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.

Proceedings SOR'09

Proceedings of the 10th International Symposium on OPERATIONAL RESEARCH

SOR '09

Nova Gorica, Slovenia
September 23-25, 2009

Edited by:
L. Zadnik Stirn • J. Žerovnik • S. Drobne • A. Lisec

SOR '09 Proceedings

The 10th International Symposium on Operational Research in Slovenia

Nova Gorica, SLOVENIA, September 23 - 25, 2009

Edited by:

L. Zadnik Stirn, J. Žerovnik, S. Drobne and A. Lisec

Slovenian Society Informatika (SDI)

Section for Operational Research (SOR)

Proceedings of the 10th International Symposium on Operational Research
SOR'09 in Slovenia, Nova Gorica, September 23 - 25, 2009.

Organiser : Slovenian Society Informatika – Section for Operational Research, SI 1000 Ljubljana, Vožarski pot 12, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

Under the auspices of the Slovenian Research Agency

First published in Slovenia in 2009 by Slovenian Society Informatika – Section for Operational Research, SI 1000 Ljubljana, Vožarski pot 12, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

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

519.8(082)

INTERNATIONAL Symposium on Operational Research in Slovenia (10 ; 2009 ; Nova Gorica)
SOR '09 proceedings / The 10th International Symposium on Operational Research in Slovenia, Nova Gorica, Slovenia, September 23-25, 2009 ; [organiser Slovenian Society Informatika, Section for Operational Research] ; edited by L. Zadnik Stirn ... [et al.]. - Ljubljana : Slovenian Society Informatika (SDI), Section for Operational Research (SOR), 2009

ISBN 978-961-6165-30-3
1. Zadnik Stirn, Lidija 2. Slovensko društvo Informatika. Sekcija za operacijske raziskave
247453440

Proceedings of the 10th International Symposium on Operational Research in Slovenia (SOR'09) 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 : Studio LUMINA and Samo Drobne
Printed by : Birografika BORI, Ljubljana, Slovenia

Preface

Operations Research comprises a large variety of mathematical, statistical and informational theories and methods to analyze complex situations and to contribute to responsible decision making, planning and the efficient use of the resources. In a dynamic and changing world of increasing complexity and scarce natural resources, in particular during the financial crisis, there is a growing need for such approaches in many fields of our society.

The 10th International Symposium on Operations Research, called SOR'09, which was held in Nova Gorica, Slovenia, from September 23 through September 25, 2009, was the tenth event in the traditional biannual series of international conferences organized by Slovenian Society INFORMATIKA, Section of Operational Research.

As traditionally, also SOR'09 was an international forum for scientific exchange at the frontiers of Operations Research in mathematics, statistics, economics, engineering, education, environment and computer science. We believe that the presentations reflected the state of the art in Operations Research as well as the actual challenges. Besides contributions on recent advances in the classical fields, the presentations on new interactions with related fields as well as an intense dialogue between theory and the numerous applications, were delivered at the symposium. The program consisted of parallel sessions with contributed talks selected and by several plenary lectures given by invited speakers.

The first part of the Proceedings of SOR'09 includes invited papers or extended abstracts of invited lectures, presented by 7 prominent scientists: Zoran Babić, University of Split, Faculty of Economics, Split, Croatia; Tibor Csendes, University of Szeged, Institute of Informatics, Szeged, Hungary; Robert W. Grubbström, Linköping Institute of Technology, Linköping, Sweden; Marion Rauner, University of Vienna, Faculty of Business, Economics and Statistics, Vienna, Austria; Moshe Sniedovich, University of Melbourne, Department of Mathematics and Statistics, Australia; Tadeusz Trzaskalik, The Karol Adamiecki University of Economics in Katowice, Department of Operations Research, Katowice, Poland and Di Yuan, Department of Science and Tehnology, Linköping University, Norrköping, Sweden.

The second part of the Proceedings includes 54 papers written by 105 authors and co-authors. These papers were accepted after a review process carried out by the members of the Program Committee assisted by a few additional reviewers appointed by the Committee members. Most of the authors of the contributed papers came from Slovenia (48), others are from Croatia (29), Serbia (5), Slovak Republic (5), the Netherlands (5), Germany (4), Poland (3), United Kingdom (3), Austria (2), Spain (2), Sweden (2), Australia (1), Bosnia and Herzegovina (1), Hungary (1), and Russia (1). The contributed presentations were divided into the following sections (the number of papers in each section is given in parentheses): Discrete Mathematics and Optimization (9), Multicriteria Decision Making (6), Production and Inventory (5), Scheduling and Control (5), Finance and Investments (5), Location and Transport (6), Environment and Human Resources (6), OR Perspectives (2), Statistics (10). The contributed papers in this volume appear in alphabetical order of the first author within the section.

The Proceedings of previous nine International Symposia on Operations Research organized by Slovenian Section of Operations Research are cited in the following secondary and tertiary publications: Current Mathematical Publications, mathematical Review, MathSci, *Zentralblatt fuer Mathematik/Mathematics Abstracts, MATH on STN International, CompactMath, INSPEC. This volume is expected to be cited in the same publications. For the first time, a special issue of the Central European Journal of Operational Research with selected full length reviewed papers based on presentations given at the symposium will be published after the conference.*

The 10th International Symposium on Operations Research SOR'09 stood under the auspices of the Slovenian Research Agency, and was granted by sponsors cited in these Proceedings. The opening address was given by Prof. Dr. L. Zadnik Stirn, the President of the Slovenian Section of Operations Research, Mr. N. Schlamberger, the President of Slovenian Society INFORMATIKA, Mr. B. Nemec, M.Sc., the chair of the SOR'09 Organizing Committee and the representative of HIT, Nova Gorica, and representatives of different professional institutions and Operations Research Societies from other countries.

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 Slovenian Section of Operations Research. Many thanks to them. Further, we would like to express our deepest gratitude to the members of the Program and Organizing Committees, to the reviewers, chairpersons, to the sponsors, especially HIT, Nova Gorica and Austrian Science and Research Liaison Office, Department of Ljubljana, and to all the numerous people - far too many to be listed here individually - who helped in carrying out The 10th International Symposium on Operations Research SOR'09 and in putting together these Proceedings. At last, we appreciate the authors' efforts in preparing and presenting the papers, which made The 10th Symposium on Operational Research SOR'09 successful. The success of the appealing scientific events at SOR'09 and the present proceedings should be seen as a result of our joint effort.

Nova Gorica, September 23, 2009

Lidija Zadnik Stirn
Janez Žerovnik
Samo Drobne
Anka Lisec
(Editors)

Foreword

Symposium SOR'09 is tenth in a row and as such offers an opportunity to reflect on the history of operations research, to touch its present achievements, and to see what is in it in the future. Operations research, although according to some less orthodox views which place it further back in history, being no older than fifty years is one of more recent mathematical disciplines that is more and more important and acknowledged as an independent field of research and application. It is namely disputable if isolated results that would be today classified into operations research could be attributed to it as they were supported neither by its theory nor with its methods. The new branch of mathematics which started as an effort to solve purely practical problems of logistic nature is today generally recognized as an important and independent science. Ties with mathematics that still exist are theories and methods and it is likely to remain so in the future. Operations research has produced methods and theories of which probably the most famous and applied to very different activities is theory of games; however, the importance and scope of operations research go much further than that. Methods such as modeling and simulation that have been used in different branches of engineering in the past have been thanks to operations research provided with a firm theoretical basis which upgraded them from a matter of choice based on feeling and guesswork to an indispensable and objective problem solving tool where computer as an instrument is a must.

A proof for such statement is the SOR'09 Proceedings. A few papers deal with development of theory and methods which is understandable. Theories are rather few as to be produced they require much time and hard work if they are to be consistent and impeccable. Methods are more diverse and numerous as they are about application of theories to various subjects. Most of the papers therefore deal with applications of theory and methods to problems in most various fields: manufacturing, finance, business, sociology, politics, economy (including transition economy), environment, management, disasters, logistics, education, tourism, transportation, employment, livestock, medicine, property, wood industry, by which however the list is far from exhausted. New problems arise daily in all fields of human endeavor. The common denominator of operations researchers are solutions arrived at through operations research, its theories, theorems, and methods. That looks like not very much but there is more than meets the eye. Problems that first need to be clearly formulated do not come just from the above fields but same or similar appear elsewhere. The solutions thereof that are presented as accomplishments of today will be sooner or later a part of commodities or services in general use. All users will benefit from such solutions without even realizing it. Does today anyone thinks of operations research when for example sitting in a car that corrects mistakes of the driver; when taking a medicine; when a courier service delivers a shipment in an astonishingly short time?

While new problems and challenges arise by the minute, resources – be they private, public or natural - are final and limited. A consequence is that users of services and commodities require best value for their money and that suppliers try to deploy their resources to develop best services and goods. The bottom line is competitiveness: buyers expect best value for their money whereas suppliers must use their resources most efficiently. If buyers can rely on market as their supporting mechanism, provided that supply is ample enough, suppliers can no more rely upon guesswork and intuition. Mistakes are too expensive and present too big a risk not only for a correctness of a solution but even for the existence of the supplier. In the modern world where quality and price are main criteria for competitiveness it would *be naïve to rely on a second chance. This is just the element that emphasizes the role of operations research today and forecasts it's even more important role in the future. The biannual scientific symposium is a proof that we believe in the future of operations research. It is also an opportunity to express an appreciation and our gratitude to our colleagues that have had several years ago the vision, energy and courage to make the first step – SOR'93.*

Nova Gorica, September 23, 2009

Niko Schlamberger
(President of Slovenian Society INFORMATIKA)

On the Occasion of the 10th OR Symposium in Slovenia

The present symposium on Operations Research (OR), called The 10th International Symposium on Operations Research, SOR'09, represents the jubilee symposium in the series of international OR meetings organized by Slovenian Section of Operational Research (SOR). As all nine previous OR symposia were characterized as successful in the scientific as well as organizational sense, the members of SOR can not only be proud of their past activities, but are also obliged to use this moment for the analysis of the past achievements and future perspectives.

SOR was established in December 1992 and has today 77 members, mostly coming from universities, (research) institutes as well as from management (firms). The goal of SOR is to pursue, support and facilitate research, development, application and education in the OR area where mathematics, economics, computer science, statistics, environmental economics and system theory as well as some other disciplines come together. Therefore, interdisciplinarity, and applied science of OR are one of the main concerns of the SOR. Thus, SOR is a forum for scientists and practitioners in all areas of OR and neighbouring fields, across the disciplines and programs in resource management, networks, tools, information and education. The main concern of SOR is to increase visibility and influence, especially closer collaboration with educational institutions (universities) and practice (industry), international cooperation and publicity.

In view of these objectives, since 1993 SOR has been organizing international symposia on OR, known as SOR'93, SOR'94, SOR'95, SOR'97, SOR'99, SOR'01, SOR'03, SOR'05, SOR'07, and SOR'09. At first, these symposia were organized annually, later, after the reached agreement with the Croatian Operational Research Society, they were organized biannually, each year in one country, alternating Slovenia and Croatia. OR symposia in Slovenia consist of plenary sessions, where several (usually 6 or 7) prominent scientists present invited papers, and of about 60 contributed papers in sessions: networks, stochastic and combinatorial optimization, algorithms, multicriteria decision making, scheduling and control, location theory and transport, environment and human resource management, duration models, finance and investment, production and inventory, education, statistics, OR communications, etc. The presented papers are reviewed by two independent reviewers according to the international standards and are published in the Proceedings. The editors, leading members of SOR, are proud of all ten SOR Proceedings, especially of the last five, which are cited in the international publication basis, like INSPEC, MathSci, etc. (see the list of SOR Proceedings on the back cover page of this SOR'09 Proceedings).

*SOR has published three monographs, the last two in English: ***Rupnik, V., Teorija faktorjev integrabilnosti gospodarstva in njihovo praktično modeliranje, SDI-SOR Series, No. 1, Ljubljana, Slovenia, 1996, 720 pages; *** Solutions to production problems, Rupnik V., Zadnik Stirn, L., Drobne, S. (eds.), SDI-SOR Series, No. 2, Ljubljana, Slovenia, 2000, 290 pages; *** Selected decision support models for production and public policy problems, Zadnik Stirn, L., Indihar Štemberger, M., Ferbar, L., Drobne, S. (eds.), SDI-SOR Series, No. 3, Ljubljana, Slovenia, 2005, 243 pages.*

Since 1999 members of SOR have been co-editing Central European Journal of Operations Research (CEJOR). Further, SOR was in the year 2007 accepted as the 49th member of IFORS, and in 2008 as the 30th member of EURO.

Over the years, the SOR activities, conferences, publications and international contacts have not only grown in volume, but have also expanded into new ideas, new areas of OR theory and applications. Thus, the future objectives of SOR are to develop and promote the theory of OR, management science, decision analysis and computer science, as well their applications in: business, management, economics and finance, engineering, natural resources and environment management, energy planning, e-business and e-government, accounting and auditing, computer science and information technology, human resource management, information systems management, logistics and supply chain management, marketing, operations management, risk analysis and management, telecommunications, transportation, in general principles of practice such as problem structuring for multicriteria, dynamic and fuzzy problems, implementation of group decision and negotiation support systems, etc. Additionally, SOR plans to continue with the organization of biannual international conferences on OR in agreement with Croatian OR Society (alternating Slovenia and Croatia), with editing and publishing conference proceedings and monographs, with collaboration on a bilateral basis with Austrian, German, Polish, Czech, Hungarian, Slovak, etc., societies, with active participation in IFORS, EURO and IFIP TC7, with incorporation into broader community, with taking care for the education in the field of OR, and finally taking care for dissemination of OR achievements and OR terminology.

In this sense, the main aim of SOR is to present the insights on the latest OR theoretical developments, to direct the methodologies, models and techniques, to deal with decision making problems in various fields of applications, and to increase the awareness of importance regarding the exploitation of the OR models.

We hope that these plans, ideas and future perspectives are the principal factors of SOR power, and the base for SOR future accomplishments.

It is with great sadness to inform you that two distinguished members of SOR are no longer with us. This summer, Prof. DDr. Ludvik Bogataj in Prof. Dr. Stanislav Indihar, two internationally acclaimed and respected scientists in the field of OR left SOR forever. They were excellent and influential members of SOR, the founding members of SOR, regular and active participants and organizers of SOR events, including SOR'09, both with great aspirations for SOR future activities, and above all, great friends. SOR members are grateful to them for their enormous contribution to OR. We will miss them in our future endeavors; we will try to accomplish their ideas, and we will keep them in our minds forever.

At this point, many thanks to the SOR members, all OR colleagues and friends over the world who helped to organize all ten OR symposia in Slovenia, presented papers, discussed at sessions, published their scientific OR achievements in SOR publications, promoted and supported SOR in many different ways, who believed in SOR activities, and shared their knowledge and ideas. There are so many that it is impossible to list them here. Their names appear on separate pages of SOR publications. Here, we acknowledge Mr. Samo Drobne, M.Sc. (cofounding member of SOR, secretary of SOR, co-organizer of all ten symposia, co-editor and technical editor of SOR proceedings and monographs, writer of numerous OR papers in SOR publications, etc.) for his invaluable contribution to SOR deeds.

Last but not least, I wish great scientific success to SOR'09, and an enjoyable stay to the participants so that everyone will leave the symposium feeling personally enriched, and I express hope for SOR further prosperity in the general sense of the word.

Nova Gorica, September 23, 2009

Lidija Zadnik Stirn
(President of Section of Operational Research)

In memoriam
Ludvik Bogataj (1949-2009)

On August 23, at the age of 60, Prof. DDr. Ludvik Bogataj died from a sudden heart attack. He was a professor of mathematics at the Faculty of Economics, University of Ljubljana and a prominent member of several international professional associations in the field of operations research.

At the University of Ljubljana he graduated in mathematics in 1972 and then in 1975 took his M. Sc. degree in operations research and a Ph. D. in operations research in 1979. He also took his M. Sc. degree in mathematics at the University of Zagreb in 1985 and in 1990 a Ph. D. in mathematics at the University of Ljubljana.

Starting with his dissertations, the thread of Prof. Bogataj's research concerns translations in time, in particular time delays. In the problem of optimal control, delays may occur in the state or in the control. Prof. Bogataj derived some new results concerning the general case, when delays occur both in the state as well as in the control, in particular he obtained some estimates of the influence of certain perturbations on the objective functional.

The central interest of Prof. Bogataj's research was the MRP (Material Requirenment Planning) theory, as developped by R. W. Grubbström. Prof. Bogataj with his associates extended this theory and interconnected it with the decision theory and game theory; various stochastic issues were also considered. In recent time the research of Prof. Bogataj focused on extending the MRP theory to the distribution part of the supply chain. In this context he addressed also the issue of reverse logistics.

These results were published in SCI-indexed journals and in numerous proceedings of international conferences and symposia on operations research. Prof. Bogataj also had invited lectures at numerous conferences and universities.

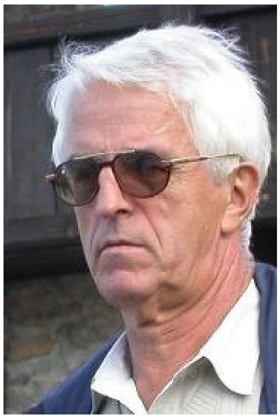
Comparable to his research achievements was also his engagement in professional associations, particularly those in the field of operations research. He was a founding member of the Section for Operations Research of Slovenian Society Informatika and from the start a member of the Program Committee of SOR. He served as chair of various symposium sessions and, of course, also participated with his contributions.

Prof. Bogataj was also a founding member of ISIR (International Society for Inventory Research) and a regular and active participant at their meetings. In 1998 he and his wife, Prof. Dr. Marija Bogataj, who was all the time his closest associate, were presented the ISIR Service Award for "being regular participants of the ISIR Symposia, with high level papers, organizing a successful ISIR Summer School, followed with a two volume publication, and serving as chairpersons and referees at most Symposia." Furthermore, Prof. Bogataj was a Programme Committee Member of the International Working Seminars on Production Economics held in Igls, Innsbruck and an associate editor of the Central European Journal of Operations Research.

We will remember him with admiration.

Dušan Hvalica

In memoriam
Stanislav Indihar (1941-2009)



In this summer night of July 23 Prof. Dr. Stanislav Indihar quietly passed away.

He was born in a small village Žimarice near Ribnica in Slovenia on August 9 of the year 1941. He attended primary school in the nearby village Sodražica. Even then, he was very interested in mathematics. As a high school student in Kočevje he actively participated in the Mathematical-physical journal, then the central Yugoslavian mathematical journal for younger students. He studied mathematics in Ljubljana and graduated in 1965. He became the assistant of Mathematics at the Faculty of Mechanical Engineering in Ljubljana.

In 1970 he moved with his family to Maribor and began teaching at the Economic and Commercial College in Maribor. In 1975, he received the doctorate degree in mathematical sciences at the University of Ljubljana. In 1985 he became a full professor of mathematics at the Faculty of Economics and Business of the University of Maribor, where he worked until his retirement in spring 2007.

Scientific work and research of Prof. Dr. Indihar covers a wide range of areas and activities. In conjunction with the doctoral thesis his early work was devoted to problems of mathematical programming, first bilinear and later multilinear. With his scientific work which was characterized by rigorous derivations of results Prof. Dr. Indihar early joined the international scientific community. His main results include combinatorial methods for bilinear programming and convex biquadratic maximization. His later scientific work concerns the analysis of forecasting methods and their critical evaluation. With his work Prof. Dr. Indihar successfully demonstrated the importance of the mathematical methods and operations research for the analysis of economic problems.

Most of his results were presented at numerous international scientific conferences and symposia on operations research. He was a member of various professional associations such as the Slovenian mathematical society, the American mathematical society, the German and Austrian society of operations research. He was a member and cofounder of the Slovenian society for operations research. He actively participated with papers, was a chair of different sessions and was a member of the program committee at all of the international symposia on Operations research organized in Slovenia (SOR) by the Slovenian society of operations research.. This year, he was again a member of the program committee of the tenth international symposium on operations research (SOR09). Unfortunately the untimely death prevented him to be an active participant at the SOR09. Since the year 2007 Prof. Dr. Indihar was an honourable member of Slovenian society for operations research. He was a member of the editorial board of the Central European Journal for Operations Research since 1999.

We will miss Prof. Dr. Indihar at future events of Slovenian society for operations research. However his thoughts and deeds will remind us of his personality also in the future.

Miklavž Mastinšek

Sponsors of the 10th International Symposium on Operational Research in Slovenia (SOR'09)

The following institutions and companies are gratefully acknowledged for their financial support of the symposium:

- *Austrian Science and Research Liaison Office, Department of Ljubljana, Ljubljana*
- *Faculty of Information Studies, Novo mesto*
- *HIT, d. d. Nova Gorica*
- *Slovenian Research Agency, The Republic of Slovenia*

Contents

Section II: Multicriteria Decision Making 137

Section III: Production and Inventory 195

Section IV: Scheduling and Control 247

Section V: Finance and Investments 299

Section VI: Location and Transport 345

Section VII: Environment and Human Resources 405

Author index

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Plenary Lectures

DE NOVO PROGRAMMING – A BLUFF OR NOT A BLUFF

Zoran Babić
Faculty of Economics
Matice hrvatske 31
Split – Croatia
e-mail: babic@efst.hr

Abstract: De Novo presents a special approach to optimization. Instead of "optimizing a given system" it suggests a way of "designing an optimal system". The basis of the standard approach is a production model with resources defined in advance, so that the constraints and the feasible set are fixed. However, de Novo approach does not limit the resources as most of the necessary resource quantities can be obtained at certain prices. Resources are actually limited because their maximum quantity is governed by the budget which is an important element of de Novo. Using de Novo approach most various cases can be handled more effectively than by the standard programming models. Changes in prices, technological coefficients, increasing costs of raw materials, quantity discounts and other similar and real production situations can be easily incorporated in de Novo model and provide very satisfactory solutions.
In the area of multi-criteria programming de Novo also offers a variety of new ideas and approaches. Nevertheless, in spite of numerous possibilities offered by de Novo programming, in the last two decades only a small number of authors have used it in their work. This paper will attempt to present the advantages of de Novo approach.

Keywords: De Novo programming, multi-criteria programming, multiple prices

1 INTRODUCTION

More than twenty years ago Milan Zeleny, one of the founders of multicriteria programming and decision making, introduced a new concept into mathematical programming – the so called De Novo programming [20]. Since then De Novo programming has been being developed, but at a quite slow pace. Most studies and new ideas related to De Novo come from the Far East (China, Korea, Taiwan), whereas the few studies and fresh ideas in Europe and the USA come mainly from researchers originating from the Far East (see [7] – [18], [26]). The reason for that must be in the fact that Professor Zeleny was for a long time working and lecturing in these countries thus leaving a deep trace of his research. Nevertheless, Zeleny has not abandoned the idea of De Novo programming, but has recently included it (single and multicriteria approach) into the eight concepts of optimization ([23], [24], [25]) among which classic optimization is only a special case.

What is the reason for the relatively weak interest in De Novo programming? We may wonder whether there is a mistake or a wrong hypothesis built in the approach. My intention, and the goal of this paper, is not to show you anything new about De Novo. Although most of you have heard a lot or at least something about De Novo, my goal is to popularize this, in my opinion, very sophisticated and useful way of decision making and to popularize De Novo programming among researchers and students that up to now have not heard at all, or have heard only little about it.

The purpose of “system optimization” cannot be just to improve the performance of a given, pre-configured system (search for efficiency), but rather to find the best system configuration, itself, i.e., design an optimal system search for effectiveness). System boundaries in real world are not conveniently given and fixed a priori, but are themselves to be determined in an optimal way. The message is clear: optimizing a “badly” configured system is ineffective and wasteful while optimally designed systems hardly need any further “optimization” [21].

In real world problems technological coefficients and parameters are not precisely known. Due to this uncertain nature, the fuzzy set theory can also be used in De Novo programming problems. In that way uncertain parameters can be represented as fuzzy, interval, or stochastic numbers, resulting in uncertain De Novo programming. Li and Lee [11] extended Zeleny's basic method to identify fuzzy system designs for De Novo problems by considering the fuzziness in coefficients. In 1993 the same authors [12] treated fuzzy goals and fuzzy coefficients simultaneously, depending on a numerical approach which could be solved as either linear or nonlinear problems. Later, in 2006, Chen and Hsieh [7] presented a fuzzy multi-stage De Novo programming which was viewed as a fuzzy dynamic programming problem. Finally, in 2009 Zhang at al. [26] present an interval De Novo programming approach for the design of optimal water resources-management systems under uncertainty. The model is derived by incorporating the existing interval programming and De Novo programming, allowing uncertainties represented as intervals within the optimization framework. This approach is then applied to design an inexact optimal system with budget limit instead of finding the optimum in a given system with fixed resources in the water resources planning case.

One of the important issues in multi-criteria De Novo programming is to determine an optimum-path ratio for enforcing a particular budget level of resources so as to establish the optimal system design. In Zeleny's basic method, the optimum-path ratio $r^* = B/B^*$ (see [20]) for achieving the multi-criteria performance of the ideal system design related to a given budget level (B) was used to determine the optimal system design. Shi [13] defines six possible types of optimum-path ratios for finding optimal system designs in the multi-criteria De Novo programming. Each optimum-path ratio is chosen to enforce a given budget level to reach a realized performance of the system design. As optimum-path ratios are different, one can establish different optimal system designs as options for the decision maker.

A few years later Shi [14] proposes an approach of multi-criteria De Novo programming to deal with system designs problem with multiple decision makers and a possible debt. Namely, in many real world situations, an important decision has to be made by a group of responsible people such as the members of the board of directors or trustees. A framework of multiple decision makers thus should be considered in the system design process. Because of investment and cash flow limitations, a firm may not always have an adequate budget for production. When a possible budget deficit or debt occurs at the design time, the designed system may become infeasible and Shi proposes a way how to overcome these difficulties.

In 1999 Taguchi and Yokota [17] formulated a De Novo nonlinear integer programming problem of system reliability with interval coefficients. It is used for estimating and designing optimal reliability of an incomplete fault detecting and switching (FDS) system. They discuss and compare the efficiency between the proposed method and the former one.

In their paper at the Innovative Computing Conference in 2007, Yu and Wang [18] used multi-criteria De Novo programming in one possible model of portfolio selection problem. In order to eliminate the trade-off between return and risk, the concept of De Novo programming is applied which can show the least budget needed to achieve these two conflicting goals at the same time.

2 SINGLE CRITERION DE NOVO PROGRAMMING

The traditional resource allocation problem in economics is modeled via standard criterion linear programming formulation of the single-objective product-mix problem as follows (see Tabucanon [16]):

$$\text{Max } z = c_1 x_1 + c_2 x_2 + \ .....\ + c_n x_n \qquad (1)$$

$$\begin{aligned} \text{s.t.} \quad & a_{11} x_1 + a_{12} x_2 + \ .....\ + a_{1n} x_n \le b_1 \\ & a_{21} x_1 + a_{22} x_2 + \ .....\ + a_{2n} x_n \le b_2 \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (2) \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & a_{m1} x_1 + a_{m2} x_2 + \ .....\ + a_{mn} x_n \le b_m \end{aligned}$$

$$x_j \ge 0 \ , j = 1, 2, ..., n \qquad (3)$$

where

x_j, (j = 1, 2, ..., n) - production levels of n products;
z - value of product mix which has to be maximized;
c_j - objective function coefficients;
a_{ij} - level of usage of resource i by activity j;
b_i - (i = 1, 2, …, m) – given availability of resource i.

Because all components of $b = (b_1, b_2, \ldots, b_m)$ are determined a priori, the problem (1) – (3) deals with the optimization of a given system.

In De Novo formulation the purpose is to design an optimal system and the following formulation is of interest:

$$\text{Max } z = c_1 x_1 + c_2 x_2 + \ .....\ + c_n x_n \qquad (4)$$

$$\begin{aligned} \text{s.t.} \quad & a_{11} x_1 + a_{12} x_2 + \ .....\ + a_{1n} x_n = \beta_1 \\ & a_{21} x_1 + a_{22} x_2 + \ .....\ + a_{2n} x_n = \beta_2 \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (5) \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & a_{m1} x_1 + a_{m2} x_2 + \ .....\ + a_{mn} x_n = \beta_m \end{aligned}$$

$$p_1 \beta_1 + p_2 \beta_2 + \ .....\ + p_m \beta_m \le B \qquad (6)$$

$$x_j, \beta_i \ge 0 \ , j = 1, 2, ..., n; \ i = 1, 2, ... , m \qquad (7)$$

where

$\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ – set of decision variables representing the level of resource
i to be purchased,
p_i - unit price of resource i,
B - total available budget for the given system.

Now the problem is to allocate the budget so that the resulting portfolio of resources maximizes the value of the product mix (with given unit prices of m resources, and with given total available budget).

The main difference of the two models lies in the treatment of the resources which become decision variable β_i in the De Novo formulation. The model determines the best mix of not only the output, but also the combination of inputs still to be acquired. In that way it is a design of an optimal system as against the optimization of a given system.

Solving the De Novo model can be made simpler by the substitution of equations (5) into the budget equation (6). Given the market price, p_i for resource i, we have

$$p_1 a_{1j} + p_2 a_{2j} + \ .....\ + p_m a_{mj} = v_j, \quad j = 1, 2, ..., n,$$

or

$$v_j = \sum_{i=1}^{m} p_i a_{ij} \ , j = 1,2,....,n \qquad (8)$$

where

a_{ij} - level of usage of resource i per unit of activity j;
p_i - unit price of resource i.

Therefore, v_j represents the unit variable cost of producing product j.

Using that we can reformulate the De Novo model into:

$$\text{Max } z = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n \quad (9)$$

$$\text{s.t.} \quad v_1 x_1 + v_2 x_2 + \ldots + v_n x_n \leq B \quad (10)$$

$$x_j \geq 0, \; j = 1, 2, \ldots, n \quad (11)$$

If there are no other constraints, solving the simplified model is easy with one objective and one constraint involved. The solution procedure would be:

(i) Find $\underset{j}{Max} \; (c_j / v_j)$

The ratio c_j / v_j represents the profitability of product j (if the objective function is to maximize profit). This step searches for the most profitable product.

(ii) For $\underset{j}{Max} \; (c_j / v_j)$, say (c_k / v_k) corresponding to x_k, the amount of x_k to be produced would be $x_k{}^* = B / v_k$.

This implies that all of the resources will be used to produce the most profitable product x_k, the amount of which will be dictated only by the budget in the absence of any other constraints.

If demand limits apply for product k, then we produce product k in such amount that it does not exceed the demand limit, or the maximum allowed by the budget, and then go to the next profitable product until the budget is used up.

Optimal solution for the resources (β_i) to be purchased can be easily found from relations (5).

3 INCREASING COST OF RAW MATERIALS

The linear programming model is sometimes difficult to apply in real business situations due to its assumption of proportionality. A frequent phenomenon arising in practice is the varying price of the same resource. Namely, if a company needs additional quantities of raw materials it is possible to buy them from another supplier but at a different (usually higher) price. Let us assume that i raw material can be purchased at the price p_i, but only for the quantity lower than Q. To purchase i raw material above that quantity it is necessary to take another supplier whose price is $p_i' > p_i$. Then the relation for the i raw material is transformed into:

$$a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{in} x_n = b_i + d_i, \quad (12)$$

with additional constraint $b_i \leq Q$, where d_i is the additional quantity of the i raw material with the unit price p_i'.

Since the same raw material has different price variable, the income from end product unit is not constant anymore. Therefore, maximizing the sum of $c_j x_j$, would not be an accurate measure of net income. Net income equation (6) should be recalculated as the difference between sales and total cost of materials, where the objective function will include materials at both prices. Consequently, if s_j is the sales price of j product, the objective function has the following form:

$$\text{Max } z = \sum_{j=1}^{n} s_j x_j - \sum_{i=1}^{m} p_i b_i - \sum_{k=1}^{l} p_k' d_k \qquad (13)$$

where d_k (k = 1, ...,l) are those materials which in additional quantities can be bought only at a higher price (p_k').

In the budget equation (10) it is also necessary to introduce costs for additional quantities of raw materials, so that it now takes the following form:

$$\sum_{i=1}^{m} p_i b_i + \sum_{k=1}^{l} p_k' d_k \leq B \qquad (14)$$

There is no need to specify that b_i should reach the maximum value of Q first, before allowing d_i greater than zero. The optimization model ensures b_i reaching the maximum value of Q because of the lower penalty, i.e. lower price p_i.

4 QUANTITY DISCOUNTS

Let us now consider such production situation when there are quantity discounts granted for bulk orders of raw materials. Therefore, in addition to the increasing cost effect we have to introduce this possibility into the model. Let us assume that for the k resource (b_k) the valid price is p_k as long as the purchased quantity is below Q, and the discounted price p_k' is valid for the entire quantity if the purchased quantity is higher than Q. Consequently the assumption is opposite to the one in the previous model, i.e. $p_k' < p_k$.

The previous formulation is not applicable since the optimization model will prefer using the less expensive material without satisfying the quota (Q). A different model has to be formulated with a slightly more complicated procedure.

Let

- b_k, p_k - the amount and price of k raw material if it is purchased at less than the quantity discount volume;
- d_k, p_k' - the amount and price of k raw material if it is purchased at the quantity discount.

The new model, in that case, instead of one equation for k raw material has two relations, and those are:

$$a_{k1} x_1 + a_{k2} x_2 + \ ..... \ + a_{kn} x_n \leq b_k + M y_1 \qquad (15)$$
$$a_{k1} x_1 + a_{k2} x_2 + \ ..... \ + a_{kn} x_n \leq d_k + M y_2 \qquad (16)$$

and, according to this, two relations as budget constraints:

$$p_1 b_1 + p_2 b_2 + ...+ p_k b_k + ... + p_m b_m \leq B + M y_1 \qquad (17)$$
$$p_1 b_1 + p_2 b_2 + ...+ p_k' d_k + ... + p_m b_m \leq B + M y_2 \qquad (18)$$

where M is a very large positive number ($M >> 0$).

Besides that new variables y_1 and y_2 are integer 0 - 1 variables, for which is:

$$y_1 + y_2 = 1 \qquad (19)$$
$$y_1, y_2 = 0 \text{ or } 1 \qquad (20)$$

In the above model there are two 0-1 variables y_1 and y_2, where due to the relation (19) only one of them always equals 1, and the other equals zero. Naturally, if the model comprises a number of resources that can be purchased at a discounted price then there are more 0-1 variables.

The problem of mutual exclusiveness of the variables b_k and d_k can be introduced in the following way:

If $d_k = 0$ (there is no quantity discount) then the relation $b_k < Q$ is valid i.e. the needed quantity of raw materials is below the quantity required to obtain the discount.

Similarly, if $b_k = 0$ then the relation should be $d_k \geq Q$ (the quantity required for the discount is reached). Due to that, let us introduce two additional constraints:

$$b_k + M\, d_k \leq Q + N\, y_1 \qquad (21)$$
$$d_k + N\, y_2 \geq Q + M\, b_k \qquad (22)$$

where $N >> M >> 0$, i.e. N and M are very great positive numbers, but N is also much greater than M.

The model objective function remains the same, while p_k' stands for raw material prices, which are either discounted or rising.

5 MULTIPLE CRITERIA DE NOVO PROGRAMMING

Multi-criteria optimization is premised by the conflict of objectives. The individual optimum for each objective simultaneously cannot be achieved in a feasible set. This is because the resources in the traditional product mix model are determined a priori and hence the constraint set is fixed. The compromise solution, therefore, is one which gives the individual objectives lesser values than their individual optimum or ideal solutions. The so-called “ideal“ solution is, in fact, infeasible (Figure 1).

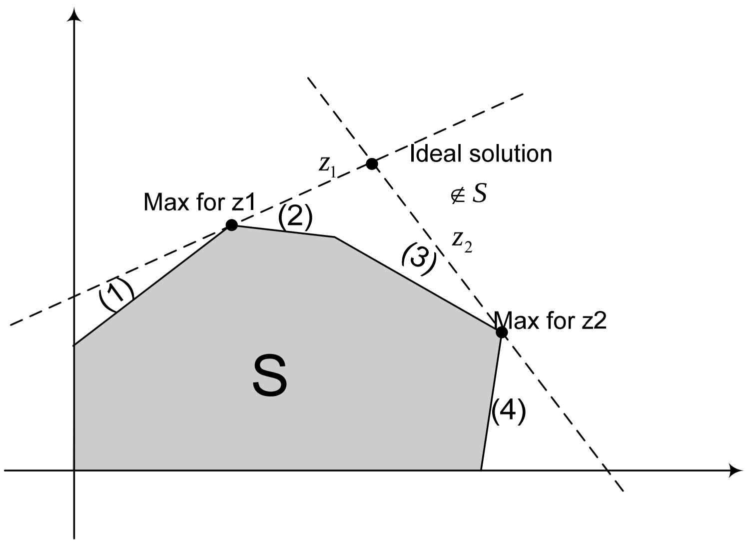


Figure 1.

In the case of De Novo formulation where the resources are not fixed at the outset, it is possible to readjust the resource constraints in such a way that the originally infeasible ideal solution becomes feasible as depicted in Figure 2. If the budget restriction does not allow readjustment of the resource constraints to cause full achievement of the ideal solution, it is enough to aspire for a near-satisfaction of the ideal solution.

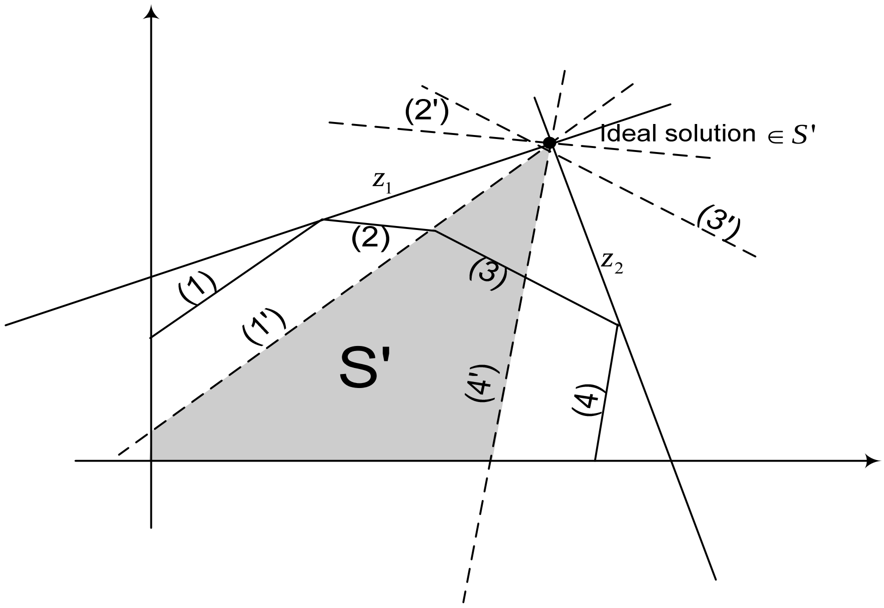


Figure 2.

It is obvious that the ideal z^* is only relative to a particular given system and its value is therefore as arbitrary as the system generating it. Each system, including the optimally designed one, will be characterized by its own ideal. In other words, system ideal will depend of the available amount of money, i.e. of the available budget.

Let us formulate a multi criteria linear programming problem (see Zeleny [25]):

$$\text{Max } z = Cx \quad \text{s.t.} \quad Ax - b \leq 0,\ pb \leq B,\ x \geq 0 \tag{23}$$

where C is the matrix whose rows are the coefficients of q objective functions, and A is the matrix of dimension $m\times n$, b is the m-dimensional unknown vector of resources, x is n-dimensional vector of decision variables, p is the vector of the unit prices of m resources, and B is the given total available budget.

Solving the problem (12) means finding the optimal allocation of B so that the corresponding resource portfolio b maximizes simultaneously the values $z = Cx$ of the product mix x.

As in the single criteria case we can transform this problem into:

$$\text{Max } z = Cx \quad \text{s.t.} \quad \text{V}x \leq \text{B},\ x \geq 0 \tag{24}$$

where $z = (z_1, z_2, \ldots, z_q)$ and $V = (V_1, V_2, \ldots, V_n)$.

Let $z_{k^*} = max\ z_k, k = 1,\cdots,q$, be the optimal value for the k objective of problem (24) subject to $Vx \leq B$, $x \geq 0$. Let $z^* = \left(z_{1^*},\cdots,z_{q^*}\right)$ be the q-objective value for the ideal system with respect to B. Then, a metaoptimum problem can be constructed as follows:

$$\text{Min } Vx \ \text{ s.t. } Cx \geq z^*,\ x \geq 0 \tag{25}$$

Solving the problem (14) yields x^*, B^* ($= Vx^*$) and b^* ($= Ax^*$). The value B^* identifies the minimum budget to achieve z^* through x^* and b^*.

Since $B^* \geq B$, the optimum-path ratio for achieving the ideal performance z^* for a given budget level B is defined as.

$$r^* = \frac{B}{B^*} \tag{26}$$

Optimal design now is (***x, b, z***), where $\boldsymbol{x} = r^*x^*$, $\boldsymbol{b} = r^*b^*$ and $\boldsymbol{z} = r^*z^*$. the optimum-path ratio provides an effective and fast tool for the efficient optimal design of linear systems.
Shi [13] introduces six types of optimum path ratio dealing with two additional types of budgets (other than *Bi and* B*). The first is B_j^k, the budget level for producing the optimal x_j^k with respect to the k objective, referring back to the single-objective De Novo programming problem. The other, B^{**}, refers to the case $q \leq n$ (the number of objectives is less than the number of variables). It can be shown that $B^{**} \geq B^* \geq B \geq B_j^k$, for $k = 1, \ldots, q$.
These six types of optimum-path ratios are:

$$r_1 = \frac{B^*}{B^{**}}, \quad r_2 = \frac{B}{B^{**}}, \quad r_3 = \sum \lambda_k \frac{B_j{}^k}{B^{**}}, \quad r_4 = r^* = \frac{B}{B^*},$$

$$r_5 = \sum \lambda_k \frac{B_j{}^k}{B^*}, \quad r_6 = \sum \lambda_k \frac{B_j{}^k}{B}$$

They lead to six different policy considerations and optimal system designs. Comparative economic interpretations of all optimum-path ratios are dependent on the decision maker's value complex (Zeleny [23]).

6 CONCLUSION

De Novo presents a special approach to optimization. Instead of "optimizing a given system" it suggests a way of "designing an optimal system". The basis of the standard approach is a production model with resources defined in advance, so that the constraints and the feasible set are fixed.

However, De Novo approach does not limit the resources as most of the necessary resource quantities can be obtained at certain prices. Resources are actually limited because their maximum quantity is governed by the budget which is an important element of De Novo. Using De Novo approach, most various cases, especially in multi-criteria optimization, can be handled more effectively than by the standard programming models. Changes in prices, technological coefficients, increasing costs of raw materials, quantity discounts and other similar and real production situations can be easily incorporated in De Novo model and provide very satisfactory solutions.

The new idea in this research is the way how to behave in such cases where we have additional constraints (which may make our meta-optimal problem infeasible), and how to behave in the situation when decision maker has some additional preferences above the chosen objective functions. We show how to behave in such situations, but we must have in mind that every such problem has its specific characteristics and its decision maker. We must not give the final solution without the decision maker, but we hope that De Novo concept of optimization can give the decision maker many new views of his (hers) problem and that he (she) can optimize his (hers) preferences in mutual competition of the chosen criteria.

References

[1] Babić, Z., I. Pavić (1992): *De Novo Programming in MCDM***,** 16th Symposium on Operations Research, Trier, Germany, Physica-Verlag, Heidelberg, 579-582.

[2] Babić, Z., I. Pavić (1996): *Multicriterial Production Programming by De Novo Programming Approach*, International Journal of Production Economics, Vol. 43, No.1, 59-66.

[3] Babić, Z. (2005): *Production Planning via De Novo Programming*, Global Business & Economics Anthology (Selected Papers from Business & Economics Society International Conference, Flagstaaf, Arizona), Worcester, USA, 476-484.

[4] Babić, Z., T. Hunjak (2005): *Multiple Prices and De Novo Programming*, Proceedings of the 8th International Symposium on Operational Research - SOR 2005 - Nova Gorica, Slovenia, 311-316.

[5] Babić, Z., T. Hunjak, I. Veža (2006): *Optimal System Design with Multi-criteria Approach,* Global Business & Economics Anthology (Selected papers from the 2006 Business & Economics Society International Conference, Firenza, Italy), Worcester, USA, 493-502.

[6] Babić, Z., T. Hunjak (2006): *The Use of Multicriteria De Novo Programming in the Production Planning Problem*, Proceedings of the11th International Conference on Operational Research KOI 2006, Pula, Croatia, 257-266.

[7] Chen, Y.W., H. E. Hsieh (2006): *Fuzzy Multi-stage De Novo Programming Problem*, Applied Mathematics and Computation 181, 1139-1147.

[8] Chen, J.K.C., K.C.Y. Chen, S.Y. Lin, B.J.C. Yuan (2008): *Build Project Resources Allocation Model for Infusing Information Technology into Instruction with De Novo Programming*, International Conference on Business and Information - BAI 2008, Seoul, South Korea.

[9] Huang, J.J., G.H Tzeng, C.S. Ong (2005): *Motivation and Resource Allocation for Strategic Alliances through the De Novo Perspective*, Mathematical and Computer Modeling, 41 (6-7), 711-721.

[10] Huang, J.J., G.H Tzeng, C.S. Ong (2006): *Choosing Best Alliance Partners and Allocating Optimal Alliance Resource Using the Fuzzy Multi-objective Dummy Programming Model*, Journal of the Operational Research Society, 57, 1216-1223.

[11] Li, R.J., E.S. Lee (1990): *Multi-criteria De Novo Programming with Fuzzy parameters*, Computers and Mathematics with Applications Vol. 19, No. 5, 13-20.

[12] Li, R.J., E.S. Lee (1993): *De Novo Programming with Fuzzy Coefficients and Multiple Fuzzy Goals*, Journal of Mathematical Analysis and Applications 172, 212-220.

[13] Shi, Y. (1995): *Studies on Optimum Path Ratios in Multi-criteria De Novo Programming Problems,* Computers & Mathematics with Application, Vol. 29, No. 5, 43-50.

[14] Shi, Y. (1999): *Optimal System Design with Multiple Decision Makers and Possible Debt: A Multi-criteria De Novo Programming Approach*, Operations Research, 47, 723-729.

[15] Shi, Y., D. L. Olson, A. Stam (Eds.) (2007): *Advances in Multiple Criteria Decision Making and Human Systems Management: Knowledge and Wisdom*, IOS Press, Amsterdam, Berlin, Oxford, Tokyo, Washington.

[16] M.T. Tabucanon (1988): *Multiple Criteria Decision Making in Industry*, Elsevier, Amsterdam.

[17] Taguchi, T., T. Yokota (1999): *Optimal Design Problem of System Reliability with Interval Coefficient Using Improved Genetic Algorithms*, Computers & Industrial Engineering 37, 145-149.

[18] Yu, J.R., H.H. Wang (2007): *Multiple Criteria Decision Making and De Novo programming in Portfolio Selection*, Second International Conference on Innovative Computing, Information and Control, ICICIC 2007, Kumamoto, Japan, p. 194.

[19] Zeleny, M. (1982): *Multiple Criteria Decision Making,* McGraw-Hill, New York

[20] Zeleny, M (1986): *Optimal System Design with Multiple Criteria: De Novo Programming Approach*, Engineering Costs and Production Economics, No. 10, 89-94.

[21] Zeleny, M (1990): *Optimizating Given Systems vs. Designing Optimal Systems: The De Novo programming approach*, International Journal of General Systems, 17, No.4, 295-307.

[22] Zeleny, M. (1997): *The Fall of Strategic Planning (Editorial)*, Human Systems Management, 16, 77-79.

[23] Zeleny, M. (1998): *Multiple Criteria Decision Making: Eight Concepts of Optimality*, Human Systems Management, 17(2), 97-107.

[24] Zeleny, M. (2005): *The Evolution of Optimality: De Novo Programming*; C.A. Coello Coello et al. (Eds.): EMO 2005, LNCS 3410, 1-13.

[25] Zeleny, M. (2009): *On the Essential Multidimensionality of an Economic Problem: Towards Tradeoffs-Free Economics*, AUCO Czech Economic Review 3, 154-175.

[26] Zhang, Y.M., G.H. Huang, X.D. Zhang (2009): *Inexact De Novo Programming for Water Resources Systems Planning*, European Journal of Operational Research 199, 531-541.

RELIABLE OPTIMIZATION TECHNIQUES AND THEIR APPLICATION

Tibor Csendes
University of Szeged, Institute of Informatics
H-6720 Szeged, Árpád tér 2, Hungary
csendes@inf.u-szeged.hu

Abstract: The talk provides an introduction to interval arithmetic based techniques for the verification of mathematical models. Illustrative examples are described from the fields of circlepacking, chaotic behaviour dynamical systems, and process network synthesis.

Keywords: interval methods, theorem proving, computer assisted, circle packing, dynamical systems, process network synthesis.

1 INTRODUCTION

Model verification, optimization, and especially global optimization are sensitive on the reliability of the numerical computations. There exist practical problems where good approximate solutions are more or less accepted as the true solutions. Still there remain important application fields where the guaranteed reliability of the provided solution is of ample importance. The uncertainties are mostly caused by the rounding errors. These are necessarily part of the calculations when the algorithms are coded with floating point arithmetic - which allows quick computation. To provide a remedy for these problems, we shall apply interval arithmetic based inclusion functions [11]. These offer a theoretically reliable and computationally tractable means of locating a feasible suboptimal interval.

Denote the real numbers by x, y, …, the set of compact intervals by I := { [a, b] | a ≤ b, a, b in R}, and the set of n-dimensional intervals (also called simply intervals or boxes) by I^n. Capital letters will be used for intervals. For real vectors and interval vectors the notations

$$x = (x_i),\ x_i \text{ in } R, \text{ and } X = (X_i),\ X_i \text{ in } I$$

are applied, respectively.

The idea of interval calculations is to extend the basic operations and the elementary functions from the real numbers to intervals. Finding the range for a function over an n-dimensional interval has in general the same complexity as an optimization problem, because we have to find the extreme values of the function over the interval. By using interval arithmetic it is possible to find bounds on the function values more efficiently. The interval operations can be carried out using only real operations. For the argument intervals [a, b] and [c, d] the following expressions hold:

$$[a, b] + [c, d] = [a+c, b+d]$$
$$[a, b] - [c, d] = [a-d, b-c]$$
$$[a,b] * [c, d] = [\min \{ac, ad, bc, bd\}, \max \{ac, ad, bc, bd\}]$$
$$[a, b] / [c, d] = [a, b] * [1/d, 1/c] \quad \text{if } 0 \text{ is not in } [c, d].$$

As an example, consider the range of $x-x^2$: it is [-2, 0.25] on the argument interval of [0, 2]. In contrast to that, the above interval arithmetic will provide the inclusion of [-4, 2], which is much wider. This too conservative estimation can be improved at the cost of more computation with sophisticated numerical techniques.

If outwardly directed rounding is also applied, then the interval calculated by a computer contains every real number that can be a result of the given function on real numbers inside the original intervals. The technique of producing an inclusion function by replacing the real variables and operations by their interval equivalent is called *natural interval extension* [11, 12, 14].

The interval arithmetic based inclusion functions (denoted by capital letters) can be utilized in the procedure given below to provide a guaranteed reliability solution for the global minimization problem.

Interval arithmetic based Branch-and-Bound optimization algorithm

---

0. Set Y = X and y = min F(X). Initialize the list L = ((Y,y)) and the cut-off level z = max F(X).
1. Choose a coordinate direction k in {1,2,...,n}.
2. Bisect Y in direction k: $Y = V^1 \cup V^2$.
3. Calculate $F(V^1)$ and $F(V^2)$, and set $v^i = \min F(V^i)$ for i = 1, 2 and $z = \min \{z, \max F(V^1), \max F(V^2)\}$.
4. Remove (Y, y) from the list L.
5. Cut-off test: discard the pair (V^i, v^i) if $v^i > z$ (where i in {1, 2}).
6. Monotonicity test: discard any remaining pair (V^i, v^i) if 0 is not in $\nabla F_j(V^i)$ for any j in {1,2,...,n}, and i= 1, 2.
7. Add any remaining pair(s) to the list L. If the list becomes empty, then STOP.
8. Denote the pair with the smallest second element by (Y,y).
9. If the width of F(Y) is less than ε, then print F(Y) and Y, STOP.
10. Go to Step 1.

Optimization algorithms are readily applicable for finding extremal values or extremal settings of discrete geometrical models, such as those in packing and covering problems. It sounds also surprising to some extent how optimization procedures can be applied for proving mathematical theorems - considering that traditional optimization methods provide usually just approximations for the aimed extremal solutions.

In the present study we show examples when both stochastic global optimization algorithms [4], [9] and interval arithmetic based branch-and-bound techniques [6], [8] [18] are successfully applied to provide rigorous statements on open theoretical mathematical problems.

2 RECYCLING IN PROCESS NETWORK SYNTHESIS

The problem is to find that separation network setting for the divisors x_i in [0.0, 1.0], (i=1, 2, 3, 4) that allow a minimal cost functioning (that depends on the amount of material flowing through the separators). The folklore engineering rule of thumb was that the optimal solution cannot contain cycles where the material would flow forever. We have checked a simple problem instance, given by Zoltán Kovács [3], [5] (illustrated on Figure 1 below).

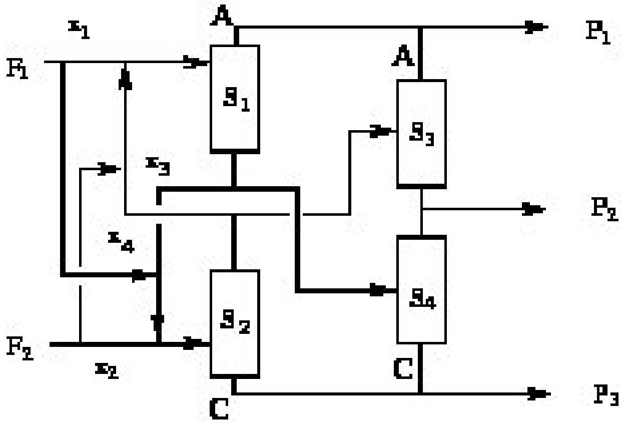


Figure 1: An example separator network design problem with two feeds and three products of given composition.

The result obtained by a stochastic, clustering global optimization method based on real function evaluations are given in Table 1.

Table 1: The numerical results obtained by an approximate global optimization algorithm on the separation network design problem.

f*	x_1*	x_2*	x_3*	x_4*	NFE	CPU (s)
62.550111	0.00296602	0.99875609	0.75149047	1.0000000	56 067	8.73
62.791640	0.01029729	0.99757082	0.84846761	1.0000000	27 232	4.45
62.851381	0.02461725	0.99482994	0.64821641	1.0000000	61 190	9.45
62.855458	0.00166749	0.99553307	0.86046189	1.0000000	51 263	8.13
62.836668	0.05248426	0.99983809	0.81663978	1.0000000	38 757	6.10
average					47 486.3	7.470

Although all the found approximate optimal solutions have a positive value for the variable x_3 and x_4, it is obviously no proof that the real optimum has positive values here too. This would ensure the cycling behavior we aim to prove.

In contrast to that, the interval arithmetic based B&B procedure have found the following result at the cost of 35,683 function evaluations, 3 minutes CPU time, and 10,000 memory units (containing one subinterval):

$F(X^*) = [62.49, 62.69]$,

$X^* =$ [0.00000000000, 0.00097656250], [0.99804687500, 0.99902343750], [0.71875000000, 0.72070312500], [0.99804687500, 1.00000000000].

Due to the fact that we have applied the critical outward rounding, and since the given intervals contain all points that can be optimal, we can conclude that the optimal solution must be positive in the variables x_3 and x_4, and hence the optimal solutions of separator networks can in fact have recycling.

3 OPTIMIZATION FOR CIRCLE PACKING

An often heard question is, how can we characterize the maximal size or difficulty of problems that still can be solved by interval inclusion function based methods [11], [12], [13]. The short answer is that the dimension of a problem is a wrong measure, since low dimensional problems can be hopeless, and larger dimensional ones can be solved in a short time. For interval techniques the most dangerous is the large excess width, a bad estimation of the range of the related function on the studied intervals. It is most affected by the dependency problem, that is caused by multiple appearances of the same variable in a complex expression. According to this, the rule of thumb says that in case all involved variables appear only a few times in the expression of the objective and constraint functions, then the overestimation will be small, and the optimization algorithms can be successful even for larger dimensional problems (the number of variables can be up to 100).

A telling example for the capabilities of interval optimization methods is the results on circle packing problems. Mihály Csaba Markót [15], [16] was able to solve the problem cases of n = 28, 29, and 30, i.e. to find the configuration of n congruent nonoverlapping maximal circles fitting into the unit square. These problems were held before as hopeless, since the expected CPU time necessary for their solutions with the last available techniques were estimated to be around decades. The problems have 56, 58, and 60 variables, respectively, and hundreds of nonlinear constraints.

The difficulty of the problems is highlighted by the facts that (due to obvious geometrical reasons) there are an astronomical number of equivalent, symmetric optimal solutions, and in the cases of n = 28, 29 there exist positive measure sets of global optimizer points. See the details of the packings on Figure 2.

Standard interval optimization algorithms could not solve the problems. Careful problem decomposition, and custom made built in acceleration devices based on the understanding of the problem structure enabled the successful solution. The running times were below 3 days, and ca. one million subintervals were stored during the solution process. The uncertainty in the position of the circles has been decreased by more than 700 orders of magnitude in each case. Due to the controlled outward rounding mode applied, the obtained results are reliable even in the sense of a rigorous mathematical theorem [20].

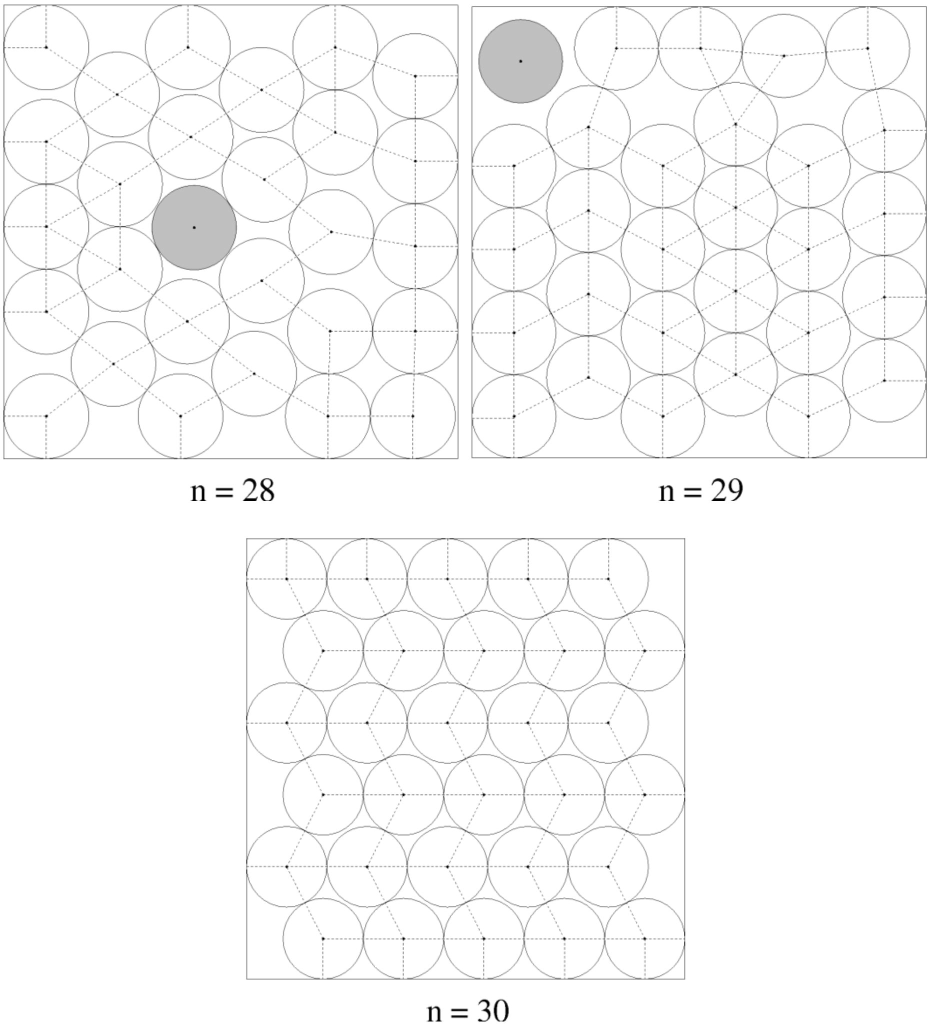


Figure 2: The proven optimal circle packings into the unit square for 28, 29, and 30 circles.

4 CHAOS LOCATION FOR DYNAMICAL SYSTEMS

We give now an optimization model and the related algorithms to locate chaotic regions of dynamic systems [8]. Computer-assisted proofs for the existence of chaos are important for the understanding of dynamic properties of the solutions of differential equations. These techniques have been intensively investigated recently, see e.g. [10], [17], [19], [21], and [22].

We study verified computational methods to check and locate regions the points of which fulfil the conditions of chaotic behaviour. The investigated Hénon mapping is $H(x,y) = (1+y-Ax^2, Bx)$. The paper [21] considered the $A = 1.4$ and $B = 0.3$ values and some regions of the two dimensional Euclidean space: $E = E_1 \cup E_2 = \{(x, y) \mid x \geq 0.4, y \geq 0.28\} \cup \{(x, y) \mid x \leq 0.64, |y| \leq 0.01\}$, $O_1 = \{(x,y) \mid x<0.4, y> 0.01\}$, $O_2 = \{(x, y) \mid y<0\}$.

According to [21], Theorem 1 below ensures the chaotic behaviour for the points of the parallelograms Q_0 and Q_1 with parallel sides with the x axis (for $y_0 = 0.01$ and $y_1 = 0.28$, respectively), with the common tangent of 2, and x coordinates of the lower vertices are $x_a = 0.460$, $x_b = 0.556$; and $x_c = 0.558$, $x_d = 0.620$, respectively. The mapping and the problem details (such as the transformed sides of the parallelograms, $H^7(a)$, $H^7(b)$, $H^7(c)$, and $H^7(d)$) are illustrated on Figure 3.

Theorem 1. Assume that the following relations hold for the given particular Hénon mapping:

$$H^7(a \cup d) \text{ is a subset of } O_2, \quad (1)$$
$$H^7(b \cup c) \text{ is a subset of } O_1, \quad (2)$$
$$H^7(Q_0 \cup Q_1) \text{ is a subset of } R^2 \setminus E, \quad (3)$$

then chaotic trajectories belong to the starting points of the regions Q_0 and Q_1.

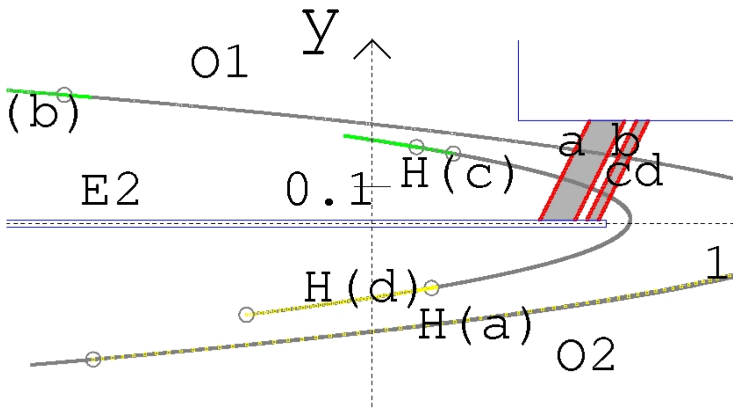


Figure 3: Illustration of the H^7 transformation for the classic Hénon parameters $A = 1.4$ and $B = 0.3$ together with the chaotic region of two parallelograms. The line segments *a, b, c*, and *d* are sides of the parallelograms Q_0 and Q_1.

The main difficulty of checking conditions (1) to (3) is that one has to prove these for a continuum of points. In [21] the author calculated the Lipschitz constant, gave an upper bound for the rounding error committed and thus reduced the whole task to investigating a finite number of points of a dense enough grid. This method works only with human interaction. To search chaotic regions an automated checking routine is more appropriate. The technique we applied combines interval arithmetic and adaptive branch-and-bound subdivision of the region of interest. It is basically a customized version of the technique introduced in Section 1.

This algorithm first encloses the sets Q_0 and Q_1 in a 2-dimensional closed interval I, the starting interval. Then to prove subset relations an adaptive branch-and-bound technique generates such a subdivision of the starting interval that either:

- for all subintervals the given conditions of chaos hold - in case they contain points of the respective sets, or
- it is shown that a small subinterval (of a user set size) exists that contains at least one point of the respective set, and it contradicts at least one of the relations.

Our algorithm is capable of recognizing that a region satisfies the conditions of chaos. We have proven the correctness of the procedure in [8].

Once we have a reliable computer procedure to check the conditions of chaotic behavior of a mapping it is straightforward to set up an optimization model that transforms the original chaos location problem to a global optimization problem.

The chaotic regions have several parameters that identify them. In the early phase of our investigation we have restricted the search to locate two parallelograms similar to that used in the article [21]: we are allowed to change the vertical and horizontal positions and also the common tangent, but the parallelograms always had two sides parallel to the x axis. It is also possible to find fitting parameter values for the Hénon mapping, i.e., for the mapping parameters A and B, and furthermore also for parameters of the aimed sets of the underlying theorem, e.g. the border coordinates of the set E.

The search for a chaotic region was modelled by a constrained global optimization problem. Subsequently the constraints were represented by a penalty function approach. The original objective function was constant, still the possibility exists to extend it to a more complex form that expresses further aims, e.g. to locate a second chaotic region, different from the known one.

The key question for the successful application of a global optimization algorithm is how to compose the penalty functions. On the basis of earlier experiences collected solving similar constrained problems, we have decided to add a nonnegative value proportional to how much the given condition was violated, plus a fixed penalty term in case at least one of the constraints was not satisfied.

As an example, consider the case when one of the conditions for the transformed region was hurt, e.g. when (2), i.e., the relation $H^k(b \text{ cup } c)$ subset O_1 does not hold for a given k-th iterate, and for a region of two parallelograms. For such a case the checking routine will provide a subinterval I that contains at least one point of the investigated region, and which contradicts the given condition. Then we have calculated the Hausdorff distance of the transformed subinterval $H^k(I)$ to the set O_1 of the right side of the condition,

$$\max_{z \text{ in } Hk(I)} \inf_{y \text{ in } O1} d(z,y),$$

where $d(z,y)$ is a given metric, a distance between two two-dimensional points. Notice that the use of maximum in the expression is crucial, with minimization instead our optimization approach could provide (and has provided) result regions that do not fulfill the given conditions of chaotic behaviour.

We have considered the following bound constrained problem for the T inclusion function of the mapping T:

$$\min_{x \text{ in } X} g(x), \tag{4}$$

where

$$g(x) = f(x) + p(\textstyle\sum_{i=1}^{m} \max\{z \text{ in } T(I(x))\} \inf_{y \text{ in } Si} d(z,y)),$$

X is the n-dimensional interval of admissible values for the parameters x to be optimized, $f(x)$ is the original, nonnegative objective function, and $p(y) = y + C$ if y is positive, and $p(y) = 0$ otherwise. C is a positive constant, larger than $f(x)$ for all the feasible x points, m is the number of conditions to be fulfilled, and Si is the aimed set for the i-th condition. In this discussion $I(x)$ is the subinterval returned by the checking routine (or the empty set). The interval $I(x)$ depends implicitly on the parameter x to be optimized.

For more complicated cases the fixed sets given in Theorem 1 should also be changed subject to certain structural constraints, e.g. the x_a, x_b, x_c, and x_d coordinates of the parallelograms have to follow this order. These new conditions can also be represented in a similar way, following the penalty function approach of (4).

