljubljana	Slovenia	marko.robnik@ fri.uni-lj.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
146.	Blaž	Rodič	Faculty of Information Studies, Laboratory of Decision Support Systems	Ulica talcev 3	8000	Novo mesto	Slovenia	blaz.rodic@fis.unm.si
147.	Lorenzo	Ros-McDonnell	Technical University Cartagena; Mediterranean Institute for Advances Studies (MEDIFAS)	S/ Dr. Fleming, s/n; Mednarodni prehod 6, Vrtojba	30202; 5000	Cartagena; Sempeter pri Gorici	Spain; Slovenia	lorenzo.ros@upct.es
148.	Črtomir	Rozman	University of Maribor, Faculty of agriculture and life science	Pivola 11	2311	Hoče	Slovenia	crt.rozman@um.si
149.	Rebeka	Rudolf	University of Maribor	Slomškov trg 15	2000	Maribor	Slovenia	rebeka.rudolf@uni-mb.si
150.	Zoran	Selinger	Wolfgang Digital Palmerston House	Denzille Lane		Dublin 2	Ireland	zselinger@gmail.com
151.	Fabio	Sgarbossa	Department of Management and Engineering, University of Padua	Stradella San Nicola 3	36100		Italy	fabio.sgarbossa@unipd.it
152.	Kateřina	Slaninová	VŠB - Technical University of Ostrava, IT4Innovations	17. listopadu 15/2172	708 00	Ostrava	Czech Republic	katerina.slaninova@vsb.cz
153.	Majid	Soleimani-Damaneh	School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran & School of Mathematics, Institute for Research in Fundamental Sciences (IPM)	P.O. Box: 19395-5746		Tehran	Iran	soleimani@khayam.ut.ac.ir
154.	Petar	Sorić	University of Zagreb, Faculty of Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	psoric@efzg.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
155.	Michaela	Staníčková	VŠB-Technical University of Ostrava, Faculty of Economics, Department of European Integration	Sokolská třída 33	701 21	Ostrava	Czech Republic	michaela.stanickova@vsb.cz
156.	Marko	Starbek	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	marko.starbek@fs.uni-lj.si
157.	Oskar	Sterle	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	oskar.sterle@fgg.uni-lj.si
158.	Bojan	Stopar	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	bojan.stopar@fgg.uni-lj.si
159.	Galina	Stoyanova Panayotova	University of Library Studies and Information Technologies	bul.Tsarigradsko shose 119	1784	Sofia	Bulgaria	panayotovag@gmail.com
160.	Alexander S.	Strekalovskiy	Matrosov Institute for System Dynamics and Control Theory of SB RAS	Lermontov St., 134	664033	Irkutsk	Russian Federation	
161.	Ervin	Strmcnik	Faculty of Economics, University of Ljubljana	Kardeljeva pl. 17	1000	Ljubljana	Slovenia	strmcnikervin@guest.arnes.si
162.	Alzbeta	Szendreyova	University of Zilina, Department of Applied Mathematics	Univerzitna 1	01026	Zilina	Slovak Republic	alzbeta.szendreyova@fri.uniza.sk
163.	Nataša	Šarlija	Faculty of Economics in Osijek, J.J. Strossmayer University of Osijek	Trg Ljudevita Gaja 7	31000	Osijek	Croatia	natasa@efos.hr
164.	Boško	Šego	University of Zagreb, Department of Mathematics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	bsego@efzg.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
165.	Slavko	Šimundić	University of Split, Faculty of Law	Domovinskog rata 8	21000	Split	Croatia	slavko.simundic@pravst.hr
166.	Tihana	Škrinjarić	University of Zagreb, Department of Mathematics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	tskrinjaric@efzg.hr
167.	Ou	Tang	Linköping University, Department of Management and Engineering Division of Production Economics		581 83	Linköping	Sweden	ou.tang@liu.se
168.	Aneta	Trajanov	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	aneta.trajanov@ijs.si
169.	Nejc	Trdin	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	nejc.trdin@ijs.si
170.	Tadeusz	Trzaskalik	University of Economics in Katowice, Department of Operations Research	ul.1 Maja 50	40-287	Katowice	Poland	tadeusz.trzaskalik@ue.katowice.pl
171.	Tea	Tušar	Jožef Stefan Institute, Department of Intelligent Systems	Jamova cesta 39	1000	Ljubljana	Slovenia	tea.tusar@ijs.si
172.	Aleksander	Vesel	University of Maribor; Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	vesel@uni-mb.si
173.	Jelena	Vidović	University of Split, University Department of Professional Studies	Kopilica 5	21000	Split	Croatia	jvidovic@oss.unist.hr
174.	Ilko	Vrankić	University of Zagreb, Faculty of Economics and Business	Trg J. F. Kennedy 6	10000	Zagreb	Croatia	ivrankic@efzg.hr
175.	Vida	Vukašinović	Jožef Stefan Institute	Jamova cesta 39	1000	Ljubljana	Slovenia	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
176.	Gerhard Wilhelm	Weber	Middle East Technical University, Üniversiteler Mahallesi; Institute of Applied Mathematics	Dumlupınar Bulvarı 1	06800	Cankaya Ankara	Turkey	gweber@metu.edu.tr
177.	Shuoguo	Wei	Linköping University, Department of Management and Engineering Division of Production Economics		581 83	Linköping	Sweden	
178.	Houmin	Yan	Department of Management Sciences, College of Business	The City University of Hong Kong			Hong Kong	cbhyan@cityu.edu.hk
179.	Ozen	Yavas	Istanbul Technical University, Industrial Engineering Department	ITU Isletme Fakultesi, Macka	34357	Istanbul	Turkey	oznyavas@gmail.com
180.	Omer Faruk	Yilmaz	Istanbul Technical University, Industrial Engineering Department	Macka	34367	Istanbul	Turkey	ofyilmaz@itu.edu.tr
181.	Tulug Figen	Yilmaz	METU Computer Engineering Department			Ankara	Turkey	
182.	Lidija	Zadnik Stirn	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	lidija.zadnik@bf.uni-lj.si
183.	Marijana	Zekić-Sušac	University of Josip Juraj Strossmayer in Osijek, Faculty of Economics	Gajev trg 7	31000	Osijek	Croatia	marijana@efos.hr
184.	Darko	Zelenika	Faculty of Information Studies, Laboratory of Data Technologies	Ulica talcev 3	8000	Novo mesto	Slovenia	darko.zelenika@ fis.unm.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
185.	Jovana	Zoroja	University of Zagreb, Faculty Economics and Business, Department of Informatics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	jzoroja@efzg.hr
186.	Giulio	Zotteri	Politecnico di Torino	C.so Duca degli Abruzzi 24	10129	Torino	Italy	giulio.zotteri@polito.it
187.	Hugo	Zupan	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 4	1000	Ljubljana	Slovenia	hugo.zupan@fs.uni-lj.si
188.	Bernard	Ženko	Jožef Stefan Institute, Department of Knowledge Technologies	Jamova cesta 39	1000	Ljubljana	Slovenia	bernard.zenko@ijs.si
189.	Janez	Žerovnik	University of Ljubljana, Faculty of Mechanical Engineering; Institute of Mathematics, Physics and Mechanics	Aškerčeva 4; Jadranska 19	1000	Ljubljana	Slovenia	janez.zerovnik@ fs.uni-lj.si
190.	Jaka	Žgajnar	University of Ljubljana, Biotechnical faculty, Department for Animal Science	Groblje 3	1230	Domžale	Slovenia	jaka.zgajnar@bf.uni-lj.si
191.	Berislav	Žmuk	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	bzmuk@efzg.hr