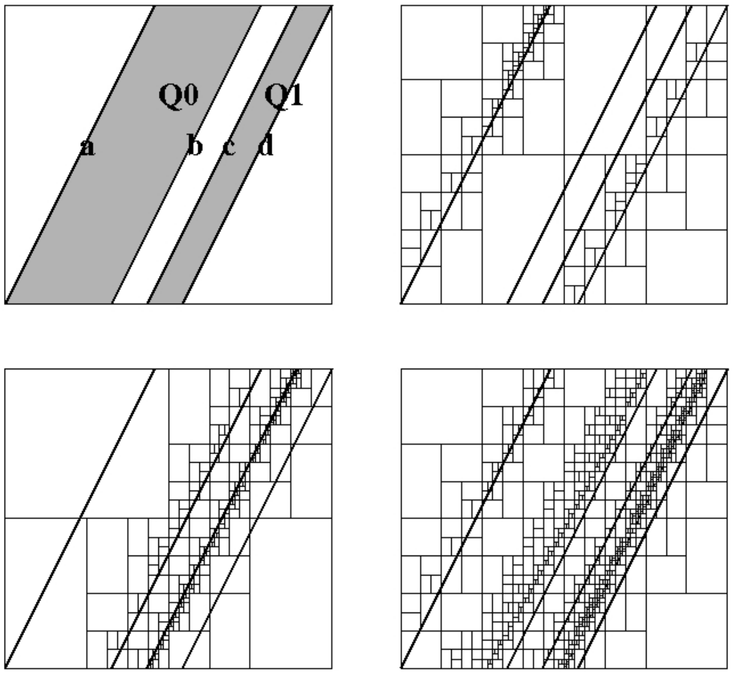


Figure 4: The parallelograms and the starting interval covered by the verified subintervals for which the given condition holds (in the order of mentioning in Theorem 1).

For the computational experiments we have applied the C-XSC programming language [14] supporting interval arithmetic. The results were obtained both in Linux and in the Cygwin environment, on an average personal computer. In the present section we just provide some demonstrative examples for the functioning of the introduced technique. First we have checked the reported chaotic region [21] by our checking routine. The illustration for that is Figure 3.

The papers [1], [2], [7] provide additional new chaotic regions located by the present method. Summarizing the results, we were able to prove with an acceptable amount of computation and human overhead that the published system is chaotic in some given regions.

The optimization algorithms applied to obtain the results presented here can be downloaded for academic purposes from the web page of the author:

`www.inf.u-szeged.hu/~csendes`

Acknowledgements. The presented research has been supported among others by the project TÁMOP–4.2.2/08/1/2008–0008.

References

[1] B. Bánhelyi, T. Csendes, and B.M. Garay: Optimization and the Miranda approach in detecting horseshoe-type chaos by computer. Int. J. Bifurcation and Chaos 17(2007) 735-748

[2] B. Bánhelyi, T. Csendes, B.M. Garay, and L. Hatvani: A computer-assisted proof for Sigma_3-chaos in the forced damped pendulum equation. SIAM J. on Applied Dynamical Systems 7(2008) 843-867

[3] A.E. Csallner: Global optimization in separation network synthesis. Hungarian J. Industrial Chemistry 21(1993) 303–308

[4] T. Csendes: Nonlinear parameter estimation by global optimization - efficiency and reliability. Acta Cybernetica 8(1988) 361-370

[5] T. Csendes: Optimization methods for process network synthesis - a case study, In: Christer

Carlsson and Inger Eriksson (eds.): Global & multiple criteria optimization and information systems quality. Abo Academy, Turku, 1998, pp. 113-132

[6] T. Csendes: New subinterval selection criteria for interval global optimization. J. Global Optimization 19(2001) 307-327

[7] T. Csendes, B. Bánhelyi, and L. Hatvani: Towards a computer-assisted proof for chaos in a forced damped pendulum equation. J. Computational and Applied Mathematics 199(2007) 378-383

[8] T. Csendes, B.M. Garay, B. Bánhelyi: A verified optimization technique to locate chaotic regions of a Hénon system. J. of Global Optimization 35(2006) 145-160

[9] T. Csendes, L. Pál, J.O.H. Sendín, J.R. Banga: The GLOBAL Optimization Method Revisited. Optimization Letters 2(2008) 445-454

[10] Z. Galias and P. Zgliczynski: Abundance of homoclinic and heteroclinic orbits and rigorous bounds for the topological entropy for the Hénon map. Nonlinearity 14(2001) 909-932

[11] E. Hansen: Global optimization using interval analysis, Marcel Dekker, New York (1992)

[12] R. Hammer, M. Hocks, U. Kulisch, and D. Ratz: Numerical Toolbox for Verified Computing I., Springer, Berlin (1993)

[13] R.B. Kearfott: An interval branch and bound algorithm for bound constrained optimization problems. J. Global Optimization 2(1992) 259-280

[14] R. Klatte, U. Kulisch, A. Wiethoff, C. Lawo, and M. Rauch: C-XSC - A C++ Class Library for Extended Scientific Computing, Springer, Heidelberg (1993)

[15] M.Cs. Markót: Optimal Packing of 28 Equal Circles in a Unit Square - the First Reliable Solution. Numerical Algorithms 37(2004) 253-261

[16] M.C. Markót and T. Csendes: A new verified optimization technique for the "packing circles in a unit square" problems. SIAM J. on Optimization 16(2005) 193-219

[17] A. Neumaier, and T. Rage: Rigorous chaos verification in discrete dynamical systems, Physica D 67(1993) 327-346

[18] L. Pál and T. Csendes: INTLAB implementation of an interval global optimization algorithm. Optimization Methods and Software 24(2009) 749-759

[19] T. Rage, A. Neumaier, and C. Schlier: Rigorous verification of chaos in a molecular model, Phys. Rev. E 50(1994) 2682-2688

[20] P.G. Szabó, M.Cs. Markót, T. Csendes, E. Specht, L.G. Casado, and I. García: New Approaches to Circle Packing in a Square – With Program Codes. Springer, Berlin (2007)

[21] P. Zgliczynski: Computer assisted proof of the horseshoe dynamics in the Hénon map. Random & Computational Dynamics 5(1997) 1-17

[22] P. Zgliczynski: On smooth dependence on initial conditions for dissipative PDEs, an ODE-type approach. J. Differential Equations 195(2003) 271-283

RECENT DEVELOPMENTS IN MRP THEORY - AND FUTURE OPPORTUNITIES

Robert W. Grubbström FVR RI
Linköping Institute of Technology, SE-581 83 Linköping, Sweden
Mediterranean Institute for Advanced Studies, SI- 5290, Šempeter pri Gorici, Slovenia
robert@grubbstrom.com

Abstract: *Material Requirements Theory (MRP Theory)* combines the use of *Input-Output Analysis* and *Laplace transforms*, enabling the development of a theoretical background for multi-level, multi-stage production-inventory systems together with their economic evaluation, in particular applying the *Net Present Value* (NPV) principle. The time scale may be continuous ("bucketless" in MRP terminology) or discrete. Since the average cost measure may be viewed as an approximation of NPV, the theory may be applied to cases, when average costs are preferred as the objective function.

Central in this theory are the fundamental equations explaining the time development of available inventory, allocated component stock (allocations), and backlogs. These are balance equations, in which the *generalised input matrix* plays a predominant rôle. The input matrix (next assembly quantity matrix) from Input-Output Analysis describes the Bill-Of-Materials in terms of amounts of materials and subcomponents needed on each level in the product structures. In MRP *lead times* are taken into account, describing the advanced timing these amounts are needed in comparison to the time of completion. Assuming the lead times to be constants, as is the case in classical MRP, using Laplace transforms, they may be modelled as operators in a diagonal lead time matrix. Multiplying the lead time and input matrices, results in the generalised input matrix, which captures all requirements as well as the needs of their advanced timing. The vector-valued balance equation for *available inventory* (total inventory less earmarked materials reserved for use according to the Master Production Schedule, MPS) then become very compact and captures any MRP case with arbitrary product structures and arbitrary MPS.

Also services in the sense of capacity requirements may be included as amounts in the input matrix by introducing a further row for each type of capacity. The generalised input is then extended accordingly, taking into consideration when capacities are needed compared to the completion dates.

A basic constraint, when formulating the problem of optimising the MPS, is the *available inventory constraint* requiring that available inventory never may be negative if the MPS is to be feasible. A corresponding constraint concerns capacities, whenever capacities are assumed limited. If backlogs are allowed in the problem, these only concern items which have an external demand from outside customers. Internally demanded items may never be backlogged.

Recently, MRP Theory has been extended in new directions. Hitherto, it has mainly dealt with *assembly structures* by which items produced downstream contain one or more sub-items on lower levels, but at each stage, the assembly activity produces only one type of output. The material flow is thus convergent. Attention has now turned to *arborescent systems,* in which one input may create more than one type of output. Here the material flow becomes divergent. Typical examples of arborescent systems are transportation (distribution) and recycling.

For arborescent systems the input time, rather than the completion time, will be the natural reference of each process. Then outputs will be available according to a planned delay after input time. Introducing a new diagonal matrix, the *output delay matrix*, with

operators as elements in its diagonal corresponding to these delays, a *generalised output matrix* is formed capturing output amounts as well as their delayed timing.

As a second example, we choose the classical problem of *optimal dynamic lotsizing* in MRP. The optimal batch sizes of various produced amounts and their timing are to be determined. New formulations within MRP Theory, using a binary representation of the decision variables (emanating from the *Triple Algorithm*), has enabled the development of optimal lotsize solution procedures for any deterministic assembly system. Since lead times are present in this class of problems, they also cover the so called *Multi-Level Lotsizing* problem (MLLS), in which all lead times are assumed zero.

This presentation intends to provide an historical overview over the progress of MRP Theory, highlighting certain recent results, and pointing at future opportunities for analysing more general sets of problems.

Selected references

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

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

[3] Grubbström, R. W., Bogataj, L., Bogataj, M., *A Compact Representation and Optimisation of Distribution and Reverse Logistics in the Value Chain*, WP, (Mathematical economics, operational research and logistics, 5), Ljubljana: Faculty of Economics, KMOR, 2007.

[4] Grubbström, R. W., Bogataj, M., Bogataj, L., Optimal Lotsizing within MRP Theory, *Preprints of the 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'09)*, Moscow, June 3-5, 15-30.

[5] Grubbström, R. W., Molinder, A., Further Theoretical Considerations on the Relationship between MRP, Input-Output Analysis and Multi- Echelon Inventory Systems, *International Journal of Production Economics*, 35, 1994, 299-311.

[6] Grubbström, R. W., Ovrin, P. Intertemporal Generalization of the Relationship between Material Requirements Planning and Input- Output Analysis, *International Journal of Production Economics*, 26, 1992, 311-318.

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

[8] Grubbström, R.W. Wang, Z., A Stochastic Model of Multi-Level/Multi-Stage Capacity-Constrained Production-Inventory Systems, *International Journal of Production Economics*, 81-82, 2003, 483-494.

Keywords: Material requirements planning, MRP, Laplace transform, input–output analysis, net present value.

***Keynote Lecture**, Slovenian Operational Research Society Conference ’09, Nova Gorica, September 23-25, 2009.*

OPERATIONS RESEARCH AND MANAGEMENT GAMES WITH A SPECIAL FOCUS ON HEALTH CARE GAMES

Marion Rauner
University of Vienna, Faculty of Faculty of Business, Economics, and Statistics
Bruenner Str. 72, A-1210 Vienna, Austria
marion.rauner@univie.ac.at

Markus Kraus
Institute for Advanced Studies, Department of Economics and Finance
Stumpergasse 56, A-1060 Vienna, Austria
kraus@ihs.ac.at

Abstract: Operations Research, internet, and e-learning highly increased the potential of management games for teaching and policy making in the last years. First, we give a short overview on general management games. Due to increased health care costs, health care games play a key role in teaching, policy, and research. We review and classify health care management games according to general classification attributes, application areas, target groups, and players' decisions. Health care games also illustrate the application of Operations Research techniques such as simulation, queuing theory, resource allocation models, assignment strategies, stock keeping methods, staff scheduling approaches, and principle-agent theory.

Keywords: management games, hospital games, simulation, decision support, education and distance learning

1 INTRODUCTION

Dempsey et al. [1] define games as follows: "A game is a set of activities involving one or more players. It contains goals, constraints, payoffs, and consequences. A game is rule-guided and artificial in some respects. Finally, a game involves some aspects of competition, even if that competition is with oneself." Chinese people used games as early as 3000 B.C. Battle games such as Wie-Hai were played especially in China (Geilhardt and Mühlbrandt [8]). In the 17th and 18th centuries, war chess variants (e.g., The King's Game, War Chess) became popular. During World War II, war business games were increasingly used (e.g., Barbarossa, Sea Lion, Total War in Pasific). This was also the time when Operations Research (OR) became popular in optimizing war-related strategic and logistic problems. OR emerged in the United Kingdom and was applied to air defense and spread to other aspects of the military machine such as antisubmarine warfare and strategic bombing (Rau [21]).

On the strength of the war experience, the importance of games for management was disclosed. In management games, knowledge can be applied to an artificial setting with a high learning effect. Confucius told "I hear and I forget, I see and I remember, I do and I understand." (Feinstein [7]). This is why, numerous management games have been developed in the fields of general management, accounting, finance, marketing, and production & logistics since the 1950s (Faria and Wellington [6]) and enjoy a great popularity until today. Early most popular games include (Faria [4], Watson and Blackstone [23]): 1) Top Management Decision Simulation, developed by the American Management Association 1956, USA; 2) Top Management Decision Game, developed by Schneider (University of Washington) 1957, USA; 3) Business Management Game, developed by Andlinger and Greene (McKinsey) 1958, USA; and 4) The Carnegie Tech Management Game, developed by the Carnegie Institute of Technology, end-1950s, USA.

In former days, management games were offline games using cards, dices, boards etc. (Faria and Wellington [6]). The development of the computer technology and internet has added a new dimension to the world of gaming and online (computer-based) games were developed. Especially, online learning has become an accepted and popular learning method in comparison to face-to-face education. Online games are mainly based on simulation techniques. Compared to pure simulation human beings play a significant part in games. They make decisions at various stages and thereby improve their decision making skill. Management games help players better understand the reality that is simulated in a simplified way (Watson and Blackstone [23]).

The application of management games provides many advantages (Dieleman and Huisingh [2], Faria and Dickinson [5], Graf [9]). For example, participants learn about different areas of management. All decisions can be repeated and feedback is given immediately. Wrong decisions are not punished compared to a real-world-setting. In addition, people with different social positions, academic education, cultural backgrounds, and different values interact with each other. The players learn to get along with each other and to build teams. Internet technologies enable to overcome the distance and time problem by bringing together participants from all over the world. Furthermore, information from internet and communication platforms can be used by the players to improve their decision making.

However, along with advantages of management games also several disadvantages have to be considered (Lane [15]). It is often the case that goals and learning effects are not clear to all participants. Briefing and debriefing are sometimes neglected. It is difficult to determine the optimal level of complexity of a game and to compress time which are both necessary to model a simplified model of the reality. Thus, a systematic design process and a validity check by experts and future players as well as an extensive testing are crucial (Peters et al. [20]).

2 HEALTH CARE GAMES

Health care games were developed in the late 1960s and two types of games can be distinguished: 1) medical education games and 2) health care management games (Panosch [19]).

Medical education games occurred during the late 1960s and early 1970s. They mainly concentrate on teaching medical students and/or staff how to best treat a patient. This can be compared to a flight simulator situation when a pilot has to first train on the simulator before he/she flights an airplane in difficult situations. Such games have a high potential to decrease medical failures and errors. The slogan “learning by doing” becomes less acceptable when high risk care is needed. Different simulation tools in medical education are used including (Ziv et al. [24]): low-tech simulators, simulated/standardized patients, screen-based computer simulators, complex task trainers, and realistic patient simulators. Thus, games are alternative teaching methods to gain experience and to learn medical knowledge.

Health care management games emerged in the early 1970s. The importance of such games has been increasing because of rising health care costs due to expensive technologies, ageing patients, and more demanding patients (OECD Health Data [18]). Hospitals consume about 40% of the total health care costs (Hofmarcher and Rack [12]). Thus, especially hospital games play a key role for educating students, staff, and health care policy makers to best plan for scarce resources.

We review and classify health care management games published in the international literature or main practice applications according to the following criteria (Kraus et al. [14]): 1) general classification attributes, 2) application areas, 3) target groups, and 4) decisions (Fig. 1).

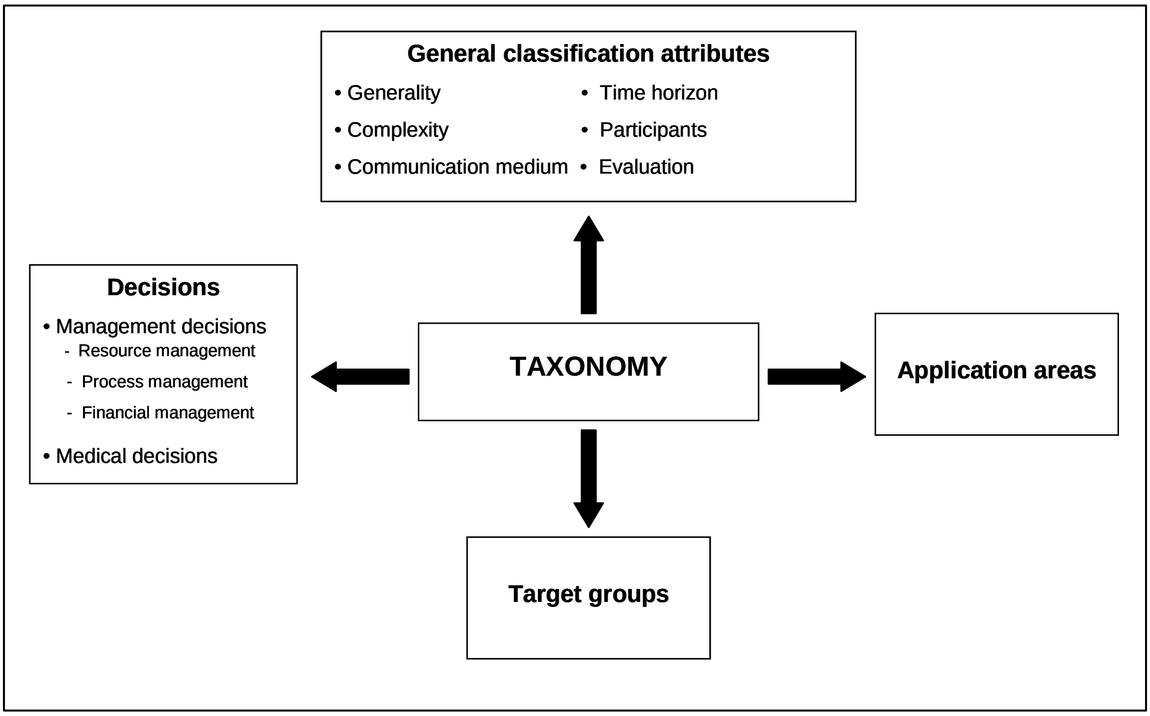


Figure 1: Taxonomy for Health Care Games

2.1 General classification attributes

Health care games can be distinguished by general classification attributes that contain: 1) generality, 2) complexity, 3) communication medium, 4) time horizon, 5) participants, and 6) evaluation.

2.1.1 Generality

Health care games can be either general or functional. While general health care games focus on playing the entire or main functions of health care institutions, functional games are applied to special areas or departments of a health care institution. Most health care games concentrate on running a hospital. We found that offline (non-computer-based) games are mainly functional games. For example, Kennedy et al. [13] develop a game to actively investigate health-related needs of older patients based on a game board, while Meterissian et al. [16] play a game for surgeons based on questions and answers. In contrast, online (computer-based) games have a higher proportion of general games running an entire hospital (e.g., Rauner et al. [22]). As an example for a functional online game, Hans and Nieberg [10] illustrate a variety of OR methods that can be applied to improve operating room management.

2.1.2 Complexity

In general, offline games are rather simple compared to the more complex online games. Many health care games use simulation techniques, especially discrete event simulation. For example, Rauner et al. [22] have implemented a general internet-based hospital game named Coremain hospital. They have chosen the object-oriented programming language C# and the ASP.NET 2.0 (Active Server Page.NET) Web Programming Framework based on the .NET Framework to realise a distributed and server-centric web application with four tiers: 1) client, 2) web-server, 3) application server, and 4) database server. All simulation events are stored in a database for further evaluation in contrast to other hospital games in which simulation events and players' decisions cannot be traced. Additionally, in most other hospital games the parameters of the games are hard-coded and are not stored in a database.

2.1.3 Communication medium

Health care games can be either played offline or online. Historically, offline games were developed that applied cards, dices, game boards, paper and pencil, or question and answer techniques for playing the game. The majority of games exerts game boards such as the Pain Game of Morton and Tarvin [17] in which the players have to answer questions concerning pain assessment and pain management.

Only a few games use the internet for playing health care games (e.g., Rauner et al. [22], Hans and Nieberg [10]). If the game is played centralized in one room (e.g., PC laboratory), then the internet is less important compared to a decentralized game situation (e.g., different PC laboratories, PCs at work or home) in which the players can not communicate face to face. The decentralized game situation requires the internet-based approach and enables to overcome local and temporal restrictions.

2.1.4 Time horizon, participants, and evaluation

Nearly all health care games play several periods with multiple players. For almost all online games, the evaluation of results takes place after each period, while many offline games are evaluated during the period. For example in Coremain hospital [22], players that run a hospital compete with other players of regional hospitals for patients and budget. This is an unique and realistic feature of this hospital game, as in other games players manage a hospital without being influenced by the decisions of the other hospitals' players.

2.2 Application areas

We found that the main application area for health care games is teaching. However, an important focus is also given to decision making. Some games want to also improve research. For example, Rauner et al. [22] plan to investigate Coremain hospital in detail by experimental economics with both teaching (e.g., game situation) and policy purposes (e.g., impact of reimbursement systems). By analyzing reimbursement systems, basics of principle-agent theory can be explained to the players.

2.3 Target groups

Health care games target at different player groups such as students of different majors (e.g., management/economics, medicine, nursing, pharmacy), hospital staff (e.g., nurses, doctors, other medical staff, technical staff, administrative and management staff), and general health care policy makers (e.g., health policy leaders, health care practitioners/professionals). Most health care games are designed to be played by hospital staff and students with majors in management, economics, or medicine.

2.4 Decisions

A key attribute of hospital games comprises the decisions to be taken by the players. We categorize two types of decisions: 1) management decisions and 2) medical decisions.

2.4.1 Management decisions

Players have to manage: 1) resources, 2) processes, and 3) financial issues.

2.4.1.1 Resource management

All online health care games account for resource management decisions. For example, players plan for beds, equipment (e.g., computer systems, radiology rooms, diagnostic devices, picture diagnostic units, laboratory machines, operating rooms, pharmaceutical equipment), staff (e.g., nurses, doctors, radiology assistants), therapeutic and diagnostic procedures, and additional agencies, centres, clinics, and facilities. Staff planning is incorporated by the majority of online health care games. Then, either beds or equipment are managed by the players. Furthermore, stock keeping issues arise and can be solved by using OR techniques. However, a main focus is on the planning of equipment while the number of beds is fixed. For example, Hartmann et al. [11] incorporate the management of pharmaceutical equipment.

2.4.1.2 Process management

The illustration of scheduling issues constitutes a main target of health care games. Players make scheduling decisions for the following areas: 1) human (e.g., patients, staff) and resource allocation (e.g., beds), 2) human scheduling, and 3) service planning.

Only a few health care games account for human or resource allocation. Human scheduling denotes a key feature, especially patient and staff scheduling. As patients have to queue for a lot of services, it is essential that the host of the game teaches some basics of queuing theory and priority rules for scheduling to the players. Furthermore, OR basics on staff scheduling should be explained to the players.

Services have also to be planned such as in emergency departments, laboratories, operating rooms, pharmacies, radiology rooms, and transportation departments. For example, players of Coremain hospital [22] decide on priority rules for queuing and types of queues in operating and radiology rooms. Furthermore, critical decisions on admission and discharge of patients are also made by the players. Different assignment strategies can be studied by the players.

2.4.1.3 Financial management

Players have to learn that their decisions do not only have logistical and medical consequences, but mainly financial ones. Especially in times of scarce budgets, this feature of health care games becomes crucial. Only well-managed health care facilities using OR techniques (e.g., simulation, queuing theory, resource allocation models, assignment strategies, stock keeping methods, staff scheduling approaches, and principle-agent theory) can survive in the long run even if they are publicly-owned.

Therefore, all health care games include financial decisions to be made by their players such as: 1) budget planning, 2) investment planning, 3) external financial investment planning, 4) reimbursement planning, 5) reimbursement negotiation, and 6) additional services and charges determination. For example, a main target of Coremain hospital [22] consists of illustrating the financial, managerial, and medical consequences of different patient reimbursement strategies for hospitals, staff, and patients. Therefore, it is essential to explain basics of principle-agent theory to the players.

2.4.2 Medical decisions

Medical decisions include diagnosis, emergency handling, and treatment of patients. For example, in the Lactation Game [3] players have to choose different breastfeeding strategies.

3 CONCLUSION

Since World War II, both OR and management games have gained importance to better plan for scarce resources. By using management games, knowledge can be applied to an artificial setting with a high learning effect, especially computer-based games. The internet opened new possibilities to overcome the distance and time problem by bringing together participants from all over the world.

The majority of health care games are complex online games applied to an entire health care institution, mostly hospitals. For example, in Coremain hospital [22] players run an entire hospital competing with players of other hospitals for patients and budget in a region. Although the main focus on health care management game lay on teaching and sometimes on policy making in the past, the incorporation of research issues was added in the last years. Experimental economics can be applied to investigate both teaching (e.g., game situation) and policy issues (e.g., impact of reimbursement systems) in the future.

Health care games support players to understand decision making in the management field (resources, processes, and finance) as well as in the medical field. Health care games mainly focus on managerial decisions. Financial decisions especially became crucial in the last years due to scarce budgets, increasing health care demand, and technology costs. To optimize reimbursement for health care institutions, policy makers have to apply more intensively OR techniques such as simulation, queuing theory, resource allocation models, assignment strategies, stock keeping methods, staff scheduling approaches, and principle-agent theory. Management games can be a starting point to get in touch with such techniques and to overcome barriers in practice to apply OR methods, especially in fields with publicly-owned institutions such as hospitals.

References

[1] Dempsey, J., Haynes, L., Lucassen, B., Casey, M., 2002. Forty simple computer games and what they could mean to educators. Simulation & Gaming, Vol. 33, No. 2, pp. 157 – 168.

[2] Dieleman, H., Huisingh, D., 2006. Games by which to learn and teach about sustainable development: exploring the relevance of games and experiential learning for sustainability. Journal of Cleaner Production, Vol. 14, No. 9-11, pp. 837 – 847.

[3] Elder, S., Gregory, C., 1996. The "Lactation Game": An Innovative Teaching Method for Health Care Professionals. Journal of Human Lactation, Vol. 12, No. 2, pp. 137 – 138.

[4] Faria, A., 1998. Business Simulation Games: Current Usage Levels – An Update. Simulation & Gaming, Vol. 29, No. 3, pp. 295 – 308.

[5] Faria, A., Dickinson, J., 1994. Simulation Gaming for Sales Management Training. The Journal of Management Development, Vol. 13, No. 1, pp. 47 – 59.

[6] Faria, A., Wellington, W., 2004. A Survey of Simulation Game Users, Former-Users, and Never-Users. Simulation & Gaming, Vol. 35, No. 2, pp. 178 – 207.

[7] Feinstein, A., Mann, S., Corsun, D., 2002. Charting the experiential territory: Clarifying the experiential territory – Clarifying definitions and uses of computer simulation, games, and role play. The Journal of Management Development, Vol. 21, No. 9/10, pp. 732 – 744.

[8] Geilhardt, T., Mühlbradt, T. (eds.), 1995. Planspiele im Personal- und Organisationsmanagement. Verlag für Angewandte Psychologie, Göttingen .

[9] Graf, J. (ed.), 1992. Planspiele – simulierte Realitäten für den Chef von morgen. Gabal Verlag GmbH, Speyer.

[10] Hans, E., Nieberg, T., 2007. Operating Room Manager Game. INFORMS Transactions on Education, Vol. 8, No. 1, pp. 25 – 36.

[11] Hartmann, A., Cruz, P., 1998. Interactive Mechanisms for Teaching Dermatology to Medical Students. Archives of Dermatology, Vol. 134, No. 6, pp. 725 – 728.

[12] Hofmarcher, M., Rack, H., 2006. Gesundheitssysteme im Wandel. MWV Medizinsich Wissenschaftliche Verlagsgesellschaft OHG, Berlin.

[13] Kennedy, D., Fanning, K., Thornton, P., 2004. The Age Game: An Interactive Tool to Supplement Course Material in a Geriatrics Elective. American Journal of Pharmaceutical Education, Vol. 68, No. 5, Article 115.

[14] Kraus, M., Rauner, M., Schwarz, S., 2009. The Potential of Health Care Games for Teaching and Policy Decision Making. Working paper, University of Vienna.

[15] Lane, D., 1995. On a Resurgence of Management Simulations and Games. The Journal of the Operational Research Society, Vol. 46, No. 5, pp. 604 – 625.

[16] Meterissian, S., Liberman, M., McLeod, P., 2007. Games as teaching tools in a surgical residency. Medical Teacher, Vol. 29, Vol. 9, pp. e258 – e260.

[17] Morton, P., Tarvin, L., 2001. The Pain Game: Pain Assessment, Management and Related JCAHO Standards. The Journal of Continuing Education in Nursing, Vol. 32, No. 5, pp. 223 – 227.

[18] OECD Health Data, July 2009.

[19] Panosch, B., 2008. Management Games: A powerful tool to teach competence and knowledge? Master thesis, University of Vienna.

[20] Peters, V., Vissers, G., Heijne, G., 1998. The Validity of Games. Simulation & Gaming, Vol. 29, No. 1, pp. 20 – 30.

[21] Rau, E., 2005. Combat science: the emergence of Operational Research in World War II. Endeavour, Vol. 29, No. 4, pp. 156 – 161.

[22] Rauner, M., Kraus, M., Schwarz, S., 2008. Competition under different reimbursement systems: The concept of an internet-based hospital management game. European Journal of Operational Research, Vol. 185, No. 3, pp. 948 – 963.

[23] Watson, H., Blackstone, J., 1981. Computer Simulation, 2nd edition, John Wiley&Sons, New York.

[24] Ziv, A., Wolpe, P., Small, S., Glick, S., 2003. Simulation-Based Medical Education: An Ethical Imperative. Academic Medicine, Vol. 78, No. 8, pp. 783 – 788.

MODELING OF ROBUSTNESS AGAINST SEVERE UNCERTAINTY

Moshe Sniedovich
University of Melbourne, Department of Mathematics and Statistics
VIC 3010, Australia
moshe@unimelb.edu.au

Abstract: Decision-making in the face of severe uncertainty is an onerous task. The difficulty here is that the severity of the uncertainty hinders the quantification of the likelihood of future events, thereby exposing decision-making to unforeseen risks. Unsurprisingly, the state of the art is such that no fully satisfactory formula for dealing with problems involving decision-making under severe uncertainty exists. This paper examines some of the conceptual, methodological and practical aspects of this important area of decision theory, with the emphasis being on the modeling aspects of robustness against severe uncertainty.

Keywords: severe uncertainty, robust decisions, decision theory.

1 INTRODUCTION

The ongoing crisis in the global economy and financial markets has put decision-making under uncertainty in the spotlight. For one thing, it has highlighted - as have Nassim Taleb's best-selling books [18,19] on "Black Swans" – the need to examine philosophically and technically the ability of formal mathematical models to adequately represent and cope with high-impact-low-likelihood events. Add to this the almost daily appearance in the press and on the web of headlines such as these:

"US to reveal **robustness** of its top banks' financial health"
"**Stress test** details to be disclosed for 19 biggest US banks"

and the picture is complete. The expectation is that decision under severe uncertainty be based on some sort of robustness criterion.

But the point to note here is that however intuitive and compelling the term "robustness" may be, its exact definition is not necessarily straightforward. Indeed, its definition and usage can vary significantly depending on the situation under consideration, and the modeling paradigm one uses to describe this situation formally.

In this paper we examine the basic issues that arise in the formulation of the concept "robustness" in the context of non-probabilistic mathematical models that are used to describe decision-making problems subject to severe uncertainty. In Operations Research, such models are associated primarily with the fields of classical Decision Theory [3,5,11,24] and Robust Optimization [2,4,6,8,9,10,12,14,15,16,25]. But the point is that, notwithstanding its obvious importance for the field of Operations Research, treatment of the concept "robustness" in the "non-specialized" Operations Research literature is, shall we say, marginal. Indeed, it is treated as a rather specialized concept to be taken up by the specialized literature. So, a formal discussion of this concept would, for instance, be absent from popular undergraduate textbooks [5,24].

We begin, in Section 2, with a brief discussion of the term "severe uncertainty" and its quantification via non-probabilisitc models. Then, in Section 3, we look at the treatment given the concept "robustness" in Operation Research. This is followed by a technical discussion in Section 4 on the classification of various types of robustness and the relationship between them. In Section 5 we consider some mathematical modeling aspects of robustness with the view to illustrates some of the subtle considerations that need to be reckoned with in the definition, modeling and interpretation of this concept. Certain issues in scenario generation relevant to this discussion are examined in Section 6.

2 SEVERE UNCERTAINTY

Recall that classical decision theory [3,11] distinguishes between three types of conditions that pertain to a state-of-affairs, namely

✯ Certainty ✯ Risk ✯ Uncertainty

The "Risk" category refers to situations where the uncertainty is taken to be amenable to quantification by standard probabilistic constructs such as probability distributions. In contrast, the "Uncertainty" category refers to situations where our knowledge about a parameter of interest is so meagre that the uncertainty cannot be quantified even by means of an "objective" probability distribution.

The point is then that "Uncertainty" eludes "measuring". It is simply impossible to provide a means by which we would "measure" the level, or degree, or intensity of "Uncertainty" to thereby indicate how great or daunting it is. To make up for this difficulty, a tradition has developed whereby the level, or degree, of "Uncertainty" is captured descriptively, that is informally, through the use of "labels" such as:

Strict, Severe, Extreme, Deep, Substantial, Essential, Hard,
High, True, Fundamental, Wild, Knightian, True Knightian,

and … more.

The trouble is, however, that all too often, these terms are used rather carelessly and imprecisely. Whereas, if they are to function properly in the context of a mathematical model depicting a given problem, it is imperative to be clear on the precise meaning of the term that is used in the framework concerned.

In this discussion we understand the term "severe uncertainty" to designate situations where the uncertainty under consideration holds with regard to a given parameter of interest, call it u, whose "true" value, call it u*, is unknown. All we know is that u* is an element of some given set **U**. But, we have no clue as to how likely it is that a given u in **U** is the true value. In short, the uncertainty is non-probabilistic and likelihood-free.

Note that methodologically we impose no conditions whatsoever on what u is, or what it represents. From a mathematical point of view u can be a scalar, a vector, a matrix, a set, a function and so on. All we have is a set **U** and all we know is that the true value of u is an element of **U**.

It is against this kind of severe uncertainty that our investigation of robustness of decisions is conducted. Meaning that we seek to determine what decisions will perform (relatively) well, despite the true value of the parameter of interest being unknown and despite our ignorance as to the likelihood of which element of **U** is the true value of u. We are completely in the dark here.

3 ROBUSTNESS

To make clear what "Robustness" connotes in this discussion, consider the first paragraph of the entry *Robustness* in WIKIPEDIA. It reads as follows:

> ***Robustness** is the quality of being able to withstand stresses, pressures, or changes in procedure or circumstance. A system, organism or design may be said to be "robust" if it is capable of coping well with variations (sometimes unpredictable variations) in its operating environment with minimal damage, alteration or loss of functionality.*

So in the context of this discussion, we say that a *decision is robust* if its outcomes or consequences cope well with changes in the decision-making environment under consideration caused by severe uncertainty in the true value of the parameter of interest $u \in \mathbf{U}$.

We consider two paradigms for modeling robustness: the one used in *classical decision theory* [3,11] and the one used in *robust optimization* [2,8,14].

3.1 Classical decision theory

In this framework, which was formulated by Wald [21,22,23], decision-making under severe uncertainty is cast as a *game* between a decision maker and antagonistic *Nature*, where the parameter whose true value is subject to severe uncertainty is represented by the *state* of Nature. In this framework, the decision maker seeks the best (relative to her decisions) worst-case (relative to the state of nature) outcome.

In contrast to the set-up in *Game Theory* [5,20,24], here it is assumed that the decision-maker plays first. Namely, the decision-maker is first to make a decision, whereupon Nature responds by selecting her state accordingly. This game is captured in the famous *Maximin model* [3,11,15,21,22,23]. For our purposes here it will be convenient to formulate it as follows:

$$\textit{Maximin Model:} \quad z^* := \max_{d \in D} \min_{s \in S(d)} f(d,s) \qquad (1)$$

where D denotes the *decision space*; S(d) denotes the set of *states* associated with decision d, and f denotes the *objective function*. Note that in this framework the state s embodies the uncertain parameter of interest whose true value is subject to severe uncertainty.

The robustness of decision d is stipulated by

$$z(d) := \min_{s \in S(d)} f(d,s) \,,\ d \in D \qquad (2)$$

which is commonly called *the security level* of decision d. Also note that, by definition, if the decision maker selects decision d then the reward generated by this decision will be at least as large as z(d) no matter what state $s \in S(d)$ is selected by Nature. The Maximin model prescribes that the decision maker select a decision with the highest security level possible.

There are other variations on this theme: Savage's *Minimax regret* model and Hurwicz' *Pessimisim-Optimizim model* [3,7,11,13].

3.2 Robust optimization

In this framework, a parameter whose true value is subject to severe uncertainty is incorporated in the conventional optimization model. To explain the difficulties encountered in the definition of robustness in this context, consider the following parametric optimization problem:

$$\text{Problem P(u):} \quad x^*(u) := \arg \mathop{opt}_{x \in X(u)} g(x,u) \ ,\ u \in \mathbf{U} \qquad (3)$$

where x denotes the decision variable, g denotes the objective function and X(u) denotes the decision space associated with u. Typically X(u) is defined by *constraints* on the decision variable x.

By definition, x*(u) denotes the decision that optimizes g(x,u) over x in X(u) for a *given* value of u. So, a robust solution to this parametric problem should be a solution x that "performs well" with respect to both X(u) and g(x,u) for all u in **U**.

Ideally then, a robust decision is a decision that is optimal for Problem P(u) for all u in **U**. We shall refer to such a decision as a *super-robust decision*. The difficulty obviously is that super-robust decisions rarely exist. But even in cases where they do exist, they are rather innocuous hence of little interest. In any event, no super-robust decisions typically exists for (3), so we need to settle for a less demanding definition of robustness.

The question is then: what formulations would be suitable for the robust-optimization counterparts of Problem P(u), u $\in \mathbf{U}$?

The classification discussed in the next section provides a partial answer to this important question and sheds some light on the difficulties associated with this task.

4 CLASSIFICATION OF ROBUSTNESS

We examine two classification schemes: one distinguishes between robustness with respect to the objective function g and robustness with respect to the constraints of the optimization problem. The other distinguishes between various subsets of the uncertainty set **U** over which robustness is sought.

4.1 Robust satisficing vs robust optimizing

There are situation where g(x,u) is independent of u, in which case we can define the robust counter-part of Problem P(u), u $\in \mathbf{U}$ as follows:

$$\textit{Robust-satisficing Problem}: \ Z(U) := \underset{\substack{x \in X(u) \\ \forall u \in \mathbf{U}}}{opt} \ g(x) \tag{4}$$

Note that this definition assumes that the $\cap_{u \in U} X(u)$ is not empty. If this set is empty, then the robust-satisficing problem has no solutions. In such cases it may prove necessary to relax the requirement $\forall u \in \mathbf{U}$ under the opt operation.

In other situations, the sets X(u), u $\in \mathbf{U}$ are independent of u**,** in which case we can define the robust counter-part of Problem P(u), u $\in$U as follows:

$$\textit{Robust-optimizing Problem}: \ z^*(U) := \min_{x \in X} \max_{u \in U} \frac{|g^*(u) - g(x,u)|}{|g^*(u)|} \ , \ g^*(x) \neq 0 \tag{5}$$

where $g^*(u) := \underset{x \in X}{opt}\, g(x,u)$.

Alternatively, we can consider the following simplified approach, where we declare a decision $x^* \in X$ to be robust-optimizing iff

$$|g(x^*,u) - g^*(u)| \leq \varepsilon \, , \forall u \in \mathbf{U} \tag{6}$$

for some pre-specified positive value of ε. Note that this problem can be formulated as follows:

$$z^*(U) := \max_{x \in X} \min_{u \in \mathbf{U}} h(x,u) \tag{7}$$

where

$$h(x,u) := \begin{cases} 1 \ , & |g(x^*,u) - g^*(u)| \leq \varepsilon \\ 0 \ , & \text{otherwise} \end{cases} \tag{8}$$

Also, if in this case z*(U)=0 then the implication is that there are no robust-optimizing solutions for the specified value of ε. Other definitions of robustness can be found in [2,4,6,8,9,10,12,17,25].

4.2 Complete vs partial vs local robustness

Thus far, our definitions of robustness imposed some performance conditions on *all* u in **U**. We refer to such robustness as *complete*, meaning that *every element of **U** must satisfy the robustness conditions*.

Often such definitions of robustness are too demanding in the sense that either no decision can satisfy them, or that tracking down decisions possessing this kind of robustness proves too difficult, analytically and/or computationally.

In such cases one can weaken the robustness conditions by dropping the requirement that they be satisfied for all u in **U**. Instead, these conditions would be required to hold only for some subset **U'** of **U**.

For example, suppose that the uncertainty space is $\mathbf{U} = \{(u_1,u_2): 0 \le u_1,u_2 \le 1000\}$. Instead of requiring that the robustness conditions be satisfied for all u in U, we can let say $\mathbf{U}' = \{u \in \mathbf{U} : u \text{ integer}\}$, that is **U'** is the subset of **U** consisting of the integer elements of **U**.

Furthermore, we do not have to specify **U'** a priori. Indeed, we can allow it to be the largest subset of **U** over which the robustness conditions are satisfied. For example, let $\rho(V)$ denote the "size" of set $V \subseteq \mathbf{U}$ assuming that $\rho(V) > 0$ for any non-empty subset V of **U** and $\rho(V) > \rho(W)$ for any proper subset W of V. Also let,

$$\varphi(x,u,V) := \begin{cases} \rho(V) & , |\mathrm{g(x,u)} - \mathrm{g}^*(\mathrm{u})| \le \varepsilon \\ 0 & , \quad \text{otherwise} \end{cases} \tag{9}$$

and consider

$$\rho^*(U) := \max_{\substack{x \in X \\ V \subseteq U}} \min_{u \in V} \varphi(x,u,V) \tag{10}$$

By definition, an optimal solution (x*,V*) to this problem has the following property: V* is the largest subset V of **U** over which the robustness condition $|g(x,u) - g^*(u)| \le \varepsilon$ is satisfied by some decision for all u in V.

And if instead we let

$$\varphi(x,u,V) := \begin{cases} \rho(V) & , c(x,u) \le c^* \\ 0 & , \text{ otherwise} \end{cases} \tag{11}$$

then an optimal solution (X*,V*) to this problem has the property that V* is the largest subset of **U** over which the constraint $c(x,u) \le c^*$ is satisfied by some decision x in X. Note that this is a generalization of Starr's *domain-criterion* [17].

We refer to robustness of this type as *partial* to indicate that the robustness conditions are not required to be satisfied over the entire uncertainty set **U**.

Now, a crucially important point one must be vigilant about in the definition of partial robustness under conditions of severe uncertainty is our ignorance about the true value of the parameter u. Furthermore, that the model of uncertainty is non-probabilistic and likelihood-free. So, to insure that the defined partial robustness does indeed provide an appropriate indication as to how robust the decisions are relative to the uncertainty space **U**, the subsets V of **U** considered in the robustness analysis must be defined with care.

To illustrate this important point, we now consider a special type of partial robustness, namely *local* robustness. Let $\tilde{u}$ be an element of **U** and let $\mathrm{N}(\delta,\tilde{u})$ denote a neighborhood of "size" δ around $\tilde{u}$, assuming that $N(0,\tilde{u}) = \{\tilde{u}\}$ and that $N(\delta,\tilde{u}) \subseteq N(\delta+\varepsilon,\tilde{u}), \forall \delta,\varepsilon > 0$.

Now suppose that the sets V considered in (9) are such neighborhoods around $\tilde{u}$. Then the robustness model will be formulated as follows:

$$\delta^*(U,\tilde{u}) := \max_{\substack{x \in X \\ \delta \ge 0}} \min_{u \in N(\delta,\tilde{u})} \psi(x,u,\delta) \tag{12}$$

where

$$\psi(x,u,\delta) := \begin{cases} \delta \ , \ c(x,u) \le c^* \\ 0 \ , \ \text{otherwise} \end{cases} \tag{13}$$

But, in this case, the analysis may yield a decision that will not be robust relative to the uncertainty set **U**. Indeed, it is easy to come up with examples where the decision selected by (11)-(12) is far less robust with respect to the robustness condition $c(x,u) \le c^*$ relative to **U**, then other decisions, in the sense that other decisions satisfy this constraint on much larger subsets of **U**. This is illustrated in the next section.

The inference is that, robustness models such as (11)-(12) are appropriate in situations where the uncertainty under consideration is not severe, namely in cases where $\tilde{u}$ is a good, reliable estimate of the true value of u.

5 MATHEMATICAL MODELING ISSUES IN ROBUSTNESS ANALYSIS

Over the years, Wald's Maximin model and its many variants have become the predominant non-probabilistic paradigms for robust decision making under severe uncertainty. For this reason it is important to become adept at modeling the Maximin and, to be able to identify Maximin models particularly, if they are in disguise. In a word it is important to develop healthy mathematical modeling skills. To illustrate this point consider the model

$$\alpha^* := \max_{q \in Q} \max \left\{ \alpha \ge 0 : r_c \le R(q,u), \forall u \in U(\alpha,\tilde{u}) \right\} \tag{14}$$

where Q represents the decision space, u represents a parameter whose true value is subject to severe uncertainty, u represents a point estimate of the true value of u, R(q,u) represents the reward generated by (q,u), r represents the critical value of the reward, and $U(\alpha,\tilde{u})$ represent a neighborhood of size around $\tilde{u}$. It is assumed that $U(0,\tilde{u}) = \{\tilde{u}\}$ and $U(\alpha,\tilde{u}) \subseteq U(\alpha+\varepsilon,\tilde{u})$ for all $\alpha,\varepsilon \ge 0$.

It is a straightforward exercise to show that

$$\max_{q \in Q} \max \left\{ \alpha \ge 0 : r_c \le R(q,u), \forall u \in U(\alpha,\tilde{u}) \right\} = \max_{\substack{q \in Q \\ \alpha \ge 0}} \min_{u \in U(\alpha,\tilde{u})} \sigma(q,u,\alpha) \tag{15}$$

where

$$\sigma(x,u,\alpha) := \begin{cases} \alpha \ , \ R(q,u) \ge r_c \\ 0 \ , \ \text{otherwise} \end{cases} \tag{16}$$

the implication being that (13) is a Maximin model in disguise.

Yet, it is argued [eg. 1] that (13) is not a Maximin model. What is more, the theory based on this model is hailed as novel and radically different from all current theories for decision under uncertainty. The point to note here is that the argument culminating in these erroneous conclusions stems from poor mathematical modeling skills. This is brought out by the inability to incorporate the size α as a decision variable in the Maximin model and in the inability to construct an appropriate penalty function for this purpose such as that given in (15).

Similarly, it is a straightforward exercise to show that

$$\beta^* := \min_{q \in Q} \min \left\{ \alpha \ge 0 : r_c \le R(q,u), \text{ for some } u \in U(\alpha,\tilde{u}) \right\} \tag{17}$$

is a Minimin model, namey

$$\min_{q\in Q} \min \{\alpha \geq 0 : r_c \geq R(q,u), \text{ for some } u \in U(\alpha,\tilde{u})\} = \min_{\substack{q\in Q\\ \alpha\geq 0}} \min_{u\in U(\alpha,\tilde{u})} \vartheta(q,u,\alpha) \tag{18}$$

where

$$\vartheta(x,u,\alpha) := \begin{cases} \alpha \ , \ R(q,u) \geq r_c \\ \infty \ , \ \text{otherwise} \end{cases} \tag{19}$$

It is also noteworthy that, users and proponents of these models seem to be unaware that the robustness yielded by these models is *local* in that the decisions' robustness is evaluated only in the neighborhood of the estimate u. Add to this the fact that the models are assumed to operate under conditions of severe uncertainty namely, that this estimate is a wild guess and is likely to be substantially wrong; and the inevitable conclusion is that these models in effect fail to tackle the severity of the uncertiany under consideration [14,15,16]. And yet, these models are used to seek robust decisions under conditions of severe uncertainty.

So, to illustrate why local robustness of this type cannot be countenanced for cases subject to severe uncertainty, consider the case where the uncertainty space U is the real line and the robustness condition is a constraint of the form $R(q,u) \geq r_c$. The values of R(q,u) for two decisions q' and q'' is shown in Figure 1.

Now suppose that, $U(\alpha,\tilde{u}) = \{u \in \Re : |u - \tilde{u}| \leq \alpha\}$, and as indicated in Figure 1, set $r_c=3$ and $\tilde{u} = 0$. Then according to (13) decision q'' is more robust then decision q' despite q' satisficing the robustness condition almost everywhere on U, and q'' violating this condition almost everywhere on U. Since under conditions of severe uncertainty the estimate is a "wild guess", there is no reason to believe that q'' is more robust than q' on U.

As far as partial robustness is concerned, the definition of the size of subsets of the uncertainty space U should reflect the requirements of the specific problem in question, keeping in mind that the uncertainty model is non-probabilistic and likelhood-free.

To illusrate, consider the case where U is the 4 by 4 square centered at the origin (0,0) and the decision space consists of three decisions, that is $D=\{d^{(1)},d^{(2)},d^{(3)}\}$. Now, suppose that these decisions satisfy the robustness conditions on the subsets $V^{(1)}$, $V^{(2)}$ and $V^{(3)}$ repsectivey, as shown in Figure 2, where the subsets are represented by the respective shaded areas.

Note that the areas of the "safe" regions, where the robustness conditions are satisfied, are roughly the same (equal to approximately 8). Does this mean that the "actual" robustness of these decisions is also roughly the same?

Experience has shown that it is difficult to resolve such issues in the abstract, that is, in general terms. These issues are often problem-oriented and must be resolved in accordance with the practical considerations that pertain to the specific problem considered.

6 SCENARIO GENERATION

In many situations the uncertainty space U is not given so it must be "estimated" or "approximated". In others, it is given but technical/analytical considerations impede using it "as is", so it must be "approximated". The technical term used to describe the construction of approximations of the uncertainty space is "scenario generation" [8,9].

In the case of severe uncertainty, especially in the context of non-probabilisitc models of uncertainty, scenario generation can be an exacting complicated task because it requires considerable problem-specific expertise. One of the difficulties is to determine a proper balance between two conflicting objectives.

On the one hand, it is important – perhaps even imperative – to include in the robustness analysis Black Swans [19], that is high-impact low-likelihood scenarios. The incorporation of such scenarios in the analysis is to insure that the stress-test that the system is subjected to is

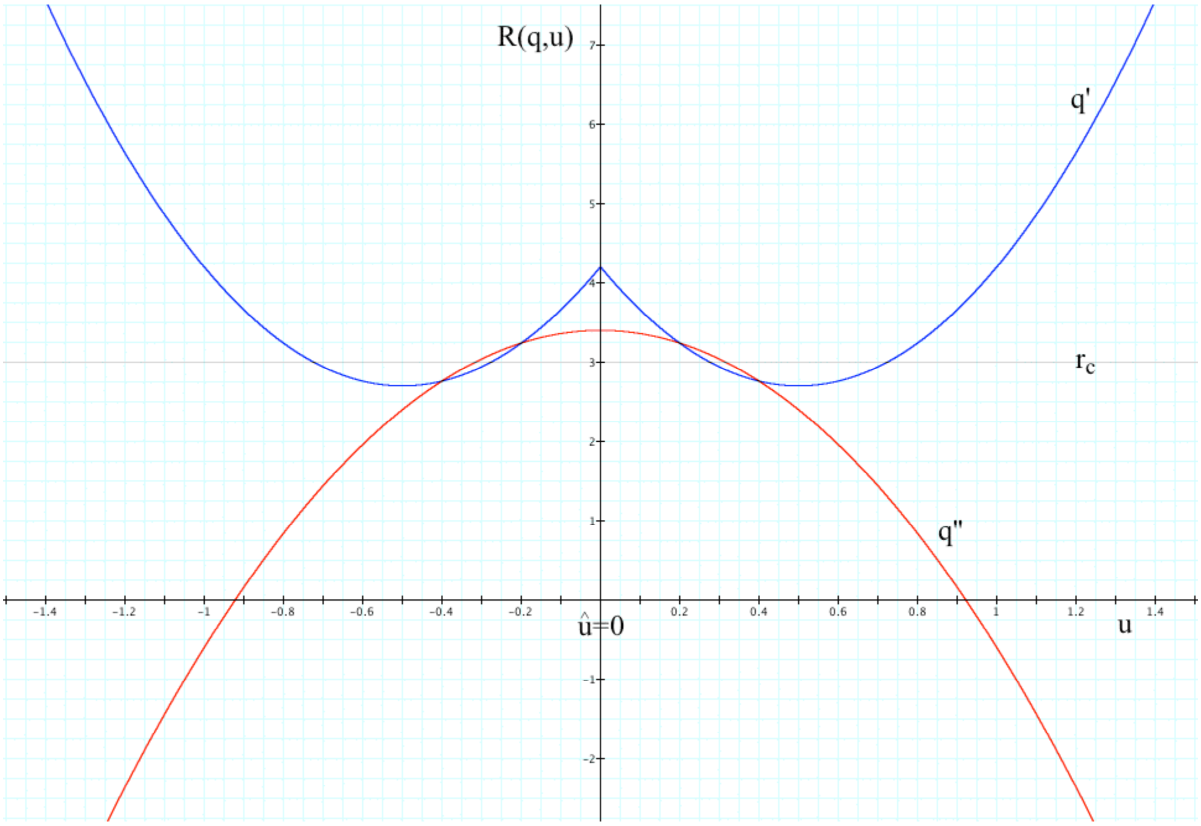


Figure 1: Performance of two decision variables.

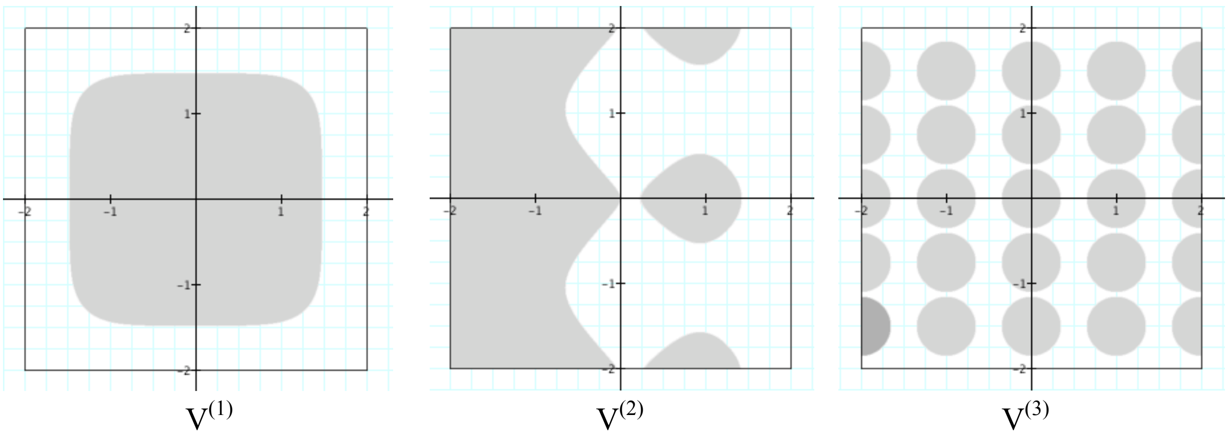


Figure 2: Partial robustness of three decisions.

sound and valid. On the other hand, it is rarely possible/desirable to incorporate in the analysis *doomsday scenarios*, as such scenarios may overwhelm the analysis and hinder a proper discrimination between the robustness of decisions of much "milder" scenarios. There are of course, exceptions:

> *Our editorial policy treats ALL doomsday scenarios equally. We don't play favorites, we don't have an agenda. Our sole purpose is to provide a comprehensive, organized and easy to use resource for matters relating to doomsday, Armageddon and other global disasters that could spell the end of the world. If you are serious about the end times and are looking for genuine resources, Doomsday Guide is a good starting point. Aren't you glad you got here before it was too late? Be sure to visit our Doomsday Videos section to see what might be headed our way.*
> http://www.doomsdayguide.org/doomsday_scenarios.htm (Downloaded: May 22, 2009)

It is fair to say that for these and other reasons, scenario generation in the famework of robust decision-making in the face of severe uncertainty is more art than science. The task is particularly difficult in cases where the uncertainty space is *unbounded.* It is therefore imperative to be skeptical of, or in the very least scrutinize carefully methodologies that offer simple recipes for handling such cases.

In particular, as indicated in the preceding section, it is important to be on guard against methodologies offering recipes that turn out to be grounded in *local robustness* as these are thoroughly unsuitable for the treatment of severe uncertainty. The picture is this:

where the shaded area represents the uncertainty space **U**, + represents the location of the estimate, and the black area represents the neighborhood of the estimate in which the robustness analysis is conducted.

A case in point is Info-Gap decision theory [1] whose purported ability to cope with *unbounded uncertainty spaces* is cited as one of its major attractions. This view about Info-Gap's capabilities stems from a failure to appreciate that Info-Gap's robustness analysis is *local* par excellence. Namely, that it takes no account whatsoever of the decision's performance in areas of the uncertainty space **U** that are distant from the estimate of the parameter of interest. Indeed, that its oblivion to decisions' perform in the overwhelming area of the uncertainty space turn most of this space into a *No Man's land* [14,16].

The point here is that the trouble with local models of robustness is that not only are they incapable of dealing with *Black Swans*, they may even fail to deal with plain, ordinary, *White Swans* that reside just outside the local neighborhood of **U** in which the robustness analysis is conducted.

Finally, as for the question whether Nassim Taleb's popular books [18,19] will have any impact on progress and developments in the area of scenario generation, the answer inevitably is: only time will tell!

7 CONCLUSIONS

We have seen that "robustness" cannot be treated as a "one size fits all" criterion. To begin with, the seemingly obvious meaning of the concept "robustness" cannot be taken for granted to allow using it "as is". Indeed, if this concept is to function properly in the framework in which it is applied, especially in the context of decision under severe uncertainty, its meaning must be made clear and it must be given a precise definition.

Likewise, the mathematical definition of robustness cannot be done in the abstract, in general terms. It must reflect the tasks, objectives, constraints etc captured by the decision-making problem under consideration. This is of the essence in the context of decision in the face of uncertainty where the difficulties posed by the severity of the uncertainty are fundamental.

As for defining robustness, one must particularly be vigilant about definitions yielding local robustness. For, as illustrated by the case of Info-Gap, definitions yielding local robustness cannot possibly be countenanced for the treatment of decision problems subject to severe uncertainty. This is so not only because local robustness has no point or merit under conditions of severe uncertainty, it is because local robustness accomplishes the precise opposite of what is sought: reliable decisions-making in the face of uncertainty.

References

[1] Ben-Haim. Y., 2006. Info-Gap Decision Theory: Decisions Under Severe Uncertainty, Elsevier, Amsterdam.

[2] Ben-Tal A., El Ghaoui, L. and Nemirovski, A., 2006. Mathematical Programming, Special issue on Robust Optimization, Volume 107(1-2).

[3] French, S., 1988. Decision Theory, Ellis Horwood, NY.

[4] Gupta, S.K., and Rosenhead, J., 1968. Robustness in Sequential Investment Decisions, Management Science, 15(2) , pp. B-18-B-29.

[5] Hillier, F.S., and Lieberman, G.L., 2001. Introduction to Operations Research, McGraw-Hill.

[6] Huber, P.J., 1981. Robust Statistics, Wiley-Interscience.

[7] Hurwicz, L., 1951. Optimality criteria for decision making under ignorance, Cowles Commission Discussion Paper No. 370.

[8] Kouvelis P. and Yu G., 1997. Robust Discrete Optimization and Its Applications, Kluwer.

[9] Mulvey, J.M., Vanderbei, R.J., Zenios, S.A., 1995. Robust Optimization of Large-Scale Systems Operations Research, 43(2), 264-281.

[10] Mutapcic, A. and Boyd, S., 2009. Cutting-set methods for robust convex optimization with pessimizing oracles, Optimization Methods and Software, 24(3), 381- 406.

[11] Resnik, M.D., 1987. Choices: an Introduction to Decision Theory, University of Minnesota Press, Minneapolis, MN.

[12] Rosenhead, J.V., Elton, M., and Gupta, S.K., 1972. Robustness and optimality as criteria for strategic decisions, Operational Research Quarterly, 23, 413- 431.

[13] Savage, L.J., 1951. The theory of statistical decision, Journal of the American Statistical Association, 46, 56-67.

[14]Sniedovich, M., 2007. The Art and Science of Modeling Decision-Making Under Severe Uncertainty, Journal of Decision Making in Manufacturing and Services, 1(1-2), 111-136.

[15] Sniedovich, M., 2008. Sniedovich M., Wald's Maximin Model: A Treasure in Disguise!, Journal of Risk Finance, 9(3), 278-291.

[16] Sniedovich, M., 2008. A Critique of Info-Gap Robustness Model. In: Martorell et al. (eds), Safety, Reliability and Risk Analysis: Theory, Methods and Applications, pp. 2071-2079, Taylor and Francis Group, London.

[17] Starr, M.K., 1966. A discussion on some normative criteria for decision-making under uncertainty, Industrial Management Review, 81(1), 71-78.

[18] Taleb, N.N., 2001. Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets. Random House.

[19] Taleb, N.N., 2007. The Black Swan: The Impact of the Highly Improbable, Random House.

[20] Von Neumann, J., 1928. Zur Theorie der Gesellsehaftsspiele, Mathematisehe Annalen, 100, 295-320

[21] Wald, A., 1939. Contributions to the theory of statistical estimation and testing hypotheses, The Annals of Mathematics, 10(4), 299-326.

[22] Wald, A., 1945. Statistical decision functions which minimize the maximum risk, The Annals of Mathematics, 46(2), 265-280.

[23] Wald, A., 1950. Statistical Decision Functions, John Wiley, NY.

[24] Winston, W.L., 2004. Operations Research: Applications and Algorithms, Thomson.

[25] Wong, H-Y., and Rosenhead, J., 2000. A rigorous definition of robustness analysis, Journal of the Operational Research Society, 51, 172-182.

VECTOR APPROACH TO CHANGEABLE IMPORTANCE OF PERIOD OBJECTIVES IN MULTIOBJECTIVE DYNAMIC PROGRAMMING

Tadeusz Trzaskalik
The Karol Adamiecki University of Economics in Katowice
Department of Operations Research
ul. Bogucicka 14, 40-226 Katowice, Poland
ttrzaska@ae.katowice.pl

Abstract. When analyzing multiple objective, multiperiod decision processes it is worth while to describe relations among objectives considered in consecutive periods and objectives for the whole process. The aim of the paper is to describe a situation in which period objectives have changeable importance in consecutive periods. The problem of α-importance of period objectives is formulated and definitions of the importance function for the considered process and of α-efficient realizations are given. Relations between sets of efficient solutions and α-efficient solutions are analyzed, and some decision recommendations are given. Numerical examples are also presented.

Keywords: multiobjective dynamic programming, multiperiod decision process, period objectives, multiperiod objectives, α-efficient realization.

1 INTRODUCTION

Several real life decision problems can be characterized by conflicting objectives and multiple stages. They are often considered as multiple criteria dynamic problems. Many research papers have been published in this field. Various theoretical and applied directions of research have been presented. Among them worth mentioning are papers devoted to multiobjective shortest path problems [5, 3, 25], time dependency [6, 7], generating efficient solutions [2], multiobjective optimal control [11, 12, 26], fuzzy approach [1,8,10], computational methods [13, 14, 24], ecological problems [4, 15, 16, 27]. Surveys of research were presented by Li and Haimes [9] and Trzaskalik [18].

In this article — which is a continuation of the author's earlier papers — multiobjective, multiperiod finite decision processes with fixed number of periods are considered.

We will consider period objectives, equivalent period criteria functions, and their compositions, called "multiperiod criteria functions".

Objectives describing the same aspect of the process in different periods are called monomial. The set of all objectives taken into consideration in evaluation of a process can be divided into groups of monomial objectives. An objective is important in the considered period if the equivalent period criterion is taken into consideration in the evaluation of the whole process realization.

The importance of considered period objectives can change during the run of the process. Such problems were considered in [19, 20, 21]. It was assumed that the importance of objectives could be characterized by means of a zero-one objective importance matrix. If an objective was important in the period in question, the corresponding value in this matrix was equal to one; if it was not important, the value was equal to zero. Two approaches were proposed. In the first [8], the objective was important (or not) depending on the values reached by the equivalent criteria in the previous periods. In the second [21], the objective was important (or not) depending on the state of the process at the beginning of the considered period.

This paper follows the first approach and presents its extension. We assume that the period objective can be important in the period in question and the weight of its importance is a number from the range [0, 1]. The value zero means that the value of the equivalent criterion is not taken into consideration in the evaluation of the whole process realization; the value one means that the considered period objective is fully important. Intermediate values are also admissible – the closer a value is to one, the more important objective is. It is assumed that all the weights are given by the decision maker before the process starts.

The problem of solving such a problem is considered as a vector maximization problem (as opposed to the assumptions in [19], where scalarization approach has been applied). The problem of definition of efficient solutions arises when changeable importance of period objective is taken into account.

To solve this problem, α-important efficient realizations are defined. When choosing the final solution, the decision maker can use two sets of realizations – the set of efficient realizations and the set of α-efficient realizations. Two questions arise: the first about the relationship between these two sets and the second about the relationship between the relation of domination in the classical sense and the relation of domination with changeable importance of objectives, which is introduced in the paper.

The aim of the paper is to give the formal description of the situation presented above and to propose a draft of a method of choosing the final solution when period objectives have changeable importance in consecutive periods.

The paper consists of six sections. In section 2 the classical dynamic vector optimization problem is formulated. Notation and basic notions relevant to multiperiod, multicriteria decision processes, the definition of separability and monotonicity and an applied version of Bellman's Principle of Optimality are presented. In section 3 the problem of α-importance of period objectives is formulated and the definitions of importance function for considered process and α-efficient realizations are given. In section 4 relations between sets of efficient solutions and α-efficient solutions are analyzed, and some decision recommendations are given. In section 5 numerical examples are presented. Concluding remarks in section 6 end the paper.

2 DYNAMIC VECTOR OPTIMIZATION PROBLEM

2.1 Notation and basic concepts

We define $\overline{1,T}$ as the set of all integer numbers from 1 to T and denote it as follows:

$$\overline{1,T} = \{1,\dots,T\}$$

We consider a discrete decision process consisting of T periods. Let y_t be the state variable in the period t, Y_t –the set af all feasible state variables for the period t, x_t – the decision variable for the period t and $X(y_t)$ – the set of all feasible decision variables for the period t and the state y_t. We assume that all sets of states and decisions are finite (or compact). A period realization is defined as follows:

$$d_t \cong (y_t, x_t)$$

D_t is the set of all period realizations in period t.

We assume that for $t \in \overline{1,T}$ the transformations

$$\Omega_t: D_t \rightarrow Y_{t+1}$$

are given. $d = (d_1,\dots,d_T)$ is called a process realization. D is the set of all the process realizations. We define

$$D_t(y_t) \cong \{(y_t, x_t): x_t \in X_t(y_t)\}$$

to be the set of all period realizations of the process which begin in the given state y_t.

We assume that for each period t, K period criteria functions $F_t^k : D_t \rightarrow R$ are defined. For the given realization d we obtain the values

$$
\begin{array}{ccc}
F_1^1(d_1) & F_1^2(d_1)\ldots & F_1^K(d_1) \\
\ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots\ldots \\
F_T^1(d_T) & F_T^2(d_T)\ldots & F_T^K(d_T)
\end{array}
$$

Each M-dimensional (M≥2) function with components

$$F^m = \Phi^m \left(F_1^1, F_1^2, \ldots, F_1^K, F_T^1, F_T^2, \ldots, F_T^K\right)$$

for $m \in \overline{1,M}$, where Φ^m is a scalar function , can be considered as a vector-valued criterion function. We have

$$F \cong [F^1,\ldots,F^M]'$$

The components F^m are called multiperiod criteria functions. We want to maximize all components of F.

Let us assume that two process realizations: $\overline{d}$, $\tilde{d}$ and vectors

$$F(\overline{d}) \cong [F^1(\overline{d}), \ldots F^M(\overline{d})]$$

$$F(\tilde{d}) \cong [F^1(\tilde{d}), \ldots F^M(\tilde{d})]$$

are given. The relation of domination ≥ is defined as follows:

$$F(\overline{d}) \geq F(\tilde{d}) \Leftrightarrow$$

$$\underset{m \in \overline{1,M}}{\forall} [F^m(\overline{d}) \geq F^m(\tilde{d})] \quad \wedge \quad \underset{i \in \overline{1,M}}{\exists} [F^m(\overline{d}) > F^m(\tilde{d})]$$

If $F(\overline{d}) \geq F(\tilde{d})$ we say that the vector $F(\overline{d})$ dominates $F(\tilde{d})$ and the realization $\overline{d}$ is better than the realization $\tilde{d}$. The realization $\overset{*}{d}$ is efficient iff

$$\sim \underset{\overline{d} \in D}{\exists} F(\overline{d}) \geq F(\overset{*}{d})$$

We denote by $\overset{*}{D}$ the set of all efficient realizations for the given criterion function F. The problem of finding $\overset{*}{D}$ is called the dynamic vector optimization problem. The set

$$\overset{*}{D}(\overline{d}) \cong \{\overset{*}{d} \in \overset{*}{D} : F(\overset{*}{d}) \geq F(\overline{d})\}$$

consists of all efficient realizations which are better than $\overline{d}$.

2.2 Separability

F^m is scalar backward separable if for $t \in \overline{1,T}$ and $m \in \overline{1,M}$ there exist functions $f_t^m(F_t^1,\ldots,F_t^K)$ and operators o_t^m such that

$$F^m = f_1^m o_1^m (f_2^m o_2^m (\ldots (f_{T-1}^m o_{T-1}^m f_T^m)\ldots))$$

The last function in the sequence $\{g_t^m : t \in \overline{T,1}\}$ defined as follows:

$$g_t^m \cong f_T^m$$
$$g_t^m \cong f_T^m \, o_T^m \, g_{t+1}^m$$
fulfills the condition
$$g_1^m \cong F^m$$
F is backward separable if each component of F^m is scalar backward separable.

Let
$$F_t \cong [f_t^1,\ldots,f_t^M],$$
$$G_t \cong [g_t^1,\ldots,g_t^M],$$
$$o_t \cong [o_t^1,\ldots,o_t^M],$$
Then
$$F = F_1 \, o_1 \, (F_2 \, o_2 \, (\ldots (F_{T-1} \, o_{T-1}) \ldots))$$
$$G_T = F_T$$
$$G_t = F_t o_t \, G_{t+1}$$
$$G_1 = F$$
Forward separability can be defined in the similar way [18].

2.3 Monotonicity

Let B be the set of such elements for which the operation a o_t b is feasible. We define:
$$a \bullet_t B \cong \{c: \exists_{b \in B} \; c = a \, o_t \, b\}$$
Let F is backward separable. Attainable sets $W(y_t)$ for $t \in \overline{T,1}$ are defined as
$$W(y_T) \cong \bigcup_{d_T \in D_T(y_T)} F_T(d_T)$$
$$W(y_t) \cong \bigcup_{d_t \in D_t(y_t)} F_t(d_t) \bullet_t W(\Omega_t(d_t))$$
F is backward monotone iff
$$\forall_{t \in \overline{1,T}} \; \forall_{y_t \in Y_t} \; \forall_{d_t \in D_t(y_t)} \; \forall_{w, \overline{w} \in W(\Omega_t(d_t))} \{ w \geq \overline{w} \Rightarrow [F_t(d_t) \, o_t \, w] \geq [F_t(d_t) \, o_t \, \overline{w}] \}$$

Forward monotonicity can be defined in the similar way [18].

2.4 Decomposition of the process

If F is backward separable and backward monotone, we can formulate and prove Backward Principle of Optimality.

Theorem 1

If
$$\overset{*}{d} \cong (\overset{*}{y}_1, \overset{*}{x}_1, \ldots, \overset{*}{y}_T, \overset{*}{x}_T) \in \overset{*}{D}$$
then
$$\sim \exists_{t \in \overline{1,T}} \; \sim \exists_{\tilde{d}(\overset{*}{y}_t) \in D(y_t)} \; G_t(\tilde{d}(\overset{*}{y}_t)) \geq G_t(\overset{*}{d}(y_t)) \qquad \blacksquare$$
The proof of the theorem can be found in [18].

For the forward separable and forward monotone function F, the forward principle of optimality can be formulated in the similar way [18].

Applying backward (or forward, if applicable) Principle of Optimality, we can formulate the optimality equations and decompose the process to find the sets $F(\overset{*}{D})$ and $\overset{*}{D}$. For a given realization $\bar{d}$ we can determine whether it is efficient or not and – if $\bar{d} \notin \overset{*}{D}$ – construct the set $\overset{*}{D}(\bar{d})$. The algorithm for testing efficiency is described in [18, p.54].

3 α-IMPORTANCE

Assume that for $t \in \overline{1,T}$, $k \in \overline{1,K}$ the matrix **α** which consists of the values $\alpha_t^k \in [0,1]$, describing the importance of period objectives, is given. The dimension of that matrix is $T \times K$. For the given process realization d we obtain the following values of the equivalent period criteria functions:

$$\alpha_1^1 F_1^1(d_1) \quad \alpha_1^2 F_1^2(d_1) \quad \ldots \quad \alpha_1^K F_1^K(d_1)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\alpha_T^1 F_T^1(d_T) \quad \alpha_T^2 F_T^2(d_T) \quad \ldots \quad \alpha_1^K F_T^K(d_T)$$

The M-dimensional function $\overset{\alpha}{F}$ with the components:

$$\overset{\alpha}{F}{}^m = \Phi^m(\alpha_1^1 F_1^1, \alpha_1^2 F_1^2, \ldots, \alpha_1^K F_1^K, \ldots, \alpha_T^1 F_T^1, \alpha_T^2 F_T^2, \ldots, \alpha_T^K F_T^K)$$

is defined as a vector-valued importance function. We have:

$$\overset{\alpha}{F} \cong [\overset{\alpha}{F}{}^1, \ldots, \overset{\alpha}{F}{}^M]'$$

Realization $\overset{\alpha}{d}$ is **α**-efficient, iff

$$\sim \underset{\bar{d} \in D}{\exists} \overset{\alpha}{F}(\bar{d}) \geq \overset{\alpha}{F}(\overset{\alpha}{d})$$

Let $\overset{\alpha}{D}$ be the set of all **α**-efficient realizations.

Theorem 2

a) If F is backward separable, then $\overset{\alpha}{F}$ is backward separable.

b) If F is forward separable, then $\overset{\alpha}{F}$ is forward separable.■

The proof of the theorem can be found in [17].

We extend the definitions of backward monotonicity. F is backward monotone (in the extended sense) iff

$$\underset{t \in \overline{1,T}}{\forall}, \underset{y_t \in Y_t}{\forall}, \underset{d_t \in D_t(y_t)}{\forall} \quad w \geq \overline{w} \Rightarrow F_t(d_t)\, o_t\, w \geq F_t(d_t)\, o_t\, \overline{w}$$

The definition of forward monotonicity can be extended in the similar way.

Corollary

a) If the function F is backward monotone (in the extended sense), then it is also backward monotone.

b) If the function F is forward monotone (in the extended sense), then it is also forward monotone.■

The proof of the corollary is trivial.

Theorem 3

a) If F is backward monotone (in the extended sense), then $\overset{\alpha}{F}$ is also backward monotone.

b) If F is forward monotone (in the extended sense), then F^{α} is also forward monotone. ■

The proof of the theorem can be found in [17].

Let

$$\overset{\alpha}{D}(\bar{d}) \cong \{\overset{\alpha}{d} \in \overset{\alpha}{D} : \overset{\alpha}{F}(\overset{\alpha}{d}) \geq \overset{\alpha}{F}(\bar{d})\}$$

denote the set of α-efficient realizations better than $\bar{d}$.

We assume that the functions F and $\overset{\alpha}{F}$ have backward difference property [18]. For the given process realization $\bar{d}$ we define the auxiliary function:

$$\overset{\alpha}{\psi}(d) \cong \overset{\alpha}{F}(d) - \overset{\alpha}{F}(\bar{d})$$

where

$$\overset{\alpha}{\psi} = [\overset{\alpha}{\psi}{}^1, \ldots, \overset{\alpha}{\psi}{}^m]$$

and

$$\overset{\alpha}{\psi}{}^m(d) \cong \overset{\alpha}{F}{}^m(d) - \overset{\alpha}{F}{}^m(\bar{d})$$

Theorem 4

The realization $\overset{\alpha}{d}$ jest α-better than the realization $\bar{d}$ iff

$$\forall_{m \in \overline{1,M}} \overset{\alpha}{\psi}{}^m(\overset{\alpha}{d}) \geq 0 \wedge \exists_{i \in \overline{1,M}} \overset{\alpha}{\psi}{}^i(\overset{\alpha}{d}) > 0$$ ■

The proof of the theorem can be found in [18].

For a chosen realization $\bar{d}$ it is possible to determine whether $\bar{d}$ is α-efficient or not and, if $\bar{d} \notin \overset{\alpha}{D}$, to find the set $\overset{\alpha}{D}(\bar{d})$ (by applying algorithm 2.1 [18, p.54]).

4 RELATIONSHIPS BETWEEN $\overset{*}{D}$ AND $\overset{\alpha}{D}$

Usually $\overset{*}{D} \neq \overset{\alpha}{D}$. We consider the set $\overset{*}{D} \cup \overset{\alpha}{D}$. It can be written as follows:

$$\overset{*}{D} \cup \overset{\alpha}{D} = (\overset{*}{D} \cap \overset{\alpha}{D}) \cup (\overset{*}{D} \setminus \overset{\alpha}{D}) \cup (\overset{\alpha}{D} \setminus \overset{*}{D})$$

We suggest that the decision maker consider the set $(\overset{*}{D} \cup \overset{\alpha}{D})$ when making a final choice. Let us assume that her/his final decision is $\overset{f}{d}$. One of the following three situations can take place:

1. $\overset{f}{d} \in (\overset{*}{D} \cap \overset{\alpha}{D})$.

The decision maker has chosen a realization which is both efficient and α-efficient. No further action is recommended.

2. $\overset{f}{d} \in (\overset{*}{D} \setminus \overset{\alpha}{D})$.

The final realization is efficient, but not α-efficient. It means that the importance structure proposed earlier has not been fully taken into account when choosing the final solution. We find the set $\overset{\alpha}{D}(\overset{f}{d})$ and present it to the decision maker. He/she is proposed to repeat the analysis of the final choice, taking into consideration the set of realizations $\overset{f}{d} \cup \overset{\alpha}{D}(\overset{f}{d})$.

3. $\overset{f}{d} \in (\overset{\alpha}{D} \setminus \overset{*}{D})$

The final solution is consistent with the assumed importance structure but it is not efficient with respect to F. The choice of the decision maker is correct; in any case the set $\overset{*}{D}(\overset{f}{d})$ can be found and (by analyzing the set $\overset{f}{d} \cup \overset{*}{D}(\overset{f}{d})$) the decision maker can explain why the non-efficient (with respect to the vector-valued criterion function F) realization $\overset{f}{d}$ has been proposed as the final one.

5 NUMERICAL EXAMPLES

We consider a three period decision process (see [7, p.57]). The numerical data are

$$Y_t = \{0, 1\} \text{ for } t = 1,2,3,4$$
$$X_t(0) = \{0, 1\},\ X_t(1) = \{0,1\} \text{ for } t = 1,2,3$$

and

$F_1^1(0,0) = 7$	$F_1^1(0,1) = 6$	$F_1^1(1,0) = 3$	$F_1^1(1,1) = 4$
$F_1^2(0,0) = 5$	$F_1^2(0,1) = 6$	$F_1^2(1,0) = 9$	$F_1^2(1,1) = 10$
$F_2^1(0,0) = 6$	$F_2^1(0,1) = 3$	$F_2^1(1,0) = 7$	$F_2^1(1,1) = 7$
$F_2^2(0,0) = 8$	$F_2^2(0,1) = 9$	$F_2^2(1,0) = 8$	$F_2^2(1,1) = 4$
$F_3^1(0,0) = 4$	$F_3^1(0,1) = 3$	$F_3^1(1,0) = 10$	$F_3^1(1,1) = 3$
$F_3^2(0,0) = 9$	$F_3^2(0,1) = 8$	$F_3^2(1,0) = 2$	$F_3^2(0,0) = 3$

The graph of the process is given in Fig. 1.

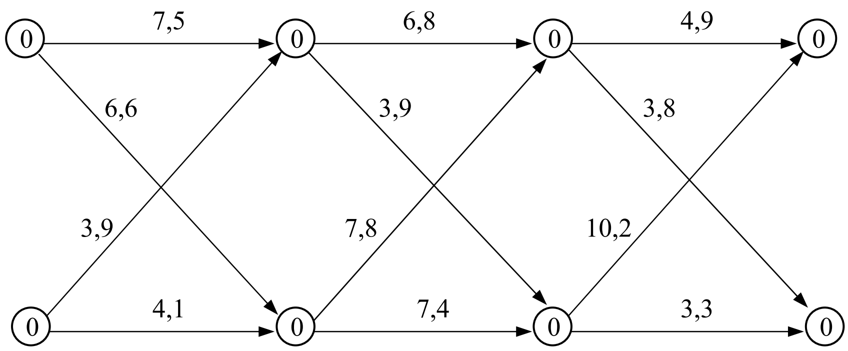


Fig. 1

We consider the additive vector-valued criterion function

$$F = [F^1, F^2]'$$

where

$$F^1(d) = F_1^1(d_1) + F_2^1(d_2) + F_3^1(d_3)$$
$$F^2(d) = F_1^2(d_1) + F_2^2(d_2) + F_3^2(d_3)$$

and K=2, T=3, M=2.

It is easily seen that the criterion function F is both backward separable and monotone and forward separable and monotone.

We will consider four objective importance matrices:

$$\alpha_0 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \alpha_1 = \begin{bmatrix} 1 & 0.5 \\ 1 & 0.8 \\ 0.5 & 1 \end{bmatrix} \quad \alpha_2 = \begin{bmatrix} 1 & 0.5 \\ 0.8 & 1 \\ 0.5 & 1 \end{bmatrix} \quad \alpha_3 = \begin{bmatrix} 1 & 0.2 \\ 0.5 & 1 \\ 0.1 & 1 \end{bmatrix}$$

For α_0 all period objectives are important in all the stages, and therefore we obtain efficient realizations. For α_1, α_2 and α_3 the importance of the first period objective (weakly) decreases and the importance of the second period objective (weakly) increases.

The number of realizations of the considered process is equal to 16. We can number them from 0 to 15 [22], and therefore we have the ordered set $D = \{\overset{0}{d}, \overset{1}{d}, \ldots \overset{15}{d}\}$. As the number of realizations is small, we can handle the full list. The results of calculations are shown in Table 1.

Table 1

Realization	α_0 F^1	α_0 F^2	α_1 F^1	α_1 F^2	α_2 F^1	α_2 F^2	α_3 F^1	α_3 F^2
0	17	22	15	17,9	13,8	19,5	10,4	18
1	16	21	14,5	16,9	13,3	18,5	10,3	17
2	20	16	15	11,7	14,4	13,5	9,5	12
3	13	17	11,5	12,7	10,9	14,5	8,8	13
4	17	23	15	18,4	13,6	20	9,9	18,2
5	16	22	14,5	17,4	13,1	19	9,8	17,2
6	23	12	18	8,2	16,6	9	10,5	7,2
7	16	13	14,5	9,2	13,1	10	9,8	8,2
8	13	26	11	19,9	9,8	21,5	6,4	18,8
9	12	25	10,5	18,9	9,3	20,5	6,3	17,8
10	16	20	11	13,7	10,4	15,5	5,5	12,8
11	9	21	7,5	14,7	6,9	16,5	4,8	13,8
12	15	27	13	20,4	11,6	22	7,9	19
13	14	26	12,5	19,4	11,1	21	7,8	18
14	21	16	16	10,2	14,6	11	8,5	8
15	14	17	12,5	11,2	11,1	12	7,8	9

It can be seen that the realizations $\overset{4}{d}, \overset{6}{d}$ and $\overset{12}{d}$. are efficient and α-efficient for all the considered importance matrices α_1, α_2 and α_3, so $\overset{4}{d}, \overset{6}{d}, \overset{12}{d} \in (\overset{*}{D} \cap \overset{\alpha}{D})$. These realizations can be chosen as final realizations without any doubts.

The realization $\overset{14}{d}$ is both efficient and α_i-efficient for i = 1 and 2, but it is not α_3-efficient. If the decision maker would like to choose this realization as the final one, we have

$\overset{14}{d} \in (\overset{*}{D} \setminus \overset{\alpha}{D})$. We find the set $\overset{\alpha}{D}(\overset{14}{d}) = \{\overset{0}{d}, \overset{4}{d}\}$ and present it to the decision maker. He/she is proposed to repeat the analysis of the final choice, taking into consideration the set $\{\overset{14}{d}, \overset{0}{d}, \overset{4}{d}\}$.

If for α_2 the final choice is $\overset{0}{d}$ or $\overset{2}{d}$, or for α_3 the final choice is $\overset{0}{d}$, the chosen realization is consistent with assumed importance structure, but is not efficient. According to the proposed procedure the sets

$$\overset{0}{d} \cup \overset{*}{D}(\overset{0}{d}) = \{\overset{0}{d}, \overset{4}{d}\} \text{ for } \alpha_2 \text{ and } \alpha_3$$

$$\overset{2}{d} \cup \overset{*}{D}(\overset{2}{d}) = \{\overset{2}{d}, \overset{6}{d}, \overset{14}{d}\} \text{ for } \alpha_2$$

are found and the decision maker is asked to explain why the non-efficient (with respect to the vector-valued criterion function F) realization $\overset{0}{d}$ (or $\overset{2}{d}$) has been chosen as the final one (alternatively, the decision maker is suggested to repeat the final choice analysis).

5 CONCLUDING REMARKS

The aim of the discussion in this paper was a formal description of a situation which can occur in the analysis of changeable importance of period objectives in multiobjective dynamic optimization. The definition of extended backward and forward monotonicity had to be formulated. The modification did not narrow the possibility of practical application of proposed solutions.

It could be important for a decision maker to be advised what to do if the higher number of elements of the set is higer. It will be elaborated in the next paper.

In the present paper only a part of the wide range of problems connected with the changeable importance of objectives was presented. The next step could be the consideration of the situation in which the importance of criteria depends on states of the process at the beginning of the consecutive periods.

A stochastic approach to the analysis of this problem, initialized in [12], is also possible.

References

[1] Abo-Sinna M., Amer A.H., Sayed H, 2006. An interactive algorithm for decomposing the parametric space in fuzzy multiobjective dynamic programming problem. Applied Mathematics and Computation, 174, 1, pp. 684-699.

[2] Abo-Sinna M., Hussein M., 1995 An algorithm for generating efficient solutions of multiobjective dynamic programming problems. European Journal of Operational Research, 80, 1, pp. 156-165.

[3] Getachew T.,1992. A recursive algorithm for multi-objective network optimization with time-variant link costs. Ph.D Dissertation, Clemson University.

[4] P.Gong, 1992. Multiobjective dynamic programming for forest resource management. Forest Ecology and Management, 48, 1-2, pp. 43-54.

[5] Henig M.,1995. The shortest path problem with two objective functions. European Journal of Operational Research, 25, 281-291.

[6] Kostreva M., Lancaster L., 2002. Multiple objective path optimization for time dependent objective functions. In: Trzaskalik T., Michnik J. (eds.) Multiple Objective Programming and Goal Programming. Recent Developments. Physica-Verlag, pp. 127-142.

[7] Kostreva M., Wiecek M.,1993. Time dependency in multiple objective dynamic programming. Journal of Mathematical Analysis and Applications, 173, 289-307.

[8] Li D., Cheng C., 2004. Stability on multiobjective dynamic programming problems with fuzzy parameters in the objective functions and in the constraints. European Journal of Operational Research, 158, 3, pp. 678-696.

[9] Li.D., Haimes Y.Y., 1989. Multiobjective dynamic programming. The state of the art. Control Theory and Advanced Technology, 5(4), pp.471-483.

[10] Li L., Lai K.K, 2001. Fuzzy dynamic programming approach to hybrid multiobjective multistage decision-making problems. Fuzzy Sets and Systems 117, 1, pp.13-25.

[11] Liao Z., Li D., 2002. Adaptive differential dynamic programming for multiobjective optimal control. Automatica, 38, 6, pp.1003-1015.

[12] D.Lozovanu, S.Pickl, 2007. Algorithms for solving multiobjective discrete control problems and dynamic c-games on networks. Discrete Applied Mathematics, 155, 14, pp.1846-1857.

[13] Sitarz S., 2009. Ant algorithms and simulated annealing for multicriteria dynamic programming. Computers & Operations Research, 36, 2, 2009, pp. 433-441.

[14] Sitarz S., 2006. Hybrid methods in multi-criteria dynamic programming. Applied Mathematics and Computation, 180, pp. 38-45.

[15] Szidarovsky F., Duckstein L., 1986. Dynamic multiobjective optimization: A framework with application to regional water and mining management. European Journal of Operational Research, 24, 2, pp.305-317.

[16] Szidarovsky F., Gershon M., Bardossy A., 1987. Application of multiobjective dynamic programming to regional natural management resources. Applied Mathematics and Computation, 24, 4, pp. 281-301.

[17] Trzaskalik T., 2008. Zmienna, ważona istotność celów w wielokryterialnym programowaniu dynamicznym (Changeable, weighted importance of objectives in multiobjective dynamic programming). In: j.Biolik (ed.) Dylematy ekonometrii. The Karol Adamiecki Univestity of Economics in Katowice, pp.223-235 (in Polish).

[18] Trzaskalik T., 1998. Multiobjective Analysis in Dynamic Environment. The Karol Adamiecki University of Economics in Katowice.

[19] Trzaskalik T., 1997. Hierarchy depending on state in multiple objective dynamic programming. Operations Research and Decision, 2/1997, pp.65-73.

[20] Trzaskalik T., 1996 Fixed and changeable hierarchical problems in multiple objective dynamic programming. La Faculté des sciences de l'administration de l'Université Laval. Document de travail.

[21] Trzaskalik T., 1995. Hierarchy depending on value in multiple criteria dynamic programming. Foundations on Computing and Decision Science, 20, 2, pp.139-148.

[22] Trzaskalik T., 1990. Wielokryterialne dyskretne programowanie dynamiczne. teoria i zastosowania w praktyce gospodarczej (Multicriteria Discrete Dynamic Programming. Theory and economic applications). The Karol Adamiecki University of Economics in Katowice (in Polish).

[23] Trzaskalik T., Sitarz S., 2007 Discrete dynamic programming with outcomes in random variable structures. European Journal of Operational Research 177, pp.1535-1548,.

[24] Wang Y., Dang C., 2008. An evolutionary algorithm for dynamic multi-objective optimization. Applied Mathematics and Computation, 205, 1, pp. 6-18.

[25] Warburton A., 1987 Approximation of Pareto optima in multiple objective shortest path problems. Operations Research, 354(1), pp.70-79.

[26] Wei Q., Zhang H., Dai J., 2009. Model-free multiobjective approximate dynamic programming for discrete time nonlinear systems with general performance index function. Neurocomputing, 72, 7-9, 2009, pp. 1893-1848.

[27] Zhang Z. Multiobjective optimization immune algorithm in dynamic environments and its application to greenhouse control. Applied Soft Computing, 8, 2, 2008, pp.959-971.

OPERATIONS RESEARCH IN MOBILE CELLULAR NETWORK DESIGN

(Extended Abstract)

Di Yuan
Department of Science and Technology, Linköping University,
SE-601 74 Norrköping, Sweden.
diyua@itn.liu.se

Keywords: mobile networks, base station location, radio resource management, discrete optimization.

1. BACKGROUND

In this talk we summarize experiences, achievements, recent advances, and new challenges in applying operations research (OR), in particular discrete optimization, to the design of mobile cellular networks. A cellular radio access infrastructure consists in a set of base stations (BSs). A BS may have one or several cells (aka sectors), each of which is served by a radio antenna. In the last few decades, the underlying radio technology has evolved from the first generation (1G) analog systems, to second generation (2G) digital systems (e.g., GSM) that reached tremendous popularity, and then to third generation (3G) networks (aka as UMTS) that extended the service domain to multimedia. Beyond-3G systems for mobile broadband Internet consists in high speed packet access (HSPA), aka 3.5G, followed by long term evolution (LTE), a fourth generation (4G) technology. These generations of networks differ in their way of utilizing the radio resource. As a result, types and characteristics of the network optimization problems change along with the technological evolution. Two closely-related optimization tasks are infrastructure planning (i.e., BS location), and radio resource management (RRM). Typical optimization objectives are cost, coverage, capacity, and interference.

2. FREQUENCY ASSIGNMENT IN GSM

The talk reviews GSM frequency assignment. GSM uses frequency and time division multiple access; the spectrum is divided into a set of frequencies with a number of time slots on each. GSM network planning admits a decomposition approach. First, BS location is performed, a problem tightly connected to classical facility location and set covering. Given the BS locations, the next major planning step deals with the assignment of frequencies to the cells to satisfy the capacity requirement, which directly translates into the number of frequencies. Reuse of frequencies may cause interference between cells.

Two representative versions of the frequency assignment problem (FAP) are minimum-order FAP and minimum-interference FAP. In the former, the objective is to minimize the total number of frequencies required to avoid interference. Minimum-interference FAP, on the other hand, amounts to minimizing the total interference caused by frequency reuse for a given number of frequencies. Note that minimum-order FAP is a generalization of the vertex coloring problem. In real-life networks, however, minimum-interference frequency assignment is more adequate since the number of available frequencies is fixed.

Literature on OR applied to FAP is very rich. See, for example, [1, 5] for some surveys. Many optimization methods based on integer programming, meta heuristics, and constraint programming have been developed, with great success in leveraging the power of OR to optimize GSM network performance. Today, FAP is routinely solved by optimization methods embedded into network planning tools.

3. KEY PERFORMANCE CONSIDERATIONS IN 3G

The next part of the talk examines 3G/UMTS network planning. 3G/UMTS adopts code division multiple access – a user's channel is identified by a code sequence. The channels use a common spectrum band, thus frequency assignment is not present. The key parameter is transmission power. UMTS uses power control, that is, the power is dynamically adjusted to maintain the signal-to-interference and noise ratio (SIR) at a target value. Let P_i denote the power used by cell antenna i, and let p_{ij} be the part of P_i for serving user j. A simplified SIR equation of j reads

$$\frac{p_{ij} g_{ij}}{\sum_{h \neq i} P_h g_{hj} + \eta_j} = \gamma_j,$$

where g_{ij} is the radio propagation factor between i and j, η_j is the noise effect, and γ_j is the SIR threshold. Given a snapshot of the user locations, the SIR equations form an equation system, which can be transformed into a compact form, resulting in an interference coupling matrix that captures the interference relation between cells [4]. For RRM, it is essential that the total power requirement (i.e., P_i) stays below the antenna's power limit. A similar consideration applies to the uplink direction.

There is a growing amount of literature on 3G/UMTS network optimization. See, for example, [2, 4] for BS location and antenna configuration considering SIR, and [7] for coverage planning.

4. RADIO RESOURCE OPTIMIZATION IN BEYOND-3G NETWORKS

The final part of the talk deals with optimization of beyond-3G networks: HSPA (3.5G) and LTE (4G). For HSPA, SIR remains the key consideration. However, power control is not present at the downlink. Instead of having a fixed SIR threshold, the performance target is to achieve as high SIR as possible to maximize data throughput, and the total antenna output power can be regarded as a constant. Within this optimization framework, finding optimal antenna tilting angels has been studied in [6].

LTE uses orthogonal frequency division multiple access (OFDMA). The spectrum is divided into a large number of sub-channels, each can carry a data stream. Although interference avoidance in LTE resembles FAP to some extent, there are major differences. First, LTE admits much more flexible resource allocation. In addition to frequency, resource parameters include power and time. Thus the characteristics of the resulting optimization problem differ from FAP. Second, reuse of all sub-channels in all cells actually tends to maximize the overall data throughput. Channel separation may however be useful for improving data rate at interference-sensitive regions. In the talk we present some recent results on this topic [3]. We conclude the talk by envisioning forthcoming research on OR in LTE network engineering.

References

[1] K. Aardal, C. P. M van Hoesel, A. M. C.A, Koster, C. Mannino, and A. Sassano, Models and solution techniques for the frequency assignment problem, *4OR*, vol. 1, pp. 261–317, 2003.

[2] E. Amaldi, A. Capone, and F. Malucelli, Planning UMTS base station location: Optimization models with power control and algorithms, *IEEE Transactions on Wireless Communications*, vol. 2, pp. 939–952, 2003.

[3] L. Chen and D. Yuan, Soft frequency reuse in large networks with irregular cell pattern: how much gain to expect? *Proceedings of IEEE PIMRC '09*, 2009.

[4] A. Eisenblätter, T. Koch, A. Martin, T. Achterberg, A. Fügenschuh, A. Koster, O. Wegel, and R. Wessäly, Modelling feasible network configurations for UMTS, in G. Anandalingam and S. Raghavan (eds.), *Telecommunications Network Design and Management*, Kluwer Academic Publishers, 2002.

[5] FAP Website, `http://fap.zib`, 2009.

[6] F. Gunnarsson, I. Siomina, and D. Yuan, Automated optimization in HSDPA radio network planning. In: B. Furht and S. Ahson (eds.), Handbook of HSDPA/HSUPA Technology, CRC press, in press.

[7] I. Siomina, P. Värbrand, and D. Yuan, Automated optimization of service coverage and base station antenna configuration in UMTS networks, *IEEE Wireless Communications Magazine*, vol. 13, pp. 16–25, 2006.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section I:

Discrete Mathematics and Optimization

OPTIMIZATION OF WASTE COLLECTION IN MARIBOR

Darko Bečaj[1], **Janez Žerovnik**[2,3]
[1] Snaga, javno podjetje d.o.o., Nasipna 64, 2000 Maribor, Slovenia
[2] IMFM/DTCS, Jadranska 19, 1000 Ljubljana, Slovenia
[3] Univerza v Ljubljani, Fakulteta za strojništvo, Aškerčeva 6, 1000 Ljubljana, Slovenia
Emails: darko.becaj@snaga-mb.si, janez.zerovnik@imfm.uni-lj.si

Abstract: The aim of the planning guidance of vehicles in the collection and removal of waste is to minimize transportation costs. The costs are valued in the time spent for the execution of the working process. This type of tasks is usually modeled by the Chinese postman problem. Optimization is done by application of algorithms from graph theory. Here we report on practical application of the approach by the company SNAGA d.o.o. in the town of Maribor. Considerable savings were obtained on the area observed.

Keywords: Waste collection, route planning, graph theory, Euler tour, Chinese postman problem

1 INTRODUCTION

In this paper, weighted directed graph G=(V, A, u) is a triple, where V is an arbitrary set of vertices, A is a set of arcs or directed edges (i.e. ordered pairs of vertices) and u is a weight function which assigns a positive real number to arcs.

In graph theory, the Chinese postman problem, postman tour or route inspection problem is to find a shortest closed circuit that visits every edge of a (connected) undirected graph. When the graph has an Eulerian circuit, that circuit is an optimal solution. Alan Goldman of NIST first coined the name 'Chinese Postman Problem' for this problem, as it was originally studied by the Chinese mathematician Mei-Ku Kuan in 1962 [4].

If a graph is Eulerian, then an Eulerian path visits every edge, and so the solution is to choose any Eulerian path. An Eulerian path is a path in a graph which visits each edge exactly once. Similarly, an Eulerian circuit is an Eulerian path which starts and ends on the same vertex. They were first discussed by Leonhard Euler while solving the famous Seven Bridges of Königsberg problem in 1736. Graphs which allow the construction of Eulerian circuits are called Eulerian graphs. Euler observed that a necessary condition for the existence of Eulerian circuits is that all vertices in the graph have even degree [3]. Carl Hierholzer published the first complete characterization of Eulerian graphs in 1873, by proving that in fact the Eulerian graphs are exactly the graphs which are connected and where every vertex has an even degree [5].

If the graph is not Eulerian, it must contain vertices of odd degree. By the handshake lemma, there must be an even number of these types of vertices. Note that we must revisit edges that come out of these vertices for the solution. We make the graph Eulerian by doubling the paths that connect these vertices in pairs. We choose the pairs such that the total distance covered of all paths that connect these vertices are as small as possible. This means that we have to find a minimal matching in the auxiliary complete weighted graph where the vertices are the vertices of original graph of odd degree and edge weights are the distances in the original graph. Now the solution is an Eulerian path for this new graph.

Given a graph G = (V, A, u), a matching M in G is a set of pairwise non-adjacent edges; that is, no two edges share a common vertex. The cost of a matching is the sum of its edge weights. A minimum matching can be found in polynomial time [6]. For further reading on graph theory and applications we refer to [13], see also [11].

Standard applications of the Chinese postman problem and its variants include street sweeping, solid waste collection, salt gritting, snow plowing, road inspection, but also testing of computer systems, DNA fragment assembly and others. For more details on particular models and further examples see papers [7], [1], [2].

Here we report on practical application of the approach by the company SNAGA in the town of Maribor. We first give some details of the model used, and then briefly show the application of the standard method on a practical example. Considerable savings were obtained on the area observed which provides good motivation for future work on more accurate models involving more additional constraints that will on one hand request further studies that may lead to development of new algorithms and, on the other hand, will allow even more savings in practical application.

2 PROBLEM DESCRIPTION

In general problem, we are limited by the number of vehicles and their capacity. There may be more types of vehicles with different capacities. Some larger types of vehicles can not enter narrow streets. The service of each vehicle starts and ends at the earlier defined initial node, which is usually also the final node (position). Working time has a ceiling. Usually, the total collecting area is decomposed into smaller areas that can be served by individual routes. During the travel, the vehicle is collecting waste which may never exceed the capacity of the vehicle. Time spent at the collection area was measured during certain time period. It is divided into the time spent for waste collection and the rest is useless time for connecting journeys.

The area of treatment is limited due to the complexity of the agenda at the assembly area in the City of Maribor. Logistic company, depending on the capacity of the vehicle and the compression factor of waste in the vehicle, gather different capacities of containers such as weight of waste that will be enough to drive at one route. The area belongs to the synthesis of one individual settlement and double family houses.

In this preliminary study, several assumptions that simplify the general problem were adopted. In the example elaborated here, we consider a small area that can be covered by one route of one vehicle (see Figure1). For each link in the area, we define the weight corresponding to the total time the vehicle needs to traverse it while collecting the waste.

Figure 1: Road links on the area

The network has 12 vertices of degree 1. The incident edges represent the impasse, which will clearly be traversed twice by the vehicle. Therefore the edges are duplicated.

Impasses are marked by: BC, DE, HI, MO, MN, PR, PQ, WX, YZ, AA-AB, AC-AD, AI-AJ.

After this transformation we get a graph with the following node degrees dg (v), see Table 1.

Table 1: Degrees and the number of nodes in the graph after the transformation

Degree [dg(*v*)]	Number of nodes	Marks of the nodes
2	15	A, C, E, G, I, N, O, R, Q, X, Z, AB, AD, AJ, AK
3	10	F, K, L, T, U, V, AE, AF, AG, AH
4	8	B, H, J, S, Y, AA, AC, AI
5	2	D, W
6	2	M, P

3 THE ALGORITHM

We apply the standard method for optimal solution of the Chinese postman:

1. count and identify the subset of nodes in odd degree,
2. find the shortest paths between all pairs of nodes of odd degree by Dijkstra algorithm,
3. find a minimum matching on the complete subgraph on odd degree nodes, where the edge weights are distances computed in step 2,
4. double edges corresponding to the shortest paths, to obtain a new graph G' which is Eulerian,
5. find an Eulerian tour with Fleury's algorithm.

For the details we refer to textbooks, for example [13], [11] or [8], [9], [10].

In our example, the algorithm of Dijkstra was used to calculate the distances between pairs of odd degree nodes: D, F, K, L, T, U, V, W, AE, AF, AG, AH. Note that it is also easy to obtain the corresponding shortest paths from the Dijkstra's algorithm. The actual distances between odd degree nodes are given in Table 2.

Table 2: The matrix of time [s] runs between connecting nodes odd levels

	D	**F**	**K**	**L**	**T**	**U**	**V**	**W**	**AE**	**AF**	**AG**	**AH**
D	**0**	18	48	44	28	39	45	21	47	55	69	60
F		**0**	46	26	46	57	63	39	65	73	96	58
K			**0**	20	20	25	32	39	63	42	50	12
L				**0**	40	45	52	35	61	62	70	32
T					**0**	11	18	19	45	28	42	32
U						**0**	7	30	38	17	31	27
V							**0**	24	31	10	24	34
W								**0**	26	34	48	51
AE									**0**	21	35	65
AF										**0**	14	44
AG											**0**	38
AH												**0**

For calculations we have used "CPP ver. 1.0" [14]. The graph after doubling of edges is given in Figure 2.

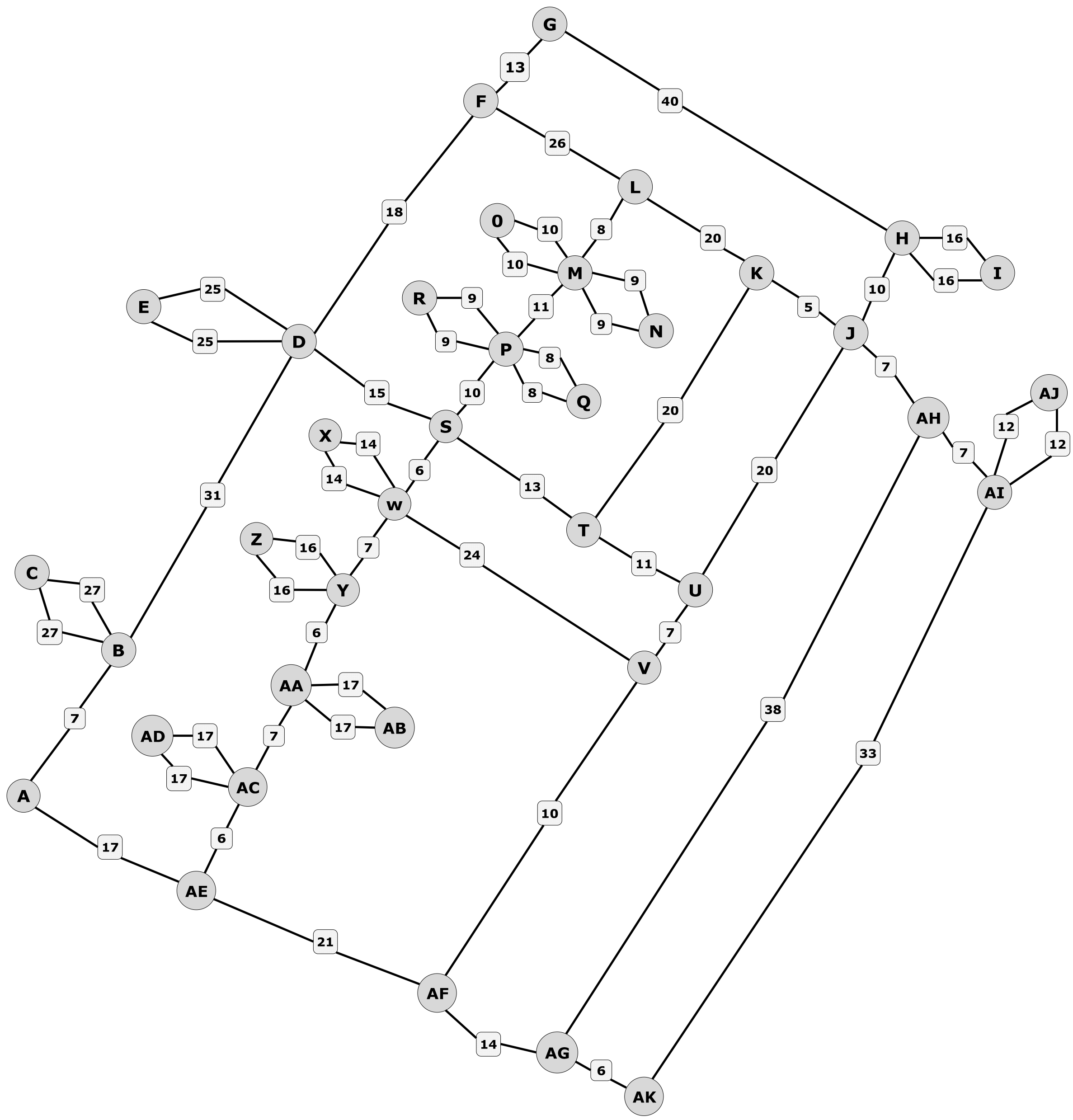


Figure 2: The new network graph. Impasses are doubled.

The next step in solving the Chinese postman problem is finding a minimum matching [6].

In our case, the minimum weight on the connections can be quickly found by searching all possible pairs of nodes odd levels (Table 3).

Table 3: Additional Links

Additional links pairs of odd degree nodes	Weights of links
D-F; L-K; AH-AG; AE-AF; V-U; T-W;	123
D-F; L-K; AH-AG; V-AF; T-U; AE-W;	123
D-T; F-L; K-AH; AF-AG; AE-W; U-V;	**113**
D-F; L-K; V-W; T-U; AH-AG; AE-AF;	132
D-F; L-W; K-T; U-V; AE-AF; AG-AH;	143
D-W; F-L; K-T; U-V; AE-AF; AG-AH;	137

Minimum weight matching is: DT, FL, K-S, AF-AG, AE-W, UV; with a total weight 113.

On Figure 4 is a network graph where the shortest paths determined by the minimum matching are added which results in doubling some additional edges.

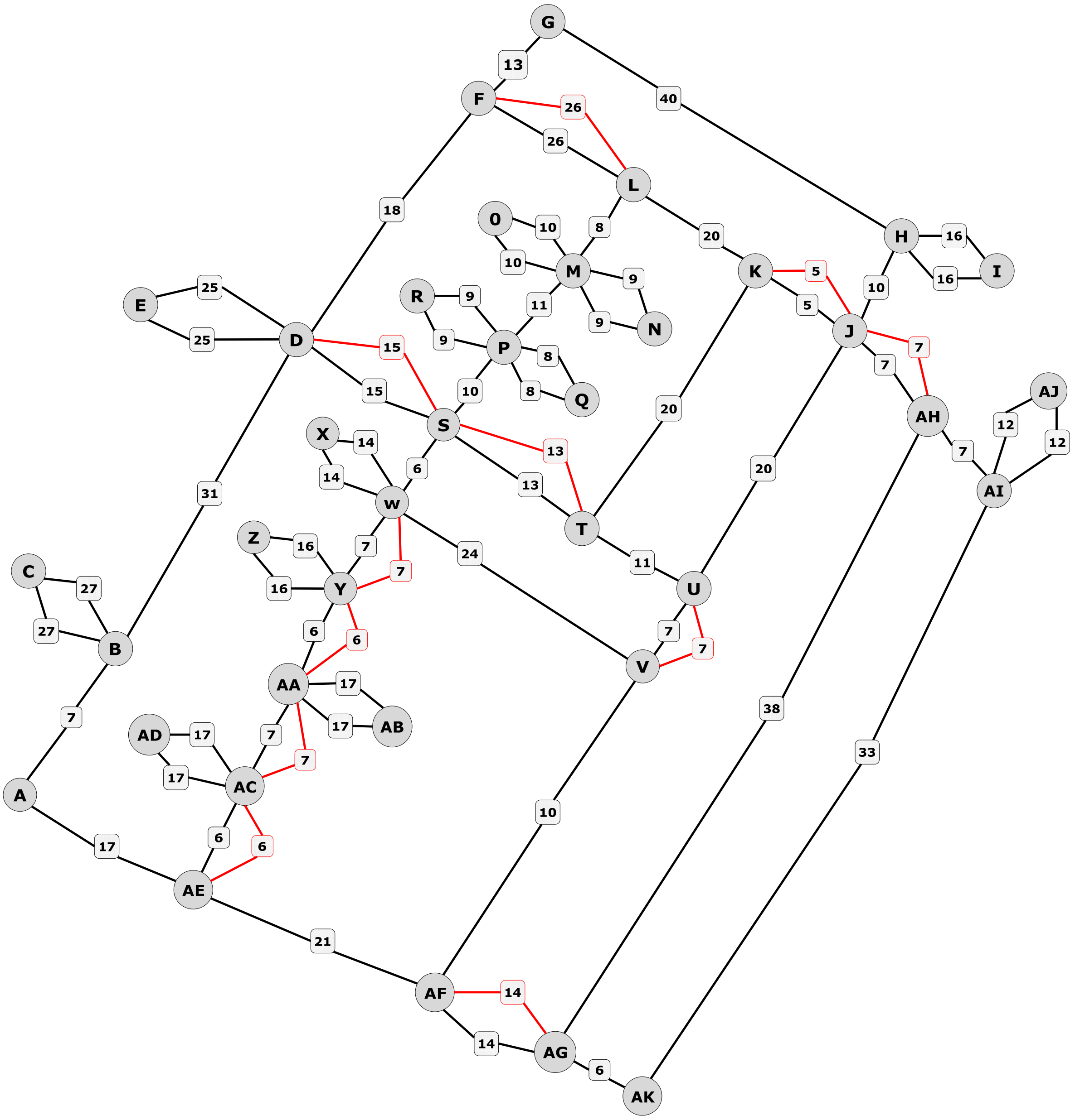


Figure 4: Eulerian graph after doubling of the shortest paths given by minimum matching

An Eulerian tour in the graph can be found, for example, with Fleury's algorithm. In our case one of the solutions is the sequence:

A-B-C-B-D-E-D-F-G-H-I-H-J-K-L-F-L-M-O-M-N-M-P-R-P-Q-P-S-D-S-T-K-J-AH-J-U-T-S-W-X-W-Y-Z-Y-AA-AB-AA-AC-AD-AC-AE-AC-AA-Y-W-V-U-V-AF-AG-AH-AI-AJ-AI-AK-AG-AF-AE-A.

The minimum weight of the whole network represents the best result on tour. We compare the minimum weight with actual time spent for collecting and driving in the area.

4 CONCLUSION, THE SAVINGS

The solution was evaluated by collecting real measurements on the ground. Data was obtained on days with continuous monitoring of a vehicle equipped with a GPS device. The time data was monitored at a small interval in seconds. Time spent driving to the individual connections was 7226 seconds, which is around 120 minutes. The minimum weight of additional connections in our solution is 113 seconds (DT: 28; FL: 26, K-S: 12, AF-AG: 14, AE-W: 26; UV: 7) or 1.9 minutes. Hence the optimal tour would take 122 minutes, which is about 18 minutes less than the previously used tour that is 140 minutes long. The redundant part of the tour was thus reduced by 90%.

We note that the cost of qualifying hours conducted refuse collection vehicles, represents an unnecessary lost of time, which results in nearly 25 € cost on the small example elaborated here. This means around 5.60 kilometers.

According to the partial capture of the daily areas we note that 18 minutes represents approximately 15 percent of the total time required for the synthesis route of the area. Given the whole time collecting on a daily route, which is an average of 430 minutes of effective work, saved 18 minutes means approximately 4 per cent of the total working time.

We plan to continue the work in several directions:

1. the optimization procedure will be applied to larger area including a heuristics for partition of the area to single routes. Here, different types of vehicles will have to be taken into account.

2. When a part of the street is traversed for the second time, the time needed is shorter as there is no need to collect waste. This should be included in the upgraded model.
3. Due to continuous turnover of certain restrictions or urban situations that affect the removal process it may be necessary to provide some additional restrictions or priorities. For example, time windows for waste collection at some streets may be introduced, or priority can be given to certain streets by introducing penalties for late waste removal. See, for example [12] and the references there.

References

[1] Eiselt, H. A., Gendreau, M., Laporte, G., 1995. Arc routing problems, Part I: The Chinese postman problem, Operations Research 43(2), pp. 231-242.

[2] Eiselt, H. A., Gendreau, M., Laporte, G., 1995. Arc routing problem, Part II: The Rural postman problem, Operations Research 43(3), pp. 399-414.

[3] Euler, L., 1736. Solutio problematis ad geometriam situs pertinentis, *Comment. Academiae Sci. I. Petropolitanae* 8, pp. 128-140.

[4] Guan, M., 1962. Graphic Programming using odd and even points, Chinese Mathematics 1, pp. 273-277.

[5] Hierholzer, C., 1873. Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechnung zu umfahren, Mathematische Annalen 6, pp. 30-32.

[6] Korte, B.,Vygen, Y., 2002. Combinatorial optimization: Theory and Algorithms, Second Edition, Research Institute for Discrete Mathematics, University of Bonn, 231 p.

[7] Lemieux, P. F., Campagna, L., 1984. The snow ploughing problem solved by a graph theory algorithm, Civil Engineering Systems 1, pp. 337-341.

[8] Pevzner, P. A., Tang, H., Waterman, M. S., 2001. An Eulerian path approach to DNA fragment assembly, Proceedings of the National Academy of Sciences, 98(17), pp. 9748-9753.

[9] Thimbleby, H., 2003. The directed Chinese Postman Problem, Software Practice & Exp., 33(11), pp. 1081-1096.

[10] Thimbleby, H.W., Witten, I. H., 1993. User Modelling as Machine Identification: New Methods for HCI, Advances in Human-Computer Interaction, H. R. Hartson & D. Hix, eds., IV, pp. 58-86.

[11] Žerovnik, J., 2005. Osnove teorije grafov in diskretne optimizacije, Univerza v Mariboru, Fakulteta za strojništvo, Maribor, pp. 40-41, 116-119.

[12] Kramberger, T., Žerovnik, J, 2007. Priority Constrained Chinese postman problem: Logistics and sustainable transport, 22-05-07, vol.1, no.1, 15.

[13] Wilson, R.J.,Watkins, J.J, 1990. Graphs, An Introductionary Approach, John Wiley and Sons, New York, pp.157-195.

[14] http://www.cs.sunysb.edu/~algorith/implement/chinese-postman/distrib/

CONVEX AND POSITIVE NONCOMMUTATIVE POLYNOMIALS[1]

Kristijan Cafuta*, Igor Klep†, Janez Povh‡

*University in Ljubljana, Faculty of Electrical Engineering, Slovenia
†University in Maribor, Faculty of natural science and mathematics, Slovenia
†University in Ljubljana, Faculty of mathematics and physics, Slovenia
‡Institute of mathematics, physics and mechanics Ljubljana, Slovenia
‡Faculty of information studies in Novo mesto, Slovenia

email: kristijan.cafuta@fe.uni-lj.si, igor.klep@fmf.uni-lj.si, janez.povh@fis.unm.si

Abstract
This paper provides an overview of convexity and positivity for polynomials in noncommuting variables. Special emphasis is given to algorithmic and computational aspects which have been implemented in `NCSOStools`, a MATLAB toolbox for

- symbolic computation with polynomials in noncommuting variables;
- constructing and solving convexity and positivity programs for polynomials in noncommuting variables.

`NCSOStools` can be used in combination with semidefinite programming software, such as SDPT3 or SeDuMi to solve these constructed programs.

Keywords: noncommutative polynomial, convex polynomial, positive polynomial, semidefinite programming, MATLAB toolbox.

Math. Subj. Class. (2010): Primary 11E25, 90C22; Secondary 08B20, 13J30, 90C90

1 INTRODUCTION

Starting with Helton's seminal paper [6], *free semialgebraic geometry* is being established. Among the things that make this area exciting are its many facets of applications. A nice survey on applications to control theory, systems engineering and optimization is given in [5], while applications to mathematical physics and operator algebras have been given by the second author [12, 13]. Emphasis in this note is not so much on applications but rather on explaining the basic framework.

Unlike classical semialgebraic (or real algebraic) geometry where real polynomial rings in *commuting* variables are the objects of study, free semialgebraic geometry deals with real polynomials in *noncommuting* (NC) variables and their matrix images. Of interest are various notions of *convexity* and *positivity* induced by these. For instance, positivity via positive semidefiniteness or the positivity of the trace leading to notions of convexity and trace convexity of polynomials. These can be reformulated and effectively studied using sums of hermitian squares (with commutators) and semidefinite programming.

We developed `NCSOStools` as a consequence of this recent interest in noncommutative convexity, positivity and sums of (hermitian) squares (SOHS). `NCSOStools` is an open source MATLAB toolbox for constructing and then solving such SOHS problems using semidefinite

[1]Project was partially supported by the Slovenian research agency under contracts 1000-08-210518 and Z1-9570-0101-06.

programming. As a side product our toolbox implements symbolic computation with noncommuting variables in MATLAB. It can be downloaded from [17].

There is a certain overlap in features (mainly on the level of symbolic computation) with Helton's `NCAlgebra` package [8] for Mathematica, however our primary interest is with the different notions of convexity, positivity and sum of hermitian squares (with commutators) problems which often operate at the numerical level, so we feel MATLAB is the optimal framework for studying these. Readers interested in solving sums of squares problems for commuting polynomials are referred to one of the many great existing packages, such as SOSTOOLS [20], GloptiPoly [10], or YALMIP [14].

The main contribution of the paper is a brief overview of convexity and positivity of NC polynomials and to describe how to check for these properties using `NCSOStools`.

This paper is organized as follows. The first subsection fixes notation and introduces terminology. Then in Section 2 we introduce convex and trace convex NC polynomials, and present some of their basic properties and the connection to positivity of NC polynomials. The latter is then studied in detail in Section 3, where the natural correspondence between sums of hermitian squares (SOHS) and semidefinite programming (SDP) is explained thoroughly. Section 4 is brief, works on the symbolic level and introduces commutators and cyclic equivalence. These notions are used in Section 5 to study trace positive NC polynomials using sums of hermitian squares and commutators. Such representations can again be found using semidefinite programming. Throughout the paper many examples are presented illustrating the theory, main ideas and our MATLAB toolbox `NCSOStools`.

1.1 Notation

We write $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{R}$ for the sets of natural and real numbers. Let $\langle \bar{X} \rangle$ be the monoid freely generated by $\bar{X} := (X_1, \dots, X_n)$, i.e., $\langle \bar{X} \rangle$ consists of *words* in the n noncommuting letters $X_1, \dots, X_n$ (including the empty word denoted by 1).

We consider the algebra $\mathbb{R}\langle \bar{X} \rangle$ of polynomials in n noncommuting variables $\bar{X} = (X_1, \dots, X_n)$ with coefficients from $\mathbb{R}$. The elements of $\mathbb{R}\langle \bar{X} \rangle$ are linear combinations of words in the n letters $\bar{X}$ and are called *NC polynomials*. The length of the longest word in an NC polynomial $f \in \mathbb{R}\langle \bar{X} \rangle$ is the *degree* of f and is denoted by $\deg f$. We shall also consider the degree of f in X_i, $\deg_i f$. Similarly, the length of the shortest word appearing in $f \in \mathbb{R}\langle \bar{X} \rangle$ is called the *min-degree* of f and denoted by $\operatorname{mindeg} f$. Likewise, $\operatorname{mindeg}_i f$ is introduced. If the variable X_i does not occur in some monomial in f, then $\operatorname{mindeg}_i f = 0$. For instance, if $f = 6X_1^2X_2X_4^2 - X_2X_1X_2X_3 + 7X_2X_1X_3X_1^2$, then $\deg f = 5$, $\deg_1 f = 3$, $\deg_2 f = 2$, $\deg_3 f = 1$, $\deg_4 f = 2$, and $\operatorname{mindeg} f = 4$, $\operatorname{mindeg}_1 f = \operatorname{mindeg}_2 f = 1$, $\operatorname{mindeg}_3 f = \operatorname{mindeg}_4 f = 0$.

An element of the form aw where $0 \neq a \in \mathbb{R}$ and $w \in \langle \bar{X} \rangle$ is called a *monomial* and a its *coefficient*. Hence words are monomials whose coefficient is 1. We equip $\mathbb{R}\langle \bar{X} \rangle$ with the *involution* $*$ that fixes $\mathbb{R} \cup \{\bar{X}\}$ pointwise and thus reverses words, e.g. $(2X_1X_3^3X_2 - 3X_3^2)^* = 2X_2X_3^3X_1 - 3X_3^2$. Hence $\mathbb{R}\langle \bar{X} \rangle$ is the $*$-algebra freely generated by n symmetric letters. Let $\operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$ denote the set of all *symmetric elements*, that is, $\operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle = \{f \in \mathbb{R}\langle \bar{X} \rangle \mid f = f^*\}$. The involution $*$ extends naturally to matrices (in particular, to vectors) over $\mathbb{R}\langle \bar{X} \rangle$. For instance, if $V = (v_i)$ is a (column) vector of NC polynomials $v_i \in \mathbb{R}\langle \bar{X} \rangle$, then V^* is the row vector with components v_i^*. We shall also use V^t to denote the row vector with components v_i.

Recall: a symmetric matrix $A \in \mathbb{R}^{s \times s}$ is positive semidefinite if and only if it is of the form B^tB for some $B \in \mathbb{R}^{s \times s}$. Equivalently, all of its eigenvalues are nonnegative. If A is also invertible, then we call it positive definite. We write $A \succeq B$, respectively $A \succ B$, to denote that $A - B$ is positive semidefinite, respectively positive definite.

2 CONVEX AND TRACE CONVEX NC POLYNOMIALS

Although most of our computations evolve around *positivity*, the initial motivation stems from the study of *convexity* in dimension-free control theory.

2.1 Convex NC polynomials

Motivated by considerations in engineering system theory (cf. [5] for a modern treatment), Helton and McCullough [7] studied convex NC polynomials. An NC polynomial $p \in \mathbb{R}\langle \bar{X} \rangle$ is *convex* if it satisfies

$$p(tA + (1-t)B) \preceq tp(A) + (1-t)p(B)$$

for all $0 \leq t \leq 1$ and for all tuples A, B of symmetric matrices of the same size.

In applications convexity is a very desirable property, so the first question is how to test for convexity of a given NC polynomial. To this end we rephrase convexity using second directional derivatives. Given $p \in \mathbb{R}\langle \bar{X} \rangle$, consider

$$r(\bar{X}, \bar{H}) = p(\bar{X} + \bar{H}) - p(\bar{X}) \in \mathbb{R}\langle \bar{X}, \bar{H} \rangle.$$

The *second directional derivative* $p''(\bar{X}, \bar{H}) \in \mathbb{R}\langle \bar{X}, \bar{H} \rangle$ is defined to be twice the part of $r(\bar{X}, \bar{H})$ which is homogeneous of degree two in $\bar{H}$. Alternatively,

$$p''(\bar{X}, \bar{H}) = \frac{d^2 p(\bar{X} + t\bar{H})}{dt^2}\Big|_{t=0}.$$

For example, if $p(\bar{X}) = X_1 X_2 X_1$, then $p''(\bar{X}, \bar{H}) = 2(X_1 H_2 H_1 + H_1 X_2 H_1 + H_1 H_2 X_1)$. (Note p'' has double the number of variables of p in general.)

Theorem 2.1 (Helton-McCullough [7, Theorem 2.4]) *$p \in \mathbb{R}\langle \bar{X} \rangle$ is convex if and only if $p'' \in \mathbb{R}\langle \bar{X}, \bar{H} \rangle$ is positive.*

An NC polynomial $p \in \mathbb{R}\langle \bar{X} \rangle$ is *positive* if $p(A) \succeq 0$ for all tuples of symmetric matrices A of the same size. In Subsection 3.2 below we will see that testing for positivity of NC polynomials is readily done using sums of hermitian squares programs and semidefinite optimization. Thus convexity is easily tested e.g. using `NCSOStools` and has been implemented under `NCisConvex0`.

Example 2.2 *Surprisingly, $p = X^4$ is* not *convex. While this can be seen from the definition, it is even easier to use Theorem 2.1. First of all,*

$$p'' = 2(X^2H^2 + XH^2X + H^2X^2 + HX^2H + XHXH + HXHX).$$

Now with $x = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ and $h = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, $p''(x,h) = \begin{bmatrix} 34 & -56 \\ -56 & 82 \end{bmatrix}$ is clearly not positive semidefinite.

However, the convexity test can be simplified and greatly improved using [7, Theorem 3.1]: every convex NC polynomial $p \in \mathbb{R}\langle \bar{X} \rangle$ is of degree ≤ 2. Hence only variables $\bar{H}$ will appear in the homogeneous degree 2 polynomial p''. This simplifies the convexity test significantly, see `NCisConvex`. For yet another type of convexity test we refer the reader to [3]. Also, the generalization to *NC rational functions* has been given by Helton, McCullough and Vinnikov [9] but is outside the scope of this note.

2.2 Trace convex NC polynomials

Given the restrictive nature of convexity for NC polynomials, weaker notions need to be considered. One of them uses the trace and is presented in this subsection. Recall: the trace of a square matrix is the sum of its diagonal entries.

An NC polynomial $p \in \mathbb{R}\langle \bar{X} \rangle$ is *trace convex* if it satisfies

$$\operatorname{tr} p(tA + (1-t)B) \leq t \operatorname{tr} p(A) + (1-t) \operatorname{tr} p(B)$$

for all $0 \leq t \leq 1$ and for all tuples A, B of symmetric matrices of the same size.

As with convexity, tracial convexity can be rephrased using second directional derivatives:

Proposition 2.3 *$p \in \mathbb{R}\langle \bar{X} \rangle$ is trace convex if and only if $p'' \in \mathbb{R}\langle \bar{X}, \bar{H} \rangle$ is trace positive, i.e., $\operatorname{tr} p''(\bar{X}, \bar{H}) \geq 0$ for all tuples of symmetric matrices $\bar{X}$ and $\bar{H}$.*

Proof: The proof of this proposition is essentially the same as that of Theorem 2.1 and also resembles the classical proof from calculus showing that a smooth function is convex whenever its Hessian matrix is positive definite. The details are left as an exercise for the interested reader. ■

However, trace positivity turns out to be hard to check, so we consider a relaxation instead, the so-called trace-sohs condition; see Section 5. This can be checked using semidefinite optimization and sums of hermitian squares (with commutators) programs.

3 POSITIVE NC POLYNOMIALS, SUMS OF HERMITIAN SQUARES

Given the inexplicable connection between convexity and positivity of NC polynomials, we now turn to positivity which is easier to handle from a computational point of view.

An NC polynomial of the form g^*g is called a *hermitian square* and the set of all sums of hermitian squares will be denoted by Σ^2. A polynomial $f \in \mathbb{R}\langle \bar{X} \rangle$ is *a sum of hermitian squares (SOHS)* if it belongs to Σ^2. Clearly, $\Sigma^2 \subsetneq \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$. For example, $X_1X_2 + 2X_2X_1 \notin \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$, $X_1^2 X_2 X_1^2 \in \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle \setminus \Sigma^2$,

$$2 + X_1X_2 + X_2X_1 + X_1X_2^2X_1 = 1 + (1 + X_2X_1)^*(1 + X_2X_1) \in \Sigma^2.$$

If $f \in \mathbb{R}\langle \bar{X} \rangle$ is SOHS and we substitute symmetric matrices $A_1, \ldots, A_n$ of the same size for the variables $\bar{X}$, then the resulting matrix $f(A_1, \ldots, A_n)$ is positive semidefinite. Helton [6] proved (a slight variant of) the converse of the above observation: if $f \in \mathbb{R}\langle \bar{X} \rangle$ and $f(A_1, \ldots, A_n) \succeq 0$ for *all* symmetric matrices A_i of the same size, then f is SOHS.

The following proposition (cf. [6, §2.2] or [11, Proposition 2.1]) is the noncommutative version of a classical result due to Choi, Lam and Reznick [4, §2].

Proposition 3.1 *Suppose $f \in \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$ is of degree $\leq 2d$. Then $f \in \Sigma^2$ if and only if there exists a positive semidefinite matrix G satisfying $f = W_d^* G W_d$, where W_d is a vector consisting of all words in $\langle \bar{X} \rangle$ of degree $\leq d$.*

Conversely, given such a positive semidefinite matrix G with rank r, one can construct NC polynomials $g_1, \ldots, g_r \in \mathbb{R}\langle \bar{X} \rangle$ of degree $\leq d$ such that $f = \sum_{i=1}^r g_i^* g_i$.

The matrix G is called *a Gram matrix* for f.

Example 3.2 *In this example we consider an NC polynomial in 2 variables which we denote by X, Y. Let*

$$f = 1 + X^2 + X^3Y + XY + YX + YX^3 + 2YX^4Y + 2YX^2Y.$$

Let V be the subvector $\begin{bmatrix}1 & X & XY & X^2Y\end{bmatrix}^t$ of W_3. Then the Gram matrix for f with respect to V is given by

$$G(a) := \begin{bmatrix} 1 & 0 & 1 & a \\ 0 & 1 & -a & 1 \\ 1 & -a & 2 & 0 \\ a & 1 & 0 & 2 \end{bmatrix}.$$

(Hence $f = V^*G(a)V$.) This matrix is positive semidefinite if and only if $a \in [-1, 1]$ as follows easily from the characteristic polynomial of $G(a)$ or by considering a Schur complement of $G(a)$. When $a = -1$ we have $G(-1) = C^tC$ for $C = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$. From $CV = \begin{bmatrix}1 + XY - X^2Y & X + XY + X^2Y\end{bmatrix}^t$ it follows that

$$f = (1 - XY + X^2Y)^*(1 - XY + X^2Y) + (X + XY + X^2Y)^*(X + XY + X^2Y) \in \Sigma^2.$$

The problem whether a given polynomial is SOHS is therefore a special instance of a semidefinite feasibility problem. This is explained in detail in the following two subsections.

3.1 Semidefinite programming

Semidefinite programming (SDP) is a subfield of convex optimization concerned with the optimization of a linear objective function over the intersection of the cone of positive semidefinite matrices with an affine space. More precisely, given symmetric matrices $C, A_1, \ldots, A_m$ of the same size over $\mathbb{R}$ and a vector $b \in \mathbb{R}^m$, we formulate a *semidefinite program in standard primal form* (in the sequel we refer to problems of this type by PSDP) as follows:

$$\begin{array}{lrcl} \inf & \langle C, G\rangle & & \\ \text{s. t.} & \langle A_i, G\rangle & = & b_i, \quad i = 1, \ldots, m \\ & G & \succeq & 0. \end{array} \tag{PSDP}$$

Here $\langle \cdot, \cdot \rangle$ stands for the standard scalar product of matrices: $\langle A, B\rangle = \operatorname{tr}(B^*A)$. The dual problem to PSDP is the *semidefinite program in the standard dual form*

$$\begin{array}{ll} \sup & \langle b, y\rangle \\ \text{s. t.} & \sum_i y_iA_i \preceq C. \end{array} \tag{DSDP}$$

Here $y \in \mathbb{R}^m$ and the difference $C - \sum_i y_iA_i$ is usually denoted by Z.

The importance of semidefinite programming was spurred by the development of efficient methods which can find an ε-optimal solution in a polynomial time in s, m and $\log \varepsilon$, where s is the order of matrix variables G and Z and m is the number of linear constraints. There exist several open source packages which find such solutions in practice. If the problem is of medium size (i.e., $s \leq 1000$ and $m \leq 10.000$), these packages are based on interior point methods, while packages for larger semidefinite programs use some variant of the first order methods (see [16] for a comprehensive list of state of the art SDP solvers and also [19, 15]). Our standard reference for SDP is [22].

3.2 Sums of hermitian squares and SDP

In this subsection we present a *conceptual algorithm* based on SDP for checking whether a given $f \in \operatorname{Sym}\mathbb{R}\langle \bar{X}\rangle$ is SOHS. Following Proposition 3.1 we must determine whether there exists a positive semidefinite matrix G such that $f = W_d^*GW_d$, where W_d is the vector of all words of degree $\leq d$. This is a semidefinite feasibility problem in the matrix variable G, where the constraints $\langle A_i, G\rangle = b_i$ are implied by the fact that for each product of monomials $w \in \{p^*q \mid p, q \in W_d\}$ the following must be true:

$$\sum_{\substack{p,q\in W_d\\ p^*q=w}} G_{p,q} = a_w, \tag{1}$$

where a_w is the coefficient of w in f ($a_w = 0$ if the monomial w does not appear in f).

Any input polynomial f is symmetric, so $a_w = a_{w^*}$ for all w, and equations (1) can be rewritten as

$$\sum_{\substack{u,v\in W_d\\ u^*v=w}} G_{u,v} + \sum_{\substack{u,v\in W_d\\ u^*v=w^*}} G_{u,v} = a_w + a_{w^*} \quad \forall w \in \{p^*q \mid p, q \in W_d\}, \tag{2}$$

or equivalently,

$$\langle A_w, G\rangle = a_w + a_{w^*} \quad \forall w \in \{p^*q \mid p, q \in W_d\}, \tag{3}$$

where A_w is the symmetric matrix defined by

$$(A_w)_{u,v} = \begin{cases} 2; & \text{if } u^*v \in \{w, w^*\},\ w^* = w,\\ 1; & \text{if } u^*v \in \{w, w^*\},\ w^* \neq w,\\ 0; & \text{otherwise.}\end{cases}$$

Note: $A_w = A_{w^*}$ for all w.

As we are interested in an arbitrary positive semidefinite $G = [G_{u,v}]_{u,v\in W_d}$ satisfying the constraints (3), we can choose the objective function freely. However, in practice one prefers solutions of small rank leading to shorter SOHS decompositions. Hence we minimize the trace, a commonly used heuristic for matrix rank minimization. Therefore our SDP in the primal form is as follows:

$$\begin{array}{lrcl} \inf & \langle I, G\rangle & & \\ \text{s.t.} & \langle A_w, G\rangle & = & a_w + a_{w^*} \quad \forall w \in \{p^*q \mid p, q \in W_d\} \\ & G & \succeq & 0. \end{array} \tag{SOHS$_{\text{SDP}}$}$$

(Here and in the sequel, I denotes the identity matrix of appropriate size.) To reduce the size of this SDP (i.e., to make W_d smaller), we may employ the following simple observation:

Proposition 3.3 *Let $f \in \operatorname{Sym}\mathbb{R}\langle\bar{X}\rangle$, let $m_i := \frac{\operatorname{mindeg}_i f}{2}$, $M_i := \frac{\deg_i f}{2}$, $m := \frac{\operatorname{mindeg} f}{2}$, $M := \frac{\deg f}{2}$. Set $V := \{w \in \langle\bar{X}\rangle \mid m_i \leq \deg_i w \leq M_i \text{ for all } i,\ m \leq \deg w \leq M\}$. Then $f \in \Sigma^2$ if and only if there exists a positive semidefinite matrix G satisfying $f = V^*GV$.*

Proof: This follows from the fact that the highest or lowest degree terms in a SOHS decomposition cannot cancel. ■

Example 3.4 (Example 3.2 revisited) *Let us return to*

$$f = 1 + X^2 + X^3Y + XY + YX + YX^3 + 2YX^4Y + 2YX^2Y.$$

We shall describe in some detail (SOHS$_{\text{SDP}}$) for f. From Proposition 3.3, we obtain

$$V = \begin{bmatrix}1 & X & Y & XY & YX & X^2 & X^2Y & YX^2 & XYX\end{bmatrix}^t.$$

Thus G is a symmetric 9×9 matrix and there will be 47 matrices A_w, as $|\{u^*v \mid u, v \in V\}| = 47$. In fact, there are only 32 different matrices A_w as by definition we have $A_w = A_{w^*}$. Here is a sample:

$$A_{YX} = A_{XY} = \begin{bmatrix} 0&0&0&1&1&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \end{bmatrix}, \quad A_{X^2YX} = A_{XYX^2} = \begin{bmatrix} 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \end{bmatrix}.$$

These two give rise to the following linear constraints in ($SOHS_{\mathrm{SDP}}$):

$$\begin{aligned} G_{1,XY} + G_{X,Y} + G_{XY,1} + G_{1,YX} + G_{Y,X} + G_{YX,1} &= \langle A_{XY}, G\rangle \\ &= a_{XY} + a_{YX} = 2, \\ G_{X^2,YX} + G_{X^2Y,X} + G_{XY,X^2} + G_{X,YX^2} = \langle A_{X^2YX}, G\rangle &= a_{X^2YX} + a_{XYX^2} = 0, \end{aligned}$$

where we have used a_w to denote the coefficients of f and the entries of V enumerate the columns, while the entries of V^* enumerate the rows of G.

Comparing with Example 3.2 we see that the Gram matrix obtained by using the vector V from Proposition 3.3 is larger than necessary. A closer look into the semidefinite program reveals that among the 32 equations we also have the following:

$$2G_{XYX,XYX} \;=\; \langle A_{XYX^2YX}, G\rangle \;=\; a_{XYX^2YX} \;=\; 0.$$

Hence the diagonal entry $G_{XYX,XYX}$ must be equal to 0, so the whole row $G_{XYX,:}$ and column $G_{:,XYX}$ must equal zero, since we are looking for a positive semidefinite matrix G. Therefore we can eliminate XYX from V and reduce the complexity of our semidefinite program. Similarly we realize that $G_{X^2,X^2} = 0$, $G_{Y,Y} = 0$, $G_{XY,YX} = 0$ and $G_{X^2Y,YX^2} = 0$, hence the Gram matrix finally reduces to a 4×4 matrix and our semidefinite program contains only 9 linear constraints.

A procedure of constructing a vector V eliminating all "obviously" redundant words is called the *Augmented Newton Chip method* and is presented in detail [11]. It is implemented in `NCSOStools` in the procedure `NCsos`. We review the theoretical underpinning below and demonstrate it on an example.

Define for $i \in \mathbb{N}_0$ the ith *right chip function* $\mathrm{rc}_i : \langle \bar{X}\rangle \to \langle \bar{X}\rangle$ by

$$\mathrm{rc}_i(w_1 \cdots w_n) := w_{n-i+1}w_{n-i+2}\cdots w_n$$

if $i \le n$ and $\mathrm{rc}_i(w) = w$ otherwise. (In case $i = 0$, the empty product is defined to be the empty word 1.) As an example, $\mathrm{rc}_4(X_4X_3^6X_1X_2^2X_1) = X_1X_2^2X_1$.

Algorithm 1 below (the Augmented Newton chip method) reduces the word vector needed in the Gram matrix test for a sum of hermitian squares decomposition of a symmetric NC polynomial f.

Theorem 3.5 ([11, Theorem 3.1, Lemma 4.2]) *Suppose $f \in \operatorname{Sym}\mathbb{R}\langle \bar{X}\rangle$. Then $f \in \Sigma^2$ if and only if there exists a positive semidefinite G satisfying*

$$f = W^*GW, \tag{4}$$

where W is the output in vector form given by the Augmented Newton chip method.

INPUT: $f \in \operatorname{Sym}\mathbb{R}\langle \bar{X} \rangle$ *with* $\deg f \leq 2d$, $f = \sum_{w \in \langle \bar{X} \rangle} a_w w,$, *where* $a_w \in \mathbb{R}$.

STEP 1: *Define the support of* f *as* $\mathcal{W}_f := \{w \in \langle \bar{X} \rangle \mid a_w \neq 0\}$.

STEP 2: $W := \varnothing$.

STEP 3: *Let* $m_i := \frac{\operatorname{mindeg}_i f}{2}$, $M_i := \frac{\deg_i f}{2}$, $m := \frac{\operatorname{mindeg} f}{2}$, $M := \frac{\deg f}{2}$. *The set of admissible words is defined as*

$$\mathcal{D} := \{w \in \langle \bar{X} \rangle \mid m_i \leq \deg_i w \leq M_i \text{ for all } i,\, m \leq \deg w \leq M\}.$$

STEP 4: *For every* $w^*w \in \mathcal{W}_f$:

Substep 4.1 *For* $0 \leq i \leq \deg w$: *if* $\operatorname{rc}_i(w) \in \mathcal{D}$, *then* $W := W \cup \{\operatorname{rc}_i(w)\}$.

STEP 5: *Compute the matrix* $Z = W^*W$.

Substep 5.1 *While there exists* $z_{ii} = w^*w$ *such that* $a_{z_{ii}} = 0$ *and* $z_{ii} \neq z_{j,k}$ *for all* j, k *with* $j, k \neq i$: *delete* w *from* W *and recompute* Z.

OUTPUT: W.

Algorithm 1: The Augmented Newton chip method

Example 3.6 *Let us consider the following NC polynomial*

$$f = X^{20}Y^{30}X^{30}Y^{30}X^{20} - X^{20}Y^{30}X^{16}Y - YX^{16}Y^{30}X^{20} + YX^2Y$$

The vector W *given by Proposition 3.3 contains all possible monomials with degree* ≤ 65 *and there are exactly* $2^{66} - 1$ *such monomials. Without any reduction this would give rise to a semidefinite program in a matrix variable of order* $2^{66} - 1$ *which is certainly out of reach for any state-of-the-art SDP solver. However, the Augmented Newton chip method, implemented in the procedure* `NCsos` *in our package* `NCSOStools` *[2], yields exactly two monomials* $W = \begin{bmatrix} XY & X^{15}Y^{30}X^{20} \end{bmatrix}^t$ *and the optimal solution:*

$$f = (XY - X^{15}Y^{30}X^{20})^*(XY - X^{15}Y^{30}X^{20}).$$

4 COMMUTATORS AND ZERO TRACE NC POLYNOMIALS

It is well-known and easy to see that trace zero matrices are sums of commutators. Less obvious is the fact (not needed in this paper) that trace zero matrices are commutators.

In this section we present the corresponding theory for NC polynomials and describe how it is implemented in `NCSOStools`. Most of the results are taken from [12] and we refer the reader to [1] for some extensions.

Definition 4.1 *An element of the form* $[p, q] := pq - qp$ *for* $p, q \in \mathbb{R}\langle \bar{X} \rangle$ *is called a* commutator. *Two NC polynomials* $f, g \in \mathbb{R}\langle \bar{X} \rangle$ *are called* cyclically equivalent *(*$f \overset{\text{cyc}}{\sim} g$*) if* $f - g$ *is a sum of commutators:*

$$f - g = \sum_{i=1}^{k} (p_i q_i - q_i p_i) \text{ for some } k \in \mathbb{N} \cup \{0\} \text{ and } p_i, q_i \in \mathbb{R}\langle \bar{X} \rangle.$$

Example 4.2 $2X^3Y + 3XYX^2 \overset{\text{cyc}}{\sim} X^2YX + 4YX^3$ *as*

$$2X^3Y + 3XYX^2 - (X^2YX + 4YX^3) = [2X, X^2Y] + [X, XY^2] + [4X, YX^2].$$

The following remark shows that cyclic equivalence can easily be tested.

Remark 4.3

(a) For $v, w \in \langle \bar{X} \rangle$, we have $v \overset{\text{cyc}}{\sim} w$ if and only if there are $v_1, v_2 \in \langle \bar{X} \rangle$ such that $v = v_1v_2$ and $w = v_2v_1$. That is, $v \overset{\text{cyc}}{\sim} w$ if and only if w is a cyclic permutation of v.

(b) Two polynomials $f = \sum_{w \in \langle \bar{X} \rangle} a_w w$ and $g = \sum_{w \in \langle \bar{X} \rangle} b_w w$ ($a_w, b_w \in \mathbb{R}$) are cyclically equivalent if and only if for each $v \in \langle \bar{X} \rangle$,

$$\sum_{\substack{w \in \langle \bar{X} \rangle \\ w \overset{\text{cyc}}{\sim} v}} a_w = \sum_{\substack{w \in \langle \bar{X} \rangle \\ w \overset{\text{cyc}}{\sim} v}} b_w.$$

Given $f \overset{\text{cyc}}{\sim} g$ and an n-tuple of symmetric matrices A of the same size, $\operatorname{tr} f(A) = \operatorname{tr} g(A)$. In particular, if f is a sum of commutators, i.e., $f \overset{\text{cyc}}{\sim} 0$, then $\operatorname{tr} f(A) = 0$ for all such A. The converse of this easy observation is given by the following tracial Nullstellensatz:

Theorem 4.4 (Klep-Schweighofer [12]) *Let $d \in \mathbb{N}$ and $f \in \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$ be of degree $\leq d$ satisfying $\operatorname{tr}(f(A_1, \ldots, A_n)) = 0$ for all symmetric $A_1, \ldots, A_n \in \mathbb{R}^{d \times d}$. Then $f \overset{\text{cyc}}{\sim} 0$.*

To see what happens when the size of the matrices A_i is smaller than the degree of f, we refer the reader to [1].

The cyclic equivalence test has been implemented under `NCSOStools` - see `NCisCycEq`.

5 TRACE POSITIVE NC POLYNOMIALS

A notion of positivity of NC polynomials weaker than that via positive semidefiniteness considered in Section 3, is given by the trace: $f \in \mathbb{R}\langle \bar{X} \rangle$ is called *trace positive* if $\operatorname{tr} f(A) \geq 0$ for all tuples of symmetric matrices A of the same size. Clearly, every $f \in \Sigma^2$ is trace positive and the same is true for every NC polynomial cyclically equivalent to SOHS. However, unlike in the positive semidefinite case, the converse fails. That is, there are trace positive polynomials which are not cyclically equivalent to SOHS, see [12, Example 4.4] or [13, Example 3.5]. Nevertheless, the obvious certificate for trace positivity has been shown to be useful in applications to e.g. operator algebras [12] and mathematical physics [13], so deserves further attention.

Let

$$\Theta^2 := \{f \in \mathbb{R}\langle \bar{X} \rangle \mid \exists g \in \Sigma^2 : \ f \overset{\text{cyc}}{\sim} g\}$$

denote the convex cone of all NC polynomials cyclically equivalent to SOHS. By definition the elements in Θ^2 are exactly the polynomials which are sums of hermitian squares and commutators. In particular, every $f \in \Theta^2$ is trace positive.

Testing whether a given $f \in \mathbb{R}\langle \bar{X} \rangle$ is an element of Θ^2 can be done again using SDP (the so-called *Gram matrix method*) as observed in [13, §3]. A slightly improved algorithm reducing the size of the SDP needed is given by the following theorem.

Theorem 5.1 *Let $f \in \mathbb{R}\langle \bar{X} \rangle$, let $m_i := \frac{\operatorname{mindeg}_i f}{2}$, $M_i := \frac{\deg_i f}{2}$, $m := \frac{\operatorname{mindeg} f}{2}$, $M := \frac{\deg f}{2}$. Set*

$$V := \{w \in \langle \bar{X} \rangle \mid m_i \leq \deg_i w \leq M_i \text{ for all } i, \, m \leq \deg w \leq M\}.$$

Then $f \in \Theta^2$ if and only if there exists a positive semidefinite matrix G satisfying $f \overset{\text{cyc}}{\sim} V^*GV$.

Proof: Suppose $f \overset{\text{cyc}}{\sim} V^*GV$ for some positive semidefinite G. Then $G = \sum G_i G_i^t$ for some vectors G_i and $V^*GV = \sum g_i^* g_i$, where $g_i = G_i^t V$. Thus $f \in \Theta^2$.

Conversely, suppose $f \overset{\text{cyc}}{\sim} \sum g_i^* g_i$. We claim that each g_i is in the linear span of V. Assume otherwise, say one of the g_i contains a word w with $\deg_j w < m_j$. Let $\delta < m_j$ be the minimum $\deg_j$ degree in all g_i and let h_i denote the sum of all monomials of g_i whose corresponding words have $\deg_j$ equal to δ. Let $r_i = g_i - h_i$. Then $\deg_j r_i > \delta$ and

$$f \overset{\text{cyc}}{\sim} \sum g_i^* g_i = \sum (h_i + r_i)^*(h_i + r_i) = \sum h_i^* h_i + \sum h_i^* r_i + \sum r_i^* h_i + \sum r_i^* r_i. \quad (5)$$

Since each monomial w in $h_i^* r_i$, $r_i^* h_i$ and $r_i^* r_i$ has $\deg_i w \geq 2\delta$, none of these can be cyclically equivalent to a monomial in $h_i^* h_i$. Thus

$$0 \overset{\text{cyc}}{\sim} \sum h_i^* h_i, \quad f \overset{\text{cyc}}{\sim} \sum h_i^* r_i + \sum r_i^* h_i + \sum r_i^* r_i.$$

However, this implies $h_i = 0$ for all i (see [13, Lemma 3.2]; or also the proof of Proposition 3.1), contradicting the choice of w.

The remaining cases (i.e., a word w in one of the g_i with $\deg_i w > M_i$ or $\deg w < m$ or $\deg w > M$) can be dealt with similarly, so we omit the details. ■

Testing whether a NC polynomial is a sum of hermitian squares and commutators (i.e., an element of Θ^2) has been implemented under `NCSOStools` as `NCcycSos`.

Example 5.2 *Consider the polynomials*

$$f = S_{6,2}(X,Y) = X^4Y^2 + X^3YXY + X^3Y^2X + X^2YX^2Y + X^2YXYX + X^2Y^2X^2 + XYX^3Y + XYX^2YX + XYXYX^2 + XY^2X^3 + YX^4Y + YX^3YX + YX^2YX^2 + YXYX^3 + Y^2X^4.$$

$$g = S_{8,2}(X,Y) = X^6Y^2 + X^5YXY + X^5Y^2X + X^4YX^2Y + X^4YXYX + X^4Y^2X^2 + X^3YX^3Y + X^3YX^2YX + X^3YXYX^2 + X^3Y^2X^3 + X^2YX^4Y + X^2YX^3YX + X^2YX^2YX^2 + X^2YXYX^3 + X^2Y^2X^4 + XYX^5Y + XYX^4YX + XYX^3YX^2 + XYX^2YX^3 + XYXYX^4 + XY^2X^5 + YX^6Y + YX^5YX + YX^4YX^2 + YX^3YX^3 + YX^2YX^4 + YXYX^5 + Y^2X^6.$$

Polynomial f (resp. g) is the sum of all words in two variables X, Y of degree 6 (resp. 8) with $\deg_Y f = \deg_Y g = 2$ (and therefore $\deg_X f = 4$, $\deg_X g = 6$). Trace positivity of these polynomials is closely related to the Bessis-Moussa-Villani (BMV) conjecture from quantum statistical mechanics, see [13] and references therein for details.

Using `NCSOStools` we can prove trace positivity of f, g and even the stronger statement $f, g \in \Theta^2$. Namely, we can obtain explicit (rational) SOHS decompositions:

$$f \overset{\text{cyc}}{\sim} \frac{13}{3}\left(\frac{6}{13}X^2Y + XYX + \frac{3}{13}YX^2\right)^*\left(\frac{6}{13}X^2Y + XYX + \frac{3}{13}YX^2\right) + \frac{27}{13}\left(X^2Y - \frac{44}{81}YX^2\right)^*\left(X^2Y - \frac{44}{81}YX^2\right) + \frac{524}{243}(YX^2)^*(YX^2),$$

$$g \overset{\text{cyc}}{\sim} 4\left(X^3Y + \frac{1}{2}X^2YX + \frac{1}{3}XYX^2 + \frac{1}{4}YX^3\right)^*\left(X^3Y + \frac{1}{2}X^2YX + \frac{1}{3}XYX^2 + \frac{1}{4}YX^3\right) + \frac{15}{4}\left(\frac{2}{9}X^2YX + \frac{4}{9}XYX^2 + YX^3\right)^*\left(\frac{2}{9}X^2YX + \frac{4}{9}XYX^2 + YX^3\right) + \frac{4}{27}\left(-\frac{1}{4}X^2YX + XYX^2\right)^*\left(-\frac{1}{4}X^2YX + XYX^2\right) + \frac{5}{36}\left(X^2YX\right)^*\left(X^2YX\right).$$

Finally, we return to trace convex NC polynomials. Recall from Subsection 2.2 that $p \in \mathbb{R}\langle \bar{X} \rangle$ is trace convex if and only if $p'' \in \mathbb{R}\langle \bar{X}, \bar{H} \rangle$ is trace positive. As the latter property is hard to verify in practice, `NCSOStools` instead tests for a slightly stronger property $p'' \in \Theta^2$; see `NCisCycConvex`. However, for NC polynomials in *one* variable, p is trace convex if and only if $p'' \in \Theta^2$, a result due to Chris Nelson, et al. at UCSD. Equivalently: p is convex as a polynomial of one *commuting* variable. In particular, $p = X^4$ considered in Example 2.2 *is* trace convex.

We do not know whether $p \in \operatorname{Sym} \mathbb{R}\langle \bar{X} \rangle$ is trace convex if and only if $p'' \in \Theta^2$ in general for NC polynomials p of more than one variable.

6 CONCLUSIONS

In this paper we present some of the theoretical background needed for `NCSOStools`: a computer algebra system for working with noncommutative (NC) polynomials with a special focus on methods determining whether a given NC polynomial is a sum of hermitian squares (SOHS) or is cyclically equivalent to SOHS (i.e., is a sum of hermitian squares and commutators). `NCSOStools` is an open source MATLAB toolbox freely available from our web site

`http://ncsostools.fis.unm.si/`

The package contains several extensions, like computing SOHS lower bounds and checking for convexity or trace convexity of given NC polynomials. Moreover, functions have been implemented to handle cyclic equivalence. Most of the methods rely on semidefinite programming therefore the user should provide an SDP solver. Currently SDPT3 [23] and SeDuMi [21] are supported, while other solvers will be added in the future.

`NCSOStools` can handle NC polynomials of medium size, while larger problems may run into trouble for two reasons: the underlying SDP is too big for the state-of-the-art SDP solvers or the (combinatorial) process of constructing the SDP is too exhaustive. The ongoing research will mainly concern the second issue (with improvements for sparse NC polynomials and NC polynomials with symmetries). Also methods to produce exact rational solutions from numerical solutions given by SDP solvers are being implemented, in the spirit of [18].

References

[1] M. Brešar and I. Klep. Noncommutative polynomials, Lie skew-ideals and tracial Nullstellensätze, available from `http://srag.fmf.uni-lj.si`. *Accepted for publication in Math. Res. Lett.*, 2009.

[2] K. Cafuta, I. Klep, and J. Povh. NCSOStools: a computer algebra system for symbolic and numerical computation with noncommutative polynomials, available from `http://ncsostools.fis.unm.si/downloads/ncsostools_manual.pdf`. 6 Mar 2009.

[3] J. Camino, J. Helton, R. Skelton, and J. Ye. Matrix inequalities: a symbolic procedure to determine convexity automatically. *Integral Equations Operator Theory*, 46(4):399–454, 2003.

[4] M. Choi, T. Lam, and B. Reznick. Sums of squares of real polynomials. In *K-theory and algebraic geometry: connections with quadratic forms and division algebras (Santa Barbara, CA, 1992)*, volume 58 of *Proc. Sympos. Pure Math.*, pages 103–126. Amer. Math. Soc., Providence, RI, 1995.

[5] M. de Oliveira, J. Helton, S. McCullough, and M. Putinar. Engineering systems and free semi-algebraic geometry. In *Emerging Applications of Algebraic Geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 17–62. Springer, 2008.

[6] J. Helton. "Positive" noncommutative polynomials are sums of squares. *Ann. of Math. (2)*, 156(2):675–694, 2002.

[7] J. Helton and S. McCullough. Convex noncommutative polynomials have degree two or less. *SIAM J. Matrix Anal. Appl.*, 25(4):1124–1139, 2004.

[8] J. Helton, R. Miller, and M. Stankus. NCAlgebra: A Mathematica package for doing non commuting algebra, available from `http://www.math.ucsd.edu/~ncalg/`. 1 Mar 2009.

[9] J. W. Helton, S. A. McCullough, and V. Vinnikov. Noncommutative convexity arises from linear matrix inequalities. *J. Funct. Anal.*, 240(1):105–191, 2006.

[10] D. Henrion, J.-B. Lasserre, and J. Löfberg. GloptiPoly 3: moments, optimization and semidefinite programming, available from `http://www.laas.fr/~henrion/software/gloptipoly3/`. 1 Mar 2009.

[11] I. Klep and J. Povh. Semidefinite programming and sums of hermitian squares of noncommutative polynomials, available from `http://ncsostools.fis.unm.si/`. *Accepted for publication in J. Pure Appl. Algebra*, 2009.

[12] I. Klep and M. Schweighofer. Connes' embedding conjecture and sums of Hermitian squares. *Adv. Math.*, 217(4):1816–1837, 2008.

[13] I. Klep and M. Schweighofer. Sums of Hermitian squares and the BMV conjecture. *J. Stat. Phys*, 133(4):739–760, 2008.

[14] J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB, available from `http://control.ee.ethz.ch/~joloef/yalmip.php`. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[15] J. Malick, J. Povh, F. Rendl, and A. Wiegele. Regularization methods for semidefinite programming. *SIAM Journal on Optimization*, 20(1):336–356, 2009.

[16] H. Mittelman. `http://plato.asu.edu/sub/pns.html`. 1 Mar 2009.

[17] NCSOStools. `http://ncsostools.fis.unm.si`. 1 Mar 2009.

[18] H. Peyrl and P. Parrilo. Computing sum of squares decompositions with rational coefficients. *Theoretical Computer Science*, 409(2):269–281, 2008.

[19] J. Povh, F. Rendl, and A. Wiegele. A boundary point method to solve semidefinite programs. *Computing*, 78:277–286, 2006.

[20] S. Prajna, A. Papachristodoulou, P. Seiler, and P. Parrilo. SOSTOOLS and its control applications, available from `http://www.cds.caltech.edu/sostools/`. In *Positive polynomials in control*, volume 312 of *Lecture Notes in Control and Inform. Sci.*, pages 273–292. Springer, Berlin, 2005.

[21] SeDuMi. `http://sedumi.ie.lehigh.edu/`. 29 June 2009.

[22] M. Todd. Semidefinite optimization. *Acta Numerica*, 10:515–560, 2001.

[23] K.-C. Toh, M. Todd, and R.-H. Tütüncü. SDPT3 version 4.0 (beta) – a MATLAB software for semidefinite-quadratic-linear programming, available from `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`. 1 Mar 2009.

GAUSSIAN PROCESS MODELLING OF DEPENDENCIES IN MULTI-ARMED BANDIT PROBLEMS

Louis Dorard, **Dorota Glowacka** and **John Shawe-Taylor**
University College London, Department of Computer Science
London WC1E 6BT, United Kingdom
{l.dorard,d.glowacka,jst}@cs.ucl.ac.uk

Abstract: Multi-armed bandit problems, in analogy with slot machines in casinos, are problems in which one has to choose actions sequentially (pull arms) in order to maximise a cumulated reward (gain), with no initial knowledge on the distribution of actions/arms' rewards. We propose a general framework for handling dependencies across arms, based on a new assumption on the mean-reward function which is that it is drawn from a Gaussian Process (GP), with a given arm covariance matrix. We show on a toy problem that this allows to perform better than the popular UCB bandit algorithm, which considers arms to be independent.

Keywords: Artificial Intelligence, Machine Learning, Learning and Optimisation, Bandit problems, Kernel Methods, Gaussian Processes.

1 Introduction

1.1 Background

The multi-armed bandit problem is an analogy with a traditional slot machine, known as one-armed bandit, but with multiple arms. In the classical bandit scenario, the player, after pulling (or 'playing') an arm selected from a finite number of arms, receives a reward. The player has no initial knowledge about the arms, and attempts to maximise the cumulated reward through repeated plays. It is assumed that the reward obtained when playing an arm i is a sample from an unknown distribution R_i with mean μ_i. The optimal playing strategy S^*, i.e. the strategy that yields maximum cumulated reward, consists in always playing an arm i^* such that $i^* = \text{argmax}_i \mu_i$. The expected cumulated reward of S^* at time t would then be $t\mu_{i^*}$. The performance of a strategy S is assessed by the analysis of its expected regret at time t, defined as the difference between the expected cumulated reward of S^* and S at time t.

A good strategy requires to optimally balance the learning of the distributions R_i and the exploitation of arms which have been learnt as having high expected rewards. As the number of arms is finite (and usually smaller than the number of experiments allowed), it is possible to explore all the possible options (arms) a certain number of times, thus building empirical estimates of the μ_i's, and exploit the best performing ones.

Real-world applications are varied and include for instance advertisement on the web, where pulling an arm corresponds to placing an ad on a given webpage, and rewards are visitor clicks.

1.1.1 Independent arms

The UCB (Upper Confidence Bounds) algorithm and its variants ([1]) are the algorithms of reference for multi-armed bandit problems in which arms are considered to be independent. Let us denote by $\hat{\mu}_t(\mathbf{x})$ the empirical mean reward for arm $\mathbf{x}$. In order to optimise the reward at each time step in the face of uncertainty, confidence intervals $[\hat{\mu}_t(\mathbf{x}) - \hat{\sigma}_t(\mathbf{x}); \hat{\mu}_t(\mathbf{x}) + \hat{\sigma}_t(\mathbf{x})]$ are determined for the reward value we would get for each arm $\mathbf{x}$, and the algorithm plays at time t the arm which maximises the value of the upper confidence bound $\hat{\mu}_t(\mathbf{x}) + \hat{\sigma}_t(\mathbf{x})$. The bigger $\hat{\sigma}_t(\mathbf{x})$, the more likely it is that the reward will be in the confidence interval. Choosing $\hat{\sigma}_t(\mathbf{x}) = \sqrt{\frac{2\log t}{Pulls(\mathbf{x})}}$, where $Pulls(\mathbf{x})$ is the number of pulls of arm $\mathbf{x}$ so far, enables to prove a bound on the regret of UCB in $O(\log t)$.

The algorithm is as follows:

- Initialisation: play each arm once, set $t = N + 1$.
- Loop:
  - Play $\mathbf{x}_t = \text{argmax}_{\mathbf{x}}\{\hat{\mu_t}(\mathbf{x}) + \hat{\sigma_t}(\mathbf{x})\}$
  - Get reward y_t, update $\hat{\mu_t}(\mathbf{x}_t)$ accordingly, increment $Pulls(\mathbf{x}_t)$.
  - $t = t + 1$

The amount of information used by UCB in order to make its decisions is in $O(N)$, as variables $Pulls$ and $\hat{\mu_t}$ have N entries.

1.1.2 Dependent arms

Modelling dependencies across arms – if such dependencies exist – makes the exploration phases shorter, as, when an arm is played, knowledge is gained on this particular arm but also on similar arms. Thus, at each play, we can update the μ_i estimates of several arms and not just one. There have been relatively few studies on multi-armed bandit algorithms that exploit arm dependencies in their decision making. [6] have developed an algorithm which exploits cluster structures among arms. They have shown that the cumulated reward obtained with their algorithm increases faster than that of UCB on an application to web advertisement. Other studies on modelling dependencies across arms focus on the case where there is an infinity of arms which are indexed in a metric space, and make assumptions on the regularity of the mean-reward function in that space ([4] and [2]).

1.2 Motivation

We assume in our model that the reward of an arm $\mathbf{x}$ is determined by a function f applied at point $\mathbf{x}$ to which Gaussian noise is added. The variance of the noise corresponds to the variability of the reward when always playing the same arm.

Our assumption is, loosely speaking, that the rewards of arms are correlated, i.e. that it is likely that the more similar two arms are, the more similar the rewards obtained by playing them will be. Thus, playing an arm 'close' to $\mathbf{x}$ gives information on the expected gain of playing $\mathbf{x}$. These correlations can be modelled by assuming that f is a function drawn from a Gaussian Process (GP). By default, we take the mean of the GP prior to be $\mathbf{0}$, and we incorporate prior knowledge on how correlated arms are in the GP covariance matrix. Each entry (i, j) of this matrix specifies how 'close' or 'similar' arms i and j are. If arms are indexed in $\mathbb{R}^d$, for example, the covariance function can be chosen to be a squared exponential whose smoothness is adjusted to fit the characteristic length scale which is assumed for f. We are not limited to problems where arms are indexed in $\mathbb{R}^d$, but we can also consider arms indexed by any type of structured data as long as a kernel can be defined between the data points.

Hence, whereas [6]'s model of dependencies is based on a clustering of arms, our model is based on an arms covariance/kernel matrix. Using Gaussian Processes allows a similar approach to arm selection than in UCB, where we look for the arm which maximises an upper confidence function. This function, as we saw, is a sum of a mean-reward estimate ($\hat{\mu_t}(\mathbf{x})$) and an uncertainty term ($\hat{\sigma_t}(\mathbf{x})$). In our case, the estimated reward is given by the mean of the GP posterior, and the uncertainty term by the posterior variance.

1.3 Outline

The remainder of this paper is as follows: in section 2 we first present our Gaussian Processes Bandits algorithm, then analyse its computational complexity and propose more efficient variants. We propose a toy multi-armed bandit problem in section 3 and analyse the regrets of UCB and GPB on this problem.

Finally, in section 4 we bring forward our conclusions and motivate the use of our algorithm in other problems.

2 The GPB (Gaussian Processes Bandits) algorithm

2.1 Notations

We consider space $\mathcal{X} = \mathbf{X}_1, ...\mathbf{X}_N$, whose elements will be referred to as arms. κ denotes a kernel between elements of $\mathcal{X}$. The reward after playing arm $\mathbf{x} \in \mathcal{X}$ is given by $f(\mathbf{x})+\epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2_{noise})$ and $f \sim \mathcal{GP}(\mathbf{0}, \kappa(\mathbf{x}, \mathbf{x}'))$ is chosen once and for all but is unknown. Arms played up to time t are $\mathbf{x}_1, \ldots, \mathbf{x}_t$ with rewards $y_1, \ldots, y_t$. The GP posterior at time t after seeing data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)$ has mean $\mu_t(\mathbf{x})$ with variance $\sigma^2_t(\mathbf{x})$.

Matrix C_t and vector $\mathbf{k}_t(\mathbf{x})$ are defined as follows:

$$\begin{aligned} (C_t)_{i,j} &= \kappa(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2_{noise}\delta_{i,j} \\ (\mathbf{k}_t(\mathbf{x}))_i &= \kappa(\mathbf{x}, \mathbf{x}_i) \end{aligned}$$

$\mu_t(\mathbf{x})$ and $\sigma^2_t(\mathbf{x})$ are then given by the following equations (see [7]):

$$\mu_t(\mathbf{x}) = \mathbf{k}_t(\mathbf{x})^T C_t^{-1} \mathbf{y}_t \quad (1)$$

$$\sigma^2_t(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^T C_t^{-1} \mathbf{k}_t(\mathbf{x}) \quad (2)$$

2.2 Arm selection

The algorithm plays a sequence of arms and aims at optimally balancing exploration and exploitation. For this, we select arms iteratively according to a UCB-inspired formula:

$$\mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x}\in\mathcal{X}} \{ f_t(\mathbf{x}) = \mu_t(\mathbf{x}) + B(t)\sigma_t(\mathbf{x}) \} \quad (3)$$

This can be seen as active learning where we want to learn accurately in regions where the function looks good, and do not care if we make inaccurate predictions elsewhere. The $B(t)$ term balances exploration and exploitation: the bigger it gets, the more it favours points with high $\sigma_t(\mathbf{x})$ (exploration), while if $B(t) = 0$, the algorithm is greedy. In the original UCB formula, $B(t) \sim \sqrt{\log t}$.

The objective of the bandit algorithm is to minimise the regret over time, or in other terms, to find as quickly as possible a good approximation of the maximum of f, $f(\mathbf{x}^*)$. Let us define $f_t(\mathbf{x})$ by:

$$\begin{aligned} f_t(\mathbf{x}) &= \mu_t(\mathbf{x}) + B(t)\sigma_t(\mathbf{x}) \\ &= \mathbf{k}_t(\mathbf{x})^T C_t^{-1} \mathbf{y}_t + B(t)\sqrt{\kappa(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t(\mathbf{x})^T C_t^{-1} \mathbf{k}_t(\mathbf{x})} \end{aligned}$$

Our approximation of $f(\mathbf{x}^*)$ at time t is $f(\mathbf{x}_{t+1})$ where $\mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x}\in\mathcal{X}}\{f_t(\mathbf{x})\}$.

The argmax is found by exhaustive search of $\mathcal{X}$. Here we replace the problem of finding the maximum of the function f by iterations of a simpler problem, which is to maximise the function f_t which is known. At each iteration we learn new information which enables us to improve our approximation of $f(\mathbf{x}^*)$ over time.

2.3 Complexity of iteration t of GPB

There are N arms for which we need to compute f_t. The cost of computing f_t is equal to the cost of multiplying and adding the terms of Equation 3: $O(t^2)$ since the vectors are of size t and the matrices are $t \times t$).

To this we must add the cost of computing the covariance matrix inverse. If we write:

$$C_t = \begin{pmatrix} C_{t-1} & \mathbf{k}_{t-1}(\mathbf{x}_t) \\ \mathbf{k}_{t-1}(\mathbf{x}_t)^T & \kappa(\mathbf{x}_t, \mathbf{x}_t) + \sigma^2_{noise} \end{pmatrix}$$

we can use the block inversion lemma and do the computations iteratively. The cost is then $O(t^2)$ (instead of $O(t^3)$ for a standard inversion procedure).

Hence, the total cost of iteration t is: $O(t^2) * N + O(t^2) = O(t^2 N)$. In comparison, the cost of iteration t of UCB is $O(N)$ as it is equal to N times the cost of computing the upper confidence function on a given arm (constant). UCB is more efficient because its cost is constant in time.

2.4 Optimisation: reducing the size of the training set

Actually, the cost of iteration t of the algorithm is $O(|Tr(t)|^2 N)$ where $Tr(t)$ is the training set at time t, which has t elements. We can reduce the computational cost by reducing the size of the training set. For example, at each iteration we add an arm to the training set, and we can also decide to remove the oldest arm in the training set if the size of $Tr(t)$ is bigger than a certain function $S(t)$. We may want to choose $S(t) = \sqrt{N}$ so that the amount of information that GPB uses is in $O(\sqrt{N}^2) = O(N)$ and is of the same order as the amount of information used by UCB. The cost of iteration t of GPB becomes $O(N^2)$ plus, possibly, the cost of removing an element from the training set.

Two operations are required in order to remove an element from the training set:

- remove it from the list of arms that have been tried so far: $O(1)$ or $O(|Tr(t)|)$ depending on the implementation
- update C_t^{-1}, by reverting the block inversion formula: $O(|Tr(t)|^2)$

Hence, the cost of iteration t is $O(N^2) + O(N) + O(1) = O(N^2)$. Because we limited the size of the training set to a fixed value, the cost of the algorithm is now constant in time (although the constant here is bigger than with UCB).

This modified version of GPB will be referred to as *GPB-red* in the rest of this paper. An interesting thing to notice in GPB-red is that, when removing a data point from the training set, we increase the GP posterior variance at this point and points close to it, and therefore we give them more chances to be selected as increasing the variance increases f_t.

3 Experiments

3.1 Description of toy problem

For our experiments we create a toy problem where rewards are given as values of a function f plus Gaussian noise. f is chosen to be minus the 2 dimensional Rosenbrock function[1], defined in $[-1, 1] \times [-1, 1]$ and scaled so that its values are in $[0, 1]$ and its maximum is 1 (knowing the maximum of f allows to determine the regret of arm selection policies). f is defined as follows on the unit disk:

$$f(\mathbf{x}) = (-100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 161)/161$$

This can be extended to $\mathbf{x} \in [-1, 1] \times [-1, 1]$ by setting, for $\mathbf{x}$ outside the unit disk:

$$f(\mathbf{x}) = f\left(\frac{\mathbf{x}}{\| \mathbf{x} \|}\right)$$

[1]The Rosenbrock function is usually used as a test function in optimisation. It has a shallow minimum inside a deeply curved valley.

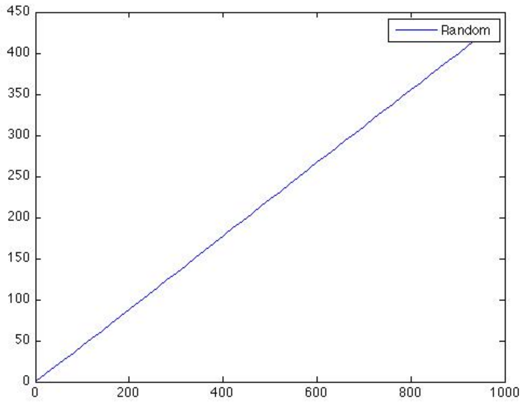


Figure 1: Regret of a random arm selection policy. The number of iterations is on the x-axis and the value of the regret is on the y-axis.

Arms are chosen to be the elements of a grid of $[-1;1]^2$ of size N. A Gaussian kernel in $\mathbb{R}^2$ is used between the elements of the grid to build the covariance matrix:

$$\kappa(\mathbf{X}_i, \mathbf{X}_j) = \exp(-\frac{||\mathbf{X}_i - \mathbf{X}_j||}{2\sigma^2})$$

In the following experiments there are $N = 81$ arms (corresponding to a $-1 : 0.25 : 1$ by $-1 : 0.25 : 1$ grid, in Matlab notation), the value of σ is set to 0.5, the added Gaussian noise has standard deviation $\sigma_{noise} = 0.3$. The estimated noise standard deviation in the GP model is taken to be the true noise standard deviation, but it could be chosen by cross validation.

3.2 Results

The regrets presented here were averaged over 100 runs of the algorithms. We experimented with three B functions ($B(t) = t$, $B(t) = \log t$ and $B(t) = 1$) and three S functions: $S(t) = t$ (Figure 2), $S(t) = N$ (Figure 3) and $S(t) = \sqrt{N}$ (Figure 4). In Figure 1 we show the regret of a random arm selection policy, for comparison with the regrets presented in the other Figures.

In particular, one can see in Figure 2 that between iterations 0 and 81 (initialisation phase), the regret of UCB is similar to the regret of the random policy, and becomes better afterwards as the algorithm starts using upper confidence bounds. We can clearly see that GPB with $B(t) = t$ does too much exploration and chooses arms quasi randomly. GPB with a smaller emphasis on exploration ($B(t) = \sqrt{\log t}$ and $B(t) = 1$), however, performs much better than UCB, and this from the very beginning. Its regret curves are also smoother than for UCB.

In Figure 3 we were able to show more iterations of the algorithms (1000) as we limited the size of GPB's training set. GPB-red(N) with $B(t) = \sqrt{\log t}$ and $B(t) = 1$ is still better than UCB, and the gap between the two variants of GPB-red(N) widens with time. However, we see in Figure 4 that further reducing the size of the training set to $\sqrt{N}$ (so that GPB-red uses as much information as UCB) does add exploration to the arm selection policy (see the remark at the end of Section 2.4), as GPB-red($\sqrt{N}$) with $B(t) = \sqrt{\log t}$ now tends to perform as a random selection policy. And, as expected, the regret of GPB with fixed $B(t)$ becomes worse when reducing the size of the training set (Figure 5).

Figures 6 and 7 show the influence of two parameters of the problem, σ_{noise} and N, on the regrets of the algorithms. Increasing σ_{noise} worsens the regrets of both UCB and GPB-red, while increasing the number of arms N worsens the regret of UCB and doesn't seem to affect GPB-red.

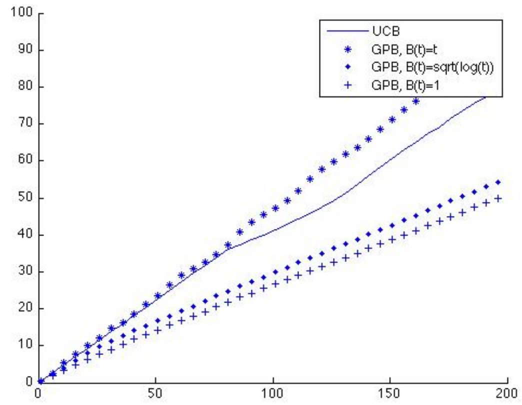


Figure 2: Comparison of the regrets of UCB and GPB.

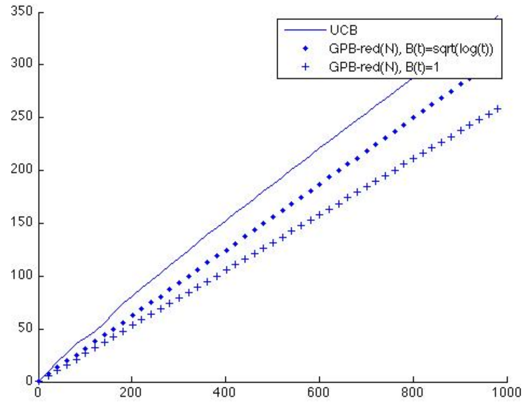


Figure 3: Comparison of the regrets of UCB and GPB-red(N).

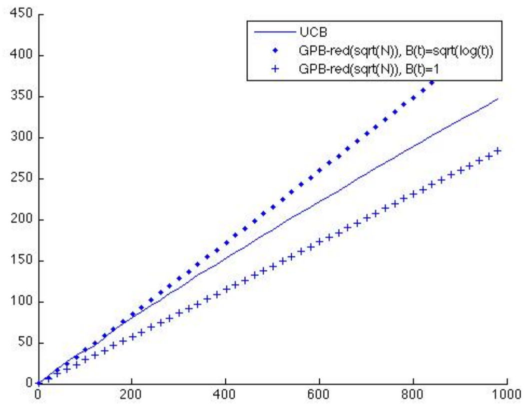


Figure 4: Comparison of the regrets of UCB and GPB-red($\sqrt{N}$).

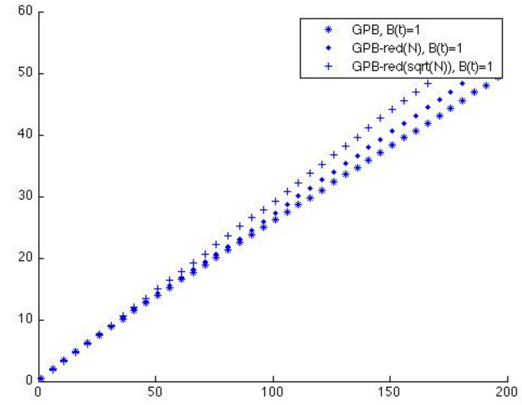


Figure 5: Comparison of the regrets of GPB-red($S(t)$) for different S functions and $B(t) = 1$.

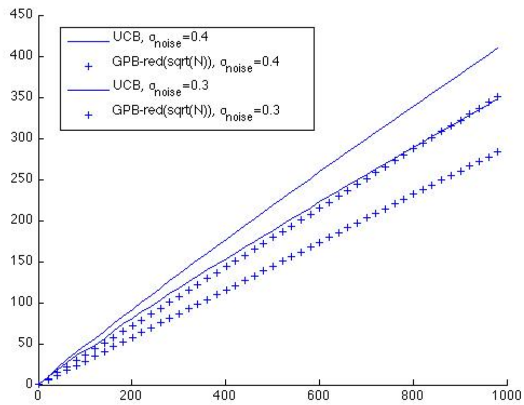


Figure 6: Regrets of UCB and GPB-red($\sqrt{N}$), $B(t) = 1$, with $\sigma_{noise} = 0.4$ (upper curves) and $\sigma_{noise} = 0.3$.

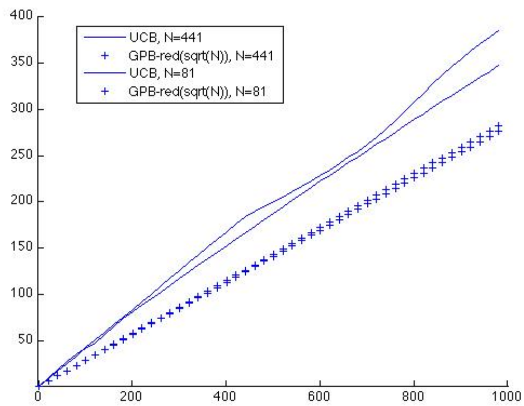


Figure 7: Regrets of UCB and GPB-red($\sqrt{N}$), $B(t) = 1$, with $N = 441$ (upper curves) and $N = 81$.

4 Conclusion

Exploiting arm dependencies allows GPB and GPB-red to have better regrets than UCB. Although GPB-red is more costly than UCB in the number of arms, using it would be beneficial in applications where real-world sample or real-world error costs are high enough to warrant the extra computational cost of exploiting arm dependencies.

One particular advantage of our approach is that we model dependencies with kernel functions. Consequently, we can consider applications where arms are characterised by sets of features, but also where arms represent objects such as text, images, music, or any structured data for which kernels exist. Thus, our framework can be applied to [6]'s web advertisement problem by defining a kernel between ads. Deciding which ads to put on which webpage is actually a content matching problem, and such problems can be modelled as multi-armed bandit problems with dependent arms. Product recommendation is also a content matching problem where we want to match elements from a set of products with elements of a set of consumers. Therefore our framework could be applied to product recommendation (for instance, if the product is music, we use a music kernel, etc.).

[5] have used UCB to devise the UCT (Upper Confidence Trees) tree search algorithm, which has been successfully applied to Go game playing ([3]). The UCT algorithm chooses paths to explore in the tree by sequentially going deeper in the tree and, at each node n, choosing the next node to go to among the children of n with the UCB formula. Thus, children nodes are considered as arms of a bandit problem, with $\hat{\mu}_t(\mathbf{x}_i)$ taken as the average of the rewards obtained by all tree paths containing node $\mathbf{x}_i$ (t is the number of iterations of the bandit problem associated to node n and is equal to the sum of the visits of all children of n). In the case of Go, nodes are labelled with Go boards and the reward of a game tree path is given by a Monte-Carlo simulation starting from the deepest node in the path: 1 if at the end of the simulation the game is won, 0 otherwise. It would be particularly interesting to see whether UCT for Go can be improved by using GPB with a Go board kernel instead of UCB. Indeed, arms are not independent since similar Go boards are likely to lead to similar rewards.

Finally, in future work we will continue with the theoretical analysis of GPB and GPB-red in order to upper-bound their expected regrets and to fix $B(t)$.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

[2] Sebastien Bubeck, Remi Munos, Gilles Stoltz, and Csaba Szepesvari. Online optimization in x-armed bandits. In *Proceedings of NIPS*, 2008.

[3] Sylvain Gelly and Yizao Wang. Exploration exploitation in go: Uct for monte-carlo go. In *Proceedings of NIPS*, 2006.

[4] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of STOC*, 2008.

[5] Levente Kocsis and Csaba Szepesvari. Bandit based monte-carlo planning. In *Proceedings of ECML*, 2006.

[6] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. Multi-armed bandit problems with dependent arms. In *Proceedings of ICML*, 2007.

[7] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

HYPERGRAPHS AND SATISFIABILITY

Dušan Hvalica
University of Ljubljana, Faculty of Economics
dusan.hvalica@ef.uni-lj.si

Abstract

A class of hypergraphs, called supporting bundles, is introduced, such that SAT translates into checking on the existence of a supporting bundle in the corresponding hypergraph.

Keywords: satisfiability, hypergraph, supporting bundle

1 Introduction

Hypergraphs have already been utilized to address the satisfiability problem. Thus, resolution has been extended to undirected hypergraphs [1, 5, 2], while in the context of directed hypergraphs it has been shown that testing of satisfiability can be reduced to searching for a zero-cardinality cut [3].

Here, another approach for translating SAT into the context of directed hypergraphs is presented. A class of hypergraphs, called supporting bundles, is introduced; it is shown that testing satisfiability of a set of clauses reduces to checking on the existence of a supporting bundle in the corresponding hypergraph.

2 Definitions and notation

A *hypergraph* G is defined as $G = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ and $\mathcal{A}$ are the sets of *nodes* and *hyperarcs*, respectively. A hyperarc $\mathcal{E}$ is defined as $\mathcal{E} = (T(\mathcal{E}), H(\mathcal{E}))$, where $T(\mathcal{E}), H(\mathcal{E}) \subset \mathcal{V}$; the sets $T(\mathcal{E})$ and $H(\mathcal{E})$ are called the *tail* and *head* of $\mathcal{E}$, respectively. A hyperarc, which head has (only) one element, is called a *B-arc* (backward (hyper)arc). A hypergraph, the hyperarcs of which are all B-arcs, is a *B-graph.*

A *subhypergraph* of a hypergraph $G = (\mathcal{V}, \mathcal{A})$ is such a hypergraph $G_1 = (\mathcal{V}_1, \mathcal{A}_1)$ that $\mathcal{V}_1 \subset \mathcal{V}$ and $\mathcal{A}_1 \subset \mathcal{A}$. When convenient, we shall denote $\mathcal{V}_1$ by $\mathcal{V}(G_1)$ and $\mathcal{A}_1$ by $\mathcal{A}(G_1)$.

The *sum* of hypergraphs G_1 and G_2 is

$$G_1 + G_2 = (\mathcal{V}(G_1) \cup \mathcal{V}(G_2), \mathcal{A}(G_1) \cup \mathcal{A}(G_2))\,.$$

For any node u its *backward star* $BS(u)$ is defined by $BS(u) = \{\mathcal{E}\,; u \in H(\mathcal{E})\}$, while its *forward star* $FS(u)$ is $FS(u) = \{\mathcal{E}\,; u \in T(\mathcal{E})\}$. A node u for which $BS(u) = \varnothing$ or $FS(u) = \varnothing$ will be called a *tip node*.

For any subhypergraph $H \subset G$ the set of its nodes v such that $BS(v) \cap \mathcal{A}(H) = \varnothing$ will be denoted by $B(H)$ while the set of its nodes v such that $FS(v) \cap \mathcal{A}(H) = \varnothing$ will be denoted by $F(H)$.

For any hypergraph $G = (\mathcal{V}, \mathcal{A})$ and hyperarc $\mathcal{E} \in \mathcal{A}$, a truth assignment $f : \mathcal{V} \to \{true, false\}$ is *incoherent* at $\mathcal{E}$ if $f(v) = true$ for every $v \in T(\mathcal{E})$ and $f(v) = false$ for every $v \in H(\mathcal{E})$. If this is not the case, f is *coherent* at $\mathcal{E}$. We shall simply say that f is *incoherent* if it is incoherent at least at one $\mathcal{E} \in \mathcal{A}$ and that it is *coherent* otherwise. Moreover, we shall say that f is *tight* at a tip node v if $f(v) = true$ for $v \in B(G)$ and $f(v) = false$ for $v \in F(G)$.

Clearly a truth assignment f is coherent at a hyperarc $\mathcal{E}$ if and only if the following applies:

- if $f(u) = true$ for every $u \in T(\mathcal{E})$ and $f(u) = false$ for every $u \in H(\mathcal{E}) \setminus \{v\}$, then $f(v) = true$,
- if $f(u) = false$ for every $u \in H(\mathcal{E})$ and $f(u) = true$ for every $u \in T(\mathcal{E}) \setminus \{v\}$, then $f(v) = false$.

Furthermore, one easily verifies the following lemma:

Lemma 1 *For every hypergraph G and any truth assignment f on G the following conditions are equivalent:*

1. *If f is tight at all tip nodes of G then f is incoherent.*
2. *If f is tight at all tip nodes of G with the exception of v and if, moreover, f is coherent, then $f(v) = true$ for $v \in F(G)$ and $f(v) = false$ for $v \in B(G)$.*

Let A be a set of propositional variables and B a set of clauses over A, i.e., of formulae of the form:

$$C_1 \wedge \cdots \wedge C_m \Rightarrow D_1 \vee \cdots \vee D_n\,, \tag{1}$$

where $C_1, \ldots, C_m, D_1, \ldots, D_n$ belong to $A \cup \{true, false\}$. We say that B is *satisfiable* if there exists a truth assignment $A \to \{true, false\}$ such that every clause in B is $true$.

The problem of determining whether a given set of clauses is satisfiable or not is called the *satisfiability problem*, SAT; it is known to be $\mathcal{NP}$-complete.

To every instance of SAT a hypergraph can be assigned in the following way: its nodes are the propositions, while its hyperarcs correspond to clauses — the matching hyperarc to the clause (1) is $(\{C_1, \ldots, C_m\}, \{D_1, \ldots, D_n\})$.

3 Tracks

A *track* is a sequence

$$u_1, \mathcal{E}_1, u_2, \mathcal{E}_2, \ldots, \mathcal{E}_{q-1}, u_q \tag{2}$$

such that

- for any triple $u_i, \mathcal{E}_i, u_{i+1}$ on the track we have $u_i \neq u_{i+1}$ and $u_i, u_{i+1} \in T(\mathcal{E}_i) \cup H(\mathcal{E}_i)$,
- for any triple $\mathcal{E}_i, u_{i+1}, \mathcal{E}_{i+1}$ on the track we have $\mathcal{E}_i \neq \mathcal{E}_{i+1}$ and one of the following holds:
  - $u_{i+1} \in H(\mathcal{E}_i)$ and $u_{i+1} \in T(\mathcal{E}_{i+1})$,
  - $u_{i+1} \in T(\mathcal{E}_i)$ and $u_{i+1} \in H(\mathcal{E}_{i+1})$.

If $u_{q-1}, \mathcal{E}_{q-1}, u_q, \mathcal{E}_1, u_2$ is a track, then track (2) is called a *hypercycle* (of course, in such case either $u_q \in T(\mathcal{E}_1)$ or $u_q \in H(\mathcal{E}_1)$; often we have $u_q = u_1$).

Clearly every path is a track and every cycle is a hypercycle and clearly the reverse is not true. One might say that at nodes a track is like a path or a "reversed" path while at hyperarcs it may be like a path or a reversed path but it may also "bounce", i.e., change from a path into a reversed path or vice versa.

For instance, $u_1, \mathcal{E}_1, u_4, \mathcal{E}_2, u_5, \mathcal{E}_4, u_6, \mathcal{E}_3, u_4, \mathcal{E}_5, u_3$ is a track in the hypergraph in Figure 1.

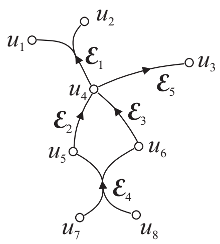


Figure 1

A track is *simple* if every element in the sequence appears only once. A hypercycle is *simple* if it is a simple track or if the first and the last node coincide while all other elements of the sequence appear only once.

If $T = u_1, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}, u_i$ and $T' = u_i, \mathcal{E}'_1, \ldots, \mathcal{E}'_{j-1}, u'_j$ are tracks and if, moreover, $\mathcal{E}_{i-1}$ and $\mathcal{E}'_1$ do not both belong to $BS(u_i)$ or to $FS(u_i)$, we shall denote

$$TT' = u_1, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}, u_i, \mathcal{E}'_1, \ldots, \mathcal{E}'_{j-1}, u'_j.$$

Clearly TT' is also a track.

For any track $T = u_1, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}, u_i$ its *reverse track* is $T^{-1} = u_i, \mathcal{E}_{i-1}, \ldots, \mathcal{E}_1, u_1$. Clearly we have $(T_1T_2)^{-1} = T_2^{-1}T_1^{-1}$.

Track T_1 is a *subtrack* of a track T, if it can be obtained by selecting two nodes on T and then removing all elements before the first node and all elements after the second node from T. If track T_1 is a subtrack of a track T, then T is an *extension* of T_1. Such a situation will be denoted by $T_1 \subset_s T$.

For any tracks $T = u_1, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}, u_i$ and $T' = u'_i, \mathcal{E}'_1, \ldots, \mathcal{E}'_{j-1}, u'_j$ and any hyperarc $\mathcal{E}$, such that $u_i, u'_i \in H(\mathcal{E}) \cup T(\mathcal{E})$ and that $\mathcal{E}$ and $\mathcal{E}_{i-1}$, resp. $\mathcal{E}$ and $\mathcal{E}'_1$, do not both belong to $BS(u_i)$ or to $FS(u_i)$, we shall denote

$$T\mathcal{E}T' = u_1, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}, u_i, \mathcal{E}, u'_i, \mathcal{E}'_1, \ldots, \mathcal{E}'_{j-1}, u'_j.$$

Clearly, $T\mathcal{E}T'$ is a track as well.

Any track $x, \mathcal{E}_1, \ldots, \mathcal{E}_2, x$ such that either $\mathcal{E}_1, \mathcal{E}_2 \in BS(x)$ or $\mathcal{E}_1, \mathcal{E}_2 \in FS(x)$ will be called a *return twist*. Clearly, if T is a return twist then T^{-1} is also a return twist. If T_1 is a return twist and $T = T'T_1$ is a track, we shall say that T *ends with a return twist*. If $T = R_1T'R_2$, where R_1 and R_2 are return twists, we say that T has return twists at both ends. It can be shown [4] that a formula is Horn renamable (for the definition of this class of formulae, see [6, 7]) if and only if the hypergraph, corresponding to it, does not contain any tracks with return twists at both ends.

4 Supporting bundles

Define *quasi supporting bundles* and *quasi opposing bundles* of nodes in a hypergraph by the following:

1. A hypergraph P of the form $P = (\{v\}, \varnothing\}$ or $P = (H(\mathcal{E}) \cup T(\mathcal{E}), \{\mathcal{E}\})$ is a quasi supporting bundle of any $u \in F(P)$ and a quasi opposing bundle of any $u \in B(P)$.

2. Let E be a quasi supporting bundle of u. For every $v \in B(E)$ let P_v be a quasi supporting bundle of v. For every $v' \in F(E) \setminus \{u\}$ let $P_{v'}$ be a quasi

opposing bundle of v'. Then $P = \left(\sum_{v\in B(E)} P_v\right) + E + \left(\sum_{v'\in F(E)\setminus\{u\}} P_{v'}\right)$ is a quasi supporting bundle of u, provided that

(a) $FS(u) \cap \mathcal{A}(P) = \varnothing$,

(b) every tip node of P_v or $P_{v'}$ (but v and v') is a tip node of P.

Similarly, $P = \left(\sum_{v\in B(E)\setminus\{u\}} P_v\right) + E + \left(\sum_{v'\in F(E)} P_{v'}\right)$, where E is a quasi opposing bundle of u, $BS(u) \cap \mathcal{A}(P) = \varnothing$, while P_v and $P_{v'}$ are as above, is a quasi opposing bundle of u.

3. Every quasi supporting bundle and every quasi opposing bundle are obtained in this way.

A quasi supporting (resp. opposing) bundle P of u is a *supporting* (resp. *opposing*) *bundle* of u if every tip node of P (except possibly u) is a tip node of the whole hypergraph. Clearly, a subhypergraph G_1 is a quasi supporting (opposing) bundle if it is a supporting (opposing) bundle as a subhypergraph of G_1.

Note that, since for every $v \in \mathcal{V}$, $(\{v\}, \varnothing)$ is a quasi supporting and a quasi opposing hypergraph of v, if for some $A_1 \subset B(E)$ and $A_2 \subset F(E) \setminus \{u\}$, quasi supporting bundles P_v, $v \in A_1$, and quasi opposing bundles $P_{v'}$, $v' \in A_2$, satisfy condition 2.b., then $Q = \left(\sum_{v\in A_1} P_v\right) + E + \left(\sum_{v'\in A_2} P_{v'}\right)$ is a quasi supporting bundle of u, provided that the tip nodes of E which do not belong to $A_1 \cup A_2 \cup \{u\}$ are also tip nodes of Q (and condition 2.a. is satisfied); analogously applies to quasi opposing bundles.

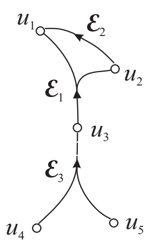


Figure 2a

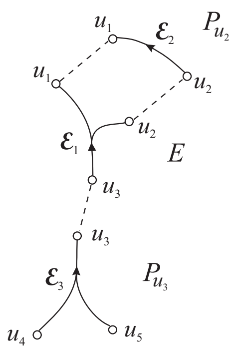


Figure 2b

The hypergraph in Fig. 2a is equal to $E + P_{u_2} + P_{u_3}$, where E, P_{u_2} and P_{u_3} are hypergraphs whose only hyperarcs are $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$, respectively (cf. Fig. 2b), so that E is a quasi supporting bundle of u_1, P_{u_3} is a supporting bundle of u_3, while P_{u_2} is an opposing bundle of u_2. Since all tip nodes of P_{u_2} and P_{u_3} are also tip nodes in the entire hypergraph, the latter is a supporting bundle of u_1.

Every quasi supporting or opposing bundle has the properties from Lemma 1:

Proposition 2 *Let P be a quasi supporting or quasi opposing bundle; if f is a truth assignment on P which is tight at every tip node of P, then f is incoherent.*

Proof. The assertion clearly applies whenever P has only one hyperarc, so let P be a quasi supporting bundle of u of the form $P = \left(\sum_{v \in B(E)} P_v\right) + E + \left(\sum_{v' \in F(E) \setminus \{u\}} P_{v'}\right)$ and suppose that the assertion applies to all proper subhypergraphs of P. Suppose that f is coherent, then it is coherent on every P_v and $P_{v'}$ as well. As the assertion is valid for every P_v and $P_{v'}$, by Lemma 1, we have $f(v) = true$ for every $v \in B(E)$ and $f(v') = false$ for every $v' \in F(E) \setminus \{u\}$, which, since $f(u) = false$ as well, by our assumption implies that f is incoherent on E, so that the assertion is valid for P as well. The same applies when P is a quasi opposing bundle of the form $P = \left(\sum_{v \in B(E) \setminus \{u\}} P_v\right) + E + \left(\sum_{v' \in F(E)} P_{v'}\right)$. Thus, the proposition follows by induction. ■

By Lemma 1, this is equivalent to:

Corollary 3 *Let P be a quasi supporting bundle of u and f a coherent truth assignment on P that is tight at every tip node of P with the exception of u. Then $f(u) = true$.*

Similarly, let P be a quasi opposing bundle of u and f a coherent truth assignment on P that is tight at every tip node of P with the exception of u. Then $f(u) = false$.

5 Satisfiability

In this section we shall assume that in our hypergraph G there are two distinguished nodes u_T (with $BS(u_T) = \varnothing$) and u_F (with $FS(u_F) = \varnothing$). Furthermore, every hyperarc of the form $(A, \varnothing)$ will be replaced by $(A, \{e_F\})$ and every $(\varnothing, B)$ will be replaced by $(\{e_T\}, B)$. Finally, we shall assume that for every truth assignment f we have $f(u_T) =$ true and $f(u_F) =$ false. Thus, for instance, f will be coherent at $(\{e_T\}, B)$ iff $f(v) =$ true for at least one $v \in B$.

It is an immediate consequence of Proposition 2 that if in G there exists a supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$, then every truth assignment on G is incoherent. It can be shown that the converse is also true:

Proposition 4 *If in G there exists no supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$, then there exists a coherent truth assignment on G.*

(Owing to space limitations we omit the proof.)

Thus, the following applies:

Theorem 5 *In G there exists a supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$, if and only if every truth assignment on G is incoherent.*

Proof. ($\Rightarrow$) If such a supporting bundle exists, by Proposition 2 no truth assignment can be coherent.

($\Leftarrow$) If there is no such supporting bundle, then, by Proposition 4, there exists a truth assignment, which is coherent on G. ■

In Section 2 we showed that to every instance of SAT a hypergraph can be assigned. Clearly for any truth assignment a clause is false if and only if that truth assignment is incoherent at the corresponding hyperarc of the hypergraph. Hence a set of clauses is satisfiable if and only if on the corresponding hypergraph there exists a coherent truth assignment. By Theorem 5, a set of clauses is therefore satisfiable if and only if in the corresponding hypergraph there exists no supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$. Thus, checking the satisfiability of a set of clauses translates into searching for a supporting bundle in a hypergraph. Since this translation can clearly be made in polynomial time, it follows that the problem of finding a supporting bundle of a node in a hypergraph is $\mathcal{NP}$-hard.

A set of clauses is *minimal unsatisfiable* if it is unsatisfiable but removing any of its clauses makes it satisfiable. Of course, minimal unsatisfiability of a set of clauses affects also the associated hypergraph.

Let S be any set of clauses and let G_S be the corresponding hypergraph. If S is unsatisfiable then in G_S there exists a supporting bundle P_S of u_F. If S is minimal unsatisfiable, then for any subset $S' = S \setminus \{c\}$ the corresponding hypergraph $G_{S'}$ contains no supporting bundle of u_F, which is only possible if $\mathcal{C} \in \mathcal{A}(P_S)$ (where $\mathcal{C}$ is the hyperarc that corresponds to c). It follows that $G_S = P_S$, the hypergraph G_S is a supporting bundle of u_F and no proper subhypergraph of G_S is a supporting bundle of u_F. Thus, G_S is a minimal supporting bundle of u_F. If G_S is a supporting bundle of u_F, then the corresponding set of clauses S is unsatisfiable. If G_S is minimal, removing any of its hyperarcs yields a hypergraph that contains no supporting bundle of u_F so that the corresponding set of clauses must be satisfiable. Hence, S is minimal unsatisfiable.

Thus, the following applies:

Proposition 6 *A set of clauses is minimal unsatisfiable iff the corresponding hypergraph is a minimal supporting bundle of u_F whose only tip nodes are u_T and u_F.*

6 Conclusion

We have introduced the concept of supporting and opposing bundles. We have demonstrated that SAT translates into checking on the existence of a supporting bundle. Thus, our result can be applied to solve SAT instances. An algorithm to this end is described in another paper.

References

[1] Cowen, R. H., 1991. Hypergraph satisfiability. Reports on Math. Logic, 25, pp. 113-117.

[2] Cowen, R. H., 2001. Property S. Reports on Math. Logic, 35, pp. 61-74.

[3] Gallo, G., Longo, G., Pallottino, S., Nguyen, S., 1993. Directed Hypergraphs and applications, Discrete Applied Mathematics, 42, pp. 177-201.

[4] Hvalica, D., 2007. Horn renamability and B-graphs. In: Zadnik Stirn, L., Drobne, S. (eds.), Proceedings of the 9th International Symposium on Operational Research, Slovenian Society Informatika, Nova Gorica, pp. 57-162.

[5] Kolany, A., 1993. Satisfiability on hypergraphs. Studia Logica, 52, pp. 393-404.

[6] Lewis, H. R., 1978. Renaming a set of clauses as a Horn set. Journal of the ACM 25, pp. 134-135.

[7] Val, A. del, 2000. On 2-SAT and Renamable Horn. In: Proceedings of AAAI/IAAI 2000, pp. 279-284.

SATISFIABILITY TESTING IN THE CONTEXT OF HYPERGRAPHS

Dušan Hvalica
University of Ljubljana, Faculty of Economics
dusan.hvalica@ef.uni-lj.si

Abstract

An algorithm for searching for supporting bundles is described, which allows for solving instances of SAT in the context of hypergraphs.

Keywords: satisfiability, hypergraph, supporting bundle

1 Introduction

The satisfiability problem has already been addressed in the context of hypergraphs. Thus, resolution has been extended to undirected hypergraphs [2, 8, 3], while in the context of directed hypergraphs it has been shown that testing of satisfiability can be reduced to searching for a zero-cardinality cut [7]. Considerable attention has also been spent on B-graphs, a subclass of hypergraphs, partly because they correspond to important subclass of SAT, the Horn-SAT, and partly because they proved to be a valuable tool in other fields — in AI and OR [6, 5].

In another contribution ("Hypergraphs and satisfiability") we introduce the concept of supporting bundles and show that testing satisfiability of a set of clauses reduces to checking on the existence of a supporting bundle in the corresponding hypergraph. Here, an algorithm for this checking is described and, based on its properties, some issues which may be advantageous with SAT are proposed.

2 Supporting bundles and satisfiability

The *satisfiability problem*, SAT — the problem of determining whether a given set of clauses is satisfiable or not, can be translated into the context of hypergraph (for definitions, see "Hypergraphs and satisfiability") by representing the clauses by hyperarcs, specifically, the clause

$$C_1 \wedge \cdots \wedge C_m \Rightarrow D_1 \vee \cdots \vee D_n \tag{1}$$

is represented by the hyperarc $(\{C_1, \ldots, C_m\}, \{D_1, \ldots, D_n\})$. The fact that for some truth assignment a clause is true is reflected in the fact that this truth assignment is coherent at the corresponding hyperarc (for details, see "Hypergraphs and satisfiability").

Supporting resp. opposing bundles are a variety of hypergrahs, defined by induction. If we assume that in our hypergraph G there are two distinguished nodes u_T (with $BS(u_T) = \varnothing$) and u_F (with $FS(u_F) = \varnothing$), that every hyperarc of the form $(A, \varnothing)$ is replaced by $(A, \{e_F\})$ and every $(\varnothing, B)$ by $(\{e_T\}, B)$ and finally, that for every truth assignment f we have $f(u_T) =$ true and $f(u_F) =$ false (clearly, this entails no loss of generality), then the following applies (cf. "Hypergraphs and satisfiability"):

Theorem 1 *In G there exists a supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$, if and only if every truth assignment on G is incoherent.*

Since a set of clauses is satisfiable if and only if there exists a coherent truth assignment on the corresponding hypergraph, it follows that a set of clauses is satisfiable if and only if in the corresponding hypergraph there exists no supporting bundle P of u_F, such that $B(P) = \{u_T\}$ and $F(P) = \{u_F\}$. Hence, checking the satisfiability of a set of clauses can be done by searching for a supporting bundle in a hypergraph. Since this translation can clearly be made in polynomial time, it follows that the problem of finding a supporting bundle of a node in a hypergraph is $\mathcal{NP}$-hard.

3 Searching for a supporting bundle

To construct a supporting bundle of u, it suffices to find, for some $\mathcal{E} \in BS(u)$, for every $v \in T(\mathcal{E})$ a supporting bundle P_v and for every $v' \in H(\mathcal{E}) \setminus \{u\}$ an opposing bundle $P_{v'}$ such that $\mathcal{A}(P_v) \cap FS(u) = \varnothing$ and $\mathcal{A}(P_{v'}) \cap FS(u) = \varnothing$ and that (with a possible exception of u) the tip nodes of P_v and $P_{v'}$ (besides v, resp. v') are also tip nodes of G. Thus, supporting and opposing bundles can be constructed recursively. This can be done in both directions, from the node whose supporting resp. opposing bundle we are looking for to the tip nodes and in the opposite direction. However, the latter may not be always successful. For instance, in the hypergraph in Fig. 2a it works fine: first opposing bundles of u_6 and u_7 are found, then opposing bundle of u_8 and then first the supporting bundle of u_4, then of u_2

and u_3 and finally of u_1. On the other hand, in the hypergraph in Fig. 2b, neither supporting nor opposing bundles can be established, although a supporting bundle of u_1 does exist.

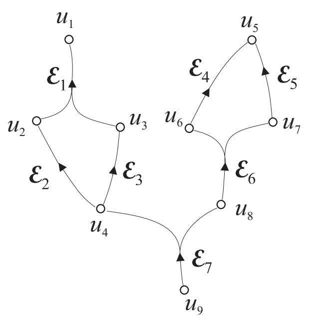


Figure 2a

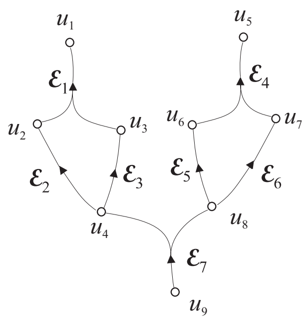


Figure 2b

Alternatively, the search can be started at the node whose supporting resp. opposing bundle we are looking for. Here, the currently constructed quasi supporting (opposing) bundle is successively extended towards tip nodes of the hypergraph G.

Consider the following procedure:

Procedure Bundle(T, ok);
begin
if T ends in a tip node or with a return twist **then**
    set $ok \leftarrow$ true and **exit**;
$ok \leftarrow$ false;
if T ends with a hypercycle then **exit**;
denote the last node on T by x;
for every hyperarc $\mathcal{E}$ such that $T_1 = x, \mathcal{E}, \ldots$ and TT_1 are tracks **do**:
    begin
    for every t such that $T_t = x, \mathcal{E}, t$ is a track **do**:
        begin
        Bundle(TT_t, ok);
        if not ok **then goto** 1;
        end;
1:     **if** ok **then**
        mark $\mathcal{E}$ and **goto** 2;
    end;
2: **end**;

Its input is a track T in the hypergraph G; the procedure constructs its extensions until they end at a tip node of G or become of the form $T = T_1T_2$, where T_1 is simple and T_2 is an elementary return twist. This is done in a depth-first manner and the tracks ending in a hypercycle are rejected, along with the tracks that split from them at a hyperarc. If extensions of T of the form $\ldots, x, \mathcal{E}, t, \ldots$, for every $t \in T(\mathcal{E}) \cup H(\mathcal{E}) \setminus \{x\}$, are successfully established, the hyperarc $\mathcal{E}$ is marked. The following applies (due to space limitations we omit the proof):

Lemma 2 *Let $T = u, \ldots, \mathcal{E}_0, x$ be a simple track in G and $\mathcal{E}_0 \in FS(x)$ (resp. $\mathcal{E}_0 \in BS(x)$). If* Bundle(T, ok) *terminates with $ok =$* true *then in G there exists a quasi supporting (resp. opposing) bundle P of x whose tip nodes, that are not also tip nodes of G, lie on T.*

P is contained in the subhypergraph G_1 of G for which $\mathcal{A}(G_1)$ consists of the hyperarcs marked by the procedure. Moreover, if T_1 is a simple track in P then TT_1 contains no hypercycle.

Now, consider also the following:

```
Procedure Supporting_bundle(u, ok);
begin
ok ← false;
for every hyperarc ℰ ∈BS(u) do:
    begin
    for every t such that T_t = u, ℰ, t is a track do:
        begin
        Bundle(T_t, ok);
        if not ok then goto 1;
        end;
1:   if ok then
        mark ℰ and goto 2;
    end;
2: end;
```

Clearly, when Supporting_bundle(u, ok) terminates with $ok = true$, by Lemma 2, there exists some $\mathcal{E} \in BS(u)$, such that there exist for every $v \in T(\mathcal{E})$ a quasi supporting bundle P_v and for every $v' \in H(\mathcal{E}) \setminus \{u\}$ a quasi opposing bundle $P_{v'}$ such that their only tip node (apart from v resp. v') that is not a tip node of G may be u. It follows that $\left(\sum_{v \in B(E)} P_v\right) + E + \left(\sum_{v' \in F(E) \setminus \{u\}} P_{v'}\right)$, where E is the subhypergraph for which $\mathcal{A}(E) = \{\mathcal{E}\}$, is a supporting bundle of u and clearly it is contained in the subhypergraph G_1 for which $\mathcal{A}(G_1)$ is the set of the hyperarcs, marked by the algorithm.

Analogously one defines procedure Opposing_bundle(u, ok), except that here hyperarcs $\mathcal{E} \in FS(u)$ are used.

The following applies:

Lemma 3 *Let G be a supporting bundle of u. Then* Supporting_bundle(u, ok) *terminates with ok =* true*.*

If G is an opposing bundle of u, then Opposing_bundle(u, ok) *terminates with ok =* true*.*

Proof. By induction. When $\mathcal{A}(G) = \{\mathcal{E}\}$, the lemma clearly applies. Otherwise, let $G = \left(\sum_{v\in B(E)} P_v\right) + E + \left(\sum_{v'\in F(E)\setminus\{u\}} P_{v'}\right)$, where E is a quasi supporting bundle of u, while P_v and $P_{v'}$ are quasi supporting resp. opposing bundles of v resp. v' whose tip nodes (with the exception of v resp. v') are also tip nodes of G. Suppose that the assertion applies to all proper subhypergraphs of G. Then, for every $v \in B(E)$, procedure Supporting_bundle(v, ok) terminates with ok = true and for every $v' \in F(E)\setminus\{u\}$, procedure Opposing_bundle(v', ok) terminates with ok = true. Now, consider any tracks T in E and T_1 in P_v, such that

- TT_1 is a track and
- T_1 ends at a tip node of P_v or with a return twist.

Then TT_1

- ends at a tip node of G or with a return twist or
- contains a subtrack TT_1' which ends with a return twist or
- is of the form $T'T''T_1'T_1''$ such that $T''T_1'$ is a hypercycle and $T'T_1''$ ends at a tip node of G or with a return twist.

Thus, if Bundle(T_1, ok), executed in P_v, terminates with ok = true, the same is true for Bundle(TT_1, ok) (or Bundle(TT_1', ok), resp. Bundle($T'T_1'', ok$)), executed in G. It follows that Bundle(T, ok), executed in E, produces the same value of ok as when executed in G and, as Supporting_bundle(u, ok) executed in E terminates with ok = true, the same is true also when executed in G. Thus, the assertion of the Lemma applies to G (the case when G is an opposing bundle can be verified analogously), so that, by induction, the assertion always applies. ■

Suppose that G_1 is a subhypergraph of G_2 and that procedure Supporting_bundle(u, ok), executed in G_1, terminates with ok = true; then, clearly, the same applies to Supporting_bundle(u, ok), executed in G_2.

Suppose now that a supporting bundle P of u exists and that procedure Supporting_bundle(u, ok) is executed in P. Since, by Lemma 3, it terminates with ok = true, the same is true also, when it is executed in G. Thus, existence of a supporting bundle of u is a necessary and sufficient condition for Supporting_bundle(u, ok) to terminate with ok = true.

Proposition 4 *Procedure* Supporting_bundle(u, ok) *terminates with* ok = true *if and only if there exists a supporting bundle of* u*. This supporting bundle is contained in the subhypergraph* G_1 *of* G *for which* $\mathcal{A}(G_1)$ *consists of the hyperarcs marked by the procedure.*

The efficiency of this algorithms can be improved. By Lemma 2, if x and $T = u, \ldots, x$ are such that, when Bundle(T, ok) terminates with $ok = true$, no tip node of the quasi supporting (opposing) bundle P of x that it has found, lies on T, then P is a supporting (opposing) bundle of x. Thus, it can be freely reused for the construction of other quasi supporting (opposing) bundles. If such nodes are labelled, then if algorithm revisits such node by some track T', the execution of Bundle(T', ok) is not necessary.

On the other hand, for the nodes that are interior to some return twist, the execution of Bundle(T, ok) for every track T ending at such a node can not be avoided (a quasi supporting (opposing) bundle of t with tip nodes on T may not be of any use when t is reached by some other track T'). It follows that the topology of return twists strongly influences the computational complexity of searching for a supporting (opposing) bundle.

4 Conclusion

We have presented an algorithm for checking on the existence of a supporting (opposing) bundle which, since SAT translates into checking on the existence of a supporting bundle, enables one to solve instances of SAT also in the context of hypergraphs. Furthermore, we have demonstrated that the computational complexity of this checking is influenced by topology of return twists in the hypergraph. Thus, it seems that further study of the influence of the topology of return twists on the computational complexity of searching for a supporting (opposing) bundles may be advantageous with SAT. Since the SAT approach turned out successful also in some OR problems [1, 4, 9, 10], this issue may be of interest also for the OR community.

References

[1] Acharyya, S., Bagchi, A., 2008. A SAT approach for solving the staff transfer problem. Ao, S.I., Castillo, O., Douglas, C., Feng, D.D., Lee, J.A. (eds.), Lecture Notes in Engineering and Computer Science, Proceedings of International MultiConference of Engineers and Computer Scientists IMECS-2008, Hong Kong, pp. 64-68.

[2] Cowen, R. H., 1991. Hypergraph satisfiability. Reports on Math. Logic, 25, pp. 113-117.

[3] Cowen, R. H., 2001. Property S. Reports on Math. Logic, 35, pp. 61-74.

[4] Ernst, M., Millstein, T., Weld, D., 1997. Automatic SAT-compilation of planning problems. Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, pp. 1169-1176.

[5] Gallo, G., Pallottino, S., 1986. Shortest path methods: a unifying approach. Math. Programming Stud., 26, pp. 45-61.

[6] Gallo, G., Urbani, G., 1989. Algorithms for testing the satisfiability of propositional formulae. J. Logic Programming, 7, pp. 45-61.

[7] Gallo, G., Longo, G., Pallottino, S., Nguyen, S., 1993. Directed Hypergraphs and applications. Discrete Applied Mathematics, 42, pp. 177-201.

[8] Kolany, A., 1993. Satisfiability on hypergraphs. Studia Logica, 52, pp. 393-404.

[9] Kundu, S., Acharyya, S., 2008. A SAT approach for solving the nurse scheduling problem. In: TENCON 2008, IEEE Region 10 Conference, Tencon, pp.1-6.

[10] Memik, S.O., Fallah, F., 2002. Accelerated SAT-based scheduling of control/data flow graphs. In: Proc. Int'l Conf. on Computer Design, pp. 395-400.

[11] Schlipf, J.S., Annexstein, F., Franco, J., Swaminathan, R., 1995. On finding solutions for extended Horn formulas. Information Processing Letters, 54, pp. 133-137.

CONTINUOUS GNSS ORBIT CONSTRUCTION USING INTERPOLATION AND NEURAL NETWORK APPROXIMATION APPROACH

Polona Pavlovčič Prešeren, Oskar Sterle, Miran Kuhar, Bojan Stopar
University of Ljubljana, Slovenija
ppavlovc@fgg.uni-lj.si osterle@fgg.uni-lj.si mkuhar@fgg.uni-lj.si bstopar@fgg.uni-lj.si

Abstract: This paper presents a comparison between polynomial and trigonometric interpolations and novel approximation methods, based on neural network approach for the GNSS precise orbit when discrete data are given. The purpose is to approximate an unknown function from scattered samples without apriori knowledge of a satellite trajectory. The aim is to expose potential usage, advantages and shortcomings of different methods, that originate from the theoretical background of a satellite motion and also from data available in different forms. Simulation results and comparison of different approximation approaches are based on normalized root mean square error function. It is shown that, adopting different interpolation/approximation techniques, is possible to reduce data amount to be transmitted and to evaluate function in any data domain. Simulation studies proved that neural network function overcomes traditional polynomial and trigonometric interpolation in better performance near the end of the given data interval and in extrapolation.

Keywords: interpolation, approximation, neural network, GNSS precise orbit, time series.

1 INTRODUCTION

Function approximation from discrete time-series is among others used on consistently reduction of data amount. This is the main challenge of real time data flow optimization with the aim of transmitted data amount reduction and consequently data flow acceleration. Secondly function construction allows us further investigation of other relations, for example in case of position data first derivatives give us velocity and second acceleration valuation. One of main challenges of function approximation is in fast and non-complicated evaluation for any data set in the function domain, although the function derives from data with complex background theory. And in special cases function construction from discrete data set leads to potential of predicting values out of the function domain.

The process of function introduction by approximation is known also as curve fitting. Interpolation is a special case of curve fitting where function passes exactly through the known data. Such interpretation disables further improvement by least-least squares collocation [11], based on principle that function only approaches known data. Approximation and interpolation techniques are based on special but exact mathematical rules and are often ranked amoung traditional techniques of function construction. They are based on Weierstrass theorem, which explains that continuous real-valued function can be well approximated using the the sum norm if the polynomial order goes to infinity.

Artificial neural networks (ANNs) in contrary to well defined mathematicial formulations provide generic black-box of function representation. Their advantage is in capability of potential pattern finding where only origin data (inputs) and unknown function results (outputs) are available. For function approximation, *feedforward neural networks* have been widely used, since theoretical results have proven their potential usage for universal approximators [2],[5]. Among various ANN models, multilayer perceptron (MLP), along with the back-propagation training algorithm, is most commonly used in practical applications. But due to certain shortcomings, which tend to have origin in complex network architecture (many hidden layers), the training process is time consuming and in case of trapping into local instead of global minima often not successful.

Simplified only three-layer networks have been proven to perform better than MLP. The training of a three-layer radial basis function networks (RBFNs) [1] is simplified, faster and more efficient for function approximation from data with local variations and discontinuities. But since basis functions in the family of RBFNs are redundant and generally not orthogonal, RBFN function representation is not unique. If the family of basis functions for the RBFN are replaced by orthonormal basis, resulting wavelet neural network (WNN) provides an unique representation and preserves the advantages of RBFNs. The idea of combining both wavelets and neural networks has been proposed also due to the similarity between discrete inverse wavelet transform and three-layer neural network [18]. It has been shown that wavelet decomposition [3], [8], which could be used as an approximation tool, has similar structure to a single hidden-layer network [18].

We first discuss the differences and similarities among well known interpolations techniques and further among different neural network approximation methods. Next, we introduce differences from numerical aspect with the main stress on equally spaced GNSS precise ephemeris data for interpolation and extrapolation.

1.2 GNSS discrete ephemeris problem formulation

There are two different forms of data, needed for a GNSS satellite position computation. Satellites themselves supply the user with navigation message, which includes broadcast ephemerides. Various agencies distribute so called precise ephemerides via internet. In case of broadcast ephemerides a representation based on *Keplerian elements plus perturbations* is used [10]. Precise ephemerides are packaged as daily solutions (0:00 to 23:45 GPS Time) and contain X, Y, Z satellite positions and satellite clock rates at 15-minute time interval [17]. Since broadcast and precise ephemerides contain different data, each uses different orbit construction methods.

GPS broadcast ephemerides are attached to a simple continuous function method, proposed in [7]. Since there are effects from the accuracy of the broadcast orbit computation procedure, satellite orbit positions from only broadcast ephemerides are not accurate enough for precise applications. Parameters are updated every two hours, but the accuracy of satellite coordinates degrades over time and can reach several meters.

For higher accuracy needs, GNSS services freely distribute post-processed or predicted precise ephemerides in the SP3 format. In SP3 files positions and clock rates for all available GPS satellites are listed at 15-minutes intervals, so precise ephemerides are often termed tabular orbits. The discrete nature of precise ephemerides requires different continuous orbit reconstruction. Traditionally, continuous orbit could be acquired from tabular SP3 data using numerical integration to solve second order differential equation [6]. Since usually only positions in the SP3 files are available, the use of interpolation/approximation techniques remains.

2 INTERPOLATION TECHNIQUES

Interpolation is a specific case of curve fitting, where the function must go exactly through the known data. Among a variety of different types of interpolation techniques the application of each specific method depends on data form available. Polynomial interpolation is sensitive to the distribution of the points being interpolated. Arbitrary selection of points on an interpolating interval could yield to polynomials with better performance to the others. But in most cases we are limited to the certain data form, often to equally spaced data, introduced in tabular form (i.e. precise GNSS ephemeris data). Numerical results present several interpolation techniques. They give the same results in case of equally spaced data, if the polynomial term is equal to the number of data. On the other hand, different set of n data on the same interval results in better specific polynomial interpolation results (Chebyshev polynomial versus Lagrange/Neville polynomials).

One of the common problem of polynomial interpolation techniques, known also as Runge's phenomenon, is the error between interpolationg polynomial and the interpolated data grows rapidly near the end-points of the interval. The problem is distinct for interpolators (for example Lagrange, Neville) with requirement, that the polynomial terms have to be equal to the number of data points. To overcome the problem of large oscillations near the end of the interval, the interpolator is used only for central subinterval interpolation. For the entire interval there are several successive subinterval polynomial constructions that overlap in time. Because the evaluation of a single polynomial interpolation is time consuming, the solution is not optimal. Cubic spline interpolation overcomes the problem of several successive subinterval constructions by solving a tridiagonal system of equations. It means that interpolation polynomials are calculated over the entire interval at one time.

If data set to be interpolated tends to have periodic trend, best results are obtained using trigonometric (or Fourier) polynomial interpolation. Similar to Runge's phenomenon in polynomial interpolation, in Fourier series Gibss phenomenon exists. In case of equally spaced data restriction the trigonometric polynomial can be calculated using Fast Fourier Transform (FFT). In FFT we have to use all available data for the trigonometric polynomial formation, since the interpolation error depends on number od data.

2.1 Lagrange polynomial interpolation

Interpolating polynomial $p_n(t)$, acquired from $n+1$ equally spaced points on a given interval, which satisfies the condition [13]:

$$p_n(t_i) = f(t_i) \text{ for i=0,1,...,n} \tag{1}$$

and is defined from polynomial coefficients $l_i(t)$:

$$l_i(t) = \frac{\prod_{k=0}^{i-1}(t-t_k)}{\prod_{k=0}^{i-1}(t_i-t_k)} \cdot \frac{\prod_{k=i+1}^{n}(t-t_k)}{\prod_{k=i+1}^{n}(t_i-t_k)}, \tag{2}$$

is called Lagrange interpolation polynomial:

$$p_n(t) = \sum_{i=0}^{n} f(t_i)\, l_i(t). \tag{3}$$

Lagrange's interpolator construction requires equally spaced data, where the polynomial terms are equal to the number of points. To avoid large oscillations near the end of the interval, several successive subinterval polynomial contructions are performed for the entire interpolation interval.

2.2 Chebyshev polynomial interpolation

Chebyshev polynomial interpolation stands for one of the alternatives to the uniform sampling grid. But since Chebyshev polynomial interpolation is not the expression of Chebyshev polynomial basis, polynomial coefficients could be computed also using equally spaced data. Chebyshev interpolation can give results indentical to Lagrange's n-term polynomial, if the same n points were taken into interpolation. It means, Chebyshev interpolating results are the same to Lagrange's, as long as the data points are tabular (i.e. equally spaced) and the number of terms in the polynomial is equally to the number of data points [4]. The advantage of Chebyshev polynomial interpolation is that the interpolation error doesn't rapidly grow near the endpoints and a single interpolation function could be used for the whole intepolating region.

Chebyshev polynomial $T_n(t)$ of degree *n* on the interval [-1,1] is defined as [13]:

$$T_n(t) = \cos\big(n \cdot \arccos(t)\big). \tag{4}$$

Chebyshev polynomials of the first kind $T_n(t)$ are given by:

$$T_n(t) = \begin{cases} 1 & n = 0 \\ t & n = 1 \\ 2t \cdot T_{n-1}(t) - T_{n-2}(t) & n = 2,3,\ldots \end{cases} \tag{5}$$

and satisfy the three recurrence relation and are orthogonal on the interval [-1,1] with respect to the weight function $w(t) = \frac{1}{\sqrt{1-t^2}}$.

Choosing distinct interpolation points t_i, acquired from (4), at the Chebyshev points, i.e. roots t_i of $T_{n+1}(t)$, that are sparse in the middle and denser near the end of the interval [-1,1],

$$t_i = -\cos\left(\frac{\pi(2i+1)}{2(n+1)}\right) \tag{6}$$

leads to a low maximum interpolation error ([9], [14]). Chebyshev interpolating polynomial could be constructed also by special choice of equally spaced data. Continuous function approximation using a finite linear combination of Chebyshev polynomials is followed by [12]:

$$f(t) = \frac{1}{2}a_0 + \sum_{i=1}^{N} a_i T_i(t) \tag{7}$$

with coefficients a_i

$$a_i = \frac{2}{\pi}\int_{-1}^{1} \frac{f(t)T_i(t)}{\sqrt{1-t^2}}\,dt. \tag{8}$$

Coefficients can be computed numerically using *M* equally spaced data points

$$t_j = -1 + j\Delta t \quad (j = 1,2,\ldots M-1), \text{ where } \Delta t = 2/M - 1 \tag{9}$$

for example using composite midpoint rule, where:

$$a_i = \frac{2}{\pi}\sum_{j=0}^{M-2} \frac{\frac{1}{2} f\big(t_{j+1/2}\big) T_i\big(t_{j+1/2}\big)}{\sqrt{1 - t_{j+1/2}^2}}\Delta t, \tag{10}$$

but by avoiding the integrand evaluation at the endpoints t_0 and t_{M-1}. $f\big(t_{j+1/2}\big)$ could be approximated by a quadratic or cubic polynomial.

The midpoint rule could be used also on the two left- and right-most panels, but on the rest we can use fourth order Simpson's $\frac{1}{3}$ rule [12], yielding to the coefficients:

$$a_i = \frac{\Delta t}{\pi}\Big(4g_2 + \frac{2}{3}g_3 + \frac{8}{3}g_4 + \frac{4}{3}g_5 + \cdots + \frac{8}{3}g_{M-3} + \frac{2}{3}g_{M-2} + 4g_{M-1}\Big), \tag{11}$$

where:

$$g_j = \frac{f(t_j)T_i(t_j)}{\sqrt{1-t_j^2}}. \tag{12}$$

If the number of panels left for the Simpson's $\frac{1}{3}$ rule is not even, a trapezoidal rule is used over one of the panels; in equation (13) trapezoidal rule is used in the third from last panel:

$$a_i = \frac{\Delta t}{\pi}\left(4g_2 + \frac{2}{3}g_3 + \frac{8}{3}g_4 + \frac{4}{3}g_5 + \cdots + \frac{8}{3}g_{M-4} + \frac{5}{3}g_{M-3} + g_{M-2} + 4g_{M-1}\right). \tag{13}$$

2.3 Trigonometric interpolation

The trigonometric (or Fourier) polynomial interpolation is used when data for function construction have periodic trend. The trigonometric approach follows the idea that interpolating polynomial is defined over the interval [0, 2π] by:

$$p_n(t_i) = a_0 + \sum_{k=1}^{n}(a_k \cos(kt) + b_k \sin(kt)) \tag{14}$$

Because trigonometric polynomial uses different function construction as regards Lagrange polynomial, the he n-term polynomial application achieves better interpolation results than use of n-data points. When dealing with almost periodic data (satellite orbit during 24 hour period), the polynomial coefficients are iteratively acquired by least-square, minimizing the interpolated versus exact value. Using a larger known data set, the interpolation is more effective also near the end of the interval.

For discrete equidistant time series data, trigonometric polynomial could be contructed using Fast Fourier Transformation (FFT). In FFT the number of nodes limitation exists; the number of nodes is of the form: $n = 2^N$, the rest of the data are used for trigonometric polynomial testing.

3 ARTIFICIAL NEURAL NETWORKS

Artificial neural networks solve problems differently from conventional algorithmic methods. The algorithmic approach is based on a cognitive problem solving; the way the problem is solved must be known and stated in small unambiguous instructions. Neural networks learn by example and cannot be programmed to perform a specific task. They contain experimentally acquired knowledge and because they find out how to solve the problem by themselves, their operation can be often unpredictable.

Neural network research is based on the idea of *connection strength*, that is, how strongly one neuron influences those neurons connected to it. The second essential element of neural connectivity is the *inhibition/excitation* distinction. The neuron *activation* will either increase the firing rates of connected neurons, or decrease the rate, respectively. The neuron's response is called the *transfer function* and explains how a neuron's firing rate varies with the input it receives.

A neural network consists of a sequence of layers, which are composed by neurons, with patterned connection bounds. There are three components in neural network formation; input nodes (units) in the input layer (known data in vector **x**), one or more layers of hidden nodes (hidden layers) and a set of nodes (units) in the output layer (known data in vector **y**). The input-output data pairs contain information. Input values are supplied as initial activation values; connection strength, inhibition/excitation conditions and transfer functions evaluate how much of the activation value is passed on further through all of connected nodes of the network. Higher the number given to a node, more activation the node contains.

The knowledge, acquired from the learning process, is stored in network connections that are presented by weights. Neural network learning is actually adjusting of connection strengths (weights). Back-propagation, which is most widely used learning process, is based on iterative learning method. One back-propagation realization consists of the input, processing, comparing the neural network output with actual known values and eventually by adjusting connection weights (Fig. 1).

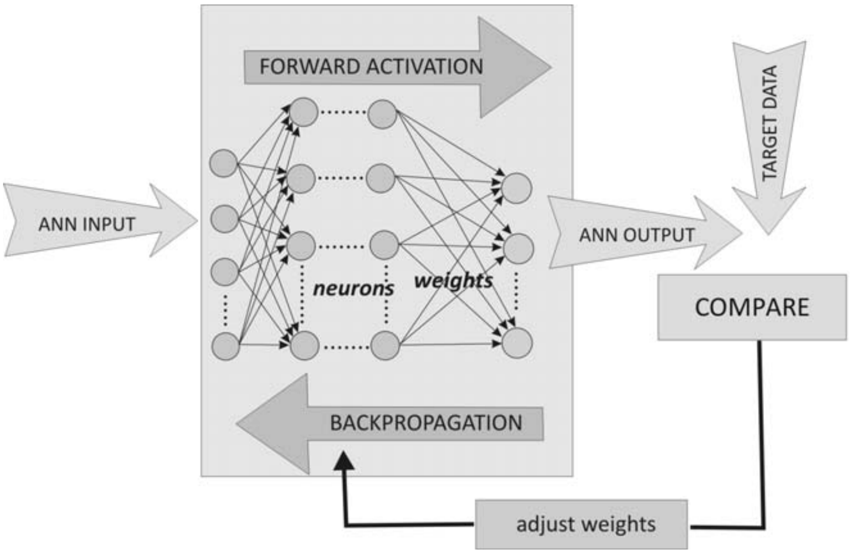


Figure 1: Feed-forward multi-layer neural network using back-propagation.

3.1 Multi-layer perceptron (MLP)

Most frequently used type of ANNs is the the multi-layer perceptron (MLP) [15], which consist of several hidden layers between the input and the output layer. In MLP each neuron i in the hidden layer k (i.e.: y_i^k) sums its input signals (the values of neurons in the *(k-1)-th* layer y_j^{k-1}) after multiplying them by connection weights w_{ij}^k and computes its output as using activation function *f*:

$$y_i^k = f\left(\sum_{j=1}^{n_{k-1}} w_{ij}^k\, y_j^{k-1}\right), \tag{15}$$

where i=1,…, n_k, and n_{k-1} number of neurons in *(k-1)-th* hidden layer is. Activation function *f* has to transform the weighted sum of all arrival signals to the neuron. Mostly a simple threshold function, sigmoidal, hyperbolic tangent or radial basis functions obtain the role of activation function *f*.

The sum of squared differences between the desired and actual values of the output neurons is defined by equation, often called cost function:

$$E = \frac{1}{2}\sum_{i=1}^{n_o}(t_i - y_i)^2, \tag{16}$$

where t_i is the target value and y_i is *i-th* output neuron in the neural network output layer with n_o number of neurons . Each weight w_{ij} is adjusted to reduce *E* in optimal way. There are different ways how w_{ij} is adjusted and depend on the training algorithm selection.

3.1.1 Back-propagation learning

Most frequently used learning algorithm is back-propagation, which is iterative gradient designed to minimize the mean square error between the actual output of a neural network and the desired value. Back-propagation algorithm has two phases: forward phase, where feed-forward propagation of input pattern signals through the network is happening with aim of functional signal computation (equation 15); in the second (backward) phase the partial derivatives of the error function with respect to the weights are computed, where-after the weights are updated (equation 17). Back-propagation is gradient descent on the cost function; in other words the weight updates in back-propagation are equivalent to:

$$\Delta w_{ij}^{k} = -\eta \frac{\partial E}{\partial w_{ij}^{k}}, \tag{17}$$

where derivatives in the equation above are evaluated by the recursive equation

$$\frac{\partial E}{\partial w_{ij}^{k}} = \frac{\partial E}{\partial y_{i}^{k}} \cdot \frac{\partial y_{i}^{k}}{\partial w_{ij}^{k}}. \tag{18}$$

Two difficulties are noted for back-propagation: difficulty in optimal step size selection, and algorithms itself often converges to some local instead of global minimum.

3.2 Radial basis function networks (RBFN)

Because of the superiority due to simple network structure (only one hidden layer), and consequently shorter learning time and better resiliency against a bad learning input-output data set, RBFNs are more suitable in function approximations with local variations and discontinuities.

In RBFN radial basis function $\phi(r)$ computations activate hidden-layer neurons. The activation depends on the distance between the input vector $\mathbf{x}$ and the center od each hidden unit $\mathbf{c}_j$ and parameter *b*, called the width. The output of the network is given by:

$$o_i = y_i(\mathbf{x}) = \sum_{j=1}^{n_H} w_{ji}\, \phi\left(\|\mathbf{x} - \mathbf{c}_j\|/b\right), \tag{19}$$

where i=1,2,...,n_o. n_o stands for the number of output neurons and n_H for the number of hidden layer neurons. $y_i(\mathbf{x})$ is the *i*-th output of the RBFN, w_{ji} is the connection weight from the hidden unit, and $\|\cdot\|$ denotes the Euclidean norm. The hidden layer performs a nonlinear transform if the input, and the output layer is by requirement linear.

A number of functions can be used as the RBF, but most usual Gaussian function is selected for the $\phi(\cdot)$. For the set of N input-output data pairs $\{(x_p, t_p)\}$ the RBFN output can be expressed in matrix form as:

$$\mathbf{O} = \mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{\Phi}, \tag{20}$$

with $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{n_H}]$ as the weight matrix $(n_H \times n_o)$, where $\mathbf{w}_i = (w_{1i}, \dots, w_{n_H i})^{\mathrm{T}}$ is. Matrix $\mathbf{\Phi} = [\phi_1, \dots, \phi_N]$ $(n_H \times N)$ is composed from outputs of the hidden layer; for the *p*th sample $\phi_p = (\phi_{p,1}, \dots, \phi_{p,n_H})^{\mathrm{T}}$, where $\phi_{p,k} = \phi(\|\mathbf{x}_p - \mathbf{c}_k\|)$:

$$\mathbf{\Phi} = \begin{bmatrix} \phi(x_1, c_1) & \cdots & \phi(x_1, c_N) \\ \vdots & \cdots & \vdots \\ \phi(x_{n_H}, c_1) & \cdots & \phi(x_{n_H}, c_{1N}) \end{bmatrix}, \tag{21}$$

$\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]$ stands for $(n_o \times N)$ matrix, where $\mathbf{y}_p = (y_{p,1}, \dots, y_{p,n_o})^{\mathrm{T}}$.

3.2.2 RBFN learning

RBFN learning is performed by error cost function minimization:

$$E = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{t}_p - \mathbf{W}^T\phi_p\|^2 = \frac{1}{N}\|\mathbf{T} - \mathbf{W}^T\mathbf{\Phi}\|_{\mathrm{F}}^2, \tag{22}$$

where $\|\cdot\|_{\mathrm{F}}$ stands for the Frobenious norm, defined as: $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{trace}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$. In RBFN learning, the network tends to find centers and the weights. The learning process is done in two stages; first weights and centers are fixed, next weights are found by solving linear equation. For the RBFN successful learning center selection is critical. It can be randomly chosen from the input data or from the cluster means. After the centers have been chosen, the weights are computed using a linear pseudoinverse solution:

$$\mathbf{W} = \mathbf{\Phi}^{+}\mathbf{T}, \tag{23}$$

where: $\mathbf{\Phi}^{+} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}$. Vector $\mathbf{T}$ contains known target values. Since $\mathbf{\Phi}$ is constant during the learning process, pseudoinversion of input patterns $\mathbf{\Phi}$ must be done only once. The existence of this linear solution means that unlike MLP networks the RBF networks have a unique local minimum (when the centers are fixed). This is so called modified backpropagation algorithm with instant training of the hidden layer [19]. Modified back-propagation learning process overcomes error back-propagation and Levenberg-Marquardt in better convergence rate.

But RBFNs have drawback in non-unique and not most efficient function representation. The problem appears because basis functions in the family are generally non orthogonal. If the family of basis functions for the RBFN is replaced by an orthonormal basis (scaling functions in the theory of wavelets), one gets WNNs, that achieve better learning success.

3.3 Wavelet neural networks

In fact, RBFN and WNN are very similar to each other. In WNN wavelet frames such as Gaussian wavelet, Morlet or Mexican hat are used for activation function. Center parameters and the width from the RBFN are replaced by translation $\mathbf{E}_j$ and dilation $\mathbf{D}_j$ vectors, thus, the output is given by:

$$\boldsymbol{o_i} = \boldsymbol{y_i}(\mathbf{x}) = \sum_{j=1}^{n_H} \boldsymbol{w_{ji}\phi_j}\left(\left\|\mathbf{x} - \mathbf{E}_j/\mathbf{D}_j\right\|\right). \tag{24}$$

Further learning is similar to RBFN, only parameters of center and width of radial basis function are replaced by translation and dilation.

4 SIMULATIONS

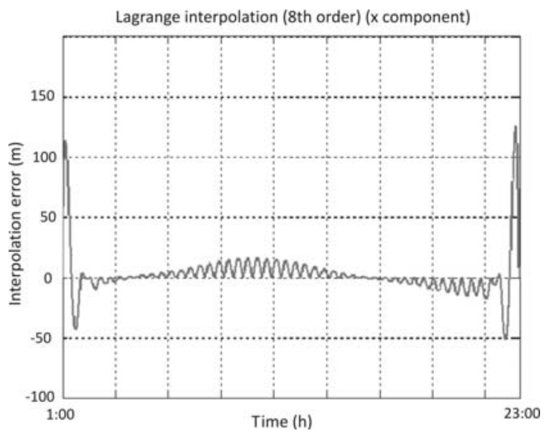
Since previous studies showed different polynomial usage in greater detail ([4], [12], [16]) we have restricted our study on showing preferences of ANNs for GNSS orbit approximation. From RBFN approximation results (Tab. 1) it can be observed, approximation results approach actual values in range of several milimeters. Network training was performed using 15-minute data (bold), 5-minute data were used for testing.

Table 1: Actual data (broadcast versus precise ephemerides differences) and RBFN approximation results (1st January 2002). Data, shown in bold, were used as training set.

Time (h)	**0:00**	0:05	0:10	**0:15**	0:20	0:25	**0:30**	0:35	0:40	**0:45**	0:50	0:55	1:00
Data [m]	**3.786**	3.681	3.570	**3.456**	3.340	3.222	**3.101**	2.978	2.854	**2.730**	2.605	2.482	2.360
RBFN[m]	**3.786**	3.679	3.568	**3.456**	3.341	3.223	**3.103**	2.972	2.858	**2.734**	2.609	2.484	2.361

Time (h)	1:05	1:10	**1:15**	1:20	1:25	**1:30**	1:35	1:40	**1:45**	1:50	1:55	**2:00**
Data [m]	2.242	2.127	**2.016**	1.910	1.809	**1.715**	1.628	1.547	**1.475**	1.413	1.362	**1.328**
RBFN[m]	2.240	2.122	**2.010**	1.903	1.804	**1.712**	1.628	1.552	**1.484**	1.423	1.370	**1.325**

Further simulations were performed to show advantages od ANN approximation and extrapolation. In interpolation (Fig. 3b) limitations of successive interpolation due to Runge's phenomena (Fig. 3a) at the end of the interval are shown. Oscillations can reach several tens of centimeters. In ANN training, by proper parameter selection, oscillations are not a subject to discuss (Fig. 3b). ANN performs well also for prediction. Prediction for 30 minutes ahead are in range of several centimeters.



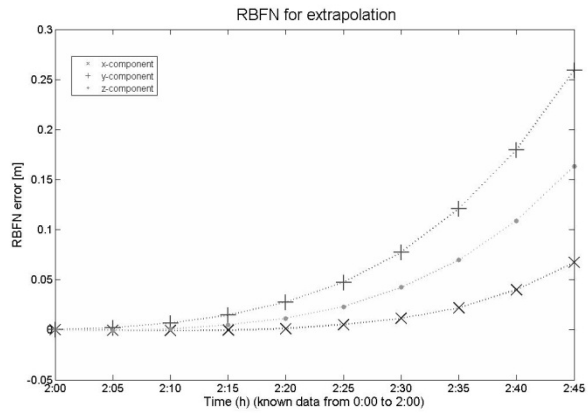


Figure 3: *a) Left:* Runge's phenomena in interpolation. Notice the oscillations near the end of the interval. *b) Right:* RBFN approximation and prediction performance in all components for the PRN02.

5 CONCLUSIONS

In this paper we have described two possible approaches to compute GNSS satellite's positions between 15-minute sampling interval, specific for the SP3 orbit files. Well known interpolation methods have disadvantage in high oscillations near the end of the interval. In cases, when satellite's coordinates close to the end of the fit interval have to be computed, new SP3 files have to be added. If new SP3 files are not available, position calculation near the end of the interval cannot be performed by intepolation. In such cases positions can be obtained using ANNs. Their capability of potential pattern finding, where only origin data in results are given, was established also for prediction.

References

[1] Broomhead, D. S., Lowe, D., 1988. Multivariable function interpolation and adaptive neworks. Complex Systems, Vol. 2, pp. 321-335.

[2] Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, pp. 303–314.

[3] Daubechies, I., 1988. Orthonormal Bases of Compactly Supported Wavelets. Communications On Pure & Applied Mathematics **41**, pp. 909-996.

[4] Feng., Y., Zheng, Y., 2005. Efficient interpolations to GPS orbits for precise wide area applications. GPS Solutions 9(4), pp. 273–282.

[5] Hornik, K., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, pp. 359–366.

[6] Hugentobler., U., Dach, R., Frides, P., 2005. Bernese GPS Software Version 5.0 Draft. Astronomical Institute University of Bern, 392 p.

[7] ICD-GPS-200C. http://www.navcen.uscg.gov/gps/geninfo

[8] Mallat, S., 1989. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation, IEEE Transactions on Pattern Analysis and Machine Intelligence*,* 11, pp. 674–693.

[9] Neagoe, V. E., 1990. Chebyshev nonuniform sampling cascaded with the discrete cosine transform optimum interpolation, IEEE Transactions on Acoustics, Speech & Signal

Processing, Vol. 38, pp. 1812-1815.
[10] Montebruck, O., Gill, E., 2000. Satellite Orbits. Springer-Verlag, 369 p.
[11] Mikhail, E. M., Ackermann, F., 1976. Observation and least squares, Harper and Row, New York, pp. 418–421.
[12] Neta. B., Sagovac, C. P, Danielson, D. A., Clynch, J. R., 1996. Fast interpolation for Global Positioning System (GPS) Satellite Orbits. In: Proc. AIAA/AAS Astrodynamics Specialist Conference, San Diego, CA, Paper Number AIAA 96-3658.
[13] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B.P., 1992. Numerical Recipes in C: the art of scientific computing, Second Edition, Cambridge University Press., 318 p.
[14] Rivlin, T. J., 1974.The Chebyshev Polynomials. Wiley, New York.
[15] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., 1986. Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing, volume 1, MIT Press., pp. 318-362.
[16] Schenewerk, M., 2003. A brief of basic GPS orbit interpolation strategies. GPS Solutions 6(4), pp. 265–267.
[17] Spofford, P.R., Remondi, B.W., 1999. The National Geodetic Survey Standard GPS Format SP3.
[18] Zhang, Q., Benveniste, A., 1992. Wavelet networks, IEEE Trans. Neural Networks, 1992, 3, pp. 889-898.
[19] Wilamowski, B. M., 1997. Modified EBP Algorithm with Instant Training of the Hidden Layer. Proceedings of Industrial Electronic Conference (IECON'97), New Orleans, November 9-14, 1997, pp. 1097-1101.

DIFFERENT APPROACHES TO THE TUNING OF THE RCPSP ALGORITHM RAR

Igor Pesek[1] and **Janez Žerovnik**[2]
[1]University of Maribor, Faculty of Natural Sciences and Mathematics
Koroška cesta 160, SI-2000 Maribor, Slovenia
[2]University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
igor.pesek@uni-mb.si
janez.zerovnik@imfm.uni-lj.si

Abstract: Tuning the algorithms can be very time consuming and not always successful. In this paper we present two different methods for determining the best values for our algorithm RaR which was developed for the well known RCPSP problem. First method is usual "brute force" method, which uses average deviation as main decision tool. Second method uses more sophisticated statistical methods for more reliable parameter tuning. We compare both methods and present the results of our experiments.

Keywords: metaheuristics, RCPSP, parameter tuning, library race.

1 INTRODUCTION

The Resource-Constrained Project Scheduling Problem (RCPSP) is a classical scheduling problem that has received a lot of attention from the community on metaheuristics (see, e.g., [1]). In addition, a large and widely-accepted dataset PSPLIB is available publicly for RCPSP [5], so that algorithms can be compared on a common ground.

The paper is organized as follows. In Section 1.1 we describe the RCPSP by providing the mathematical formulation. In Section 2.1 we outline our main local search. In Section 0we report the results of our experimental analysis and we make the comparisons. Finally, in Section 4 we draw some conclusions and we discuss future work.

1.1 RCPSP Formulation

The resource-constrained project scheduling problem (RCPSP) can be stated as follows. Given are *n* activities $a_1,\ldots,a_n$ and *r* renewable resources. A constant amount R_k of units of resource *k* is available at any time. Activity a_i must be processed for p_i time units; pre-emption is not allowed. During this time period a constant amount of r_{ik} units of resource *k* is occupied. All the values R_k, p_i, and r_{ik} are non-negative integers.

Furthermore, there are precedence relations defined between activities. That is, we are given a directed graph *G = (V,E)* with V = $\{1,\ldots,n\}$ such that if $(i,j)\in E$ then activity *j* cannot start before the end of activity *i*.

The objective is to determine starting times s_i for the activities a_i, $i=1,\ldots,n$ in such a way that:

- at each time t the total resource demand is less than or equal to the resource availability for each resource,
- the precedence constraints are fulfilled and,
- the makespan $C_{max} = \max_{i=1}^{n} c_i$, where $c_i = s_i + p_i$, is minimized.

As a generalization of the job-shop scheduling problem the RCPSP is NP-hard in the strong sense [4].

Let us illustrate the above definitions with the example on *Figure 1* which is taken from [9]. There are eleven activities and one single resource. The numbers associated with each node give the length of the activity and the units of resource it uses.

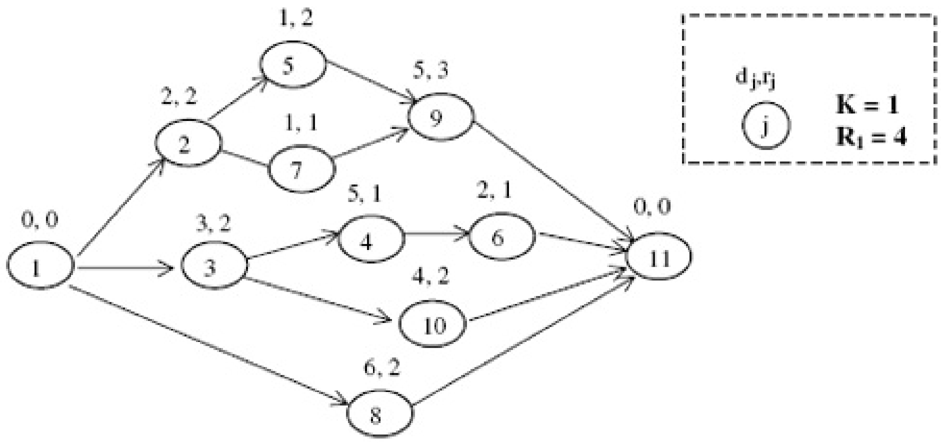


Figure 1: Activity network

Clearly, a schedule is represented by the assignment of starting times of all activities. However, it can also be represented indirectly by a *list*, which is a permutation of all the activities. An activity list is called feasible if it satisfies all precedence constraints, i.e. each activity has all its network predecessors as predecessors in the list.

It is easy to see that given a feasible activity list there is a unique way to build a feasible schedule of minimal makespan, i.e., the so-called left-justified one. This is simply obtained by placing the activities in the given order one by one at the earliest possible starting time according to precedence and resource constraints. It can also be shown that there always exists an activity list that generates an optimal schedule [4].

Given the network depicted in *Figure 1*, some of the (feasible) activity lists are: $\lambda = (1,2,3,5,7,9,8,4,6,10,11)$, $\alpha = (1,2,3,7,10,4,8,5,9,6,11)$ and $\beta = (1,2,3,7,4,10,8,5,6,9,11)$.

For example, the makespan of β is 18, as the activity list β gives rise to a schedule depicted on *Figure 2*.

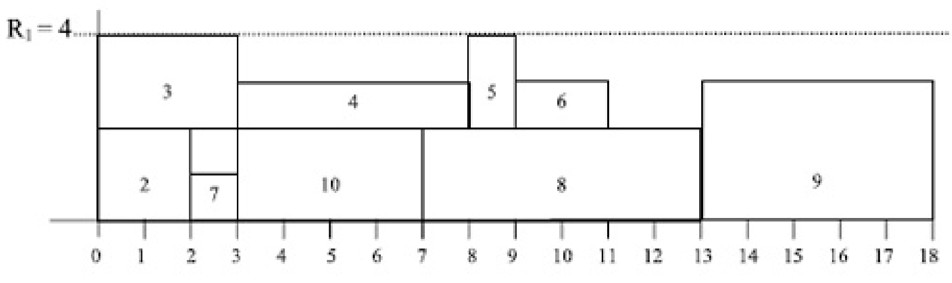


Figure 2: Feasible schedule for activity list β

1.2 Tuning the algorithms

In the vast majority of the cases, metaheuristics are tuned by hand in a *trials-and-error* procedure guided by some rules of thumb. This approach is typically adopted in most of the research papers dealing with metaheuristics. The trial-and-error approach presents many drawbacks of different nature. A major issue with this approach is that it is extremely time consuming and labor intensive. Usually it requires the attention of a particularly skilled practitioner, typically the person that implemented the algorithm.

In this paper we use a racing procedure for finding, in a limited amount of time, a configuration of a metaheuristic that performs as good as possible on a given instance class of a combinatorial optimization problem. Taking inspiration from methods proposed in the machine learning literature for model selection through cross-validation, we propose a procedure that empirically evaluates a set of candidate configurations by discarding bad ones as soon as statistically sufficient evidence is gathered against them. As the computation proceeds, if sufficient evidence is gathered that some candidate is inferior to at least another one, such a candidate is dropped from the pool and the procedure is iterated over the remaining ones. The elimination of inferior candidates speeds up the procedure and allows a more reliable evaluation of the promising ones.

More specifically, we use library RACE [2] which is part of the statistical open source framework R.

2 TUNING

In this section we first present our algorithm and then we present the results of the fine tuning of the algorithm.

2.1 The Algorithm Remove and Reinsert (RaR)

The main idea of RaR [7] is to remove a fixed number *m* of activities from the schedule and insert them back into the schedule, where *m* is a parameter of the method. A move M is thus identified by a sequence of *m* activities $M = \{a_{M_1}, a_{M_2}, \ldots, a_{M_m}\}$, and it is executed in a state S leading to state S' in the following way:

1. $S' = S - M$ // remove all activities in *M* from *S*}
2. for each $i = 1, \ldots, m$, add activity a_{M_i} to S' in the following way:
2.1. search for the position *j* in which the makespan of S' plus a_{M_i} is minimized
2.2 insert activity a_{M_i} in position *j* in S'

The neighborhood RaR, for most of the values of *m* that we consider, is a large neighborhood, and its exhaustive exploration turned out to be rather impractical. Therefore for this neighborhood, we resort to a simple hill-climbing strategy, based on a randomized non-ascending selection. In details, at each iteration, we draw a random move and compute its cost. If the move is improving or sideways, it is executed, otherwise the state is unchanged. The procedure stops after a fixed number of iterations. We call this algorithm HC(RaR) . It is worth noticing that HC(RaR) can also be seen as a form of min-conflict hill-climbing (MCHC), since activities are selected at random, but then are reinserted in the optimal way. In addition, like MCHC, it accepts sideways moves (for diversification).

2.2 Experimental setting

Experiments were performed on an Intel Pentium 4 (3.4 GHz) processor running Linux Suse v.10. The algorithms have been coded in C++ exploiting the framework *EasyLocal++* [3], and the executables were obtained using the GNU C/C++ compiler (v. 4.0.1). The statistical tests are performed using the software environment for statistical computing R. For HC(RaR) the stop criterion in our experiments is based on the number of iterations, i.e. the number of feasible schedules generated. Given that the running time of a single iteration depends on *m*, in order to make a comparison fair, we normalize the number of iterations so that the total running times are approximately the same.

3 RESULTS

In the preliminary results we noticed the importance of the parameter *m*, thus we have tried to tune this parameter with two methods on the instances with *60* jobs. First method was usual »brute force« method, where we computed all the values for the different instances of the problem and then computed the average deviation w.r.t. the lower bound (obtained through the critical-path method) and its standard deviation. In the second method we used library RACE and were computing the best value in statistically more principled way.

Table 1 shows the results for the instances of size $n = 60$, for *30* trials for each instance. The table reports the average deviation w.r.t. the lower bound and its standard deviation.

Table 1: Setting the parameter *m* with "brute force"

m	avg. diff. (%)	std. dev.
2	11.668	10.497
3	11.364	0.9448
4	11.340	10.158
5	11.252	0.9548
6	11.130	0.9713
7	**11.107**	**0.9636**
8	11.122	0.8810
9	11.242	0.8473
10	11.294	0.9244
11	11.388	0.9131
12	11.583	0.8689
13	11.742	0.9826
14	11.867	0.9330
15	11.918	0.8784
16	12.152	0.9073
17	12.345	0.8866
18	12.598	0.9920
19	12.797	0.8610
20	12.880	0.9419

The outcome of the experiments show that the best results are obtained for $m=7$ (in bold). Note that for values around *7* the results are very close and the standard deviation is relatively small for all of them.

In [6] experimental results of almost all the state-of-the-art algorithms are reported. Comparison is made with number of iterations limited to $i=1000$, *5000* and *50000*. Since we have high running times, we compare our method with results reported for *50000* iterations.
Table 2 presents the comparison of the best methods and our method where we used previously determined parameter $m=7$. We can assert that our results are well in-line with the best solvers, thus proving the fine tuning of the parameters to be very successful.

Table 2: Comparison of the best methods for RCPSP, *n=60*

Method	**Authors**	**Standard Deviation**
Scatter search	Debels et al.	10,71
GA-hybrid	Valls et al.	10,73
GA, TS	Kochetov in Stolyar	10,74
GA-FBI	Valls et al.	10,74
GA-FBI	Alcaraz et al.	10,84
GA-self-adapting	Hartmann	11,21
GA-activity list	Hartmann	11,23
HC+RAR	**---**	**11,27**
Sampling, FBI	Tormos and Lova	11,36
Sampling-LFT, FBI	Tormos and Lova	11,47
TS-activity list	Nonobe and Ibaraki	11,58
Sampling-random, FBI	Valls et al.	11,94
GA—late join	Coelho and Tavares	11,94

The second method works differently. We started the procedure with 18 candidates ($m \in [3,20]$) and 480 instances of the size *n=60* of the PSPLIB library. Since the nature of proposed algorithm RaR depends heavily on the probability, we used this fact and expanded our set of instances to more than *5000* instances. This is necessary, because we can not predict how many instances we will have to use to determine the best value of the parameter *m*. In our experiments we computed *8164* trials for which we used *1440* instances. The parameter for the value *m* determined by the library was 6. Two others, that were eliminated last were *m=5* and *m=7*.

4 CONCLUSIONS

Comparison between both methods shows that in both cases the best values are either 6 or 7. As our aim was to check the decision made by the brute force method in [7], we understand the outcome as a confirmation that the decision was a good one.

Speed or time needed to take the final decision on parameters is also important. In this respect the use of library race clearly outperforms “brute force” method. As the second method uses more sophisticated statistical methods which give more reliable results we can suspect that the second method is more suitable for making such important decisions as is to determine the best parameter(s) for an optimization algorithm. However, to make a more firm conclusions the comparison should be based on a more extensive computation. Namely, instead of using all the dataset to find suitable value(s) of the parameter one should use only a part of instances for tuning and another part of dataset instances for the testing set. This will be done in continuation of this research.

References

[1] T. Baar, P. Brucker, and S. Knust. Tabu search algorithms and lower bounds for the resource-constrained project scheduling problem. In S. Voss, S. Martello, I. Osman, and C. Roucairol, editors, Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization, Kluwer academic publishers, 1998, pp. 1 – 18.

[2] M. Birattari, T. Stutzle, L. Paquete, and K. Varrentrapp, A Racing Algorithm for Configuring Metaheuristics, Gecco, 2002.

[3] L. Di Gaspero and A. Schaerf, *EasyLocal++*: An object-oriented framework for flexible design of local search algorithms. *Software-Practice and Experience*, 33 (2003) pp. 733 – 765.

[4] R. Kolisch and S. Hartmann. Heuristic algorithms for solving the resource-constrained project scheduling problem - classification and computational analysis. In J. Weglarz, editor, Handbook on recent advances in project scheduling, . Kluwer, 1999 pp. 147 – 178.

[5] R. Kolisch and A. Sprecher, PSPLIB -- a project scheduling library, European Journal of Operational Research, 96 (1997) pp. 205 – 216,

Data available from http://129.187.106.231/psplib/ .

[6] R. Kolisch and Sőnke Hartmann, Experimental investigation of heuristics for resource-constrained project scheduling: An update, European Journal of Operational Research 174 2006 pp. 23 – 37.

[7] I. Pesek, A. Schaerf, and J. Žerovnik, Hybrid local search techniques for the resource-constrained project scheduling problem, in Hybrid Metaheuristics, T. Bartz-Beielstein, M. J. Blesa, C. B. B. Naujoks, A. Roli, G. Rudolph, and M. Sampels, Eds. LNCS, Springer, 2007, pp. 57 – 68.

[8] I. Pesek and J. Žerovnik, Best insertion algorithm for resource-constrained project scheduling problem, KOI 2006 proceedings, V. Boljuncic, Ed., 2008, pp. 169-176.

[9] V. Valls, S. Quintanilla, and F. Ballestin. Resource-constrained project scheduling: A critical activity reordering heuristic. European Journal of Operational Research, 149 (2003) pp. 282 – 301.

CONTRIBUTION OF COPOSITIVE FORMULATIONS TO GRAPH PARTITIONING PROBLEM[1]

Janez Povh

Institute of mathematics, physics and mechanics Ljubljana, Slovenia
Faculty of information studies in Novo mesto, Slovenia

email: janez.povh@fis.unm.si

Abstract
This paper provides analysis of several copositive formulations of the Graph partitioning problem (GPP) and semidefinite relaxations based on them. We prove that the copositive formulations based on results from [6, 13] are strongly related and that they both imply semidefinite relaxations which are stronger than the relaxations from the literature [17, 8].

Keywords: Graph partitioning problem, copositive programming, semidefinite relaxation

1. INTRODUCTION

Copositive programming has turned out recently to be a strong tool for handling several NP-hard problems, like the Standard quadratic programming problem [4], the Stability number problem [7], the Quadratic assignment problem [16] etc. In all cases the original problem was lifted in a higher dimensional space and the feasible set was described as an intersection of an affine space and either the cone of completely positive or copositive matrices, which are defined as follows: the cone of completely positive matrices of order n is

$$\mathcal{C}_n^* = \{X = \sum_{i=1}^{k} y_i y_i^T,\ k \geq 1,\ y_i \in \mathbb{R}_+^n,\ \forall i = 1, \ldots, k\} \tag{1}$$

and the cone of copositive matrices is

$$\mathcal{C}_n = \{X = X^T : u^T X u \geq 0,\ \forall u \geq 0\}. \tag{2}$$

The copositive formulation of a problem does not simplify the problem but only offers a new, more elegant view to the problem. All difficulties of the original problem are moved into the copositive or completely positive constraint. It is known that the decision problem whether given matrix is not copositive is NP-complete, hence the separation problem for the copositive cone is in co-NP and unless co-NP=NP there is no polynomial certificate for copositivity in general.

Once we have a copositive formulation of an NP-hard problem we can solve it approximately by solving a tractable relaxation (e.g. linear programming or semidefinite programming relaxation) [7, 15, 16] or by running any heuristics to approximately solve the copositive formulation [3, 5].

In this paper we review the most important copositive representation results and present the contribution of copositive programming to the Graph partitioning problem (GPP). The main contribution of this paper consists of two results:

- We show that the feasible sets for copositive formulations of GPP, obtained using Burer's representation technique [6] and Povh's representation technique [13] are in bijective correspondence.

[1]Project was partially supported by the Slovenian research agency under contracts 1000-08-210518.

- We study semidefinite relaxations for GPP and show that the $\mathcal{K}_0^*$ relaxations of both copositive formulations are equivalent and imply stronger lower bounds than the strongest semidefinite relaxation from the literature.

In Section 2 we review the copositive results for the Standard quadratic programming problem, the Stability number problem and arbitrary non-convex quadratic program where the feasible set is described by linear and binary constraints.

In Section 3 we define the Graph partitioning problem (GPP) and state explicitly two copositive programming formulations for GPP. We also provide the analysis of these formulations and relations between different semidefinite relaxations.

In Appendix we review some results about transportation matrices.

1.1 Notation

We denote by δ_{ij} the Kronecker δ. The ith standard unit vector is e_i, the vector of all ones is $u_n \in \mathbb{R}^n$ (or u if the dimension n is obvious) and the vector of all zeros is 0 or 0_n. The square matrix of all ones is denoted by J_n or J, the identity matrix by I and $E_{ij} = e_i e_j^T$.

In this paper we refer to various sets of matrices. Beside the cones of copositive and completely positive matrices defined in (1)–(2) we also use the vector space of real nonnegative $n \times n$ matrices: $\mathcal{N}_n = \{X \in \mathbb{R}^{n\times n} : x_{ij} \geq 0\}$, the vector space of real symmetric matrices of order n: $\mathcal{S}_n = \{X \in \mathbb{R}^{n\times n} : X = X^T\}$ and the cone of positive semidefinite matrices of order n: $\mathcal{S}_n^+ = \{X \in \mathcal{S}_n \colon y^T X y \geq 0, \forall y \in \mathbb{R}^n\}$.

We use $X \succeq 0$ for $X \in \mathcal{S}_n^+$. A linear program over $\mathbb{R}_+^n$ is called a linear program, a linear program over $\mathcal{S}_n^+$ is called a semidefinite program while a linear program over $\mathcal{C}_n^*$ or $\mathcal{C}_n$ is called a copositive program.

The sign $\otimes$ stands for the Kronecker product. When we consider the matrix $X \in \mathbb{R}^{m\times n}$ as a vector from $\mathbb{R}^{mn}$, we write this vector as $\operatorname{vec}(X)$ or x. For $u, v \in \mathbb{R}^n$ we define $\langle u,\, v\rangle = u^T v$ and for $X, Y \in \mathbb{R}^{m\times n}$ we set $\langle X,\, Y\rangle = \operatorname{trace}(X^T Y)$. If $a \in \mathbb{R}^n$, then $\operatorname{Diag}(a)$ is an $n \times n$ diagonal matrix with a on the main diagonal and $\operatorname{diag}(X)$ is the main diagonal of a square matrix X.

For a matrix $Z \in \mathcal{S}_{kn+1}$ with $k, n \geq 1$ we often use the following block notation:

$$Z = \left[\begin{array}{c|ccc} Z^{00} & Z^{01} & \cdots & Z^{0k} \\ \hline Z^{10} & Z^{11} & \cdots & Z^{1k} \\ \vdots & \vdots & \ddots & \vdots \\ Z^{k0} & Z^{k1} & \cdots & Z^{kk} \end{array}\right], \tag{3}$$

where $Z^{i0} \in \mathbb{R}^n$, $1 \leq i \leq k$ and $Z^{ij} \in \mathbb{R}^{n\times n}$, $1 \leq i, j \leq k$. Since $Z^{00} \in \mathbb{R}$, we denote it also by Z_{00}. Similarly we address components of a matrix $Z \in \mathcal{S}_{kn}$ via

$$Z = \begin{bmatrix} Z^{11} & \cdots & Z^{1k} \\ \vdots & \ddots & \vdots \\ Z^{k1} & \cdots & Z^{kk} \end{bmatrix}, \tag{4}$$

where $Z^{ij} \in \mathbb{R}^{n\times n}$.

Given an optimization problem P, we denote its optimal value by OPT_P.

2. REVIEW OF COPOSITIVE REPRESENTATION RESULTS

In last decade several authors presented results how to rewrite particular NP-hard problems as copositive programs.

Bomze et al. [4] proved that the Standard quadratic programming problem can be rewritten as a linear program over the cone of completely positive matrices. They showed that

$$\begin{aligned} &\min \quad \{x^TQx\colon \sum_i x_i = 1, x \in \mathbb{R}^n_+\} \ = \\ &\min \quad \{\langle Q,\, X\rangle \colon \langle J,\, X\rangle = 1,\ X \in \mathcal{C}^*_n\}. \end{aligned}$$

De Klerk and Pasechnik [7] reformulated the Stability number problem as a copositive program:

$$\begin{aligned} \alpha(G) &= \max \quad \{u^t x\colon x_i x_j = 0\ \forall ij \in E(G),\ x \in \{0,\, 1\}^n\} \\ &= \max \quad \{\langle J,\, X\rangle \colon \mathrm{trace}(X) = 1,\, X_{ij} = 0 \text{ for } ij \in E,\, X \in \mathcal{C}^*_n\}. \end{aligned}$$

Povh and Rendl [15, 16, 13] proved that the graph partitioning problem and the quadratic assignment problem also have a copositive representation.

Although Povh [14] extended results from [15, 16, 13] to all quadratic problems over the set of transportation matrices, currently the most general representation theorem belongs to Burer [6]. He showed that any quadratic problem with linear and binary constraints can be rewritten as a copositive program. More precisely, the optimal value of the quadratic problem

$$\begin{aligned} OPT_{LCQP} &= \min \quad x^TQx + c^Tx \\ \text{s. t.} \quad a_i^T x &= b_i, \\ x_j &\in \{0,\, 1\}, \forall j \in \mathcal{J} \\ x &\geq 0 \end{aligned}$$

is under some mild assumptions equal to the optimal value of the following copositive program:

$$\begin{aligned} OPT_{LCQP} &= \min \quad \langle Q, X\rangle + c^Tx \\ \text{s. t.} \quad a_i^T x &= b_i, \quad \langle a_i a_i^T,\, X\rangle = b_i^2 \\ X_{jj} &= x_j, \forall j \in \mathcal{J}, \quad Y \ = \ \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \in \mathcal{C}^*_{n+1} \end{aligned}$$

While Burer [6] mentioned that it is not obvious how to deal with general quadratic constraint, we point out that any quadratic constraint of the type $x^tPx = 0$ with $P \in \mathcal{N}_n$ can be added without breaking the equality, yielding the equation $\langle Q, X\rangle = 0$ in the copositive program. The proof is easy.

3. APPLICATION TO THE GRAPH PARTITIONING PROBLEM

The Graph partitioning problem (GPP) is a classical problem from combinatorial optimization. Given a simple undirected graph $G = (V,\, E)$ with $|V| = n$, a number of partitions $k > 1$ and a vector $m = (m_1,\, m_2,\, \ldots, m_k) \in \mathbb{N}^k$ with $1 \leq m_1 \leq m_2 \leq \ldots \leq m_k$, $\sum_i m_i = n$, we are interested in a partition $(S_1,\, S_2, \ldots, S_k)$ of the vertex set V such that $|S_i| = m_i$ and the total number of cut edges (i.e. edges between different sets) is minimum. We can find in the literature GPP also under the name Min-k-cut problem ([10]). A special instance of this problem is the Graph bisection problem. It is known to be NP-hard [9] and only $\mathcal{O}(\log n)$ approximation algorithms are currently known for the general instances. However, when the graph instance is dense, there exists a PTAS for the Graph bisection problem as well as for a different variant of Min-k-cut problem [2].

Graph partitioning problem has several applications: floor planning, analysis of bottlenecks in communication networks, partitioning the set of tasks among processors in order to minimize the communication between processors etc. A comprehensive survey with results in this area up to 1995 is contained in [1].

Due to proven complexity of the problem there exist several approaches to compute the optimal solution of GPP. Many of them are based on Branch and Bound technique where good lower bound for the optimal value are essential. Several authors [11, 12, 17] have shown that semidefinite programming based lower bounds are among the strongest but also very expensive to compute.

We may represent any partition into k blocks with prescribed sizes by a (partition) matrix $X \in \{0,1\}^{n\times k}$, where $x_{ij} = 1$ if and only if the ith vertex belongs to the jth set. With this notation the total number of cut edges is exactly $0.5\langle X, AXB\rangle$, where A is the adjacency matrix of the graph (i.e. $a_{ij} = 1$ if (ij) is an edge and $a_{ij} = 0$ otherwise) and $B = J_k - I_k$. If L is Laplacian matrix of a graph, then it holds $0.5\langle X, AXB\rangle = 0.5\langle X, LX\rangle$.

The partition matrices can be defined using different sets of equations. If we want to apply the Burer's theorem we shall formulate GPP as binary linear program. The feasible set for GPP consist exactly from the partition matrices, which are completely described by the following constraints: $X \in \{0,1\}^{n\times k}$, $Xu_k = u_n$, $X^T u_n = m$. We can therefore write the graph partitioning problem as

$$\text{(GPP)} \qquad \begin{aligned} OPT_{GPP} &= \min \ \tfrac{1}{2}\langle X, AXB\rangle \\ \text{s. t.} \quad Xu_k &= u_n, \\ X^T u_n &= m, \\ X &\in \{0,1\}^{n\times k}. \end{aligned}$$

Note that we can write constraint $Xu_k = u_n$ as $(u_k \otimes I_n)x = u_n$ and $X^T u_n = m$ as $(I_k \otimes u_n^T)x = m$, where $x = \text{vec}(X)$. Using Burer's result we obtain

$$(GPP_{CP1}) \qquad \begin{aligned} OPT_{GPP} &= \min \ \frac{1}{2}\langle B^T \otimes A, X\rangle \\ \text{s. t.} \quad (u_k^T \otimes I_n)x &= u_n, && (5)\\ (I_k \otimes u_n^T)x &= m, && (6)\\ \langle J_k \otimes E_{ii}, X\rangle &= 1, \ 1 \le i \le n && (7)\\ \langle E_{ii} \otimes J_n, X\rangle &= m_i^2, \ 1 \le i \le k && (8)\\ x &= \text{Diag}(X) && (9)\\ Y &= \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \in \mathcal{C}^*_{kn+1} \end{aligned}$$

Povh [13] recently reformulated GPP as a non-convex quadratic program

$$\begin{aligned} OPT_{GPP} &= \min \ \tfrac{1}{2}\langle X, AXB\rangle \\ \text{s. t.} \quad X &\in \mathbb{R}_+^{n\times k}, \\ X^T X &= M := \text{Diag}(m) \\ \text{diag}(XX^T) &= u_n \\ \langle X, J_n X J_k\rangle &= n^2 \end{aligned}$$

and by Lagrange relaxation technique obtained the following (alternative) copositive formulation of GPP

$$
(GPP_{CP2}) \qquad \begin{array}{rcll}
OPT_{GPP} & = & \min \ \frac{1}{2}\langle B^T \otimes A,\, V\rangle & \\
V & \in & \mathcal{C}^*_{kn},\ W \ \in \ \mathcal{S}^+_n & \\
\sum_i \frac{1}{m_i} V^{ii} + W & = & I_n, & (10)\\
\langle I_n,\, V^{ij}\rangle & = & m_i \delta_{ij}\ \forall i,j & (11)\\
\langle I_k \otimes E_{ii},\, V\rangle & = & 1\ \forall i, & (12)\\
\langle J_{kn},\ V\rangle & = & n^2. & (13)
\end{array}
$$

In the following subsections we explore the relations between these copositive formulations of GPP and between their semidefinite relaxations.

3.1 Both copositive formulations are strongly related

The feasible sets for (GPP_{CP1}) and (GPP_{CP2}) are in bijective correspondence.

Theorem 1

(a) If $V \in \mathcal{C}^*_{kn}$ is feasible for GPP_{CP2}, then $Y_V = \begin{bmatrix} 1 & v^T \\ v & V \end{bmatrix}$, $v = \mathrm{diag}(V)$, is feasible for GPP_{CP1} and gives the same objective value.

(b) If $Y = \begin{bmatrix} 1 & v^T \\ v & V \end{bmatrix} \in \mathcal{C}^*_{kn+1}$ is feasible for GPP_{CP1} then V is feasible for GPP_{CP2} and gives the same objective values.

Proof:

Theorem is a straightforward corollary of the following results from [13] and [6]:

$$
\begin{array}{rcl}
\mathrm{Feas}(GPP_{CP1}) & = & \mathrm{Conv}\{\begin{bmatrix} 1 \\ x \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x \end{bmatrix}^T : x = \mathrm{vec}(X),\ X \text{ feasible for GPP}\} \\
\mathrm{Feas}(GPP_{CP2}) & = & \mathrm{Conv}\{xx^T : x = \mathrm{vec}(X),\ X \text{ feasible for GPP}\},
\end{array}
$$

since there is no nonzero matrix feasible for linear constraints from GPP.

□

3.2 $\mathcal{K}^*_0$ relaxations for GPP

Copositive formulation for GPP does not make the problem tractable, but only offers a new view on this problem. Since the problem is still NP-hard, we may need good lower or upper bounds for OPT_{GPP}. Relaxation of the completely positive constraints in GPP_{CP1} and GPP_{CP1} yield lower bounds for the optimal value. We study the problems, obtained by relaxing $Y \in \mathcal{C}^*$ to $Y \in \mathcal{S}^+ \cap \mathcal{N}$, the so-called $\mathcal{K}^*_0$ relaxations [16]:

$$
\begin{array}{rcl}
OPT_{SDP1} & = & \min \ \left\{ \frac{1}{2}\langle B^T \otimes A,\, X\rangle : Y = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \in \mathcal{S}^+_{kn+1} \cap \mathcal{N}_{kn+1},\ Y \text{ feasible for } (5)-(9) \right\} \\
OPT_{SDP2} & = & \min \ \left\{ \frac{1}{2}\langle B^T \otimes A,\, V\rangle :\ \in \mathcal{S}^+_{kn} \cap \mathcal{N}_{kn},\ W \in \mathcal{S}^+_n, V, W \text{ feasible for } (10)-(13) \right\}
\end{array}.
$$

This lower bounds turn out to be very tight, in particular they are tighter than SDP lower bounds from [17] and Donath-Hoffman spectral lower bound [8], as we prove in the following subsection.

3.3 Relating SDP lower bounds for GPP

Several SDP approaches to the graph partitioning problem have been studied in the last decade. Karisch and Rendl [11] studied semidefinite lower bound for graph equipartitioning problem, Povh and Rendl [15] studied semidefinite lower bound for the special instance of GPP, the so-called min-cut problem, while Wolkovicz and Zhao [17] formulated semidefinite relaxations for GPP, based on projections onto the minimal face of the cone of positive semidefinite matrices which contains the feasible set of original problem. They proposed two semidefinite models, one in the cone $\mathcal{S}^+_{1+kn}$ and the other in $\mathcal{S}^+_{1+(k-1)(n-1)}$. Using the projection idea also made the Slater condition to hold and therefore enabled efficient employment of the interior-point methods.

We quote here only the strongest SDP model - the one with matrices from $\mathcal{S}^+_{1+(k-1)(n-1)}$. We denote it by GPP_{WZ}:

$$(GPP_{WZ}) \qquad \begin{array}{rrcl} \min & \langle L_A,\, \hat{V}Z\hat{V}^T\rangle & & \\ \text{s. t.} & \mathcal{G}_J(\hat{V}Z\hat{V}^T) & = & 0,\ \ (\hat{V}Z\hat{V}^T)_{00} \ = \ 1, \\ & \text{Arrow}(\hat{V}Z\hat{V}^T) & = & 0,\ \ Z \ \in \ \mathcal{S}^+_{1+(k-1)(n-1)}. \end{array}$$

where

$$L_A = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}I\otimes L \end{bmatrix},\quad \hat{V} = \begin{bmatrix} e_0^T \\ W \end{bmatrix},\quad W = \begin{bmatrix} \frac{1}{n}m\otimes u_n \mid V_k\otimes V_n \end{bmatrix} \text{ and } V_p = \begin{bmatrix} I_{p-1} \\ -u_{p-1}^T \end{bmatrix}.$$

Operator $\mathcal{G}_J(\cdot)$ forces the zero pattern, i.e. all non-diagonal square blocks must have zeros on the main diagonal. The operator Arrow is similar to (9), i.e. guarantees that the main diagonal is equal to the first row and therefore represents the 0-1 constraint in the original problem.

3.4 $\mathcal{K}_0^*$ relaxations are the strongest

In this section we compare lower bounds OPT_{SDP1}, OPT_{SDP2} and OPT_{WZ}. They are based on semidefinite programs in different dimensions and also the objective functions seem to be different, therefore relations are not obvious. We first show that the objective functions are equal.

Lemma 2 *Let* $Y = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix}$ *be feasible for* GPP_{SDP1}. *Then*

$$\langle I_k\otimes L,\, X\rangle = \langle B^T\otimes A,\, X\rangle.$$

Proof: From $Y\succeq 0$ it follows that $X - xx^T\succeq 0$. Therefore $\langle J,\, X\rangle \geq \langle J,\, xx^T\rangle = (\sum_i x_i)^2 = n^2$, as follows from (5) or (6).

The matrix $\hat{Y} = (u_k^T\otimes I)\, X\, (u_k\otimes I) = \sum_{1\leq i,j\leq k} X^{ij}$ is positive semidefinite and (8) implies that $\text{diag}(\hat{Y}) = u$, hence $\hat{Y}_{ij}\leq 1,\ 1\leq i,j\leq n$. Therefore

$$n^2 \leq \langle J,\, X\rangle = \sum_{1\leq i,j\leq n} \hat{Y}_{ij} \leq n^2 \tag{14}$$

and equality holds throughout.

Lemma 7 implies

$$\hat{Y} = \sum_{i,j} X^{ij} = J. \tag{15}$$

We know that $Y\geq 0$. Constraints (5) and (7) imply that

$$\text{diag}(X^{ij}) = 0_n, \quad \text{for } i\neq j. \tag{16}$$

Therefore we get

$$\begin{aligned}\langle I_k \otimes L,\, X\rangle &= \langle L,\, \textstyle\sum_i X^{ii}\rangle \;=\; \langle \mathrm{Diag}(Au) - A,\; J - \textstyle\sum_{i\neq j} X^{ij}\rangle \\ &= -\langle \mathrm{Diag}(Au) - A,\, \textstyle\sum_{i\neq j} X^{ij}\rangle \;=\; \langle A,\, \textstyle\sum_{i\neq j} X^{ij}\rangle \\ &= \langle B^T \otimes A,\, X\rangle.\end{aligned}$$

□

Let us consider the set, which contains the feasible set for GPP_{WZ}.

$$\hat{\mathcal{P}} := \left\{ Y \in \mathcal{S}^+_{1+kn} \colon \exists Z \in \mathcal{S}^+_{1+(k-1)(n-1)} \text{ s. t. } Z^{00} = 1 \text{ and } Y = \hat{V} Z \hat{V}^T \right\}. \tag{17}$$

The following characterization of $\hat{\mathcal{P}}$ has been partially shown in [17, 18].

Lemma 3 *A matrix $Y \in \mathcal{S}^+_{1+kn}$ is in $\hat{\mathcal{P}}$ if and only if Y satisfies*

(i) $Y^{00} = 1$, $Y^{0i} u = m_i$, *for* $1 \le i \le k$, $\sum_{i=1}^k Y^{0i} = u^T$.

(ii) $m_i Y^{0j} = u^T Y^{ij}$, *for* $1 \le i, j \le k$.

(iii) $\sum_{i=1}^k Y^{ij} = u Y^{0j}$.

Proof: The "only if" part is done in [18] for the case $m = u$ and in [17] for a general m. We add it here also because of the "if" part. Let $Y = \hat{V} Z \hat{V}^T$ with $Z^{00} = 1$ and $Z \in \mathcal{S}^+_{1+(k-1)(n-1)}$. Obviously $Y^{00} = 1$ and $Y \succeq 0$. Let us define operator $\mathcal{T} \colon \mathbb{R}^{(1+kn)\times(1+kn)} \to \mathbb{R}^{(k+n)\times(1+kn)}$ as $\mathcal{T}(X) = \hat{T} X$, where

$$\hat{T} = \begin{bmatrix} -m & I_k \otimes u_n^T \\ -u_n & u_k^T \otimes I_n \end{bmatrix} \in \mathbb{R}^{(k+n)\times(1+kn)}.$$

A short exercise shows that $\hat{T}\hat{V} = 0$, hence $\hat{T} Y = 0$. The second and third property from (i) are just $\hat{T} Y_{:,0}$ in explicit form. The equations from (ii) are exactly block matrix formulation of constraint $[-m \mid I_k \otimes u_n^T] \cdot Y(:, 1:kn) = 0$ and similarly are equations from (iii) obtained by writing $[-u_n \mid u_k^T \otimes I_n] \cdot Y(:, 1:kn) = 0$ in a block form.

Let us consider the converse direction. Let $Y \in \mathcal{S}^+_{1+kn}$ satisfies (i)–(iii). Then we have $\hat{T} Y = 0$, hence columns of Y are in $\mathrm{Ker}(\hat{T})$. Since $\mathrm{Ker}(\hat{T})$ is spanned by columns of $\hat{V}$ (see [17, Theorem 3.1]), it follows that $Y = \hat{V}\Lambda$, for some $\Lambda \in \mathbb{R}^{(1+(k-1)(n-1))\times(1+kn)}$. The columns of $\hat{V}$ are linearly independent, hence we can find a matrix $\hat{V}^{-1}$ such that $\hat{V}^{-1}\hat{V} = I_{1+(k-1)(n-1)}$. Matrix Y is also symmetric, therefore we have $\hat{V}\Lambda = \Lambda^T \hat{V}^T$. This implies $\Lambda = \hat{V}^{-1}\Lambda^T \hat{V}^T$ and $Y = \hat{V}\Lambda = \hat{V}(\hat{V}^{-1}\Lambda^T)\hat{V}^T$, which means that Y is equal to $\hat{V} Z \hat{V}^T$ for $Z = \hat{V}^{-1}\Lambda^T$. From $Y \succeq 0$ and $Y^{00} = 1$ it follows $Z \succeq 0$ and $Z^{00} = 1$. □

Remark 1 *In Lemma 3 we used $Y \succeq 0$ only to prove $Z \succeq 0$. Hence if $Y \in \mathcal{S}_{1+kn}$ is not positive semidefinite and satisfies (i)–(iii) from Lemma 3, then we can still write it as $Y = \hat{V} Z \hat{V}^T$ for $Z \in \mathcal{S}_{1+(k-1)(n-1)}$ with $Z_{00} = 1$.*

Moreover, if we write $Y = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \in \mathcal{S}_{1+kn}$ *and we know that*

- $X \succeq 0$ *and*
- Y *delivers properties (i)–(iii)*

then $Y \succeq 0$, because Lemma 3 implies

$$Y = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} = \begin{bmatrix} e_0^T Z e_0 & e_0^T Z W^T \\ W Z e_0 & W Z W^T \end{bmatrix}.$$

The columns of W are linearly independent therefore $X = W Z W^T \succeq 0$ implies $Z \succeq 0$, hence $Y \succeq 0$.

Now we can relate optimal values of GPP_{SDP1} and GPP_{WZ}.

Lemma 4 *If* $Y = \begin{bmatrix} 1 & x^T \\ x & X \end{bmatrix} \in \mathcal{S}^+_{1+kn}$ *is feasible for* GPP_{SDP1} *then* Y *is feasible for* GPP_{WZ}.

Proof: Suppose that $Y \in \mathcal{S}^+_{1+kn}$ is feasible for GPP_{SDP1}. We already know that $\mathcal{G}_J(Y) = 0$ - this is exactly the property (16). Since $Y^{00} = 1$ and Y is feasible for (9), we only need to prove that there exists $Z \in \mathcal{S}^+_{1+(k-1)(n-1)}$ such that $Z^{00} = 1$ and $Y = \hat{V} Z \hat{V}^T$. It is sufficient to show that Y satisfies properties (i)–(iii) from Lemma 3. The second and the third part of property (i) are exactly (5) and (6).

The next property that we are going to prove is (iii). First we point out that we can express X also as $X = \sum_{i,j} \hat{X}^{ij} \otimes E_{ij}$, where

$$\hat{X}^{ij} = \begin{bmatrix} X^{11}_{ij} & \cdots & X^{1k}_{ij} \\ \vdots & \ddots & \vdots \\ X^{k1}_{ij} & \cdots & X^{kk}_{ij} \end{bmatrix}$$

is a $k \times k$ matrix, which contains all elements, which are on the (i, j)th position in all blocks of X. Note that $\sum_{1 \le i,j \le n} E_{ij} \langle J_k, \hat{X}^{ij} \rangle = \sum_{1 \le i,j \le k} X^{ij} = J_n$, as follows from (15). Moreover, (15) implies that $\hat{X}^{ii}$ is a diagonal matrix.

Let us define

$$\hat{X} = \sum_{i,j} E_{ij} \otimes \hat{X}^{ij} = \begin{bmatrix} I \otimes e_1^T \\ I \otimes e_2^T \\ \vdots \\ I \otimes e_n^T \end{bmatrix} \cdot X \cdot \begin{bmatrix} I \otimes e_1^T \\ I \otimes e_2^T \\ \vdots \\ I \otimes e_n^T \end{bmatrix}^T.$$

Obviously we have $\hat{X} \succeq 0$. Let us fix $1 \le i \le j \le k$. Matrix Y has property (iii) from Lemma 3 if and only if $u_k^T \hat{X}^{ij} = \operatorname{diag}(\hat{X}^{jj})^T$. The last assertion follows if we apply Lemma 8 to the matrix

$$\begin{bmatrix} \hat{X}^{ii} & \hat{X}^{ij} \\ \hat{X}^{ji} & \hat{X}^{jj} \end{bmatrix}.$$

Showing (ii) is a bit involved. First we point out that $\langle J, Y^{ij} \rangle = m_i m_j$. Indeed, since $\sum_{i,j} E_{ij} \langle J, Y^{ij} \rangle \in \mathcal{S}^+_k$ has numbers m_i^2 on the main diagonal (follows from (8)) and satisfies assumptions of Lemma 7, we have $\langle J, Y^{ij} \rangle = m_i m_j$.

Let us fix $1 \le i < j \le k$ and define

$$U = \begin{bmatrix} I_n & 0 \\ 0 & u_n^T \end{bmatrix} \begin{bmatrix} X^{ii} & X^{ij} \\ X^{ji} & X^{jj} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & u_n^T \end{bmatrix}^T = \begin{bmatrix} U^{11} & U^{12} \\ U^{21} & U^{22} \end{bmatrix} \in \mathcal{S}^+_{n+1}.$$

Note that $U^{11} = X^{ii}$, $U^{12} = (U^{21})^T = X^{ij} u_n$, $U^{22} = u^T X^{jj} u_n = m_j^2$ and $u_n^T U^{12} = m_i m_j$. From $[u_n^T \mid -m_i/m_j]\, U\, [u_n^T \mid -m_i/m_j]^T = 0$ follows $U\, [u_n^T \mid -m_i/m_j]^T = 0$. This yields that $X^{ii} u_n = m_i/m_j \cdot U^{12}$ or equivalently $r_k(X^{ii}) = m_i/m_j \cdot r_k(X^{ij})$, where $r_k(\cdot)$ is the sum of the kth row, $1 \le k \le n$.

We already know that Y is feasible for (iii), hence we have also $nX^{ii}_{kk} = \sum_j r_k(X^{ij})$. By using the results from the previous paragraph we get $nX^{ii}_{kk} = \sum_j m_j/m_i \cdot r_k(X^{ii}) = n/m_i \cdot r_k(X^{ii})$ and from this we conclude that $r_k(X^{ii}) = m_i X^{ii}_{kk}$ and $r_k(X^{ij}) = m_j X^{ii}_{kk}$. This together with symmetry of X implies that Y is feasible for (ii) and Y is indeed feasible for (i)–(iii) from Lemma 3.

□

Remark 2 *The reverse direction in Lemma 4 is in general not true since feasible matrices for* GPP_{WZ} *are not in* $\mathcal{N}_{1+kn}$ *in general.*

We can also relate the semidefinite lower bounds OPT_{SDP1} and OPT_{SDP2}.

Lemma 5 *The feasible sets of GPP_{SDP1} and GPP_{SDP2} are in bijective correspondence, i. e.*

(a) If $V \in \mathcal{S}^+_{kn}$ is feasible for GPP_{SDP2}, then $Y_V = \begin{bmatrix} 1 & v^T \\ v & V \end{bmatrix}$, $v = \mathrm{diag}(V)$, is feasible for GPP_{SDP1} and gives the same objective value.

(b) If $Y = \begin{bmatrix} 1 & v^T \\ v & V \end{bmatrix} \in \mathcal{C}^*_{kn+1}$ is feasible for GPP_{SDP1} then V is feasible for GPP_{SDP2} and gives the same objective values.

Proof: (a) The $n \times n$ matrix

$$\hat{V} = \sum_{i,j} V^{ij} = (u_k \otimes I_n)^T V (u_k \otimes I_n)$$

is positive semidefinite and satisfies $\hat{V}_{ii} = \langle I_k \otimes E_{ii},\, V\rangle = 1,\ \forall i$. Since we have $\mathrm{diag}(\hat{V}) = u_n$ and $\langle J_n,\, \hat{V}\rangle = n^2$ it follows by Lemma 7 that $\hat{V} = J_n$ or equivalently $\langle J_k \otimes E_{ij}, V\rangle = 1,\ \forall i,j$. Therefore Y_V is feasible for (5) and (7). Constraint $\langle I_n,\, V^{ii}\rangle = m_i$ implies (6). To prove feasibility of (8) we consider the following $k \times k$ matrix $\tilde{V} = \sum_{i,j}\langle J_n, V^{ij}\rangle E_{ij} = (I_k \otimes u_n)^T V (I_k \otimes u_n)$. It is positive semidefinite with $\tilde{V}_{ii} = m_i^2$ and $\langle J_k,\, \tilde{V}\rangle = n^2$, hence Lemma 7 again implies that $\tilde{V}_{i,j} = m_i m_j\ \forall i,j$, or equivalently $\langle E_{ij} \otimes J_n,\, V\rangle\langle J_n,\, V^{ij}\rangle = m_i m_j$. Constraint 9 follows by construction.

It remains to prove that $Y_V \in \mathcal{S}^+_{1+kn}$. Since Y_V is feasible for GPP_{SDP1} it is feasible for GPP_{WZ} by Lemma 4, hence $Y \in \hat{\mathcal{P}}$ and therefore delivers the properties (i)–(iii) from Lemma 3. By Remark 1 it follows that $Y_V \succeq 0$. It is easy to see that V and Y_V give the same objective value.

(b) Obviously we have $V \succeq 0$ and V and Y give the same objective value. In the proof of Lemma 2 we proved that $\langle J_{kn},\, V\rangle = n^2$ and $\langle I,\, V^{ij}\rangle = 0,\ i \neq j$. From (5), (6) and (9) it follows $\langle I_k \otimes E_{ii},\, V\rangle = 1,\ \forall i$ and $\langle I,\, V^{ii}\rangle = m_i,\ \forall i$. It remains to show that $\sum_i \frac{1}{m_i} V^{ii} \preceq I_n$. It is sufficient to prove that the largest eigenvalue of $S = \sum_i \frac{1}{m_i} V^{ii}$ is 1. Since V shares the property (ii) from Lemma 3 it follows from (5) that $S u_n = \sum \frac{1}{m_i} V^{ii} u_n = \sum_i \frac{1}{m_i} m_i \mathrm{diag}(V^{ii}) = u_n$, hence u_n is eigenvector of S with eigenvalue 1. Matrix S is non-negative hence Peron–Frobenious theorem implies that the largest eigenvalue has a non-negative eigenvector. This is possible only if 1 is the largest eigenvalue. □

The following theorem is a straightforward corollary of Lemmas 4 and 5.

Theorem 6

$$OPT_{SDP1} = OPT_{SDP2} \geq OPT_{WZ}.$$

Remark 3 *Povh [13] proved that OPT_{new2} lower bound is stronger than Donath–Hoffman eigenvalue lower bound OPT_{DH} from [8]. It is easy to see that OPT_{SDP2} is stronger than OPT_{new1}, hence we also have the following relation:*

$$OPT_{SDP1} = OPT_{SDP2} \geq OPT_{DH}.$$

4. CONCLUSIONS

In the paper we explore the contribution of copositive formulations to the Graph partitioning problem. We show that the copositive formulations suggested by Burer [6] and Povh [13] are very strongly related and that they imply strong and equivalent $\mathcal{K}_0^*$ semidefinite relaxations. These relaxations are in particular stronger than other semidefinite relaxations from the literature.

References

[1] C. J. Alpert and A. B. Kahng. Recent directions in netlist partition: A survey. *Integr., VLSI J.*, 19:1–81, 1995.

[2] S. Arora, D. Karger, and M. Karpinski. Polynomial time approximation schemes for dense instances of NP-hard problems. *J. Comput. System Sci.*, 58(1):193–210, 1999.

[3] I. Bomze, F. Jarre, and F. Rendl. Quadratic factorization heuristics for copositive programming. *manuscript*, 2009.

[4] M. Bomze, M. Duer, E. de Klerk, C. Roos, A. J. Quist, and T. Terlaky. On copositive programming and standard quadratic optimization problems. *J. Global Optim.*, 18(4):301–320, 2000.

[5] S. Bundfuss and M. Dür. An adaptive linear approximation algorithm for copositive programs. *SIAM J. Optim.*, 20(1):30–53, 2009.

[6] S. Burer. On the copositive representation of binary and continuous nonconvex quadratic programs. *submitted, available on Optimization Online* (`http://www.optimization-online.org`), July 2007.

[7] E. de Klerk and D. V. Pasechnik. Approximation of the stability number of a graph via copositive programming. *SIAM J. Optim.*, 12(4):875–892, 2002.

[8] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[9] M. R. Garey and D. S. Johnson. *Computers and Intractability: a guide to the Theory of NP-Completeness.* Freeman, 1979.

[10] N. Guttmann-Beck and R. Hassin. Approximation algorithms for minimum K-cut. *Algorithmica*, 27(2):198–207, 2000.

[11] S. E. Karisch and F. Rendl. Semidefinite programming and graph equipartition. In *Topics in semidefinite and interior-point methods (Toronto, ON, 1996)*, volume 18 of *Fields Inst. Commun.*, pages 77–95. Amer. Math. Soc., Providence, RI, 1998.

[12] A. Lisser and F. Rendl. Graph partitioning using linear and semidefinite programming. *Math. Program.*, 95(1, Ser. B):91–101, 2003. ISMP 2000, Part 3 (Atlanta, GA).

[13] J. Povh. Semidefinite approximations for quadratic programs over orthogonal matrices. *Submitted*, 2009.

[14] J. Povh. *Towards the optimum by semidefinite and copositive programming.* VDM Verlag Dr. Müller, Saarbrücken, 2009.

[15] J. Povh and F. Rendl. A copositive programming approach to graph partitioning. *SIAM J. Optim.*, 18:223–241, 2007.

[16] J. Povh and F. Rendl. Copositive and semidefinite relaxations of the quadratic assignment problem. *Discrete Optimization*, 6:231–241, 2009.

[17] H. Wolkowicz and Q. Zhao. Semidefinite programming relaxations for the graph partitioning problem. *Discr. Appl. Math.*, 96-97:461–479, 1999.

[18] Q. Zhao, S. E. Karisch, F. Rendl, and H. Wolkowicz. Semidefinite programming relaxations for the quadratic assignment problem. *J. Comb. Optim.*, 2:71–109, 1998.

Appendix: Some technical results

Lemma 7 *Let $Y \in \mathcal{S}_k^+$ with $Y_{ii} = a_i$ and $\sum_{i,j} Y_{ij} = (\sum_i \sqrt{a_i})^2$. Then $Y_{ij} = \sqrt{a_i a_j}$, for $1 \le i, j \le k$, or equivalently, $Y = yy^T$ for $y = (\sqrt{a_1}, \sqrt{a_2}, \dots, \sqrt{a_n})^T$.*

Proof: Since $Y \succeq 0$ we know that $|Y_{ij}| \le \sqrt{Y_{ii}Y_{jj}} = \sqrt{a_i a_j}$ and $\sum_{i,j} Y_{ij} \le \sum_{i,j} |Y_{ij}| \le \sum_{i,j} \sqrt{a_i a_j} = (\sum_i \sqrt{a_i})^2$. The equality holds throughout if and only if $Y_{ij} = \sqrt{a_i a_j}$. □

Lemma 8 *Let*

$$Y = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \in \mathcal{S}_{m+n}^+$$

with $A = \mathrm{Diag}(a) \in \mathbb{R}^{m\times m}$, $B = \mathrm{Diag}(b) \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$. If $u_m^T a = \alpha^2$, $u_n^T b = \beta^2$ and $u_m^T C u_n = \alpha\beta$, then $Cu_n = \beta/\alpha \cdot a$ and $u_m^T C = \alpha/\beta \cdot b^T$.

Proof: Without loss of generality we can assume $a_i > 0$ and $b_i > 0$, for all i. From $Y \succeq 0$ follows using Schur complement that $A - CB^{-1}C^T \succeq 0$, hence $u_m^T(A - CB^{-1}C^T)u_m \ge 0$. But

$$\begin{aligned} u_m^T(A - CB^{-1}C^T)u_m &= \alpha^2 - \sum_{i=1}^n \frac{(u_m^T C(:,i))^2}{b_i} \\ &= \alpha^2 - \sum_{i=1}^n \Big(\frac{u_m^T C(:,i)}{b_i}\Big)^2 b_i \le \alpha^2 - \frac{\Big(\sum_{i=1}^n u_m^T C(:,i)\Big)^2}{\sum_i b_i} \\ &= \alpha^2 - \frac{\alpha^2\beta^2}{\beta^2} = 0 \end{aligned}$$

with equality holding if and only if $u_m^T C(:,i)/b_i = u_m^T C(:,j)/b_j,\ \forall i,j$. Since

$$\begin{aligned} \alpha\beta &= \sum_{i=1}^n u_m^T C(:,i) = \sum_{i=1}^n \frac{u_m^T C(:,i)}{b_i} b_i \\ &= \frac{u_m^T C(:,1)}{b_1} \sum_i b_i = \frac{u_m^T C(:,1)}{b_1} \beta^2 \end{aligned}$$

it follows $u_m^T C(:,1)/b_1 = \alpha/\beta$ and consequently $u_m^T C = \alpha/\beta \cdot b^T$. The second part of the lemma follows using $B - C^T A^{-1} C \succeq 0$. □

EXPERIMENTAL COMPARISON OF CONSTRUCTIVE HEURISTICS FOR THE CANISTER FILLING PROBLEM

Gašper Žerovnik
Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
gasper.zerovnik@ijs.si
and
Janez Žerovnik
Institute of Mathematics, Physics and Mechanics, Jadranska 19, SI-1111 Ljubljana, and
University of Ljubljana, FME, Aškerčeva 6, SI-1000 Ljubljana, Slovenia
janez.zerovnik@imfm.uni-lj.si

Abstract: The canister filling problem that arises when optimizing the spent nuclear fuel repository in hard rock. Constructive heuristics followed by remove and reinsert local optimization are considered. Several variants of the algorithm are compared on random and realistic datasets.

Keywords: Canister filling problem, spent nuclear fuel, heuristics, optimization

1 INTRODUCTION

One of the possible policies in spent nuclear fuel (SNF) management [1] is direct deposition. Currently, in Slovenia this option is seriously regarded for pressurized water reactor (PWR) nuclear power plant (NPP) Krško decommissioning program [2]. Swedish concept of deep repository in hard rock [3] is considered. In the reference design for PWR SNF, up to $n = 4$ spent fuel assemblies (SFA) are filled into metal canisters which are placed within the rock. Immediately after irradiation in the NPP the spent fuel emits large amounts of heat and has to be actively cooled down in the NPP spent fuel storage pool for several years before terminal storage in deep rock repository. Otherwise, structural firmness of the rock can be compromised by overheating which implies another natural capacity limit of the canisters. In this discussion we restrict to fixed capacity limit of C_{max} for all canisters. Naturally, we try to minimize the required number of canisters to reduce expenses and storage space by means of waiting. Of course, interim storage with active cooling is more expensive than passive terminal storage, so we aim to reduce the required cooling time by clever combining of the SFAs in canisters, considering both abovementioned restrictions. Motivated by this application, we define formally *The Canister Filling problem* below.

The Canister Filling problem. Let $S = \{P_i \mid i = 1, \ldots, m\}$ be a finite list of positive real numbers, C_{max} a positive real number, and n a positive integer. Make a partition R of S so that each subset consists of no more than n elements and sum of its elements' values does not exceed the capacity C_{max}. Minimize the number M of the subsets of the partition R!

Clearly, if $P_i > C_{max}$ for some i, there is no solution to the problem. We will avoid this complication in further discussion by assuming P_i is always smaller or equal to C_{max} for all i. With this assumption adopted, there always exists a feasible solution, namely a partition in which every subset consists of one element.

It is known NP-hard Minimum Bin Packing problem (Ref. [4], p. 407-422) can be polynomially reduced to the Canister Filling problem [5].

For a later reference let us mention a natural lower bound for the solution

$$M_{min} = \max\{M_{min,1}, M_{min,2}\} = \max\left\{\left\lceil\frac{\sum_{i=1}^{m} P_i}{C_{max}}\right\rceil, \left\lceil\frac{m}{n}\right\rceil\right\} \quad (1)$$

where the $M_{min,1} = \left\lceil \frac{\sum_{i=1}^{m} P_i}{C_{max}} \right\rceil$ is obtained if the elements can be cut, and $M_{min,2} = \lceil \frac{m}{n} \rceil$ follows from the fact that at most n elements can be put in one canister. (Here $\lceil x \rceil$ denotes the ceiling of x.) Apparently, when we compare our solution obtained by one of the algorithms to condition (1), we know that our solution is at least as good if compared to the optimal solution (which we do not know).

In the following section, constructive algorithms that run in time $O(m \log m)$ are described. Basically, we construct the basic solution with a chosen heuristic and then refine the solution using local search. One of the (promising) heuristics is partially randomized and therefore suitable to be run in multistart mode, and in this way further improvements are possible.

2 HEURISTIC ALGORITHMS

Generic algorithm. Our basic heuristic is a simple and intuitive constructive algorithm:

1. set $R = \emptyset$;
2. quicksort S;
3. **repeat**
4. set $aux = \emptyset$, $C = C_{max}$;
5. **for** $k = n$ **down to** 1 **do begin**
6. **if** there is an element $P_i \in S$ that satisfies condition (*)
7. **then** $S = S \backslash \{P_i\}$, $C = C - P_i$, $aux = aux \cup \{P_i\}$ **endif**
8. **endfor;**
9. aux to $R = R \cup \{aux\}$;
10. **until** $S = \emptyset$;
11. return R.

The initial construction is followed by local optimization, which is explained in more detail at the end of this section.

The variants we study depend only in condition (*) that is used in line 6. The variant with condition

find the largest element P_i of S that satisfies $P_i \leq C/k$; (*1)

will be called **algorithm1**. In this paper we use two more variants, in **algorithm2** the condition is

if $k = n$ **then** take the largest element P_i of S;
else find the largest element P_i of S that satisfies $P_i \leq C/k$ **endif**. (*2)

while **algorithm3** applies condition

if $k > \lfloor n/2 \rfloor$ **then** take a random element P_i of S
if it satisfies $P_i \leq C$ **endif**;
else find the largest element P_i of S that satisfies $P_i \leq C/k$ **endif**. (*3)

Several other similar heuristics have been tested but are omitted here due to worse behavior on the tested datasets.

Local search. In each algorithm we additionally integrated several steps of local optimization (of the type 'remove and reinsert' described for example in [6, 7]) in order to refine the solution.

One step of local search is performed as follows:

1. take obtained solution R;
2. set $T = \emptyset$;

3. move all elements of subsets containing less than n elements from R into T;
4. quicksort T;
5. set $aux = 0$;
6. take element x_{ij} of element R_i of R, find the largest element y of T that satisfies $y \leq C_{max} - P(R_i) + x_{ij}$;
7. **if** $y > x_{ij}$ **then** substitute x_{ij} and y, $aux = aux + 1$ **endif**;
8. **jump to** 6 (repeat loop consecutively for each element x_{ij} of every element R_i of R);
9. **if** $aux > 0$ **jump to** 5 **endif**;
10. for T use original heuristic to construct the solution;
11. append T to R;
12. return R.

The local search is performed as long as we observe some difference (note variable aux). Let us define k as the number of the first step we observe no effect of the local search ($aux = 0$ at step 9). The time consumption of the above described method is $O(m' \log m' + mk)$, where $m' \leq m$ and typically $k \ll m$. Actually, k is experimentally observed to be of order of $\log m$. We may approximate $O(m' \log m' + mk) \sim O(m \log m)$ to see the time consumption of the local search is comparable to the time of the construction of the original solution.

The essence of this local search algorithm is that every time Step 7 is executed, value one of the subsets gets closer to the capacity C_{max}, meaning that the residual elements are lower and thus presumably less problematic. At the end (Step 11) the residual elements are combined into subsets using the same original heuristic used for the construction of the basic solution. In most of cases, the refined solution is equal or better compared to the basic solution. However, it may happen the opposite.

3 EXPERIMENTAL COMPARISON

As this work has been motivated by the study of terminal spent nuclear fuel disposal, carried out on Jožef Stefan Institute for ARAO (Agency for Radwaste Treatment) and summarized in [8], the test data we use for experiments are both randomly generated and instances that are obtained by a simulation that provides realistic data which may appear as a real instance in some near future time.

Also, from the practical point of view, from now on, we will limit our study to the case of $n = 4$ elements per subset.

3.1 Uniformly distributed random generated datasets

Maybe the first and easiest efficiency check for a heuristic is to test it on some randomly generated datasets. In this subsection, all the input data sequences were generated by Mathematica 6 default random number generator. The data are uniformly distributed real numbers in the range between 0 and 1, and the default 'canister' capacity is set to $C_{max} = 2$.

First, lets consider the benefit of the local search. For the case of random number sequences of length of $m = 100$, we compared the average number M of subsets (Table 1) for all three heuristics with (M_2), without using local search (M_1), and taking the better of the two options. For each combination of algorithms, the calculation was repeated 10000 times to reduce the error. In Table 1, relative gain (of the local search) is defined as the difference gained using local search, normalized to the initial difference to the theoretical lower bound $M_{min} = 25$.

Since the local search apparently gives some quantitative improvement on the existing solutions and is not very time consuming, further results discussed in this paper will always be obtained using local search.

Table 1: Difference in average number of subsets by using local search for all 3 algorithms.

heuristic	**no l. s.** (M_1)	**l. s.** (M_2)	**best of** (M)	$M_1 - M$	$\frac{M_1-M}{M_1-M_{min}}$
algorithm1	26.52 ± 0.01	26.30 ± 0.01	26.30 ± 0.01	0.23 ± 0.01	**15%**
algorithm2	27.17 ± 0.02	26.77 ± 0.02	26.77 ± 0.02	0.39 ± 0.01	**18%**
algorithm3	26.52 ± 0.01	26.32 ± 0.01	26.27 ± 0.01	0.25 ± 0.01	**16%**

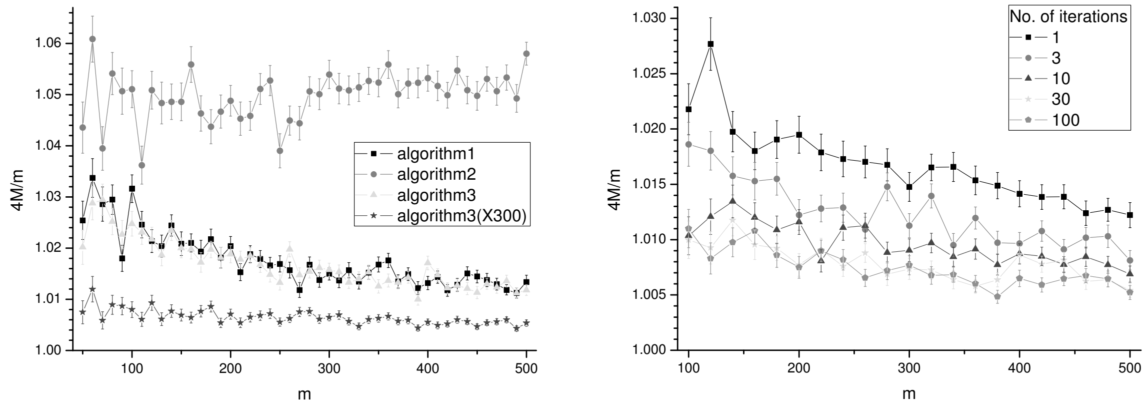


Figure 1: Normed average number of subsets $4M/m$ as a function of the input data length m for different heuristics (left) and different number of iterations of the multistart **algorithm3** (right). For each point on graph, calculation has been repeated 100 times using newly generated random input data. Mean and standard deviation have been calculated accordingly.

The **algorithm1** is very well suited for randomly distributed input data (see Fig. 1, compare to the **algorithm2**) since it 'naturally' constructs the solution as symmetrical as possible starting from the middle, whereas for data of specific type (for example nuclear power plant spent fuel assemblies) it may not always give the best results. On average, random number based **algorithm3** gives results comparable to **algorithm1** but it has one big advantage - it enables multistart because of its non-determinism. Fig. 1 shows progress of the solution by repeating **algorithm3** several times on the same data. Clearly, CPU time consumption increases linearly with number of repetitions. For illustration, points on Fig. 1 belonging to **algorithm3(X300)** have been calculated for about 12 hours on a single 2.0 GHz Dual Core AMD Opteron processor.

3.2 Realistic datasets

Input data for Fig. 3 are predicted data for NPP Krško individual SFAs thermal power as a function of cooling time for the SFAs from the last considered (22., 2007) fuel cycle. In 2007, there were a total of 872 SFAs in spent fuel storage pool from 22 fuel cycles. Number of required canisters for a fixed canister capacity of $C_{max} = 1600\,\mathrm{W}$ were calculated as a function of cooling time after the last considered NPP fuel cycle using all three reference algorithms. Note that SFA thermal power is a very complex function of cooling time [9]. In this paper, however, the reader should just accept the fact that it is (apparently) monotonously decreasing over time. In fact, especially **algorithm2** was designed to work well with NPP SFAs. The thermal power is not randomly and/or uniformly distributed over the SFAs (see Fig. 2). It seems to be distributed in separate 'clusters' belonging to different NPP operating cycles. Only within these clusters the thermal power of SFAs may resemble uniform distribution. The motivation for **algorithm2** is that the SFAs from last NPP operating cycles are most problematic because they can deviate significantly in thermal power so we try to 'get rid of them' immediately (put them into canisters as soon as possible). The actual situation is not that simple, of course. As it can be seen from Fig. 3 even for that type of data, the 'basic' **algorithm1** may give a better solution in some cases. If iterated several (e.g. 1000) times, **algorithm3** performs better than **algorithm1**. However, **algorithm2** may perform even better in some cases (e.g. left side of graph on Fig. 3). Similar relations between the algorithm performance are observed on other realistic datasets, too.

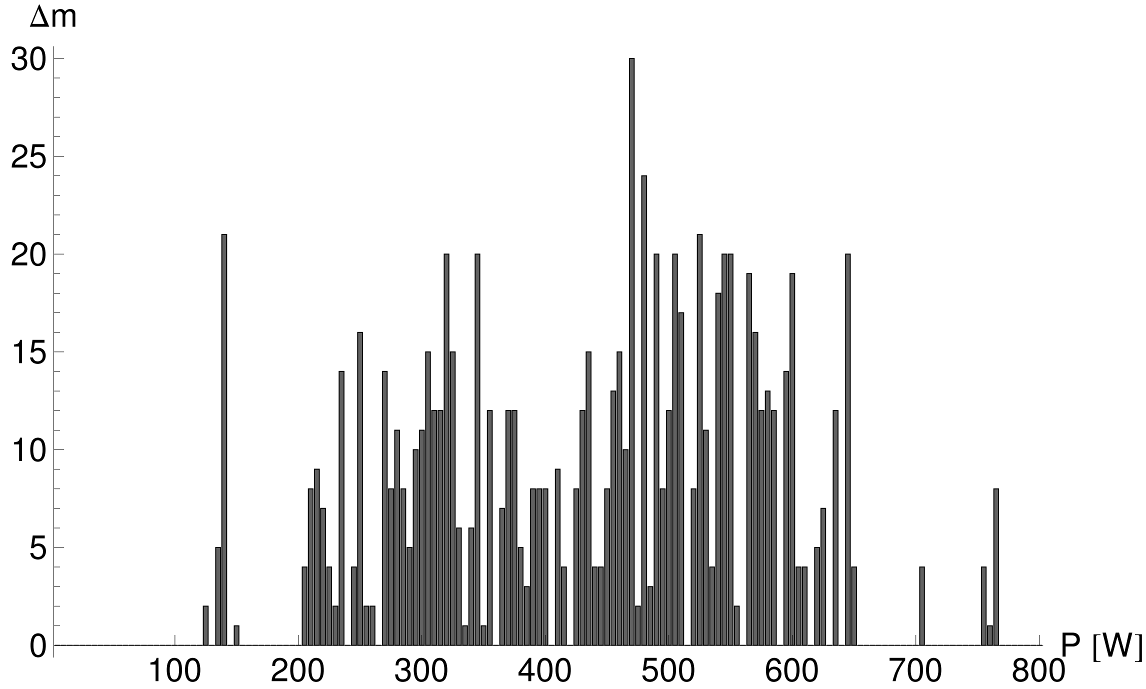


Figure 2: Distribution of the SFA over thermal power P in 2020 for SNF of 22 fuel cycles of NPP Krško. Δm denotes the number of SFAs having power within the corresponding interval of width 5 W.

4 FURTHER IMPROVEMENTS

Since our local search procedure only changes one element at once, the optimization of the algorithm is really local. Consequently, one of the possibilities for further improvements of the results is in 'delocalization' of the local search, e.g. by replacing more than one element at once. In that case, the space of feasible solutions and local minima for chosen heuristic could be considerably expanded and hence we could get a higher probability of finding a better solution.

It may be interesting to test the performance of **algorithm3** by increasing the number of repetitions. Another avenue of further research is to develop similar or completely different procedures to possibly further optimize the problem.

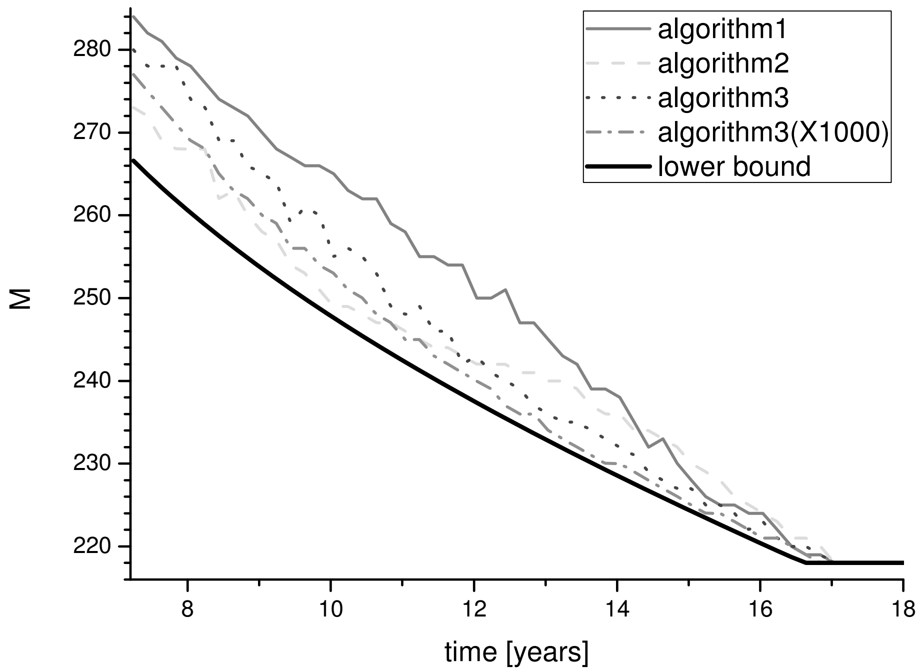


Figure 3: Number of canisters M as a function of time after last NPP fuel cycle for all three algorithms. The solutions are compared to the theoretical lower bound (1). Time resolution is 0.2 years.

5 ACKNOWLEDGEMENTS

The work on optimization of the spent nuclear fuel deposition was funded by the Slovenian Agency for Radwaste Management (ARAO) under the project number ARAO 083-08. We would also thank Ignasi Sau from Inria Sophia - Antipolis for a fruitful discussion on algorithm complexity, and colleagues Marjan Kromar, Matjaž Ravnik, and Luka Snoj from Jožef Stefan Institute for providing us data, useful information and advice about the topic.

References

[1] P. Högselius, "Spent nuclear fuel policies in historical perspective: An international comparison," *Energ. Policy*, **37**, 254 (2009) (Available online: 18 October 2008)

[2] N. Železnik, I. Mele, T. Jenko, V. Lokner, I. Levanat and A. Rapić, *Program razgradnje NEK in odlaganja NSRAO in IJG (Program of NPP Krško Decommissioning and SF & LILW Disposal)*, ARAO-T-1123-03, ARAO (Agency for Radwaste Management), Ljubljana, Slovenia, and APO (Agency for Hazardous Waste), Zagreb, Croatia (2004)

[3] A. G. Milnes, *Guide to the documentation of 25 years of geoscientific research (1976-2000)*, TR-02-18, VBB Anlggning AB, Swedish Nuclear Fuel and Waste Management Co (2002)

[4] B. Korte and J. Vygen, *Combinatorial Optimization, Theory and Algorithms*, Springer-Verlag (2000)

[5] G. Žerovnik, L. Snoj, and M. Ravnik, "Optimization of Spent Nuclear Fuel filling in Canisters for Deep Repository", *Nucl. Sci. Eng.* (October 2009), to appear.

[6] J. Brest and J. Žerovnik, "An approximation algorithm for the asymmetric traveling salesman problem," *Ricerca Operativa*, **28**, 59 (1999)

[7] I. Pesek, A. Schaerf and J. Žerovnik, "Hybrid local search techniques for the resource-constrained project scheduling problem," *Lecture notes computer science*, **4771**, 57-68 (2007).

[8] G. Žerovnik, M. Ravnik, L. Snoj, and M. Kromar, *Optimizacija polnitve vsebnikov z izrabljenim jedrskim gorivom iz NEK glede na največjo dovoljeno toplotno moč*, IJS-DP-10075, Rev. 0, Contract No.: ARAO-083-08 (March 2009)

[9] M. Ravnik, L. Snoj, G. Žerovnik, and M. Kromar, *Izračun izotopske sestave in sproščene toplote iz izrabljenega jedrskega goriva iz NEK*, IJS-DP-9841, Rev. 2 (2008)

The 10th International Symposium on
Operational Research in Slovenia

SOR ’09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section II:

Multicriteria Decision Making

EFFICIENCY OF PROBLEM LOCALIZATION IN GROUP DECISION-MAKING

Andrej Bregar
Informatika d.d., Vetrinjska ulica 2, 2000 Maribor
andrej.bregar@informatika.si

Abstract: A common approach to decision analysis is to classify/sort alternatives. A dilemma appears whether in the context of group decision-making it suffices to classify/sort actions into two instead of multiple predefined categories. A dichotomic sorting procedure for consensus seeking is hence evaluated with a simulation based experimental model. Several quality factors are considered – ability to reach a compromise, autonomous guidance and conflict resolution, convergence of opinions, and robustness of the derived solution. It is statistically proven that problem localization results in efficient, rational and credible decisions.

Keywords: Multiple criteria decision analysis, Dichotomic sorting, Consensus seeking, Simulation experiments

1 INTRODUCTION

The aim of multi-criteria decision analysis is to deal with one of common problematics – to rank actions from the best to the worst one, to choose a single action, to describe actions in terms of their performances, or to classify/sort actions into predefined categories/classes [12]. Classification refers to the assignment of alternatives into nominal categories, while in the case of sorting, categories are ordinal. There exist several methods for classification and sorting [17], such as PROMSORT [1], PROAFTN [2], PAIRCLASS [8], PROMETHEE TRI and PROMETHEE CLUSTER [9], ELECTRE TRI [11] and ELECTRE TRI for groups [5, 6]. These methods allow for the definition of an arbitrary number of categories, which may impose a substantial cognitive load on the decision-makers. A possible approach to reduce the complexity of the problem domain is to apply the localization principle. Preferences are accordingly modelled in the neighbourhood of a single reference profile. The consequence is that alternatives are classified/sorted into only two or three adjacent categories.

Arguably, localized classification/sorting should be treated as merely a special subtype of the common problematic. Not many methods are hence founded on it – two representative approaches include the dichotomic sorting procedure for group decision analysis [4] and the interactive trichotomy segmentation [10]. The primary concern about problem localization is whether the efficiency of the decision can deteriorate compared to classification/sorting into multiple categories, because alternatives are discriminated to a lesser extent, which results in poorer preferential information. Although some experimental investigations have addressed multi-criteria classification and sorting [7], only a few have focused on the case of problem localization. It has been, for example, proven that the dichotomic classification method for processing verbal data outperforms other approaches to verbal decision analysis, and is able to cope with larger sets of objects to be classified [14]. However, the efficiency of localized classification/sorting in group decision-making has not been studied so far.

The presented research tries to fill this void. Its goal is to experimentally prove that it is possible to define an efficient, credible and rational mechanism for group consensus seeking which is based on dichotomic sorting. For this purpose, the author's cooperative multi-agent negotitation procedure [4] is evaluated with regard to the following four variables: ability to reach a compromise, autonomous guidance and conflict resolution, convergence of opinions, and robustness of the consensual solution. All quality factors are chosen and operationalized in such a way that the experimental study may be founded on simulation, which is a highly common and appropriate approach to evaluating multi-critera decision analysis methods [13, 15].

The rest of the paper is organized as follows. Section 2 provides a brief description of the evaluated method. Section 3 defines the experimental model consisting of three independent and four dependent variables. In Section 4, experimental results are presented and analysed. Finally, Section 5 gives a resume and some directions for further work.

2 DICHOTOMIC SORTING PROCEDURE FOR GROUP DECISION ANALYSIS

To implement the localized alternative sorting analysis, a slightly modified version of the ELECTRE TRI method is used [4]. The set of alternatives is partitioned into two exclusive categories. Acceptable choices belong to the positive category C^+, while unsatisfactory ones are members of the negative category C^-. The categories are delimited with a profile of n referential values on criteria domains. Preferences are provided in the form of pseudocriteria, so that the indifference thresholds q_j and the preference thresholds p_j model compensation, while the discordance thresholds u_j and the veto thresholds v_j enable partial incompensation. Several types of distance based robustness metrics are also defined to reflect the minimum changes of weight, veto and preference vectors that cause the reassignment of an alternative to the other category. These three measures are aggregated with a fuzzy weighted averaging operator into the overall robustness degree $r(a_i)$ of the alternative a_i.

To direct the process of group decision-making, in the sense of preference unification, the consensus and agreement measures are defined. Let o be the number of decision-makers and C_k^+ the subset of alternatives that are approved by the k-th individual. The sum of votes for the i-th alternative is then:

$$v_i = \text{card}(a_i \in C_k^+, k = 1, \dots, o).$$

Let $v_i^+ = v_i$ and $v_i^- = o - v_i$ denote how many participants assign a_i to the C^+ and to the C^- category, respectively. The consensus degree reached for this alternative is:

$$z_i = \frac{v_i - \rho}{o - \rho}, \text{ where } v_i = \max(v_i^+, v_i^-) \text{ and } \rho = \left\lfloor \frac{o}{2} \right\rfloor.$$

Partial consensus indices (and similarly, partial agreement indices) of m alternatives are aggregated with the Werner's fuzzy intersection operator [16]:

$$Z = \gamma \cdot \min_{i=1..m} z_i + (1 - \gamma) \cdot \frac{\sum_{i=1..m} z_i}{m}, \text{ where } \gamma \in [0,1].$$

In contrast to the consensus index, the agreement degree is attributed to a single decison-maker. The more group members that assign an alternative to the same category as he does, the higher the level of agreement that is reached from his perspective:

$$\zeta_i^k = \begin{cases} (v_i^+ - 1)/(o - 1), & a_i \in C_k^+; \\ (v_i^- - 1)/(o - 1), & a_i \in C_k^-. \end{cases}$$

The deficiency of the defined metrics is that in the case when all alternatives obtain less than ρ votes in the initial iteration of the consensus seeking procedure, none of them can be unanimously sorted into the positive category. The complementary convergence metrics are then used:

$$z_i = \begin{cases} v_i^+/o, & a_i \in C_k^+; \\ v_i^-/o, & a_i \in C_k^-; \end{cases}$$

$$\zeta_i^k = \begin{cases} 1, & a_i \in C_k^+; \\ 0, & a_i \in C_k^-. \end{cases}$$

The decision-maker with the lowest degree of agreement is selected in each iteration. If his preferential parameters are modified so that unrobustly sorted alternatives are reassigned to the other category, someone else becomes the most contradictive group member, and the overall degree of consensus progressively increases. The inference of new parameter values is automatic, and can be done by solving the optimization program:

$$
\begin{aligned}
&\text{maximize } \min\{\tau_i^+, \tau_i^-\}_{i=1..m}\\
&\text{subject to}\\
&\quad \sigma(a_i) - \tau_i^+ = \lambda, \forall a_i \in \tilde{C}_k^+,\\
&\quad \sigma(a_i) + \tau_i^- = \lambda, \forall a_i \in \tilde{C}_k^-,\\
&\quad 0 \le q_j \le p_j \le u_j \le v_j \le b_j - D_j^-, \forall j = 1, \dots, n,\\
&\quad lw_j \le w_j \le uw_j, \forall j = 1, \dots, n,\\
&\quad \sum_{j=1..n} w_j = 1,\\
&\quad \lambda \in [0.5, 1].
\end{aligned}
$$

Here, $\sigma(a_i)$ is the credibility of the i-th alternative, b_j is the referential profile that delimits the two categories with regard to the j-th criterion, D_j^- is the lowest domain value of the j-th criterion, w_j, lw_j and uw_j are the weights and their lower respectively upper bounds, and λ is the cut-level. The category membership changes at the value of $\sigma(a_i) = \lambda$. The variables τ_i^+ and τ_i^- must be positive to ensure the assignment of each alternative to the right class. They are maximized for the purpose of achieving robustness. Conflicting alternatives are anewly sorted in the following way:

$$a_i \in C_k^+ \rightarrow a_i \in \tilde{C}_k^- \text{ or } a_i \in C_k^- \rightarrow a_i \in \tilde{C}_k^+.$$

The k-th decision-maker is asked to conform to the others with regard to the alternative a_i only if his evaluation $\sigma(a_i)$ contradicts judgements of more than half group members and if the robustness degree r_i^k does not exceed the sensitivity threshold ψ. However, if all of his assignments are robust or if the adjusted values violate the required constraints, he may be skipped, and the next most discordant decision-maker is chosen to negotiate. It is possible that the consensus seeking mechanism addresses several decision-makers in order to find the one who is willing to accept the proposed changes. In the worst case, the negotiation process terminates without reaching a consensus. A compromise is then made by ranking alternatives in the descending order according to the υ_i levels and the overall robustness degrees Γ_i. So:

$$a_i \succ a_j \Leftrightarrow (\upsilon_i > \upsilon_j) \vee \left((\upsilon_i = \upsilon_j) \wedge \begin{Bmatrix} \Gamma_i > \Gamma_j, & \upsilon_i \ge \rho \\ \Gamma_i < \Gamma_j, & \upsilon_i < \rho \end{Bmatrix} \right),$$

$$a_i \approx a_j \Leftrightarrow (\upsilon_i = \upsilon_j) \wedge (\Gamma_i = \Gamma_j),$$

where $\succ$ and $\approx$ are the relations of preference and indifference, respectively. Γ_i is obtained as the difference between the positive and negative robustness degrees:

$$\Gamma_i = |\Gamma_i^+ - \Gamma_i^-|,$$

so that

$$\Gamma_i^+ = \frac{\sum_{k \in E} r_i^k}{o}, \text{ where } E = \{\forall k = 1, \dots, o : a_i \in C_k^+\},$$

$$\Gamma_i^- = \frac{\sum_{k \in F} r_i^k}{o}, \text{ where } F = \{\forall k = 1, \dots, o : a_i \in C_k^-\}.$$

3 EXPERIMENTAL MODEL

3.1 Independent variables

In order to evaluate the efficiency of dichotomic sorting based group consensus seeking, several independent variables are defined:

- *Number of criteria* may be $n \in \{4, 7, 10\}$. *Number of observed alternatives* is fixed to $m = 8$ because only the $m : n$ ratio is significant. Thus, three fundamental situations are considered: $n > m$, $n \approx m$ and $n < m$. The Miller's number is selected as the mean since it represents an important psychological boundary. The deviation is uniform in both directions, where the presumption is made that less than four criteria are uncommon in decision-making.
- *Number of decision-makers* may also be $o \in \{4, 7, 10\}$. Situations with less than four decision-makers are irrelevant because the experimental study focuses on multilateral rather than bilateral negotiations. On the other hand, more than ten decision-makers rarely simultaneously engage in the process of problem solving.
- *Preferential parameters* of the decision model are obtained in the following way:
    - The referential profile is sampled from the normal distribution with the mean of 50 and the deviation of 15. Random values outside the [20, 80] interval are corrected to correspond to its limits.
    - The weights are sampled from the uniform distribution on the [0, 1] interval. Afterwards, they are normalized to sum to 1. Their lower and upper limits are also set. No weight should decrease or increase for more than 0.2 according to its initial value.
    - The indifference, preference, discordance and veto thresholds are calculated relatively to the profile. The deviations are sampled from the set of real values {0.2, 0.4, 0.6, 0.8, 1} corresponding to linguistic modifiers {very weak, weak, moderate, strong, very strong}. Table 1 shows the extreme thresholds that can be potentially computed in the case of some representative profiles. The limits of thresholds are determined by the following constraints:
        - $q_j^- \le q_j \le q_j^+ < p_j^- \le p_j \le p_j^+ < u_j^- \le u_j \le u_j^+ < v_j^- \le v_j \le v_j^+$,
        - $b_j - v_j^+ > D_j^- = 0$ and
        - $b_j + p_j^+ < D_j^+ = 100$.
    - Criteria-wise values of alternatives are sampled from the uniform distribution on the [0, 100] interval.
    - The cut-level is $\lambda = 0.5$, which is the most common value.
    - The sensitivity threshold is set to $\psi = 0.3$. This value is not too restrictive with regard to the highest robustness degree of 1. Consequently, it does not hinder the convergence, yet it prevents the reassignment of credibly/rationally sorted alternatives.

Table 1: The extreme relative thresholds with respect to the profile

b_j	q_j		p_j		u_j		v_j	
	min	max	min	max	min	max	min	max
20	1.430	2.440	2.963	5.179	4.605	8.251	6.365	11.698
50	3.576	6.101	7.408	12.946	11.513	20.627	15.913	29.245
80	5.722	9.761	11.852	20.714	18.422	33.003	25.461	46.791

3.2 Dependent variables

Ability to reach a compromise is determined with several metrics. The most basic is the percentage of cases in which the selection of the best alternative is not unique. This implies that at least two optimal alternatives are equivalent with regard to the number of votes and to the degree of robustness:

$$\exists i,j: (v_i = v_j) \wedge (\Gamma_i = \Gamma_j),$$

where

$$v_i = \max_{k=1..m} v_k \wedge \begin{Bmatrix} \Gamma_i > \Gamma_k, & \forall k \neq i: (\Gamma_i \neq \Gamma_k) \wedge (v_i = v_k) \wedge (v_i \geq \rho) \\ \Gamma_i < \Gamma_k, & \forall k \neq i: (\Gamma_i \neq \Gamma_k) \wedge (v_i = v_k) \wedge (v_i < \rho) \end{Bmatrix}.$$

The compromise is efficient if the distance between the best and any other alternative is large. The distance to the second best alternative shows to what extent the decision-makers should rely on the recommended solution. It is calculated with the following equation:

$$\Delta = \frac{v_i - v_j}{o},$$

so that

$$v_i = \max_{k=1..m} v_k \text{ and } v_j = \max_{\substack{k=1..m \\ k \neq i}} v_k.$$

For the maximal distance, the highest possible number of votes o is taken, because in the most favourable case, all decision-makers agree on sorting one alternative into the C^+ class and other alternatives into the C^- class. The normalized distance from the best alternative to the subset of all suboptimal choices is computed similarly:

$$\Delta = \frac{\sum_{\substack{k=1..m \\ k \neq i}} v_i - v_k}{o \cdot (m-1)}.$$

The drawbacks of the defined distance metrics are that they ignore the robustness degrees Γ_i, and do not indicate the distribution of votes among alternatives. Three additional metrics are therefore introduced:

- The average degree of compromise $\overline{v}_i$ for the i-th best alternative should be equal to o for the best alternative and should be near 0, otherwise. Then, the decision-makers reach a perfect agreement on approving a single choice and disapproving other ones.
- The average number of alternatives ranked in the k-th position is complementary to the former metric. It indicates how many alternatives are rewarded with o votes, how many with $o-1$ votes, and so on. The compromise solution is unambiguous when a single choice is ranked first.
- The average robustness of choices ranked in the k-th place aggregates the Γ_i degrees of those alternatives that receive $o-k+1$ votes. A high average robustness is required for the first and the last position because of unanimous sorting into the C^+/C^- class.

The ability of autonomous guidance and conflict resolution is operationalized with three measures:

- the progression of the compromise solution over iterations of the consensus seeking procedure, according to the votes that are received by different alternatives,
- the agreement degrees ζ_i^k of o individuals that are reached in each iteration,
- the absolute increase of the calculated agreement degrees.

In the initial compromise solution, votes are expected to be uniformly distributed among alternatives, so that few of them are unanimously sorted into either the C^+ or the C^- class.

Because many alternatives are approved by only some group members, a wide spectrum of different compromise degrees is obtained. However, if the decision-making mechanism has the capability of autonomous guidance and conflict resolution, the unification of opinions should be achieved. The compromise degrees υ_i must hence monotonously increase/decrease towards the upper limit of o or the lower limit of 0 votes. Similarly, the agreement degrees must approach the highest value of 1.

The convergence of opinions is also operationalized with three metrics:

- the average number of iterations,
- the monotonous iterative progression of the consensus degree,
- the improvement of the consensus degree compared to the initial compromise.

For the purpose of proving the convergence of opinions, it is appropriate to observe the consensus degree Z. Of essential importance are the final consensus degree and its distance to the initial degree. If the difference is large and positive, and if the last achieved value of Z is approximately 1, non-optimal solutions converge towards the optimal decision. However, it is not necessary to focus on boundary values only. It is instead sensible to investigate the progression of the consensus degree over all iterations of the group decision-making process. In this case, a monotonously increasing curve must be obtained. When its slope is steep, the convergence is quick, otherwise a slow progression results in excessive time and information complexity, and may be the cause of collaboration withdrawals. It is consequently necessary to measure the average number of iterations after which the process terminates. This number should neither be too large nor too small, as in the first case, the convergence is unacceptably slow, while the latter case indicates that it is impossible to reach consensus or that randomly sampled preferential parameters are not able to reflect the conflictness which is inherent in most actual problem situations.

The robustness of the consensual decision is measured relatively with regard to the initial compromise solution. For this reason, the same metrics are applied as for the variable which addresses the ability to reach a compromise. The values are sampled twice – in the first and in the last iteration. It is required that all efficiency indicators improve or at least remain the same. The metrics are:

- the difference in the percentage of cases in which the best alternative is not unique,
- the difference in the distance between the best and the second best alternative,
- the difference in the distance between the best alternative and the set of non-optimal alternatives,
- the change of the average number of votes for the i-th best alternative,
- the change of the average number of alternatives ranked in the k-th place,
- the change of the average robustness degrees of alternatives ranked in the k-th place.

4 EXPERIMENTAL RESULTS

4.1 Ability to reach a compromise

The selection of the best alternative is unique in 100 percent of cases for all experimental combinations. This implies that the robustness degree Γ_i efficiently discriminates alternatives with the same number of votes. However, the best alternative deviates substantially from all other alternatives even if only the agreement degree υ_i is considered. The average normalized distance to the second best alternative is 0.289, and the distance to the set of all non-optimal alternatives is as high as 0.625. The results for various numbers of decision-makers (o) and criteria (n) are presented in Table 2.

Table 2: The distance between the best alternative and non-optimal alternatives

o	distance to the second best alternative			distance to all non-optimal alternatives		
	$n=4$	$n=7$	$n=10$	$n=4$	$n=7$	$n=10$
4	0.213	0.313	0.363	0.664	0.641	0.571
7	0.314	0.264	0.243	0.714	0.566	0.555
10	0.250	0.345	0.295	0.679	0.581	0.655

Small differences between nine experimental combinations are mainly a consequence of random sampling. Yet, according to Figure 1, they should also be partially attributed to the values of o and n. Figure 1 depicts the average number of votes for the i-th best alternative. The best one is generally rewarded with almost all votes, while the second is approved by considerably less decision-makers, so that the rational group decision is unambiguous. The number of votes monotonously decreases and reaches the lower limit of 0 for several worst alternatives, which is reasonable and recommendable in most real-life situations. In addition, the complete spectrum of agreement degrees is obtained.

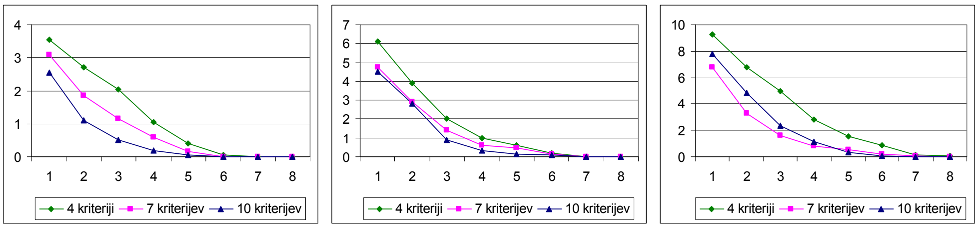


Figure 1: The average number of votes for the i-th best alternative for $o=4$, $o=7$ and $o=10$ decision-makers

Figure 2 indicates that the compromise group decision, which is based on the localized sorting, is also efficient with regard to the number of alternatives ranked in the k-th place. On the average, there is approximately one alternative at each position. Hence, one alternative receives o votes, one receives $o-1$ votes, and so on. The exception are several alternatives without any votes. Most of these alternatives are evaluated as unacceptable because of veto.

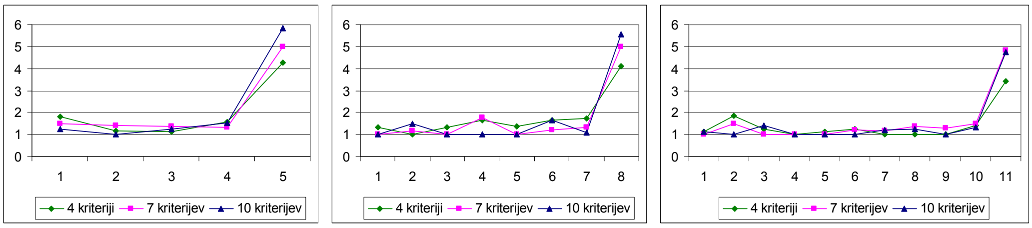


Figure 2: The average number of alternatives in the k-th place for $o=4$, $o=7$ and $o=10$ decision-makers

Finally, Figure 3 shows the average robustness of alternatives ranked at the k-th position. Because of the veto effect, alternatives that are uniformly sorted into the C^{-} class are almost totally robust. Similarly, alternatives that are approved by the majority of decision-makers are robustly sorted into C^{+}. This is essential, because such alternatives represent potentially optimal solutions. Robustness decreases from both ends of the interval. It reaches the lowest extreme point around the middle position corresponding to alternatives with approximately $o/2$ votes. Such alternatives are the source of discordance among decision-makers, so their evaluations cannot be firmly stated.

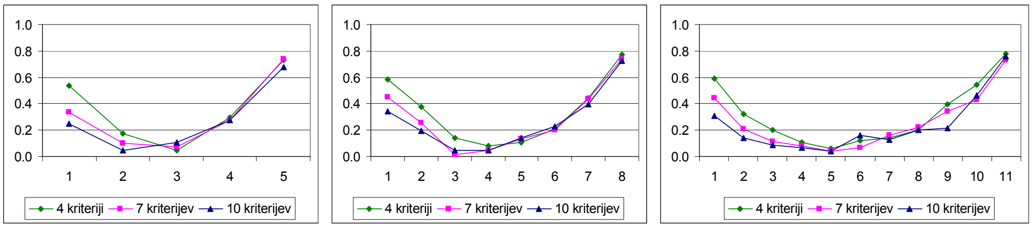


Figure 3: The robustness of alternatives in the k-th place for $o = 4$, $o = 7$ and $o = 10$ decision-makers

4.2 Ability of autonomous guidance and conflict resolution

On Figure 4, the progression of the compromise solution towards the consensual decision over several iterations of the group decision-making procedure is presented. Each alternative is denoted with a separate curved line. The upper curved line belongs to the alternative with the highest initial compromise degree. It is followed by the curve of the alternative with the second highest initial number of votes, and so on. At the bottom, there are alternatives that are sorted into the C^{-} class by most group members in the first iteration of the process. It is possible that several alternatives receive the same number of votes over all iterations. Hence, the graphs generally do not have $m = 8$ curved lines.

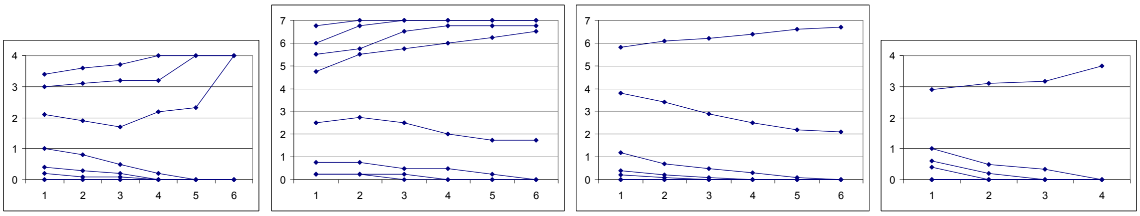


Figure 4: The progression of average compromise degrees according to $(o, n) = (4, 4), (7, 4), (7, 7)$ and $(4, 7)$

It is evident that alternatives with a lot of votes in the first iteration converge towards the highest possible number of votes. Conversely is true for alternatives with a few initial votes. It can therefore be concluded that the consensus seeking mechanism has the ability to detect discrepancies among decision-makers and to guide them in order to overcome their conflicts. However, two anomalies may be identified on Figure 4. Firstly, alternatives with the initial compromise degree of approximately ρ do not necessarily converge. And secondly, there are relatively few such alternatives, which is in contradiction to the results presented on Figures 1 and 2, where the votes are rather uniformly distributed. Both anomalies are a consequence of sampling and aggregated averages. Hence, three typical real cases are shown on Figure 5. They are more representative of the iterative progression of the compromise solution. It can be observed that total unanimity is usually reached, except when a high robustness would make conformation to the group unreasonable and irrational (third graph).

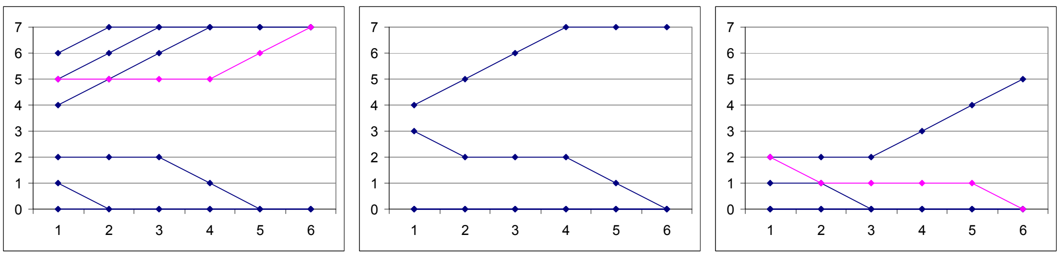


Figure 5: Typical cases of the iterative progression of the compromise solution towards the consensual decision

The ability of autonomous guidance and conflict resolution is also reflected through the agreement degrees, which increase over consecutive iterations. As is evident from Table 3, a large improvement is achieved. Moreover, the final values approach the maximum of 1.

Table 3: The difference between the final and the initial average agreement degree

o	*n*	initial agreement degree		final agreement degree		difference	
		average	deviation	average	deviation	average	deviation
4	4	0.468	0.123	0.939	0.158	0.471	0.180
	7	0.544	0.168	0.895	0.183	0.351	0.218
7	4	0.554	0.159	0.893	0.153	0.339	0.197
	7	0.563	0.141	0.946	0.137	0.383	0.165
total		0.533	0.148	0.918	0.158	0.386	0.190

4.3 Convergence of opinions

The convergence is correlated with the ability of guidance and conflict resolution. Thus, it has already been proven by statistical data presented in Subsection 4.2. In spite of this fact, three additional measurements have been performed. The difference between the final and the initial average consensus degree is very similar to the difference obtained for the average agreement degree. Statistical results therefore do not significantly deviate from data in Table 3, and are not explicitly listed. Figure 6 shows a relatively fast convergence of the consensus degree towards the maximal possible value of 1. Although the maximum is rather closely approached instead of being utterly reached, this is not a deficiency, because potentially high robustness should not allow for the total conformation of decision-makers.

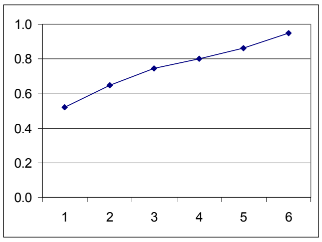


Figure 6: The iterative progression of the average consensus degree

The quickness of convergence is confirmed by data about the total number of iterations. As can be seen from Table 4, a consensus is reached after averagely 4.150 iterations. This is quick enough to prevent the group decision-making process from being subjected to a high time complexity and a substantial information load. On the other hand, more than 4 iterations suffice to reflect the initial real-life discrepancies in the decision-making group.

Table 4: The average number of iterations and the distances of the best alternative

o	*n*	average number of iterations	difference between the final and the initial distance	
			second best alternative	all non-optimal alternatives
4	4	3.900	0.050	0.132
	7	2.900	0.325	0.143
7	4	4.700	-0.086	0.049
	7	5.100	0.371	0.200
total		4.150	0.165	0.131

4.4 Robustness of the consensual decision

The robustness of the consensual decision is measured relatively according to the initial compromise. Table 4 shows that the distance between the best alternative and both the one that is second best as well as the set of all non-optimal alternatives generally increases. This means that the best alternative is to a greater extent discriminated from the others. The first graph of Figure 7 confirms this assumption. In addition, the properties of alternatives in the k-th place do not deteriorate, and the percentage of cases in which the selected alternative is not unique remains 0.

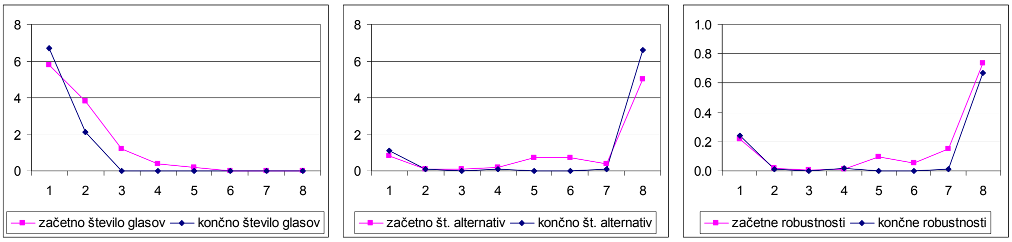


Figure 7: The changes in the number of votes for the i-th best alternative, the number of alternatives in the k-th place and the robustness of alternatives in the k-th place for $o = 7$ decision-makers and $n = 7$ criteria

5 CONCLUSION

In the presented experimental study, a dichotomic sorting procedure for group decision analysis based on the threshold model was evaluated. It was proven that problem localization results in efficient and rational consensual decisions. The research was limited to four quality factors which allow for the application of simulation. In the future, additional variables will be considered [3]: level of imprecision and uncertainty, thoroughness of problem domain analysis, cognitive load during analysis, initial cognitive load, ability of learning, ability of asynchronous interaction, time taken, and accuracy. Such a complex framework will require the combined use of two different techniques – action research and statistical experiments in a controlled laboratory environment.

References

[1] Araz, C., Ozkarahan, I. Supplier evaluation and management system for strategic sourcing based on a new multicriteria sorting procedure. International Journal of Production Economics, 106 (2), 585–606, 2007.

[2] Belacel, N. Multicriteria assignment method PROAFTN: Methodology and medical application. European Journal of Operational Research, 125 (1), 175–183, 2000.

[3] Bregar, A. Extension of the aggregation/disaggregation principle to computer-guided convergent group decision-making processes. Proceedings of Joint Workshop on Decision Support Systems, Experimental Economics and e-Participation, 95–107, 2005.

[4] Bregar, A., Györkös, J., Jurič, M. B. Interactive aggregation/disaggregation dichotomic sorting procedure for group decision analysis based on the threshold model. Informatica, 19 (2), 161–190, 2008.

[5] Demart, S., Dias, L. C., Mousseau, V. Supporting groups in sorting decisions: Methodology and use of a multi-criteria aggregation/disaggregation DSS. Decision Support Systems, 43 (4), 1464–1475, 2007.

[6] Dias, L. C., Clímaco, J. ELECTRE TRI for groups with imprecise information on parameter values. Group Decision and Negotiation, 9 (5), 355–377, 2000.

[7] Doumpos, M., Zopounidis, C. Developing sorting models using preference disaggregation analysis: An experimental investigation. EJOR, 154 (3), 585–598, 2004.

[8] Doumpos, M., Zopounidis, C. Multicriteria classification approach based on pairwise comparisons. EJOR, 158 (2), 378–389, 2004.

[9] Figueira, J., Smet, Y., Brans, J. P. MCDA Methods for Sorting and Clustering Problems: PROMETHEE TRI and PROMETHEE CLUSTER. Université Libre de Bruxelles, 2004.

[10] Jaszkiewicz, A., Ferhat, A. B. Solving multiple criteria choice problems by interactive trichotomy segmentation. EJOR, 113 (2), 271–280, 1999.

[11] Mousseau, V., Slowinski, R., Zielniewicz, P. A user-oriented implementation of the ELECTRE TRI method integrating preference elicitation support. Computers and Operations Research, 27 (7–8), 757–777, 2000.

[12] Roy, B. Multicriteria Methodology for Decision Aiding. Kluwer Academic Publishers, Dordrecht, 1996.

[13] Triantaphyllou, E., Baig, K. The impact of aggregating benefit and cost criteria in four MCDA methods. IEEE Transactions on Engineering Management, 52 (2), 213–226, 2005.

[14] Yevseyeva, I., Miettinen, K., Räsänen, P. Verbal ordinal classification with multicriteria decision aiding. EJOR, 185 (3), 964–983, 2008.

[15] Zanakis, S., Solomon, A., Wishart, N., Dublish, S. Multi-attribute decision making: A simulation comparison of select methods. EJOR, 107 (3), 507–529, 1998.

[16] Zimmermann, H.-J. Fuzzy Set Theory – and Its Applications. Kluwer Academic Publishers, Dordrecht, 1996.

[17] Zopounidis, C., Doumpos, M. Multicriteria classification and sorting methods: A literature review. EJOR, 138 (2), 229–246, 2002.

CONSIDERING INTERACTIONS IN MULTI-CRITERIA DECISION-MAKING

Vesna Čančer
University of Maribor, Faculty of Economics and Business Maribor
Razlagova 14, SI-2000 Maribor, Slovenia
vesna.cancer@uni-mb.si

Abstract: The paper introduces some possible solutions for considering interactions among criteria in the multi-criteria decision-making problems. Special attention is given to the synthesis of the local alternatives' values into the aggregate values where the mutual preferential independence between two criteria is not assumed. We delineate how to complete the additive model into the multiplicative one with synergic and redundancy elements.

Keywords: interaction, multi-criteria decision-making, preferential dependence, synergy, synthesis, weight.

1 INTRODUCTION

Multi-criteria decision-making (MCDM) methods have already turned out to be very applicable in business practice (for an overview see [5, 6]). This paper introduces some possible solutions for considering interactions among criteria[1] in the MCDM problems. They should be considered in measuring the global phenomena like globalization, sustainable development and (corporate) social responsibility, as well as in solving other complex problems like the house purchase decision [7], sensor networks [14] and software architecture [11].

If the criteria can interact with each other, not only the weights on each criterion (i.e. the criterion on the lowest hierarchy level – attribute) but also weighting on subsets of criteria should be considered (see e.g. [11, 14]). Marichal [10] defines and describes three kinds of interaction among criteria that could exist in the decision-making problem: correlation, complementary, and preferential dependency. Positive correlation can be overcome by using a weight on a subset of criteria w_{kl}, such that $w_{kl} < w_k + w_l$, where w_k and w_l are the weights of two criteria, and the sub-additive feature overcomes the overestimate during the criteria evaluation; when negative correlation occurs, the weight on a subset of criteria w_{kl} will be super-additive, given by $w_{kl} > w_k + w_l$ [14]. In a complementary type of interaction, one criterion can replace the effect of multiple criteria – the importance of the criteria pair kl is close to the importance of having a single criterion k or l [14]. In preferential dependence, the decision maker's preference for selecting an alternative is given by a logical comparison (for details see [14]). When such complex interactions exist among criteria, several authors [10, 11, 14] recommend the use of a well-defined weighting function on a subset of criteria rather than single criterion during global evaluation. For example, fuzzy logic has some suitable tools to solve MCDM problems by aggregating criteria. Grabisch and Labreuche [8] pointed out that an important feature of fuzzy integral is the ability of representing a range of interaction among criteria: from redundancy to synergy, which allows for considering both negative and positive interactions among different criteria. Since fuzzy integrals (e.g. Choquet integral) are able to model the interaction among criteria in a flexible way [8], they have already been made use of as a tool for criteria aggregation [8, 10, 11, 14].

Moaven [11] pointed out that dealing with interacting criteria was a kind of difficult issue. An overview of the most preferred and commonly used leading decision-analysis literature (see e.g. [1, 2, 7]) can let us report that decision-analysis theorists and practitioners tend to avoid interactions by constructing independent (or supposed to be so) criteria. We have already presented the frame procedure for multi-criteria decision-making by using the group of methods, based on assigning weights [6]. In synthesis, the additive model is used where the mutual preferential independence of criteria is assumed [3, 7]. However, the synthesis by the additive model may hide synergies. The frame procedure [6] can be adapted to the problem's particularities. To complete the frame procedure for multi-criteria decision making with interactions among criteria, we recommend the up-grade of the currently used traditional methods based on the top-down or bottom-up hierarchy [9, 12]: besides establishing the criteria's importance in order to define the weights of the criteria, the importance of the group of criteria should be expressed in order to assess the importance of the synergic effects of the considered group; we delineate how to complete the additive model into the multiplicative one with synergic and redundancy elements.

2 HOW TO CONSIDER INTERACTIONS IN THE MULTI-CRITERIA DECISION-MAKING, BASED ON ASSIGNING WEIGHTS

We have already developed and presented the frame procedure for multi-criteria decision-making (by using the group of methods, based on assigning weights) that complements intuition and helps us to master interdisciplinary cooperation on formal and informal principles [6]. We concluded that the problems should be approached step-by-step [6]: *Problem definition, Elimination of unacceptable alternatives, Problem structuring, Measuring local alternatives' values, Criteria's weighing, Synthesis and Ranging,* and *Sensitivity analysis*. When defining a problem, relevant criteria and alternatives should be described. Creative thinking methods (e.g. morphological analysis, brainstorming) can be used to develop alternatives. Some of them do not fulfill the requirements for the goal fulfillment and should therefore be eliminated. A complex problem which consists of a goal, criteria, very often some levels of sub-criteria, and alternatives is structured in a hierarchical model. The local values of alternatives can be measured by e.g. value functions, pair-wise comparisons or directly; the criteria's importance can be expressed by using the methods based on ordinal, interval and ratio scale [1, 2, 3]. When measuring the local values of alternatives and expressing the judgments of the criteria's importance, professionals of several fields that are capable of interdisciplinary co-operation should be involved. In synthesis, the additive model is used when calculating the global (i.e. aggregate) alternatives' values [2, 3]; such synthesis may hide synergies, and so does the alternatives' ranging. Several types of sensitivity analysis enable decision makers to investigate the sensitivity of the goal fulfillment to changes in the criteria weights (e.g. gradient and dynamic sensitivity) and to detect the key success or failure factors for the goal fulfillment (e.g. performance sensitivity).

In the Multi Attribute Value (or Utility) Theory (MAVT or MAUT) and the methodologies that were developed on its bases (e.g. Simple Multi Attribute Rating Approach – SMART [9], the Analytic Hierarchy Process – AHP [12]), the additive model is usually used when obtaining the aggregate alternatives' values in synthesis as the sum of the products of weights by corresponding local alternatives' values. When the criteria are structured in one level only, the aggregate alternatives' values are obtained by

$$v(A_i) = \sum_{j=1}^{m} w_j v_j(A_i), \quad \text{for each } i = 1, 2, \ldots, n, \tag{1}$$

where $v(A_i)$ is the value of the i^{th} alternative, w_j is the weight of the j^{th} attribute and $v_j(A_i)$ is the local value of the i^{th} alternative with respect to the j^{th} attribute.

The use of the additive model (1) is not appropriate when there is an interaction among the attributes. In order to apply the model we need to assume that mutual preferential independence exists among the attributes (see e.g. [3, 7]). The first attribute is preferentially independent of the second attribute if we prefer the alternative that is more suitable with respect to the first attribute, irrespective of the values of the second attribute; however, both alternatives have to have equal values with respect to the second attribute. If we also found that the second attribute is preferentially independent of the first attribute, then we could say that the two attributes are mutually preferential independent [7]. If mutual preferential independence does not exist, decision makers or evaluators usually return to the hierarchy (value tree) and redefine the attributes so that the attributes which are mutually preferential independent can be identified. In the occasional problems where this is not possible, other models are available which can handle the interactions among the attributes that express synergy. According to Goodwin and Wright [7], the most well known of these is the multiplicative model. Let us suppose that the MCDM problem is being solved with respect to two attributes only – this simplification is made to explain the bases of multiplicative models; the value of the i^{th} alternative $v(A_i)$ is

$$v(A_i) = w_1 v_1(A_i) + w_2 v_2(A_i) + w_{12} v_1(A_i) v_2(A_i), \quad \text{for each } i = 1, 2, \dots, n, \tag{2}$$

where w_1 is the weight of the first and w_2 is the weight of the second attribute, $v_1(A_i)$ is the local value of the i^{th} alternative with respect to the first and $v_2(A_i)$ is the local value of the i^{th} alternative with respect to the second attribute; (2) is written by following an example in [7]. The last expression in the above sum (2), which involves multiplying the local alternatives' values and the weight of the synergy between the first and the second attribute w_{12}, represents the interaction between the first and the second attribute that expresses the synergy between these attributes.

The sum of the weights in (1) equals one: $\sum_{j=1}^{m} w_j = 1$, and so does it in (2), as well. In order to complete the additive model (1) into the multiplicative one (2), one has to determine the weight of the synergy between the first and the second attribute w_{12}, and then recalculate the weights of initial factors – attributes, obtained in (1):

$$\begin{aligned} w_{1,M} &= (1 - w_{12}) w_{1,A}, \\ w_{2,M} &= (1 - w_{12}) w_{2,A}, \end{aligned} \tag{3}$$

so that the sum of the weights of initial factors and the one of the synergy between them equals one in (2):

$$w_1 + w_2 + w_{12} = 1. \tag{4}$$

In (3), $w_{1,\mathrm{A}}$ is the weight of the first and $w_{2,\mathrm{A}}$ is the weight of the second attribute in the additive model (1), whereas $w_{1,\mathrm{M}}$ is the weight of the first and $w_{2,\mathrm{M}}$ is the weight of the second attribute in the multiplicative model (2).

If the local values of alternatives with respect to each attribute are normalized (which is usual in using the computer supported multi-criteria decision making methods), (2) can be used when there is a (positive) synergic interaction between the attributes. However, there are also negative interactions between the attributes (e.g. redundancy interactions [11]). To consider both positive and negative interactions in (2), we recommend the following procedure. To simplify the synthesis, the above mentioned recalculation (3) of the weights of initial factors – attributes, obtained in (1), is not needed. Let us assume that a synergic element is an “added value” to the aggregate value, obtained by an additive model (1). When there is a positive interaction between the attributes (i.e. synergy), the product of the local alternatives’ values and the weight of the synergy between the first and the second attribute can be added to the sum (1). When there is a negative interaction between the attributes, the product of the local alternatives’ values and the weight of the interaction between the first and the second attribute can be deducted from the sum (1). Such simplification might allow decision makers to use the most preferred computer supported multi-criteria methods, based on the additive model, to obtain the aggregate values, improved for positive and negative interactions between attributes.

When the criteria are structured in two levels, the aggregate alternatives’ values are obtained by

$$v(A_i) = \sum_{j=1}^{m} w_j \left(\sum_{s=1}^{p_j} w_{js} v_{js}(A_i) \right), \qquad \text{for each } i = 1, 2, \ldots, n, \tag{5}$$

where p_j is the number of the j^{th} criterion sub-criteria, w_{js} is the weight of the s^{th} attribute of the j^{th} criterion and $v_{js}(A_i)$ is the local value of the i^{th} alternative with respect to the s^{th} attribute of the j^{th} criterion. Again, let us suppose that there is an interaction between two attributes only; to consider it, we recommend to restructure the two-level criteria hierarchy into the one-level criteria hierarchy where the global weights of the attributes are taken into consideration.

3 DIFFICULTIES OF THE CRITERIA’S AND INTERACTIONS’ WEIGHTING IN PRACTICE

In practice, the criteria’s weighting is an exacting step of the frame procedure for multi-criteria decision making (by using the group of methods, based on assigning weights), although it is supported by several computer supported methods, based on ordinal (e.g. SMARTER), interval (e.g. SMART, SWING) [9] and ratio scale (e.g. AHP) [12]. Professionals of several fields that are capable of interdisciplinary co-operation should be involved in this step. Group priorities’ establishing is well supported by the group-decision making upgrades of computer programs that have been most preferred for individual MCDM in the last – almost three – decades [5]. Because very often the decision makers are not aware of the relationships among different factors taken into account for the goal fulfillment, intuition comes into forefront when establishing the judgments on importance. Further, since decision makers are often inconsistent in criteria’s weighting and in measuring the local alternatives’ values (which means that the importance allocated to the considered criterion or the value allocated to the considered alternative is over- or undervalued), they are recommended to use the procedure for the improvement of the decision makers’ consistency [4]. Studying the corrected intensity levels, they can improve their understanding of the relationships among the criteria, and of the criteria’s meaning and importance as well.

Studying and comparing the additive model (1) and the multiplicative one (2) it can be concluded that the multiplicative model (2) does not require additional effort when measuring the local alternatives’ values with respect to each attribute. The local alternatives’ values measured directly or by making pair-wise comparisons or by using value functions [3, 6], can be used in both models.

However, the judgments’ expression about the importance of the synergies among factors requires additional efforts for decision makers to determine the appropriate weights. Since systematic procedures can not compensate for the lack of knowledge or limited abilities of decision makers, an important task is given to the requisitely holistic use of decision logic, heuristic principles, information and practical experience.

4 CONCLUSIONS

The MCDM methods are worth to develop because

- They enable a complex, integrated and logical framework that allows for interaction and interdependence among factors, structured hierarchically or like a network to deal with dependence and feedback, and
- They enable consideration of all dimensions of the so-called sustainable performance: economic, environmental, ethical and social dimensions; moreover,
- They enable consideration of the contents of social responsibility: honest behavior of influential people and organizations toward their coworkers, other business partners, broader society, and natural environment.

Some of the traditional multi-criteria methods based on assigning weights allow for consideration of several factors that can be structured in one or more criteria levels. Decision makers have to select viewpoints to be considered – from many available – and structured, including interactions among them. In the paper delineated possibilities allow decision makers to use the most preferred computer supported multi-criteria methods, based on the additive model, to obtain the aggregate values, and to improve them for positive and negative interactions between attributes.

References

[1] Belton, V., Stewart, T. J. (2002): *Multiple Criteria Decision Analysis: An Integrated Approach*. Boston, Dordrecht, London: Kluwer Academic Publishers.

[2] Bouyssou, D., Marchant, T., Pirlot, M., Perny, P., Tsoukiàs, A., Vincke, Ph. (2000): *Evaluation and Decision Models: A Critical Perspective*. Dordrecht, Boston, London: Kluwer Academic Publishers.

[3] Čančer, V. (2003): *Analiza odločanja (Decision-making Analysis. In Slovenian)*. Maribor: University of Maribor, Faculty of Economics and Business.

[4] Čančer, V. (2003): *Inconsistency in the Assessment of the Criteria Importance in the Creditworthiness Evaluation*. In: L. Zadnik Stirn, M. Bastič and S. Drobne (eds.): SOR '03 Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research.

[5] Čančer, V. (2005): *Comparison of the Applicability of Computer Supported Multi-Criteria Decision-Making Methods.* In: L. Zadnik Stirn and S. Drobne (eds.): SOR '05 Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research.

[6] Čančer, V. (2008): *A Frame Procedure for Multi-criteria Decision-making: Formulation and Applications.* Invited paper presented at KOI 2008, September 24-26, Pula, Zagreb: Croatian Operational Research Society.

[7] Goodwin, P., Wright, G. (1992): *Decision Analysis for Management Judgment*. Chichester: John Wiley & Sons.

[8] Grabisch, M., Labreuche, Ch. (2005): *Fuzzy Measures and Integrals in MCDA*. In: J. Figueira, S. Greco and M. Ehrgott (eds.), Multiple Criteria Decision Analysis. Boston, Dordrecht, London: Kluwer Academic Publishers.

[9] Helsinki University of Technology: *Value Tree Analysis.* http://www.mcda.hut.fi/value_tree/theory, consulted in June 2009.

[10] Marichal, J. L (2000): *An Axiomatic Approach of the Discrete Choquet Integral as a Tool to Aggregate Interacting Criteria*. IEEE Trans. Fuzzy Systems, 8: 800-807.

[11] Moaven, S., Ahmadi, H., Habibi, J., Kamandi, A. (2008): *A Fuzzy Model for Solving Architecture Styles Selection Multi-Criteria Problem*. In: Proceedings of the Second UKSIM European Symposium on Computer Modeling and Simulation, IEEE Computer Society.

[12] Saaty, T. L. (1994): *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. Pittsburgh: RWS Publications.

[13] Saaty, T. L. (2001): *Decision Making with Dependence and Feedback: The Analytic Network process.* Pittsburgh: RWS Publications.

[14] Sridhar, P., Madni, A. M., Jamshidi, M. (2008): *Multi-Criteria Decision Making in Sensor Networks*. IEEE Instrumentation & Measurement Magazine, February 2008: 24-29.

Remarks

[1] Let us point out that for example the Analytic Network Process [13] allows for consideration of interactions and feedbacks within clusters and between them; however, this paper deals with interactions among criteria only.

COMPARISON OF GROUP METHODS IN ANALYTIC HIERARCHY PROCESS

Petra Grošelj, Lidija Zadnik Stirn
University of Ljubljana, Biotehnical Faculty
Jamnikarjeva 101, 1000 Ljubljana, Slovenia
e-mail: petra.groselj@bf.uni-lj.si, lidija.zadnik@bf.uni-lj.si

Abstract: Group decision making is an important topic and has been studied in distinct scientific fields. Analytic hierarchy process (AHP) is a method that has often been successfully used for solving multiple criteria group decision making problems. AHP was developed by Saaty in 1980. He suggested eigenvector method for deriving a priority vector from a pairwise comparison matrix which is generated by one decision maker. According to the literature and known applications, weighted geometric mean method (WGMM) is the only appropriate and employed method when the aggregation of individual judgments (AIJ) is used and the importance of the individuals' opinions is not the same. In recent years new approaches to this topic have been studied. In this paper, we discuss the weighted least squares group method (WLSGM) and data envelopment analysis (DEA) group method, and compare them with the WGMM on a numerical example.

Key words: multiple criteria decision making; group decision making; analytic hierarchy process

1 INTRODUCTION

Group decision making is an important topic and has been studied in distinct scientific fields. Srdjevic [18] stated that there are two basically distinct methodologies in group decision-making: multiple criteria decision making, which is useful in handling structured decision-making problems and the social choice theory with its voting systems when available information is minimal, unconfident, or predominantly qualitative.

Analytic hierarchy process (AHP) is one often and successfully used method for solving multiple criteria group decision making problems. AHP was developed by Saaty [17] and was first developed for one decision maker. There are different techniques to get a group valuation: consensus between the decision makers, compromise or voting and several aggregating methods.

Forman and Peniwati [11] showed that there are two basic aggregating methods: the aggregation of individual priorities (AIP) and the aggregation of individual judgments (AIJ). The aggregation of individual priorities (AIP) is suitable when the group wants to act as separate individuals. In this case weighted arithmetic mean method (WAMM) [15] is usually used, but weighted geometric mean can also be used [11]. The aggregation of individual judgments (AIJ) is suitable when the group wants to act together as an individual. Aczel and Saaty [2] showed that the geometric mean method (GMM) is the only method that satisfies some necessary axiomatic conditions. When the importance of the individual's opinions is not the same, Aczel and Alsina [1] showed that the weighted geometric mean method (WGMM) is the only appropriate method.

Some aggregating methods depend on the prioritization procedures used to get the priority vector from the comparison matrix. The most commonly used prioritization procedure is eigenvalue method [17]. When using the row geometric mean method [8] as the prioritization procedure Escobar et al. [9] claim that it is necessary to use WGMM, and AIJ and AIP give equal group priorities.

In recent years there were also some other new approaches developed: Escobar and Moreno-Jimenez [9] asserted that AIP and AIJ demand precise judgments from the decision makers, which is not a realistic requirement in highly complex problems. They proposed another method, aggregation of individual preference structures (AIPS). Cho and Cho [7] used the concept of Taguchi's loss function, Bryson and Joseph [5] used goal programming, Mikhailov [14] used fuzzy preference programming, and Sun and Greenberg [19] used the hybrid method, which produces group priorities by directly synthesizing the comparison matrices, while minimizing some weighted deviation measure. Wang and Chin [21] presented DEA methodology for the group decision-making, which is based on DEA method prioritization procedure.

This paper discusses Sun and Greenberg [19] hybrid method and Wang and Chin [21] DEA group method. They have not yet been used in any application. So we compare them with the classic WGMM, which has been used in many applications [3], [4], [12], [13] and others.

2 GROUP DEA METHOD

Ramanathan [16] developed a DEAHP method for deriving weights from comparison matrices. This method is based on DEA, [6]. Wang and Chin [21] showed on the numerical examples that this method has some drawbacks. They proposed a new DEA methodology (called DEA method) to overcome these drawbacks.

Wang and Chin [21] also extended the new DEA methodology to the group decision making. We assume that there are n criteria and m decision makers with comparison matrices $A^{(k)} = (a_{ij}{}^{(k)})_{n\times n}$, $k=1,...,m$. The importance of k-th decision maker's opinion is denoted by α_k, for $k=1,...,m$, with $\alpha_k > 0$ and $\sum_{k=1}^{m} \alpha_k = 1$. Wang and Chin [21] developed the following LP model (1):

$$\max \quad w_0 = \sum_{j=1}^{n} \left(\sum_{k=1}^{m} \alpha_k a_{0j}{}^{(k)} \right) x_j ,$$

$$\text{subject to:} \sum_{j=1}^{n} \left(\sum_{k=1}^{m} \sum_{i=1}^{n} \alpha_k a_{ij}{}^{(k)} \right) x_j = 1, \tag{1}$$

$$\sum_{j=1}^{n} \left(\sum_{k=1}^{m} \alpha_k a_{ij}{}^{(k)} \right) x_j \geq n x_i, \; i=1,...,n,$$

$$x_j \geq 0, \; j=1,...,n.$$

3 WEIGHTED LEAST SQUARES GROUP METHOD

Sun and Greenberg [19] developed a new method for producing group priorities which minimizes the weighted Minkowski distance:

$$S = \sqrt[q]{\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \varepsilon_{ij}{}^{(k)} \right|^q} \quad , \tag{2}$$

where $\varepsilon_{ij}{}^{(k)}$ represents the deviation of the entry a_{ij} from the k-th decision maker's comparison matrix from the unknown group priority and q, which varies from 1 to ∞, characterizes the distance measure from 1-norm to ∞-norm. In this case the error structure in (2) is of additive nature, which influences the additive normalization condition in model (3).

Sun and Greenberg [19] used Minkowski distance only for *q=2* because it leads to a closed-form expression. Sun and Greenberg [19] obtained the following constrained weighted least squares optimization model (3):

$$\min\ f(w) = \sum_{k=1}^{m} \alpha_k \sum_{j=1}^{n} \sum_{i=1}^{n} \left\| a_{ij}^{(k)} w_j - w_i \right\|_2 = \sum_{k=1}^{m} \sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_k \left(a_{ij}^{(k)} w_j - w_i \right)^2 \quad (3)$$

$$\text{subject to: } \sum_{i=1}^{n} w_i = 1, \quad w_i > 0,\ i=1,...n$$

We specified the optimization model (3) as weighted least squares group method.

Sun and Greenberg [19] proved the theorem which gives the following solution of the optimization model (3). Let

$$\tilde{A} = \left(\tilde{a}_{ij}\right)_{n\times n},\ \tilde{a}_{ij} = \sum_{k=1}^{m} \alpha_k a_{ij}^{(k)},$$

$$\hat{a}_{ij} = \sum_{k=1}^{m} \alpha_k \left(\left(a_{ij}^{(k)}\right)^2 + 1 \right),\ \eta_j = \sum_{i=1}^{n} \hat{a}_{ij} \text{ and } \Lambda = diag(\eta_1, \eta_2, \ldots, \eta_n).$$

Then the group preference is given by

$$w = C^{-1}\lambda, \quad (4)$$

where

$$C = \tilde{A} + \tilde{A}^T - \Lambda,\ C^{-1} = \left(\bar{c}_{ij}\right)_{n\times n}$$

and

$$\lambda = \left(\tfrac{\lambda}{2}, \tfrac{\lambda}{2}, ..., \tfrac{\lambda}{2}\right)^T,\ \lambda = 2 \Big/ \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \bar{c}_{ij} \right).$$

4 NUMERICAL EXAMPLE AND COMPARISON OF THE METHODS

In the following example we compare the group DEA method and the weighted least squares group method. Let suppose that we have a multiple criteria decision problem, presented in Figure 1. The objective of this problem is to select the best alternative regarding the given goal, criteria and subcriteria. In our example we are interested only in the ranking of the criteria: economic (en), ecological (el), educational (ed) and social (s). Thus *n*=4. There were 3 experts (*m*=3) who pairwise compared these criteria with respect to the goal of the problem. Their comparison matrices are the following:

$$A = \begin{bmatrix} 1 & 3 & 2 & \frac{1}{2} \\ \frac{1}{3} & 1 & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{2} & 5 & 1 & 1 \\ 2 & 5 & 1 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 1 \\ 2 & 1 & 3 & 3 \\ 2 & \frac{1}{3} & 1 & 3 \\ 1 & \frac{1}{3} & \frac{1}{3} & 1 \end{bmatrix} \quad C = \begin{bmatrix} 1 & \frac{1}{2} & 4 & 1 \\ 2 & 1 & 6 & 5 \\ \frac{1}{4} & \frac{1}{6} & 1 & \frac{1}{5} \\ 1 & \frac{1}{5} & 5 & 1 \end{bmatrix}$$

The consistency ratios, as the measure of the inconsistency of the comparison matrices, presented by Saaty [17], of these three matrices were calculated by the use of the Superdecision program [20], and amount to:

$CR_A = 0.0824$, $CR_B = 0.0656$ and $CR_C = 0.0728$

Since consistency ratios of all matrices are less than 0.1, all comparison matrices are of acceptable consistency, [17].

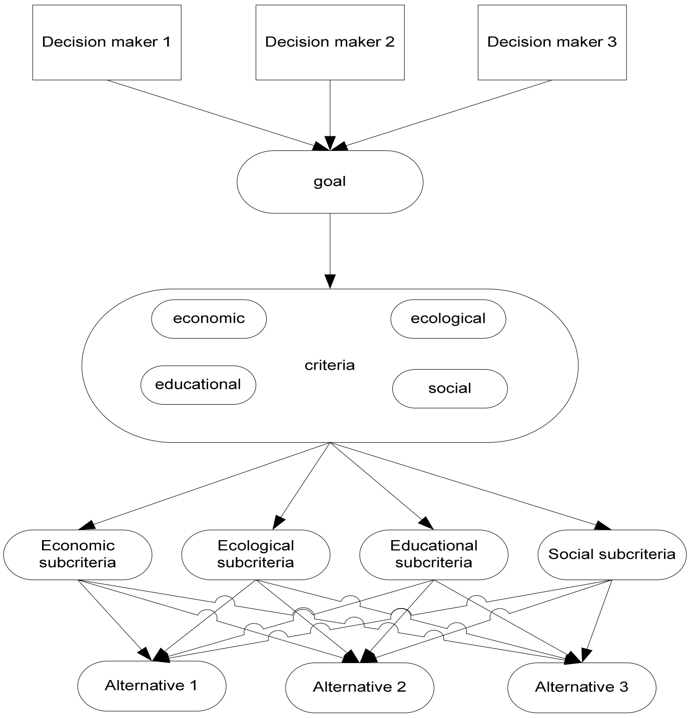


Figure 1: Hierarchical structure of the problem

The decision makers' opinions were not equally important. The relative importance of the first decision maker's opinion is *0.5*, and the relative importance of the other two decision makers' opinions is *0.25*.

Their comparison matrices were aggregated with the presented aggregating methods: the group DEA method and the weighted least squares group method.

4.1 Group DEA method

On the basis of the group DEA method, i.e., model (1), we get the following four linear programs with equal constraints and four different objective functions:

$$\begin{array}{lllllllll}
\max & w_1 & = & x_1 & + & \frac{7}{4}x_2 & + & \frac{17}{8}x_3 & + & \frac{3}{4}x_4 \\
\max & w_2 & = & \frac{7}{6}x_1 & + & x_2 & + & \frac{47}{20}x_3 & + & \frac{21}{10}x_4 \\
\max & w_3 & = & \frac{13}{6}x_1 & + & \frac{21}{8}x_2 & + & x_3 & + & \frac{13}{10}x_4 \\
\max & w_4 & = & \frac{3}{2}x_1 & + & \frac{79}{30}x_2 & + & \frac{11}{6}x_3 & + & x_4
\end{array}$$

$$\begin{array}{llllllllll} & \tfrac{215}{48}x_1 & + & \tfrac{961}{120}x_2 & + & \tfrac{877}{120}x_3 & + & \tfrac{103}{20}x_4 & = & 1 \\ & x_1 & + & \tfrac{7}{4}x_2 & + & \tfrac{17}{8}x_3 & + & \tfrac{3}{4}x_4 & \geq & x_1 \\ \text{subject to} & \tfrac{7}{6}x_1 & + & x_2 & + & \tfrac{47}{20}x_3 & + & \tfrac{21}{10}x_4 & \geq & x_2 \\ & \tfrac{13}{6}x_1 & + & \tfrac{21}{8}x_2 & + & x_3 & + & \tfrac{13}{10}x_4 & \geq & x_3 \\ & \tfrac{3}{2}x_1 & + & \tfrac{79}{30}x_2 & + & \tfrac{11}{6}x_3 & + & x_4 & \geq & x_4 \\ & x_1, x_2, x_3, x_4 \geq 0 \end{array}$$

These four linear programs were solved by Excel and the results are given in Table 1. Since the sum of the group preferences from group DEA method was not equal to *1*, we normalized the results so that they can be compared with the results of the other method.

Table 1: Priority vector and ranks obtained by group DEA method –model (1)

Group DEA method	Group preferences	Normalized group preferences	Rank
economic	0.2493	0.2150	4
ecological	0.3233	0.2788	1
educational	0.2794	0.2409	3
social	0.3077	0.2653	2

4.2. Weighted least squares group method

We calculated the group priorities on the basis of equations (4):

$$\tilde{A} = \begin{bmatrix} 1 & \tfrac{7}{4} & \tfrac{17}{8} & \tfrac{3}{4} \\ \tfrac{7}{6} & 1 & \tfrac{47}{20} & \tfrac{21}{10} \\ \tfrac{13}{6} & \tfrac{21}{8} & 1 & \tfrac{13}{10} \\ \tfrac{3}{2} & \tfrac{79}{30} & \tfrac{11}{6} & 1 \end{bmatrix}, \; \Lambda = \begin{bmatrix} 10.6962 & 0 & 0 & 0 \\ 0 & 34.6975 & 0 & 0 \\ 0 & 0 & 29.1103 & 0 \\ 0 & 0 & 0 & 16.9050 \end{bmatrix}$$

$$C^{-1} \doteq \begin{bmatrix} -0.1376 & -0.0203 & -0.0223 & -0.0319 \\ -0.0203 & -0.0365 & -0.0109 & -0.0169 \\ -0.0223 & -0.0109 & -0.0431 & -0.0159 \\ -0.0319 & -0.0169 & -0.0159 & -0.0806 \end{bmatrix}, \; \lambda \doteq \begin{bmatrix} -1.8719 \\ -1.8719 \\ -1.8719 \\ -1.8719 \end{bmatrix}$$

$$w = C^{-1}\lambda$$

The numbers in this example are rounded to four decimal places. The final group priorities of weighted least squares group method are given in Table 2:

Table 2: Priority vector and ranks for weighted least squares method

Weighted least squares group method	Group preferences	Rank
economic	0.3970	1
ecological	0.1583	4
educational	0.1726	3
social	0.2721	2

If we compare the results given in Table 1 with the ones given in Table 2, we see that there are significant differences between the group preferences and the ranking of the criteria. By the group DEA method the most important criterion with respect to our goal is ecological and the least important is economic. By the weighted least squares group method the most important criterion is economic and the least important is ecological.
Further, we compare the results with the most commonly used weighted geometric mean method (WGMM), where the prioritization method is eigenvalue method. First we combined the comparison matrices of all three decision makers with weighted geometric mean and so we got the following aggregation matrix:

$$\begin{bmatrix} 1 & \sqrt{\frac{3}{2}} & \sqrt[4]{8} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{2}{3}} & 1 & \sqrt[4]{\frac{18}{25}} & \sqrt[4]{\frac{3}{5}} \\ \sqrt[4]{\frac{1}{8}} & \sqrt[4]{\frac{25}{18}} & 1 & \sqrt[4]{\frac{3}{5}} \\ \sqrt{2} & \sqrt[4]{\frac{5}{3}} & \sqrt[4]{\frac{5}{3}} & 1 \end{bmatrix}$$

Then we used the Superdecision program [20], which uses the eigenvalue method for getting the priority vector from the comparison matrix. The results are in Table 3:

Table 3: Priority vector and ranks for WGMM

WGMM+EM method	group preferences	rank
economic	0.2741	2
ecological	0.2214	3
educational	0.2153	4
social	0.2892	1

These results also differ from the results in Table 1 and table 2. In this case the most important criterion with respect to our goal is social and the less important is educational. Since this method is most widely used, we anticipate that these results are appropriate.

5 DISCUSSION AND CONCLUSION

The comparison of three multiple criteria group decision making methods used within AHP that were discussed in the paper shows that there are significant differences in the results achieved by various group methods that we have dealt with. Our future work will be dedicated to investigating why there are such differences in results. One possible reason is that the two new methods have some imperfections that have to be corrected.

References

[1] J. Aczel, C. Alsina, On synthesis of judgments, Socio-Economic Planning Sciences, 20 (1986), pp. 333-339

[2] J. Aczel, T. Saaty, Procedures for synthesizing ratio judgments, Journal of Mathematical Psychology 27 (1983), pp. 93-102.

[3] J. Ananda, G. Herath, Multi-attribute preference modelling and regional land-use planning, Ecological Economics 65 (2008) pp. 325-335

[4] R. Aull-Hyde, S. Erdogan, J. M. Duke, An experiment on the consistency of aggregated comparison matrices in AHP, European Journal of Operational Research 171 (2006), pp. 290-295

[5] N. Bryson, A. Joseph, Generating consensus priority point vectors: a logarithmic goal programming approach, Computers & Operations Research 26 (1999), pp. 637-643

[6] A. Charnes, W.W. Cooper, E. Rhodes: Measuring the efficiency of decision making units, European Journal of Operational Research 2 (1978), pp. 429-444

[7] Y.-G. Cho, K.-T. Cho, A loss function approach to group preference aggregation in the AHP, Computers & Operations Research 35 (2008), pp. 884-892

[8] G. B. Crawford, The geometric mean procedure for estimating the scale of judgment matrix, Mathematical Modelling 9 (1987), pp. 327-334

[9] M. T. Escobar, J. Aguaron, J. M. Moreno-Jimenez, A note on AHP group consistency for the row geometric mean prioritization procedure, European Journal of Operational Research 153 (2004), pp. 318-322

[10] M. T. Escobar, J. M. Moreno-Jimenez, Aggregation of individual preference structures in AHP-group decision making, Group Decision and Negotiation 16 (2007), pp. 287-301

[11] E. Forman, K. Peniwati, Aggregating individual judgements and priorities with the Analytic Hierarchy Process, European Journal of Operational Research 108 (1998), pp. 165-169.

[12] J. Korpela, M. Tuominen, Inventory forecasting with a multiple criteria decision tool, International Journal of Production Economics 45 (1997), pp. 159–168.

[13] M.J. Liberatore, R.L. Nydick, Group decision making in higher education using the analytic hierarchy process, Research in higher education 38 (1997), pp. 593-614

[14] L. Mikhailov, Group prioritization in the AHP by fuzzy preference programming method, Computers & Operations Research 31 (2004), pp. 293–301

[15] R. Ramanathan, L.S. Ganesh, Group preference aggregation methods employed in AHP: an evaluation and intrinsic process for deriving members' weightages, European Journal of Operational Research 79 (1994), pp. 249-265.

[16] R. Ramanathan, Data envelopment analysis for weight derivation and aggregation in the analytic hierarchy process, Computers and Operations Research 33 (2006), pp. 1289-1307

[17] T.L. Saaty, Multicriteria Decision Making: The Analytic Hierarchy Process., McGraw-Hill, New York (1980)

[18] B. Srdjevic, Linking analytic hierarchy process and social choice methods to support group decision-making in water management, Decision Support Systems 42 (2007), pp. 2261–2273

[19] L. Sun, B. S. Greenberg, Multicriteria Group Decision Making: Optimal Priority Synthesis from Pairwise Comparisons, Journal of optimization theory and applications 130 (2006), pp. 317–338

[20] Superdecisions: http://www.superdecisions.com

[21] Y.M. Wang, K.S. Chin, A new data envelopment analysis method for priority determination and group decision making in the analytic hierarchy process, European Journal of Operational Research 195 (2009), pp. 239-250

FEED BLEND OPTIMIZATION BY FUZZY MULTICRITERIA LINEAR PROGRAMMING METHOD

Tunjo Perić
Bakeries Sunce, Rakitje, Rakitska cesta 98, 10437 Bestovje, Croatia
E-mail: tunjo.peric1@zg.t-com.hr

Abstract: This paper (1) considers multicriteria problem of feed blend optimization in vague conditions, (2) forms a general multicriteria programming model for feed blend optimization by MLP methods, (3) proposes and additive fuzzy asymmetrical MLP model for feed blend optimization in vague conditions, (4) applies the proposed model in solving the concrete problem of feed blend optimization, (5) points to the advantages and shortcomings of the model usage through sensitivity analysis of compromise solutions.

Keywords: multicriteria, fuzzy approach, optimization criteria, feed blend.

1 INTRODUCTION

Feed blend is a blend of ingredients that is used to feed livestock. While the feed has to meet the nutritional requirements of livestock to allow maximal weight gain its industrial production has to be economical, which can only be ensured by an optimal blending of ingredients. Optimization of feed blend in terms of both quality and economy can be carried out by use of mathematical optimization methods. Application of mathematical methods in feed blend optimization will result in a fast and efficient solution for the optimal combination of ingredients considering the nutritive needs of animals, available sources of feed, and cost reduction. The most frequently used method in feed blend optimization has been linear programming minimizing the cost function for ingredients where minimal and maximal needs for nutrients are the constraints. More about application of multicriteria programming method can be seen in the study by [3], in which the author deals with the feed blend optimization as a multicriteria problem. However, the use of deterministic methods has some shortcomings due to the vague character of the problem.

The aim of this study is to: (1) investigate the feed blend optimization criteria and point to the multicriteria character of the problem, (2) build a fuzzy MLP model for feed blend optimization, (3) on a concrete example point to the shortcomings of the deterministic methods and the advantages and limitations of the proposed model, and (4) indicate the unsolved issues in application of the fuzzy MLP in this field.

2 CHOICE OF FEED BLEND OPTIMIZAATION CRITERIA

When selecting the feed blend optimization criteria we must have in mind the following factors:
- the blend production costs,
- the needs of animals for which the feed blend is prepared
- the feed blend quality.

It would be ideal if the production costs were minimal, the needs of animals completely satisfied, and the produced blend were of maximal quality. This leads us to the optimization criteria:

1. production costs expressed in monetary units,
2. the share of ingredients necessary to satisfy the needs for maximal weight gain, and
3. the share of ingredients that negatively affect the blend quality and thus also the weight gain.

3 GENERAL MLP MODEL FOR FEED BLEND OPTIMIZATION

If we want to solve the feed blend optimization problem for a particular sort and category of livestock by application of MLP methods, we must start from the following:

- The criteria for feed blend optimization are given (they are determined in collaboration with nutritionists)
- The feed blend has to satisfy the nutritional needs of the given sort and category of livestock.
- A certain number of ingredients are available.
- The share of single ingredients in the feed blend is limited.
- There is no constraint in terms of availability of ingredients.
- The ingredients have been analyzed and it is assumed that the sample analysis corresponds to the total quantity of any ingredient.

Let us introduce the following marks:

- f_j = optimization criteria functions, $(j = 1,\ldots,k)$;
- m = number of various requirements for nutrients in a given sort and category of livestock;
- b_l = the need for a nutrient of l-kind in the blend unit, $(l = 1,\ldots,m)$;
- n = number of ingredients;
- c_{ij} = i-coefficient of j-criterion function, $(i = 1,\ldots,n;\ j = 1,\ldots,k)$;
- x_i = share of single ingredient in the blend, $(i = 1,\ldots,n)$;
- a_{il} = quantity of l-nutrient per unit of i-ingredient, $(i = 1,\ldots,n;\ l = 1,\ldots,m)$;
- u_i = maximal share of i-ingredient in the blend.

Considering the above we can formulate the MLP model:

$$(\min)\underline{f}^1 = \left[f_1, f_2, \ldots, f_p \right] = \left[c_{i1}x_i, c_{i2}x_i, \ldots, c_{ip}x_i \right] \quad (1)$$

$$(\max)\underline{f}^2 = \left[f_{p+1}, f_{p+2}, \ldots, f_k \right] = \left[c_{i,p+1}x_i, c_{i,p+2}x_i, \ldots, c_{ik}x_i \right] \quad (2)$$

$$\text{s.t.} \quad \sum_{i=1}^{n} a_{il}x_i \geq b_l,\ (l = 1,\ldots,m);\ \sum_{i=1}^{n} x_i = 1,^{1} \qquad 0 \leq x_i \leq u_i,\ \ (i = 1,\ldots,n), \quad (3)$$

where $f_1, f_2, \ldots, f_p$ are criteria functions to be minimized such as production costs and the share of ingredients that negatively affect the blend quality and thus also the weight gain of the given livestock, while $f_{p+1}, f_{p+2}, \ldots, f_k$ are criteria functions to be maximized such as the share of ingredients necessary to meet the nutritional needs and the weight gain of the livestock.

The above model is an MLP model. It is the basis for solving the feed blend optimization problems by MLP methods.

The model application (1 – 3) assumes the ability of the decision maker to select the preferred solution, which means that the decision maker is familiar with the model and able to recognize the solution that gives preferred values to criteria functions. However, in most real cases the decision maker is not able to determine precisely the contribution of criteria functions to the general goal of business operation, and he can hardly be expected to recognize the trade-off between particular criteria functions. Therefore it is necessary to

---

[1] This is set in cases when the blend recipe is drawn up. The left side of this constraint may be required to be less than 1 for the contents of any additives if they are included in the blend.

develop a fuzzy multicriteria model and to solve it as the fuzzy multicriteria programming model.

Fuzzy decision making can be symmetrical and asymmetrical. In symmetrical fuzzy decisions differences between importance of criteria functions and constraints are not introduced, while in asymmetrical multicriteria decisions criteria functions and constraints do not have equal value and have different weights [7], [8] and [5].

4 FUZZY MLP MODEL FOR FEED BLEND OPTIMIZATION

The general model of multicriteria feed blend optimization (1 – 3) may be presented as:
Find the vector $\underline{x}$ which minimizes the criteria function f_r and maximizes the criteria function f_s with

$$f_r = \sum_{i=1}^{n} c_{ri} x_i, \qquad r = 1,2,\ldots,p \tag{4}$$

$$f_s = \sum_{i=1}^{n} c_{si} x_i, \qquad s = p+1, p+2,\ldots,q \tag{5}$$

and constraints:

$$\underline{x} \in X_d, \quad X_d = \left\{ \underline{x} \mid g_l(\underline{x}) = \sum_{i=1}^{n} a_{il} \geq b_l, \sum_{i=1}^{n} x_i = 1, 0 \leq x_i \leq u_i, l = 1,2,\ldots,m, i = 1,2,\ldots,n \right\}, \tag{6}$$

where $c_{ri}, c_{si}, a_{il}, b_l$ and u_i are crisp or fuzzy values.

Zimmermann [7] solved the problem (4 – 6) by fuzzy linear programming. He formulated the fuzzy linear program by determining for each criteria function f_j its maximal value f_j^+ and its minimal value f_j^-, solving:

$$f_r^+ = \max f_r, \quad \underline{x} \in X_a, \quad f_r^- = \min f_r, \quad \underline{x} \in X_d \tag{7}$$

$$f_s^+ = \max f_s, \quad \underline{x} \in X_d, \quad f_s^- = \min f_s, \quad \underline{x} \in X_a \tag{8}$$

f_r^-, f_s^+ are obtained by solving multicriteria model as a linear programming model separately minimizing particular criteria functions and maximizing particular criteria functions, $\underline{x} \in X_d$ means that solutions have to satisfy the constraining conditions, while X_a is a set of all optimal solutions obtained by solving particular criteria functions.

As for each criteria function f_j its value is changed linearly from f_j^- to f_j^+, this value can be observed as a fuzzy number with the linear membership function $\mu_{f_j}(\underline{x})$. Consequently, the MLP model (4 – 6) with fuzzy goals and fuzzy constraints can be presented as:

$$\tilde{\tilde{f}}_r = \sum_{i=1}^{n} c_{ri} x_i \leq\approx f_r^0, \qquad r = 1,2,\ldots,p \tag{9}$$

$$\tilde{\tilde{f}}_s = \sum_{i=1}^{n} c_{si} x_i \geq\approx f_s^0, \qquad s = p+1, p+2,\ldots,q \tag{10}$$

s.t.

$$g_l(\underline{x}) = \sum_{i=1}^{n} a_{il} \geq\approx b_l, \quad l = 1,\ldots,m \tag{11}$$

$$\sum_{i=1}^{n} x_i = 1, \; 0 \leq x_i \leq u_i, \quad i = 1,\ldots,n. \tag{12}$$

In this model the sign $\approx$ indicates fuzzy environment. The symbol $\leq\approx$ indicates the fuzzy version of the sign $\leq$, and is interpreted as "essentially smaller than or equal to" while the symbol $\geq\approx$ is interpreted as "essentially greater than or equal to". f_r^0 and f_s^0 represent aspiration levels of criteria functions that would be achieved by the decision maker.

If we assume that the membership functions, which are based on preference of satisfaction, are linear, then the linear membership functions for criteria functions and constraints can be presented as follows:

$$\mu_{f_r}(\underline{x}) = \begin{cases} 1 & \text{for } f_r \leq f_r^- \\ (f_r^+ - f_r(\underline{x}))/(f_r^+ - f_r^-) & \text{for } f_r^- \leq f_r(\underline{x}) \leq f_r^+, \quad r = 1,2,\ldots,p \\ 0 & \text{for } f_r \geq f_r^+ \end{cases} \tag{13}$$

$$\mu_{f_s}(\underline{x}) = \begin{cases} 1 & \text{for } f_s \geq f_s^+ \\ (f_s(\underline{x}) - f_s^-)/(f_s^+ - f_s^-) & \text{for } f_s^- \leq f_s(\underline{x}) \leq f_s^+, \quad s = p+1, p+2,\ldots,k \\ 0 & \text{for } f_s \leq f_s^- \end{cases} \tag{14}$$

$$\mu_{g_l}(\underline{x}) = \begin{cases} 1 & \text{for } g_l(\underline{x}) \geq b_l^+, \\ (g_l(\underline{x}) - b_l^-)/(b_l^+ - b_l^-) & \text{for } b_l^- \leq g_l(\underline{x}) \leq b_l^+ \quad l = 1,2,\ldots,m \\ 0 & \text{for } g_l(\underline{x}) \leq b_l^-. \end{cases} \tag{15}$$

where $b_l^- = b_l$, and $b_l^+ = b_l + d_l$. d_l are subjectively determined constants that express the borders of allowed deviations of l inequality (tolerance interval).

In the fuzzy programming model, according to Zimmermann's approach, the fuzzy solution represents an intersection of all the fuzzy sets that represent fuzzy criteria functions and fuzzy constraints. The fuzzy solution for all the fuzzy goals and fuzzy constraints is given as follows:

$$\mu_D(\underline{x}) = \left\{ \left\{ \bigcap_{j=1}^{k} \mu_{f_j}(\underline{x}) \right\} \cap \left\{ \bigcap_{i=1}^{m} g_l(\underline{x}) \right\} \right\}. \tag{16}$$

The optimal solution ($\underline{x}^*$) is:

$$\mu_D(\underline{x}^*) = \max_{\underline{x} \in X_D} \mu_D(\underline{x}) = \max_{\underline{x} \in X_D} \min \left[\min_{j=1,\ldots,k} \mu_{f_j}(\underline{x}), \min_{l=1,\ldots,m} \mu_{g_l}(\underline{x}) \right]. \tag{17}$$

The optimal solution ($\underline{x}^*$) of the above model is equivalent to the solving of the following linear programming model [7]:

$$(\max)\lambda \tag{18}$$

s.t.

$$\lambda \leq \mu_{f_j}(\underline{x}), \quad j = 1,2,\ldots,k \tag{19}$$

$$\lambda \leq \mu_{g_l}(\underline{x}), \quad l = 1,2,\ldots,m \tag{20}$$

$$\sum_{i=1}^{n} x_i = 1; \; 0 \leq x_i \leq u_i, \quad i = 1,\ldots,n; \; \lambda \in [0,1], \tag{21}$$

where $\mu_D(\underline{x})$ represents the membership function for the optimal solution, $\mu_{f_j}(\underline{x})$ represents the membership functions for criteria functions and $\mu_{g_l}(\underline{x})$ represents the membership functions for constraints. In this model the relationship between constraints and criteria functions is completely symmetrical [7], and here the decision maker cannot express the relative value of criteria functions and constraints.

In order to express the relative importance of criteria functions and constraints we have to solve the so called weight additive model in which weights present utility functions of criteria functions and constraints [2], [5], [6] and [1].

The convex fuzzy model proposed by [2] and [5] and the weight additive model, by [6] is

$$\mu_D(\underline{x}) = \sum_{j=1}^{k} w_j \mu_{f_j}(\underline{x}) + \sum_{l=1}^{m} \beta_l \mu_{g_l}(\underline{x}), \tag{22}$$

$$\sum_{j=1}^{k} w_j + \sum_{l=1}^{m} g_l = 1, \quad w_j, \beta_l \geq 0, \tag{23}$$

where w_j and β_l are weight coefficients representing the relative importance between the fuzzy criteria functions and fuzzy constraints.

To solve the above fuzzy model we will use the following linear programming model:

$$(\max) f = \sum_{j=1}^{k} w_j \lambda_j + \sum_{l=1}^{m} \beta_l \gamma_l \tag{24}$$

s.t.

$$\lambda_j \leq \mu_{f_j}(\underline{x}), \quad j = 1,2,\ldots,k, \tag{25}$$

$$\gamma_l \leq \mu_{g_l}(\underline{x}), \quad l = 1,2,\ldots,m, \tag{26}$$

$$\sum_{i=1}^{n} x_i = 1; \ 0 \leq x_i \leq u_i, \quad i = 1,\ldots,n; \tag{27}$$

$$\lambda_j, \beta_l \in [0,1], \quad j = 1,2,\ldots,k; \ l = 1,2,\ldots,m, \tag{28}$$

$$\sum_{j=1}^{k} w_j + \sum_{l=1}^{m} g_l = 1, \quad w_j, \beta_l \geq 0, \tag{29}$$

5 CASE STUDY

5.1 Input data for determination of feed blend for pig fattening (PS-2)

Our case processes the given data required to work out the optimal feed plan (feed blend) for pig fattening PS-2. The mark PS-2 represents the blend recipe for pigs of 20-50 kg, while PS-1 stands for the recipe for pigs up to 20 kg, and PS-3 for pigs of 50-100 kg, etc. The meal has to contain minimal and maximal shares of daily nutrients. Determination of maximal and minimal share of nutrients in the blend is based on scientific research. The given data are shown in the Tables 1, 2 and 3.

The sorts of feed used to prepare the feed blend for this kind of livestock (pigs of 20-50 kg), their price per unit, and the percentage of nutrients and water per ingredient unit are shown in the Table 1. The total cost has to be minimized, the share of nutrients in the blend has to be maximised, and the share of water in the optimal meal has to be minimized.

The nutrients needed in the feed used for growing pigs and the required quantities (as suggested by nutritionists) are shown in the Table 2. Some of the ingredients are required in minimal and some in maximal quantities.

The Table 3 is a nutrition matrix and its elements a_{ik} are the contents of a particular nutrient in the feed unit.

Table 1: Sorts of feed (PS-2)

Sorts of feed		Price - c_{i1} (min)	Nutrients - c_{i2} (max)	Water - c_{i3} (min)
H1	Barley	1.75	70	11.0
H2	Maize	1.75	80	12.0
H3	Lucerne	1.65	32	6.9
H4	Powdered milk	6	86	8.4
H5	Fish meal	9	69	9.0
H6	Soya	2.7	92	10.0
H7	Soya hulls	3.5	79	11.0
H8	Dried whey	9	78	6.0
H9	Rape pellets	1.8	66	8.0
H10	Wheat	1.8	79	12.0
H11	Rye	1.8	75	11.4
H12	Millet	3.5	65	10.0
H13	Sunflower pellets	1.8	68	7.0

Table 2: Needs for nutrients

Nutrients		Constraint type	Min or max requirement - b_k
E1	Raw protein	≥	14.0
E2	Pulp	≤	7.0
E3	Calcium - Ca	≤	0.80
E4	Phosphorus – P	≥	0.50
E5	Ash	≤	7.0
E6	Metionin	≥	0.50
E7	Lizin	≥	0.74
E8	Triptofan	≥	0.11
E9	Treonin	≥	0.45
E10	Izoleucin	≥	0.52
E11	Histidin	≥	0.23
E12	Valin	≥	0.46
E13	Leucin	≥	0.77
E14	Arginin	≥	0.55
E15	Fenkalanin	≥	0.54

Table 3: Nutrition matrix (a_{ik})

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13
E1	11.5	8.9	17.0	33	61	38	42	12	36	13.5	12.6	11.0	42
E2	5.0	2.9	24.0	0.0	1.0	5.0	6.5	0.0	13.2	3.0	2.8	10.5	13.0
E3	0.08	0.01	1.3	1.25	7.0	0.25	0.2	0.87	0.6	0.05	0.08	0.1	0.4
E4	0.42	0.25	0.23	1.0	3.5	0.59	0.6	0.79	0.93	0.41	0.3	0.35	1.0
E5	2.5	1.5	9.6	8.0	24	4.6	6.0	9.7	7.2	2.0	1.45	4.0	7.7
E6	0.18	0.17	0.28	0.98	1.65	0.54	0.6	0.2	0.67	0.25	0.16	0.2	1.5
E7	0.53	0.22	0.73	2.6	4.3	2.4	2.7	1.1	2.12	0.4	0.4	0.4	1.7

E8	0.17	0.09	0.45	0.45	0.7	0.52	0.65	0.2	0.46	0.18	0.14	0.18	0.5
E9	0.36	0.34	0.75	1.75	2.6	1.69	1.7	0.8	1.6	0.35	0.36	0.28	1.5
E10	0.42	0.37	0.84	2.1	3.1	2.18	2.8	0.9	1.41	0.69	0.53	0.53	2.1
E11	0.23	0.19	0.35	0.86	1.93	1.01	1.1	0.2	0.95	0.17	0.27	0.18	1.0
E12	0.62	0.42	1.04	2.38	3.25	2.02	2.2	0.7	1.81	0.69	0.62	0.62	2.3
E13	0.8	1.0	1.3	3.3	4.5	2.8	3.8	1.2	2.6	1.0	0.7	0.9	2.6
E14	0.5	0.52	0.75	1.1	4.2	2.8	3.2	0.4	2.04	0.6	0.5	0.8	3.5
E15	0.62	0.44	0.91	1.58	2.8	2.1	2.1	0.4	1.41	0.78	0.62	0.62	2.2

5.2 Formulation and solving of the multi-criteria and fuzzy programming model for determination of the feed blend

Based on the data given in the above tables and in compliance with the requirements of the decision makers the multi-criteria linear programming model is formulated which minimises the function of blend costs and the function of the water share in the blend while maximising the function of the total nutrients in the blend.

The multi-criteria model thus includes three goal functions, 14 decision variables, and 16 constraints. Namely, the last constraint is the relation (3) from the starting model with the right side of constraint of 0.97. Namely, the diet plan always includes 3% of various vitamin additives disregarding the ingredients included in the optimal meal. The model also includes additional constraints on the quantity of particular sorts of feed. Namely, to make the diet plan as heterogeneous as possible the model limits the share of any ingredient to maximally 15%. Thus if x_i is the share of *i* feed in the optimal blend the constraints $x_i \leq 0.15, (i =$ 1,...,13) are introduced.

The multi-criteria programming model is easily established on the basis of the data from the Tables 1, 2, and 3. The coefficients of the three criteria functions are given in the Table 1, while the constraints are included by following the data from the Tables 2 and 3. The problem is first solved separately in terms of each criteria function, i.e. the so called marginal solutions are determined. The optimal solutions in terms of each criteria function are shown in the Table 4. The Table 5 is the payoff table, i.e. it shows the values of all the three criteria functions for the obtained marginal solutions.

Table 4: Marginal solutions

	Costs(min)	Nutrients(max)	Water(min)
	x_1^*	x_2^*	x_3^*
Barley	0.15	0	0
Maize	0.15	0.15	0
Lucerne	0.026	0	0.082
Powdered milk	0	0.15	0.15
Fish meal	0	0.0522	0.0192
Soya	0.1215	0.15	0.15
Soya hulls	0	0.15	0
Dried whey	0	0.15	0.15
Rape pellets	0.15	0	0.15
Wheat	0.15	0.15	0
Rye	0.0725	0.0178	0.1188
Millet	0	0	0
Sunflower pellets	0.15	0	0.15

Table 5: Payoff table

	x_1^*	x_2^*	x_3^*
Costs – f_1	**1.83645**	4.21434	3.71694
Nutrients – f_2	71.8975	**79.0368**	71.3588
Water - f_3	9.7209	9.58272	**8.00292**

The diagonal of the payoff table shows the optimal values of the single functions (ideal point). It is obvious that each of the proposed solutions is good only in terms of „its" criteria function. Thus for example the solution (blend) x_2^* involves 2.3 times higher costs, and the solution x_3^* twice higher cost than the optimally possible solution x_1^*. In order to obtain the so called best compromise solution the problem has to be solved by one of the multicriteria programming methods. In agreement with the decision maker (farm owner) it was decided to try to solve the problem by reformulating it into a corresponding fuzzy programming model.

To form the fuzzy MLP model we use the data from the payoff table (Table 5), where $f_1^- = f_1^* = 1.83645$; $f_1^+ = f_1^{\max} = 4.21434$; $f_2^+ = f_2^* = 79.0368$; $f_2^- = f_2^{\min} = 71.3588$; $f_3^- = f_3^* = 8.00292$; $f_3^+ = f_3^{\max} = 9.7209$, and Table 2, where in the constraint type „$\geq$" $b_l^- = b_l$, and $b_l^+ = 1.1 \cdot b_l$, while in the constraint type „$\leq$" $b_l^- = 0.9 \cdot b_l$, and $b_l^+ = b_l$.

Using the above data, and according to the relations (13-15), we form the membership functions for criteria functions and constraints. We will here show the criteria functions for the criteria function f_1 and the constraint E_1:

$$\mu_{f_1}(\underline{x}) = \begin{cases} 1 & \text{for } f_1(\underline{x}) \leq 1.83645 \\ (4.21434 - f_1(\underline{x}))/(4.21434 - 1.83645) & \text{for } 1.83645 \leq f_1(\underline{x}) \leq 4.21434, \\ 0 & \text{for } f_1(\underline{x}) \geq 4.21434 \end{cases} \tag{30}$$

$$\mu_{E_1}(\underline{x}) = \begin{cases} 1 & \text{for } E_1(\underline{x}) \geq 15.4 \\ (E_1(\underline{x}) - 14)/(15.4 - 14) & \text{for } 14 \leq E_1(\underline{x}) \leq 15,4 \\ 0 & \text{for } E_1(\underline{x}) \leq 14 \end{cases} \tag{31}$$

Based on the formed membership functions and considering the information on the relative importance of criteria functions and constraints obtained from the decision maker ($w_1 = 0.4$, $w_2 = 0.20$, $w_3 = 0.10$, $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$), we from the additive model of fuzzy MLP:

$$(\max) f = 0.40\lambda_1 + 0.20\lambda_2 + 0.10\lambda_3 + 0.02\gamma_1 + 0.02\gamma_2 + \ldots + 0.02\gamma_{15} \tag{32}$$

s.t.

$$\lambda_1 \leq \mu_{f_1}(\underline{x}),\ \lambda_2 \leq \mu_{f_2}(\underline{x}),\ \lambda_3 \leq \mu_{f_3}(\underline{x}),\ \gamma_1 \leq \mu_{E_1}(\underline{x}),\ \gamma_2 \leq \mu_{E_2}(\underline{x}),\ \gamma_3 \leq \mu_{E_3}(\underline{x}),$$

$$\gamma_4 \leq \mu_{E_4}(\underline{x}),\ \gamma_5 \leq \mu_{E_5}(\underline{x}),\ \gamma_6 \leq \mu_{E_6}(\underline{x}),\ \gamma_7 \leq \mu_{E_7}(\underline{x}),\ \gamma_8 \leq \mu_{E_8}(\underline{x}),\ \gamma_9 \leq \mu_{E_9}(\underline{x}),$$

$$\gamma_{10} \leq \mu_{E_{10}}(\underline{x}),\ \gamma_{11} \leq \mu_{E_{11}}(\underline{x}),\ \gamma_{12} \leq \mu_{E_{12}}(\underline{x}),\ \gamma_{13} \leq \mu_{E_{13}}(\underline{x}),\ \gamma_{14} \leq \mu_{E_{14}}(\underline{x}),\ \gamma_{15} \leq \mu_{E_{15}}(\underline{x}), \tag{33}$$

$$\sum_{i=1}^{13} x_i = 0.97,\ 0 \leq x_i \leq 0.15,\ \lambda_1, \lambda_2, \lambda_3, \gamma_1, \ldots, \gamma_{15} \in [0, 1]. \tag{34}$$

The model (32-34) is an LP model whose compromise solution is shown in the following table:

Table 6: Compromise solution: values of variables and criteria functions

Solution	λ	Values of variables	Values of criteria functions		
			f_1	f_2	f_3
$\underline{x}_1^{com}$	$\lambda_1 = 0.8943,$ $\lambda_2 = 0.4478,$ $\lambda_3 = 0.1605,$ $\lambda_4 = \lambda_5 = \ldots = \lambda_{18} = 1$	$x_2 = 0.15,\ x_4 = 0.0382,$ $x_6 = 0.15,\ x_9 = 0.15,$ $x_{10} = 0.15,\quad x_{11} = 0.15,$ $x_{13} = 0.15$	2.088 (113.70% of f_1^*)	74.7974 (94.64% of f_2^*)	9.73068 (121.59% of f_3^*)

From the Table 5 it is obvious that the function f_1 has coefficient $\lambda_1 = 0.8943$, the function f_2 coefficient $\lambda_2 = 0.4478$, and the function f_3 coefficient $\lambda_3 = 0.1605$. This is the consequence of the chosen values of weight coefficients: $w_1 = 0.4$, $w_2 = 0.20$, $w_3 = 0.10$. The values $\gamma_1 = \gamma_2 = \ldots = \gamma_{15} = 1$ are the consequence of "weak" constraints as well as the deviation from constraints defined by the decision maker when forming membership functions.

In order to show that the values λ_1, λ_2, λ_3 and f_1, f_2, f_3 reflect the values of weight coefficients determined by the decision maker, we solve several models with different values of weight coefficients. Some compromise solutions are shown in the following table:

Table 7: Compromise solutions: values for *w*, λ, f_1, f_2 and f_3

Solution	w	λ	Values of criteria functions		
			f_1	f_2	f_3
$\underline{x}_2^{com}$	$w_1 = 0.10,\ w_2 = 0.30,$ $w_3 = 0.30$ $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$	$\lambda_1 = 0.1574,$ $\lambda_2 = 0.6891,$ $\lambda_3 = 0.7099,$	3.84 (209.10% of f_1^*)	76.65 (96.98% of f_2^*)	8.60 (107.46% of f_3^*)
$\underline{x}_4^{com}$	$w_1 = 0.20,\ w_2 = 0.30,$ $w_3 = 0.20$ $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$	$\lambda_1 = 0.4003,$ $\lambda_2 = 0.8728,$ $\lambda_3 = 0.2823,$	3.2625 (177.65% of f_1^*)	78.06 (98.76% of f_2^*)	9.48 (118.46% of f_3^*)
$\underline{x}_6^{com}$	$w_1 = 0.30,\ w_2 = 0.25,$ $w_3 = 0.15$ $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$	$\lambda_1 = 0.7195,$ $\lambda_2 = 0.6852,$ $\lambda_3 = 0.1852,$	2.5035 (136.32% of f_1^*)	76.62 (96.94% of f_2^*)	9.68 (120.96% of f_3^*)
$\underline{x}_9^{com}$	$w_1 = 0.50,\ w_2 = 0.10,$ $w_3 = 0.10$ $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$	$\lambda_1 = 0.9859,$ $\lambda_2 = 0.3310,$ $\lambda_3 = 0.1123,$	1.87 (101.83% of f_1^*)	73.90 (93.50% of f_2^*)	9.83 (122.83% of f_3^*)
$\underline{x}_{11}^{com}$	$w_1 = 0.60,\ w_2 = 0.05,$ $w_3 = 0.05$ $\beta_1 = \beta_2 = \ldots = \beta_{15} = 0.02$	$\lambda_1 = 0.9859,$ $\lambda_2 = 0.3324,$ $\lambda_3 = 0.1083,$	1.869865 (101.82% of f_1^*)	73.9108 (93.51% of f_2^*)	9.8381 (122.93% of f_3^*)

Figure 1 illustrates the indicators from the Table 7.

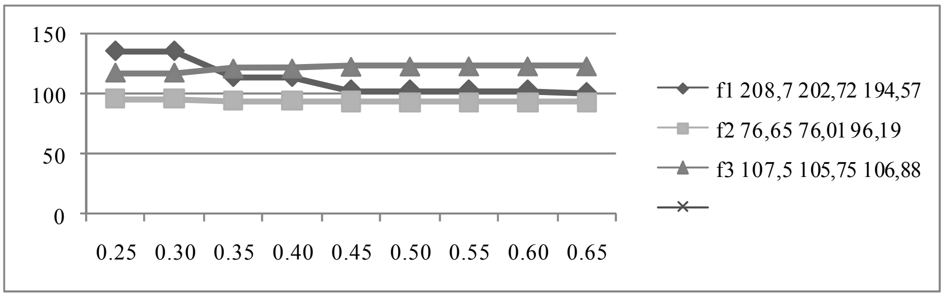


Figure 1: Graph showing the values of criteria functions and weight coefficients

From the Table 7 we can see that the increase of weight w_1 increases the value of the function f_1, therefore it is obvious that the application of the proposed model can provide us with the compromise solution which better reflects the preferences of the decision maker.

6 CONCLUSION

It can be concluded that feed blend optimization is a multicriteria problem with several conflicting goals which can be solved by different multicriteria optimization methods. As the values of criteria functions and constraints are expressed in vague conditions the application of deterministic models may not be the best one. Consequently we need to use the fuzzy MLP method. The fuzzy MLP model proposed in this work is may be used in cases when the decision maker can precisely determine his preferences in fulfillment of values of criteria functions and constraints which are essentially vague. The model is also suitable for creation of a set of compromise solutions allowing the decision maker to select the preferred solution, or a method of multi-attribute decision making can be used to select the preferred solution.

Non-linearity of membership functions and fuzzy weights of criteria functions and constraints in the proposed model are still open for further research.

References

[1] Amid, A., Ghodsypour, S. H., Brien, C. O., 2006., Fuzzy multiobjective linear model for supplier selection in supply chain, International journal of production economics, No. 104, p. 394-407.

[2] Bellman, R.G., Zadeh, L.A., 1970. Decision making in fuzzy environment, Management Sciences 17, B141-B164.

[3] Perić, T., 2008. Multi-criteria Programming - Methods and Applications (in Croatian), Alka script, Zagreb.

[4] Perić, T., Babić, Z., 2009. Determining Optimal Production Program with a Fuzzy Multiple Criteria Programming Method, Proceedings of International MultiConference of Engineers and Computer Scientists 2009, Hong Kong, p. 2006-2013.

[5] Sakawa, M., 1993. Fuzzy Sets and Interactive Multiobjective Optimization, Plenum Press, New York.

[6] Tiwari, R.N., Dharmahr, S., Rao, J.R., 1987. Fuzzy goal programming – an aditive model, Fuzzy Sets and Systems 24, 27-34.

[7] Zimmermann, H.J., 1978. Fuzzy programming and linear programming with severel objective functions. Fuzzy Sets and System 1, 45-55.

[8] Zimmermann, H.J., 1987. Fuzzy Sets, Decision Making and Expert Systems. Kluwer Academic publishers, Boston.

[9] Zimmermann, H.J., 1993. Fuzzy Sets Theory and its Applications, Kluwer Academic Publishers, Boston.

LOGISTIC REGRESSION AND MULTICRITERIA DECISION MAKING IN CREDIT SCORING

Nataša Šarlija
University of J.J. Strossmayer in Osijek, Faculty of Economics in Osijek
Gajev trg 7, Osijek, Croatia
natasa@efos.hr

Kristina Šorić, Silvija Vlah, Višnja Vojvodić Rosenzweig
University of Zagreb, Faculty of Economics
Trg J.F. Kennedya 6, Zagreb, Croatia
{ksoric,svlah,vvojvodic}@efzg.hr

Abstract: The paper aims to develop models for evaluating credit risk of small companies for one Croatian bank using two different methodologies – logistic regression and multicriteria decision making. The first method's result is the probability of default while the second method's result is the classification of the firms regarding predefined criteria for credit scoring. The paper gives the hints how to combine these two methods in order to construct an efficient strategy for achieving high performance.

Keywords: credit scoring, logistic regression, multicriteria decision making

1 INTRODUCTION

In a very aim of the banking business is providing loans to the clients. During this process and in order to make a decision whether to approve or reject a loan the bank is interested in examining a credit worthiness of a client. In the past this decision was made based on the individual judgment of bank's experts who qualitatively graded the risk after examining financial statements of the company, business plan, and interviewing the owner. During the time it has became clear that such a system was not efficient in a more complex environment and growing competition. Therefore, researchers and practitioners have started to develop quantitative models and analytical techniques in credit risk evaluation. These techniques were becoming more and more popular thanks to the information technology development and credit scoring. There are many definitions of it, but we will follow the one stating that credit scoring is the system helping the decision maker such as credit manager to determine whether or not to provide loan to clients, on the basis of a set of predefined criteria.

Credit scoring has been used in retail, corporate and small business lending. Most of the credit scoring systems vary regarding the type and quantity of the data needed for decision making. Personal and business activities have both been found relevant in a small business credit scoring system ([5], [12], [13] and [14]). In evaluating business activities researchers aim to discover financial ratios that are crucial in determining repayment prospects of the company. Most of the corporate credit scoring models use variables or financial criteria that are grouped into five categories – short-term and long-term solvency, utilization, profitability, leverage and performance. One of the first was Altman's z-score ([1]). There are authors who have developed scoring models of the companies representing a specific country. Dvoracek et al. studied bankruptcy forecasting in Czech Republic ([10]. Chancharat et al. ([6]) applied survival analysis in identifying the probability of corporate survival for Australian companies. Ciampi and Gordini ([7]) applied discriminant analysis and logistic regression in developing small enterprise default prediction model for manufacturing firms in Northern and Central Italy. Altman et al. ([2]) developed two failure prediction models for Korean companies, one for non-public entities and the other for public and private companies. Zekić et al. ([26] created small business credit scoring for Croatian micro companies. It was proved that both personal and business characteristics are relevant in small business credit scoring systems. Among personal characteristic of entrepreneurs, entrepreneur's occupation was found to be the most important one. Among small business characteristics 4 variables were found important: clear vision of the business, the planned value of the reinvested profit, main activity of the small business, and awareness of the competition.

Also there is a group of researchers who compared different methods in developing credit scoring. One of the most frequently used method is logistic regression. Altman et al. ([3]) compared linear discriminant analysis, logistic regression and neural networks in distress classification. Desai et al. ([9]) tested multilayer perception, LDA and LR. Yobas et al. ([24]) compared the predictive performance of LDA, NN, genetic algorithms, and decision trees. Galindo and Tamayo ([15]) made a comparative analysis of CART decision trees, NN, the k-nearest neighbor and probit analysis. West ([22] compared five NN algorithms with five more traditional methods.

The aim is to assess the risk of default associated with a credit product/decision. More recent papers use neural networks, evolutionary computation and genetic algorithms ([18], [20]) and support vector machine ([23]). In the last paper three link analysis algorithms based on the preprocess of support vector machine are proposed. It is shown that the genetic link analysis ranking methods have higher performance in terms of classification accuracy. There are some results in the literature obtained by linear integer programming ([16]).

Also there are some works considering the credit scoring as a multicriteria decision making problem (detailed description of the multicriteria decision making problems can be found in [11]) and applying a group decision making technique ([25]). In this study, a novel intelligent-agent-based fuzzy group decision making model is proposed as an effective multicriteria decision analysis tool for credit risk evaluation. Some artificial intelligence techniques are used to replace human experts. Thus, these artificial intelligence agents can be seen as decision members of the decision group.

In this paper we are giving some hints how to combine logistic regression and multicriteria decision making in order to construct an efficient strategy for achieving high performance. The aim is to describe how each of the method can be used in credit scoring, what kind of results can be produced and what are the assumptions, advantages and disadvantages of the methods. Furthermore, it is discussed what a bank can obtain by using logistic regression and what by using multicriteria decision making. Important characteristics in credit scoring models for small companies in Croatia are also discussed because their specific features influence the criteria choice and some input parameters for the methods used in the paper.

The structure of the paper is as follows. In Section 2 the problem of credit scoring with the data from the Croatian bank is described. Section 3.1 presents logistic regression method used for credit risk evaluation with the data given in Section 2, while Section 3.2 presents multicriteria decision making used for the same purpose. In Section 4 the discussion of the results and some concluding remarks are given. Section 5 presents some hints for future research.

2 PROBLEM DESCRIPTION

The problem of credit scoring considered in this paper is the problem of credit risk evaluation in one Croatian bank. In the present moment it is performed based on the individual judgment of bank's experts. To evaluate the applicant's judgmental credit score, 14 criteria are used and are described as follows: (1) sales; (2) profit margin; (3) total debt ratio; (4) current ratio; (5) inventory turnover; (6) repayment ratio = (EBIT +

depreciation)/outstanding debt ; (7) net cash flow; (8) client's credit history; (9) personal character in repayment; (10) business experience in the industry; (11) 1-total debt ratio; (12) average cash on business account; (13) industry risk; (14) business and marketing plan. Also, we were given the data for 60 clients which was a great problem because the procedure of collecting the data is very complicated, not available on-line, the clients have to come personally in the bank and fulfill the documentation. The data are then used in order to create so called credit card of the client. Some data are also given to the bank by Croatian Credit Bureau. The example of one credit card is given bellow:

Table 1: The credit card for Client 1

Client	***1. Sales***	***2. Profit margin***	***3. Total debt ratio***	***4. Current ratio***	***5. Inventory turnover***
Client 1	1.197	0,92%	0,14	5,79	1,62

6. Repayment ratio	***7. Net cash flow***	***8. Client's credit history***	***9. Personal character in repayment***
0,25	-178	1-2 loans	excellent

10. Business experience in the industry	***11. 1-total debt ratio***	***12. Average cash on business account***	***13. Industry risk***	***14. Business and marketing plan***
3+ years	85,66%	10000+	medium	excellent

For every of all 60 clients the same credit card was provided from the Croatian bank and used as the input for both methods used in this paper for credit risk evaluation.

3 METHODOLOGY FORMULATION

To solve the problem of credit risk evaluation presented in Section 2, we use two ways, one is the application of logistic regression methodology and the other is considering the problem as a multicriteria decision making problem. The purpose of using two methodologies is as follows. Namely the result of the first one is the probability of default for a certain client. The result of the second one is a set of groups of clients ranging from the group of best clients to the group of bad clients. In the practice the decision maker is often not so rigid making the decision. Sometimes he/she is not so sure should he/she approve the loan or not. In order to give him/her some additional information and to help him/her to make a decision the combination of different methodologies is welcome. Also, it is interesting to see what are the advantages and disadvantages of using the first or the second methodology.

3.1 Logistic regression

Previous research of methods used in credit scoring has shown that statistical methods such as the logistic regression, the linear regression, the discriminant analysis, and decision trees are mostly used. It has also been shown that the best methodology for credit scoring modeling has not been extracted yet, since it depends on the dataset characteristics. Altman et al. ([3]) showed the best result by using LDA. Desai et al. ([9]) got the best results by multilayer perception. Desai et al. ([8]) showed that LR outperformed NN. Yobas et al. ([24]) produced the best results using NN while Galindo and Tamayo ([15]) using CART decision tree.

Logistic regression modeling is widely used for analyzing multivariate data involving binary responses that we deal with in credit scoring modeling. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses. Since the likelihood function of mutually independent variables $Y_1,\ldots,Y_n$ with outcomes measured on a binary scale is a member of the exponential family with $\left(\log\left(\frac{\pi_1}{1-\pi_1}\right),\ldots,\log\left(\frac{\pi_n}{1-\pi_n}\right)\right)$ as a canonical parameter (π_j is a probability that Y_j becomes 1), the assumption of the logistic regression model is a linear relationship between a canonical parameter and the vector of explanatory variables $\mathbf{x}_j$ (dummy variables for factor levels and measured values of covariates):

$$\log\left(\frac{\pi_j}{1-\pi_j}\right)=\mathbf{x}_j^{\tau}\boldsymbol{\beta}$$

This linear relationship between the logarithm of odds and the vector of explanatory variables results in a nonlinear relationship between the probability of Y_j equals 1 and the vector of explanatory variables:

$$\pi_j=\exp\left(\mathbf{x}_j^{\tau}\boldsymbol{\beta}\right)/\left(1+\exp\left(\mathbf{x}_j^{\tau}\boldsymbol{\beta}\right)\right)$$

Detailed description of the logistic regression can be found in Harrel [17]. Logistic regression procedure is made using SAS software. Input variables for logistic regression scoring model are given in Section 2 of this paper. Those are 14 business criteria that are the basis for credit risk evaluation. As the output, we use credit scoring in the form of a binary variable with one category representing good applicants and the other one representing bad applicants. An applicant is classified as good if repayment ratio is equal or greater than 1. The data sample organized in such a way consisted of 58,3% bads and 41,7% goods. The good clients are: 2,3,12,15,22,25,26,28,32,37,40-47,49,50,54,56,57,59,60 and the bad clients are: 1,4,5-11,13,14,16,17,18,21,23,24,27,29-31,33,34,35,36,38,39,48,51-53,55,58.

The aim of logistic regression modeling is to estimate credit risk and to extract variables that are found important in credit risk prediction. We used logistic regression procedure available in SAS software, with standard overall fit measures. Variables together with their level of significance are given in Table 2.

Table 2: Variables in logistic regression model with their significance

Variable	***p-value***
net cash flow	0,0001
business experience	0,0001
total debt ratio	0,0919
inventory turnover	0,5512
industry risk	0,7191
profit margin	0,7789
personal character in repayment	0,8156
client's credit history	0,9232
average cash on business account	0,9761
sales	0,9778

In Table 2 it can be seen that the most predictive variables are net cash flow, client's business experience and total debt ratio. Looking at the regression coefficient shows that the odds of the company of being bad are increasing with the increase of total debt ratio. On the contrary, odds of the company of being good increases with the increase of the net cash flow. Concerning the client's business experience, it is shown that odds of being bad of the company with business experience of 1 to 3 years is increased in relation to those companies with more then 3 years experience in the business.

Fitting measures such as Likelihood ratio = 54,558 (p=0,0001), Wald = 8,358 (p=0,0038) and Score = 18,956 (p=0,0001) show that the model fits well. In order to test how well model classifies applicants, hit rates are calculated, good hit rate=96% and bad hit rate= 94,3%. The estimated good clients are: 2,3,5,12,13,15,22,25,26,28,32,37,40-47,50,54,56,57,59,60 and the estimated bad clients are 1,4,6-11,14,16,17-21,23,24,27,29,30,31,33-36,38,39, 48,49,51,52, 53,55,58.

3.2 Multicriteria decision making problem

We have already mentioned that there are some works considering the credit scoring as a multicriteria decision making problem and applying a group decision making technique (Yu, Wang, Lai, 2009). In that study, in order to obtain a group decision the artificial intelligence agents are used to replace human experts. This reminds us on the problem of defining the weights of criteria in mutlicriteria decision making. In the literature there are mathematical models based on multi criteria optimization, data envelopment analysis, analytic hierarchical process (AHP) and other multi-attribute rating techniques, but we do not use these approaches. We think that usually the decision maker wants to participate in the decision process, but he/she does not want to be involved too much. In order to respect this in this work we use the modification of the approach developed in [21] (partly inspired by [4]). Namely, we do not ask the decision maker to assign the weights to the criteria. The only task that he/she has to do is to group the criteria in three groups, very important criteria, less important and the least important criteria. Following this philosophy the Croatian bank which is our decision maker decided to group the criteria as follows. Very important criteria (1) sales; (2) profit margin; (3) total debt ratio; (4) current ratio; (5) inventory turnover; (6) repayment ratio = (EBIT + depreciation)/ outstanding debt ; (7) net cash flow. Less important criteria: (8) client's credit history; (12) average cash on business account; (13) industry risk; (14) business and marketing plan. The least important criteria: (9) personal character in repayment; (10) business experience in the industry.

After grouping the criteria in these three groups by the decision maker, the heuristic developed in [21] assigns the weights to the criteria inside a group based on Monte Carlo simulations.

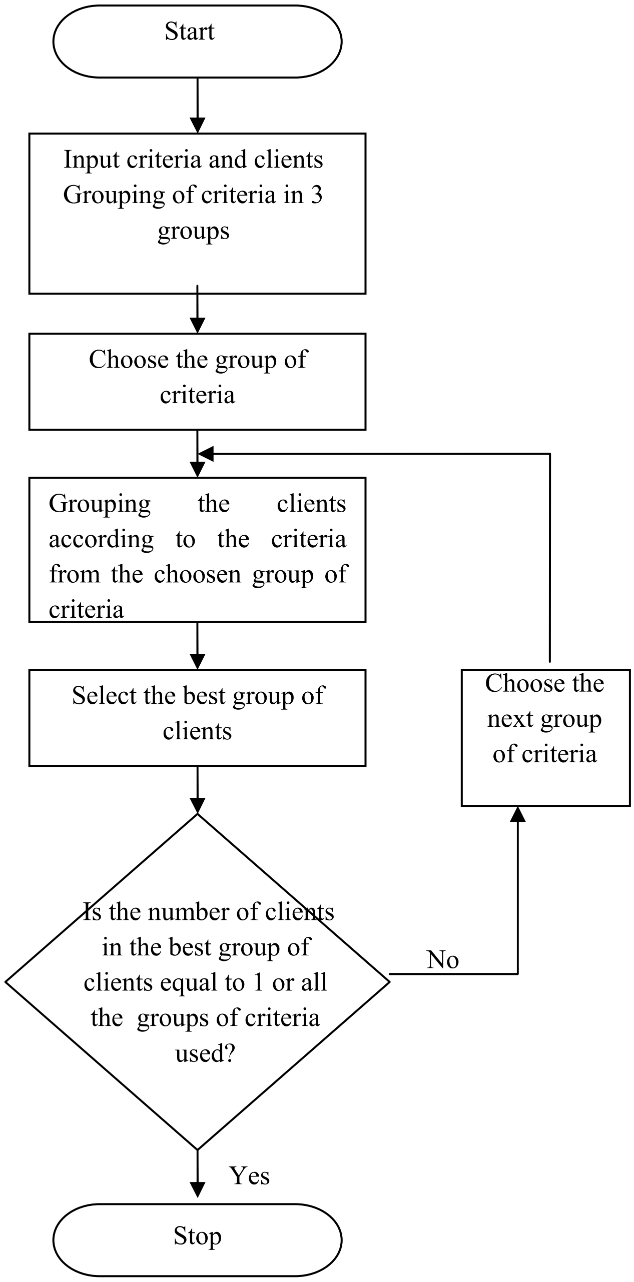


Figure 1: Block diagram of the heuristic used in [21]

Having the weights, in the first iteration the heuristic groups the clients according to their similarities with respect to the most important criteria. In the second iteration the heuristic takes the best group of clients from the first iteration and groups them according to their similarities with respect to less important criteria. Finally in the third iteration the heuristic takes the best group of clients from the second iteration and groups them according to their similarities with respect to the least important criteria. The best group is then taken as the group of clients which can be considered for approving the loan. We can notice that the heuristic from [21] gives us the relative result. Namely the result is the group of clients that is the best group but comparing with the other clients. In order to approve the loan or not the decision maker should be asked. But in order to avoid this last step of involving the decision maker in the decision process we are combining the heuristic from [21] with the method explained in 3.1.

In this work we use the modification of the heuristic developed in [21] in the sense that after obtaining the best group of clients, we eliminate them from the list and apply the whole heuristic again to the remaining clients. The result is again the best group of clients which is now the second best group. In this way in the case of the mentioned Croatian bank seven groups of clients were created starting from the best to the worst one. Also, for every client the heuristic gives the probability of belonging to the certain group which can be good information for the decision maker in the decision process of approving the loan.

The result obtained by the modification of the heuristic used in [21] is the following classification:

Table 3: Clients' classification

Group	***ID Client (probability of belonging to the group)***
1	56(1), 57(0.74), 59(0.66), 2(0.53), 6(0.2), 15(0.13), 21(0.11), 60(0.05), 1(0.03), 13(0.02), 4(0.01), 26(0.01)
2	9(1), 34(0.31), 17(0.29), 29(0.28), 14(0.16), 50(0.15), 10(0.14), 41(0.13), 38(0.09), 16(0.08), 12(0.07), 53(0.07), 52(0.06), 8(0.05), 35(0.05), 51(0.05), 37(0.04), 19(0.02), 23(0.02), 55(0.02), 27(0.01), 31(0.01)
3	36(1), 25(0.66), 45(0.46), 3(0.33), 42(0.23), 43(0.23), 5(0.14), 44(0.13), 32(0.04), 54(0.04)
4	22(1), 46(0.82), 39(0.74), 40(0.5), 48(0.34), 28(0.08)
5	20(1), 58(0.62), 11(0.45), 24(0.44), 30(0.38)
6	7(1), 18(1), 33(1)
7	47(1), 49(1)

4 CONCLUSIONS

As we have already mentioned in Section 3, to solve the problem of credit risk evaluation presented in Section 2, we use two ways, one is the application of logistic regression methodology and the other is considering the problem as a multicriteria decision making problem. The purpose of using two methodologies is as follows. Namely, every methodology has its own result and no methodology gives the optimal result. In the practice sometimes it is very hard for the decision maker to make a decision based on only one methodology because in the practice the decision makers are taking into consideration many qualitative influences. In that situation the decision maker wants to have a "quantitative based argumentation" for making a decision. In order to help him/her we proposed the combination of logistic regression and multicriteria decision making. The reason of choosing these two methods was in the fact the first one as the result gives us the default probability and the second one gives us the classification of the clients. The first one is good if the decision maker wants to be sure that the client will be able to pay a loan, while the second one is good if the decision maker has to approve the loan to some clients applied for it regardless to the default probability.

Also, in developing logistic regression credit scoring a bank has to have a data set consisted of repayment history and companies' financial data. Results of the credit scoring depend on default definition and quality of data set. If there is a small data set, quality of the scoring model could be decreased. The same could happen if there is a data set with a short repayment history. In both cases, it is justify to more rely on multicriteria decision making which as input needs just a set of defined criteria that are well know by the experts in a bank. Or, in both cases the combination of both methods is very welcome.

From the obtained results (good or bad clients in the first method and the classification from the second method) we will say that for the clients 2, 13, 26, 56, 57, 59 and 60 the decision maker can be sure that they are good clients and that they will repay the loan regularly. Since there are many good and bad clients from the first method belonging to the second and the third group from the second methods we can conclude that the problem is in the choice of the important criteria. Namely, from Table 2 it can be seen that the most predictive variables are net cash flow, client's business experience and total debt ratio. From the other side, our decision maker, the Croatian bank defined the following very important criteria: sales, profit margin, total debt ratio, current ratio, inventory turnover, repayment ratio and net cash flow. In this case the next step is to talk to the decision maker and with the information of the most predictive variables ask him/her if he/she is ready to think over his/her definition of very important criteria. In this way the new interactive method could be developed.

5 FUTURE RESEARCH

Both logistic regression and multicriteria decision making have several good reasons to be used in credit scoring modeling, some of which are described in this paper. It is clear that in an environment that is becoming more and more complex, using just judgmental systems for credit risk evaluation is just not enough so practitioners and researchers are constantly in searching for new algorithms. We believe that the further research in combining logistic regression and multicrieteria decision making has several directions.

Probability of default could be one of the criteria in multicriteria decision making. In such a way making decisions concerning credit risk in banks could be improved since other criteria usually not incorporated into the scoring model could be included in model of multicriteria decision making, such as macroeconomics indicators, trends in industry, forecasting about economy etc.

Also, we could think the other way around and result of multicrieteria decision making include as an independent variable in logistic regression. This variable can be the rank of a certain client obtained by the mutlicriteria decision making or the probability of belonging to a certain group. It is to expect that a final scoring model would have higher quality with such a strong predictor.

The third possibility is to use logistic regression result of significant variables selection as the most important and as such include them in the group of the most important criteria. Also, the significance of variables obtained by logistic regression could be used to define the weight of every criteria which is, as we know, a big problem in multicriteria decision making. This way of obtaining the weights in multicriteria decision making could substitute Monte Carlo simulations in the heuristic used in [21].

Further the new heuristic could be created as a combination of logistic regression and the modification of the heuristic used in [21] where the clients with a small probability of belonging to a certain group can be eliminated from this group and added to the list of remaining clients again for the next iteration of ranking. In order to define the lower bound for the probability of belonging to a group the logistic regression result could be used.

References

[1] Altman, E.I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, Journal of Finance Vol. 23, pp.189-209.

[2] Altman, E.I., Joung, H.E., Kim, D.W., 1995. Failure prediction: evidence from Korea, Journal of International Financial Management and Accounting, Vol. 6, No. 3, pp. 230-249.

[3] Altman, E.I., Marco, G., Varetto, F., 1994. Corporate distress diagnosis: Comparison using linear discriminant analysis and neural networks (the Italian experience), Journal of Banking and Finance 18, pp. 505-529.

[4] Alves, M.J., Climaco J., 2004. A note on a decision support system for multiobjective integer and mixed-integer programming problems, European Journal of Operational Research, pp. 258-265.

[5] Arriaza, B.A., 1999. Doing Business with Small Business. Business Credit,101(10): 33-36.

[6] Chancharat, N., Davy, P., McCrae, M.S., Tian, G.G., 2007. Firms in financial distress, a survival model analysis, 20th Australasian Finance and Banking Conference, pp.1-37.

[7] Ciampi, F., Gordini, N., 2008. Using economic-financial ratios for small enterprise default prediction modelling: an empirical analysis, Oxford Business & Economics Conference Program, Oxford, UK, pp. 1-21.

[8] Desai, V.S., Conway, D.G., Crook, J.N., Overstreet, G.A., 1997. Credit scoring models in credit union environment using neural network and generic algorithms, IMA Journal of Mathematics Applied in Business & Industry 8, pp. 323-346.

[9] Desai, V.S., Crook, J.N., Overstreet, G.A., 1996. A comparison of neural network and linear scoring models in credit union environment, European Journal of Operational Research 95, pp. 24-35.

[10] Dvoracek, J., Sousedikova, R., Domaracka, L., 2008. Industrial Entreprises Bankruptcy Forecasting, Metalurgija, Vol. 47, No. 1, pp.33-36.

[11] Ehrgott, M., 2005. Multicriteria Optimization, 2nd ed., Berlin, Springer Gamerman, D. 1997. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Boca Raton, FL: CRC Press

[12] Feldman, R., 1997. Small Business Loans, Small Banks and a Big Change in Technology Called Credit Scoring. Region 11(3), pp.18-24.

[13] Frame, W.S., Srinivasan, A., Woosley, L., 2001. The Effect of Credit Scoring on Small Business Lending. Journal of Money, Credit and Banking, 33(3), pp. 813-825.

[14] Friedland, M., 1996. Credit Scoring Digs Deeper Into Data, Credit World, 84(5), pp. 19-24.

[15] Galindo, J., Tamayo, P., 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications, Computational Economics 15, pp. 107-143.

[16] Glover, F., 1990. Improved linear programming models for discriminant analysis, Decision science 21, pp.771-785.

[17] Harrel FE Jr. 2001. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Springer: Berlin

[18] Hunag, Z., Chen, H.C., Hsu, C.J., Chen, W.H., Wu, S.S.., 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support systems 37, pp. 543-558.

[19] Shi, J., Malik, J., 2000. Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), pp. 888-905.

[20] Varetto, F., 1998. Genetic algorithms applications in the analysis of insolvency risk, Journal of Banking and Finance 22, pp. 1421-1439.

[21] Vlah, S., Šorić, K., Vojvodić Rosenzweig, V., 2008. New version of decision support system for evaluating takeover bids in privatization of the public enterprises and services, Proceedings of MESM 2008 – 9th Middle Eastern Simulation Multiconference, Amman, Jordan, pp. 74-78.

[22] West, D., 2000. Neural Network Credit Scoring Models, Computers & Operations Research 27, pp. 1131-1152.

[23] Xu, X., Zhou, C., Wang, Z., 2009. Credit scoring algorithm based on link analysis ranking with support vector machine, Expert Systems with Applications, 36, pp. 2625-2632.

[24] Yobas, M.B., Crook, J.N., Ross, P., 2000. Credit Scoring Using Evolutionary Techniques, IMA Journal of Mathematics Applied in Business & Industry, 11, pp. 111-125.

[25] Yu, L., Wang S., Lai K.K., 2009. An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring, European Journal of Operational Research, 195, pp. 942-959.

[26] Zekić-Sušac, M., Šarlija, N., Benšić, M., Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network and Decision Tree Models, Proceedings of the 26th International Conference on Information Technology Interfaces, June 7-10., 2004, Cavtat/Dubrovnik, Croatia, pp.265-270.

PRE-NEGOTIATION PHASE PROTOCOL FOR ELECTRONIC NEGOTIATION SUPPORT[1]

Tomasz Wachowicz[a] and **Paweł Wieszała**[b]
The Karol Adamiecki University of Economics in Katowice
ul. Bogucicka 14, 40-227 Katowice, Poland
[a]tomasz.wachowicz@ae.katowice.pl
[b]pawel.wieszala@ae.katowice.pl

Abstract: In the paper we discuss the issue of software support for the pre-negotiation phase. We propose to structure negotiations according to PrOACT method, which allows to define the negotiation problem and the initial negotiation space basing on negotiators' aspiration and reservation levels and the notion of intermediate solutions. Then the modified PROMETHEE algorithm is applied to create a ranking of negotiation offers. All the algorithms and procedures comprise the pre-negotiation protocol that can be implied in an electronic negotiation system the usage of which makes the pre-negotiation phase more fluent and transparent.

Keywords: negotiation support, negotiation analysis, multiple criteria decision making, preference elicitation, PROMETHEE.

1 INTRODUCTION

Negotiation is a very complex process of exchanging messages, offers and concession making in order to constitute a compromise that satisfies all the involved parties. It requires many different skills and abilities both of behavioral and formal nature like: supportive communication, conflict management, decision making or multi-objective analysis. The negotiation consists of number of stages that require different problems to solve and tasks to perform. Usually the three main phases of negotiation can be distinguished: the pre-negotiation phase (negotiation preparation), the actual conduct of negotiation and the post-negotiation phase including implementation of results or renegotiation [4]. The activities undertaken in all the stages are very important, however, many researcher emphasize the role of the pre-negotiation phase [8]. The solid negotiation preparation results then in correct selection of negotiation strategies, adequate negotiator attitude, good decisions made, wise concessions and the negotiation atmosphere. Therefore while supporting negotiation process it is very important to construct a adequate model of pre-negotiation procedure. The continuous development of information technology and web services results in many software solutions, e.g. electronic negotiation systems, that are able to support negotiation instead of the human facilitators. Negotiation support systems have also communication and coordination facilities that allow to create an e-negotiation table [6] and conduct negotiations using only the electronic platform instead of the traditional meetings. In the age of the global economy it gives the opportunity to overcome the limitations of time and space and find the business partner wherever in the World. Many such systems exist, used both for the training, research and solving the real world negotiation problems (commercial tools), like INSPIRE [7], Neggoist [10], FamillyWinner [1]. To support the pre-negotiation activities most of these systems use the utility based scoring mechanisms that require the negotiation problem to be highly structured. Moreover they require tiring and time-consuming interaction with negotiators in order to evaluate each single option within each negotiation issue by simply assigning them the utility scores (no calculations or comparisons are required within the procedure). To build the scoring system that reflects the actual

[1] This research was partially financed with the funds of Polish Ministry of Higher Education and Science for the scientific project number N N111 234936 realized in years 2009-2012.

negotiator's structure of preferences then the basic knowledge of the decision making and ratio scale interpretation is required.

In this paper we propose a new negotiation protocol for pre-negotiation phase that can be automatically realized by means of the electronic negotiation system for supporting bilateral negotiation. The protocol comprises the list of task that must be conducted by parties in order to structure the negotiation problem appropriately and construct the scoring system for the negotiation offers that could be exchanged in the next negotiation stage. We will apply the straightforward method of multiple criteria decision making called PrOACT [5] to structure the problem and then the PROMETHEE method [2], the modification of which we propose for the calculation reasons, for the offers evaluation. All the algorithms are programmed as a web based negotiation support system the work of which we illustrate with a simple example.

2 STRUCTURING THE NEGOTIATION PROBLEM FOR PRE-NEGOTIATION PROTOCOL

The pre-negotiation phase consists of many different tasks. They concern preparing the negotiation agenda, all the documents required, analyzing the strengths and weaknesses both of the negotiator and its counterparts, but also identifying the negotiation problem and creating the alternatives for the negotiation agreement [11]. In this paper we will focus on the two last activities that allow to realize the evaluation function [6] of the electronic negotiation system. In building the pre-negotiation protocol we will follow the first four steps of the PrOACT [5] method, which gives the algorithm for solving a classic multiple criteria decision making problem[2]. These steps are:

1. Defining the decision problem.
2. Identifying the objectives of both the parties.
3. Creating the alternatives,
4. Identifying the alternatives consequences.

The fifth step of PrOACT requires the implementation of the trade-off method for selecting the most satisfying alternative, which does not fit the problem we consider here and will be omitted in our pre-negotiation protocol.

Deriving from the above PrOACT algorithm we will start with identifying the negotiation problem by defining the negotiation issues under consideration. Therefore the electronic negotiation system needs to provide to the parties the communication platform (see Section 4) for exchanging the messages in order to constitute the set I of the negotiation issues. The issues are defined by a short names (e.g. cost, time of delivery) and can be both of the quantitative and qualitative nature. The next steps of the PrOACT requires creating of the negotiation offers and their consequences. We define then a particular offer (alternative) by specifying the resolution levels (options) for each negotiation issue, which can be denoted as

$$a = [x_1, x_2, ..., x_I] \quad (1)$$

where x_i is an option selected from the set X_i, of all predefined options for issue i. The classic negotiation analysis approach based on the cardinal utility theory (see [9]) require negotiators to define a priori the finite number of option values for each negotiation issue, and then score them in terms of utility. It can be troublesome, especially if we consider the quantitative issues (e.g. price) the values of which can change of very small intervals and therefore produce the huge sets of resolution levels the evaluation of which could be very

[2] The PrOACT method has already been modified and used for structuring the negotiation problem in the NegoCalc negotiation support system [12].

tiring or even impossible (taking into account e.g. the time constraints of negotiations). Therefore we propose an alternative solution for defining the negotiation space by examining the negotiators' aspiration and reservation levels. The electronic negotiation system asks both the negotiators (α and β) to declare the most preferable resolution levels (they would like to achieve) and the least preferable ones (they would hardly accept) for each negotiation issue. These values comprise the initial set of options for this issue:

$$X_i^{init} = \left\{ x_i^{most\,\alpha}, x_i^{least\,\alpha}, x_i^{most\,\beta}, x_i^{least\,\beta} \right\}. \tag{2}$$

Since the aspiration and reservation levels of the negotiators are confidential the initial sets of options X_i^{init} cannot be used directly to define the set of negotiation alternatives. We propose then to build the final sets of options X_i by adding to the initial sets of options X_i^{init} some other resolution levels (the set of intermediate levels X_i^{inter}). If the issue is quantitative the intermediate levels can be simply calculated by interpolating between the subsequent elements of the set X_i^{init}. To make the pre-negotiation analysis fluent the operations of identifying the set X_i^{inter} should be realized automatically by the analytic engine of the electronic negotiation system with the initial support of the negotiator. She or he is asked to declare the indifference threshold q_i, required primarily for the PROMETHEE algorithm (see Section 3), which reflects the maximal distance between two options for which the negotiator perceives these two options to be equally preferable. We will identify then the intermediate options moving from $x_i^{least\,\alpha}$ to $x_i^{most\,\alpha}$ with the step of q_i^{α}, which will result in fact in determining the salient options [7] for negotiator α (a similar analysis should be conducted for the negotiator β). It is also important to include into the set X_i^{inter} some values greater then $\max\left(x_i^{most\,\alpha}, x_i^{least\,\beta}\right)$ and lower then $\min\left(x_i^{least\,\alpha}, x_i^{most\,\beta}\right)$ to hide the aspiration and reservation levels of both the negotiators.

If the issue is described qualitatively the negotiators should be asked for indicating the other resolution levels they mean to consider during the negotiation process. Moreover the resolution levels need to be ordered from the most preferable to the least preferable one, to allow the electronic negotiation system conduct the calculation procedures (see Section 3, step number 2 of the algorithm). Having defined the sets of intermediate levels we can create the final sets of options for each issue $i \in I$

$$X_i = X_i^{init} \cup X_i^{inter}. \tag{3}$$

Basing on the sets of options X_i the system determines the set of feasible negotiation alternatives A by finding all the possible combinations of the options for all negotiation issues. We will consider the set A to be a preliminary set of negotiation alternatives that negotiators can modify later within the pre-negotiation phase. If they consider the negotiation space found by software system to be too narrow they can add or modify the options within each declared issue. The definition of the final set of alternatives makes the first part of the pre-negotiation phase competed. The set A is used then in the next stage of protocol to create the scoring system for the potential offers by means of PROMETHEE method and according to the negotiators' basic preference information.

3 ANALYZING THE NEGOTIATION OFFERS WITH PROMETHEE METHOD

We have decided to implement the PROMETHEE [2] in analyzing negotiation offers since it does not require subjective assigning the scoring points individually to each option within an each issue. Such assignments can be introduce in negotiations problems with relatively small negotiation space, but in large problems they can be very tiring and troublesome, especially for the negotiators without a basic mathematical knowledge of the multiple criteria decision making. Introducing PROMETHEE in offers evaluation makes the preference elicitation much fluent since the negotiators need only to declare the weights (importance) of the negotiation issues and then only few other pieces of information like the preference or indifference thresholds.

The original PROMETHEE II algorithm operates on the finite set A of alternatives $a_n, n = 1,2,...,N$ that are evaluated in terms of I different criteria. The function $f_i(a_n)$ returns the value of criterion i for alternative a_n. Comparing two alternatives a_m and a_n in terms of criterion i we determine the difference $f_i(a_m) - f_i(a_n)$, denoted as $\delta_i(a_m, a_n)$, which is used in the further calculation process for building the ranking of alternatives. In negotiation problem the set A is very specific since it consists of the alternatives being combinations of all the options for all negotiation issues. Therefore for some groups of alternatives consisted of the same values of x_i the difference δ_i will be identical. We can thus shorten our PROMETHEE calculations for negotiation problem by modifying first two steps of the algorithm. Instead of comparing the alternatives we will compare the option values for each negotiation issue separately. The modified algorithm is as follow:

1. We compare each pair of options for each negotiation issue and determine the difference $\delta_i(x_i^s, x_i^t)$, where $x_i^s, x_i^t \in X_i$ and $i = 1,...,I$; $s,t = 1,...,|X_i|$; $s \neq t$. For the qualitative issues we assume δ_i to be a difference in the ranking positions between the options under consideration.
2. We determine the values of the preference functions $P_i(x_i^s, x_i^t)$, where $i = 1,...,I$; $s,t = 1,...,|X_i|$; $s \neq t$. The preference function assigns to each difference δ_i the value from the range $[0;1]$, which reflects the negotiator's strength of preference. The value of 1 corresponds to the strong preference, while the value of 0 to the lack of preference.

   While considering the qualitative issues we recommend to apply the preference function in the following form:

$$\mathrm{P}(\delta_i) = \begin{cases} 0 & \delta_i \leq 0 \\ 1 & \delta_i > 0 \end{cases}. \tag{4}$$

   It allows to affirm the preference in case the options differ with no examining the scale of difference, which could be hard to determine taking into account a verbal character of this issue.

   For the quantitative issues the five other preference functions can be introduced (see [2]). We propose to apply for instance the following function

$$\mathrm{P}(\delta_i) = \begin{cases} 0 & \delta \leq q \\ \dfrac{\delta - q}{p - q} & q < \delta \leq p \\ 1 & \delta > p \end{cases}. \tag{5}$$

   In the above function the rules of determining the strength of the preference are straightforward and easy to explain to negotiators and require only the declaration of the indifference threshold q and the preference threshold p. However using our electronic negotiation system for pre-negotiation support (Section 4) the user can also choose other preference functions.

3. For each pair of offers we determine the overall preference indexes

$$\Pi(a_m, a_n) = \sum_i w_i P_i(x_i(a_m), x_i(a_n)), \quad (6)$$

$$\Pi(a_n, a_m) = \sum_i w_i P_i(x_i(a_n), x_i(a_m)), \quad (7)$$

where $x_i(a_m)$ and $x_i(a_n)$ denote in sequence the values of option i for negoitation offers a_m and a_n.

The next three steps of of the algorithm are the steps of the original PROMETHEE method and require:

4. Identification of the leaving and entering flow for each negotiation offer following the formulas:

$$\Phi^+(a_m) = \frac{1}{N-1}\sum_{\substack{n=1 \\ m \neq n}}^{N} \Pi(a_m, a_n), \quad (8)$$

$$\Phi^-(a_m) = \frac{1}{N-1}\sum_{\substack{n=1 \\ m \neq n}}^{N} \Pi(a_n, a_m). \quad (9)$$

5. Determination of the preference net flow for each negotiation offer according to the formula:

$$\Phi(a_m) = \Phi^+(a_m) - \Phi^-(a_m) \quad (10)$$

6. Building the ranking of the offers according to decreasing value of preference net flow Φ.

Having completed the above algorithm of modified PROMETHEE method we obtain the scoring system of negotiation offers and completed the second stage of the pre-negotiation protocol. The list of selected offers can be presented then to the negotiator including the aspiration and reservation offers and some other offers better then the aspiration one, some consisted of the intermediate options and some worse then the reservation one. Negotiator can analyze the list of offers and examine the consequences of the preference and indifference thresholds declared at the beginning of the pre-negotiation phase. She or he can freely change the thresholds or add some options and, after the scoring system recalculation, observe the changes in the offers ranking. The negotiator can also plan the concession path by analyzing the possible trade-offs between the issues. It can appear that concessions made in one issue (required by the counterpart) may be easily compensated by the slight improve for another issue and do not need to result in the moving down in the ranking levels.

While scoring the negotiation space with PROMETHEE we need to be aware of one important drawback of this method, which is a rank reversal problem [3]. It makes it very important to identify the most accurate negotiation space before starting the preference elicitation procedure and going to the actual negotiation phase. Adding or removing the options is possible and do not involve negotiator in the recalculation of the offers scoring system, however it can change the final ranking of the negotiation offers. That is because the PROMETHEE algorithm bases on the pair-wise comparison of the offers that comprises of different options. Adding (removing) options results in adding (removing) offers from the set

A, which will naturally influence the entering and leaving flows used for building the offers ranking.

The simple example of conducting prenegotiation phase according to the negotiation protocol proposed in Section 3 and 4 is presented in the next section of this paper.

4 EXAMPLE

Let us consider a simple example of business negotiation. The electronic negotiation system asks both the negotiators to describe the negotiation problem by defining the negotiation issues (Fig 1). The form includes issue name, type, values and description. When negotiator select the type (qualitative or quantitative) the system automatically activate appropriate textbox's to define the values. For each negotiation issue the negotiator needs to declare the most preferable resolution level and the least preferable one (aspiration and reservation levels). Additionally for qualitative issues she or he needs to give other intermediate resolution levels. For quantitative issues the indifference threshold is required for building the list of salient options. In our example the first negotiator defines the problem by means of two issues: price and time of delivery in days (Fig 1). The second negotiator agree on those two issues and add one more: type of payment (Fig 2).

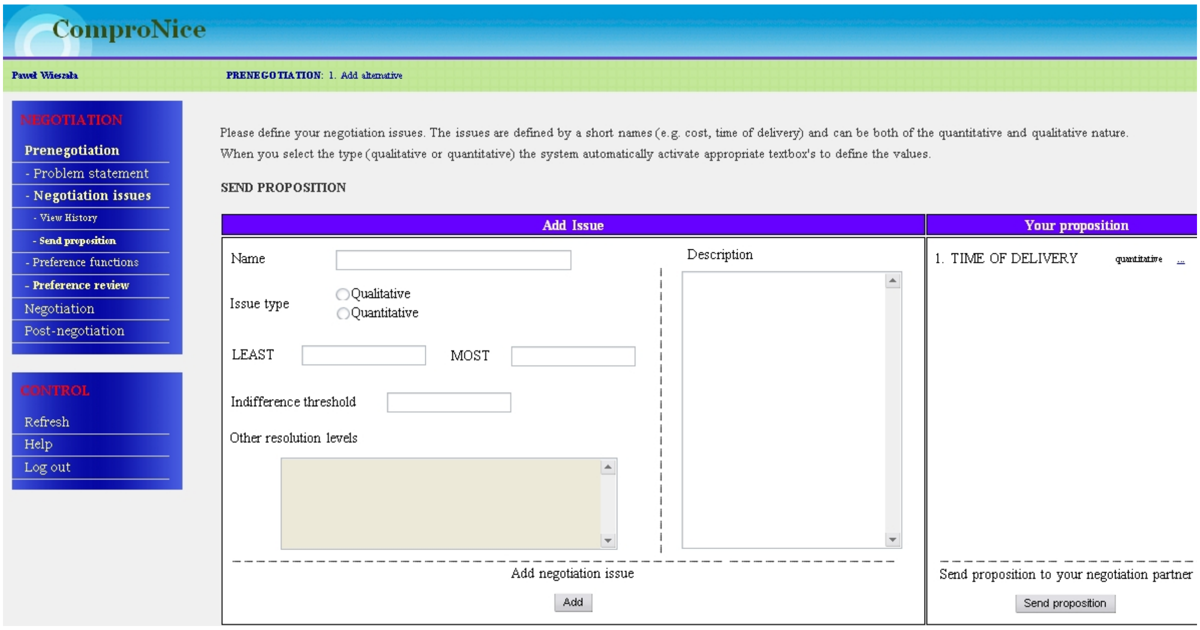


Figure 1:. The definition of the negotiation issues

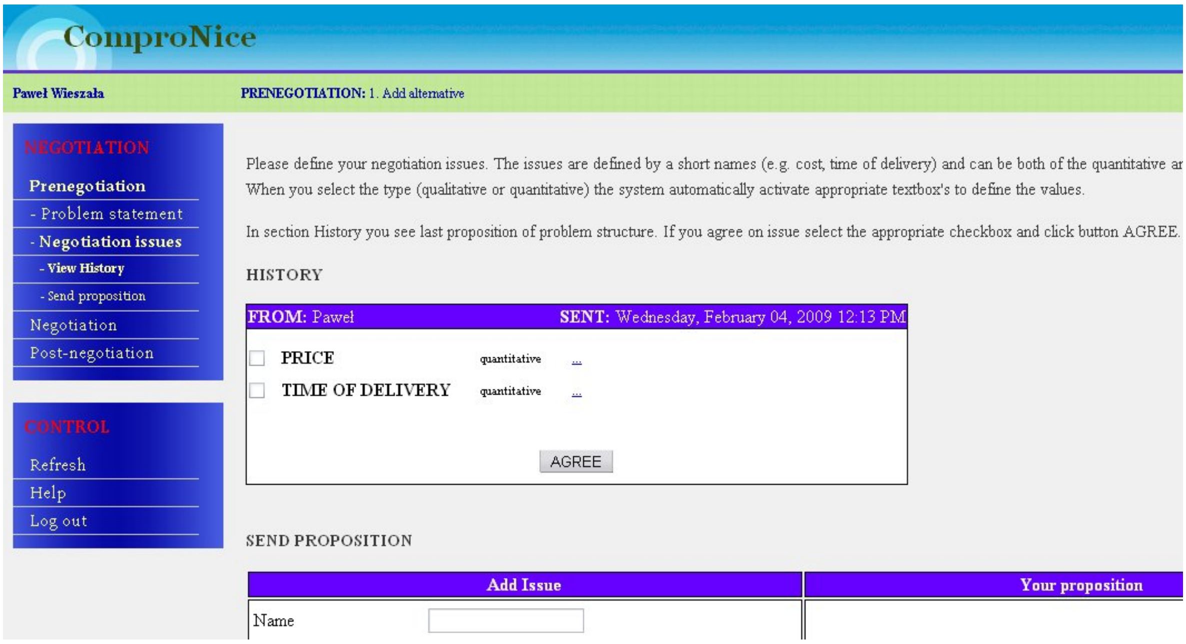


Figure 2: Agreeing for the negotiation issues proposed by the counterpart

Basing on the defined values and the indifference thresholds the system determines the negotiation space by specifying all salient options for both the negotiators (Fig 3). Then negotiators need to establish a preference function and its parameters for each negotiation issue (Fig 4). All information about PROMETHEE method, interpretation of parameters and different types of preference functions are available in documentation of the system.

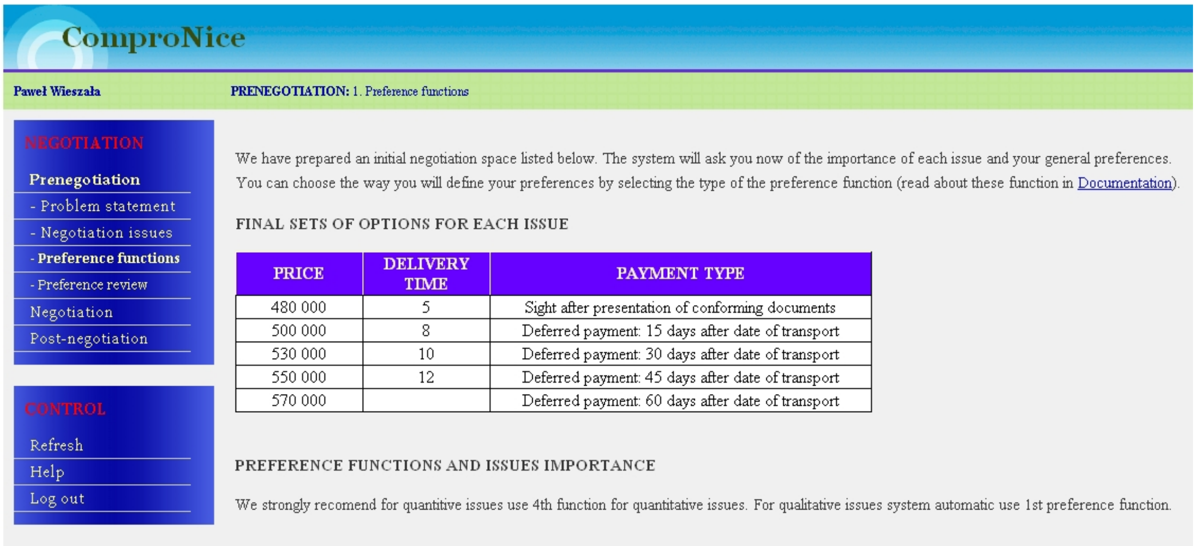


Figure 3: Specification of the negotiation space

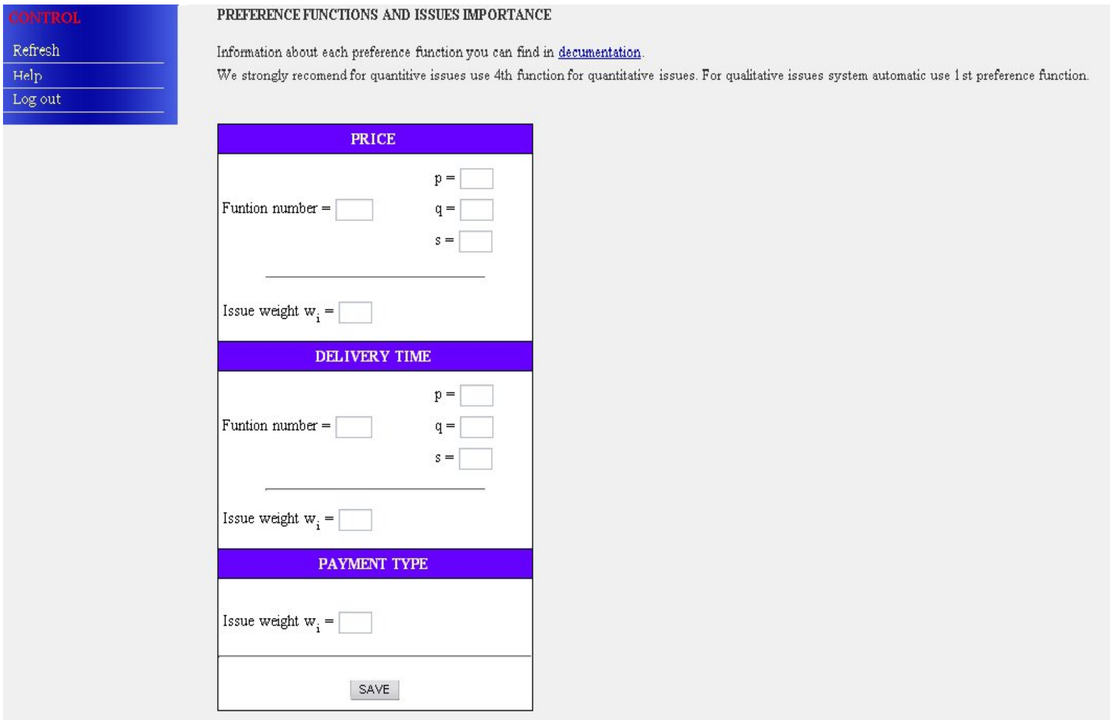


Figure 4: Defining the negotiator's preferences

The system automatically calculates value of preference for all alternatives. In our case we have 5x5x4 = 100 different alternatives. To validate negotiators preferences, some examples of their preference are shown (Fig 5). If negotiator wants to redefine the negotiation space by adding or removing options and issues she or he can now move back to the first step of the pre-negotiation phase and introduce the changes. The system will recalculate the offers ranking according to the new problem and/or new preference indexes. On the other hand the preference information and the offers ranking is saved on the system server and can be used in the next phase of the actual negotiation for comparing and analyzing the offers.

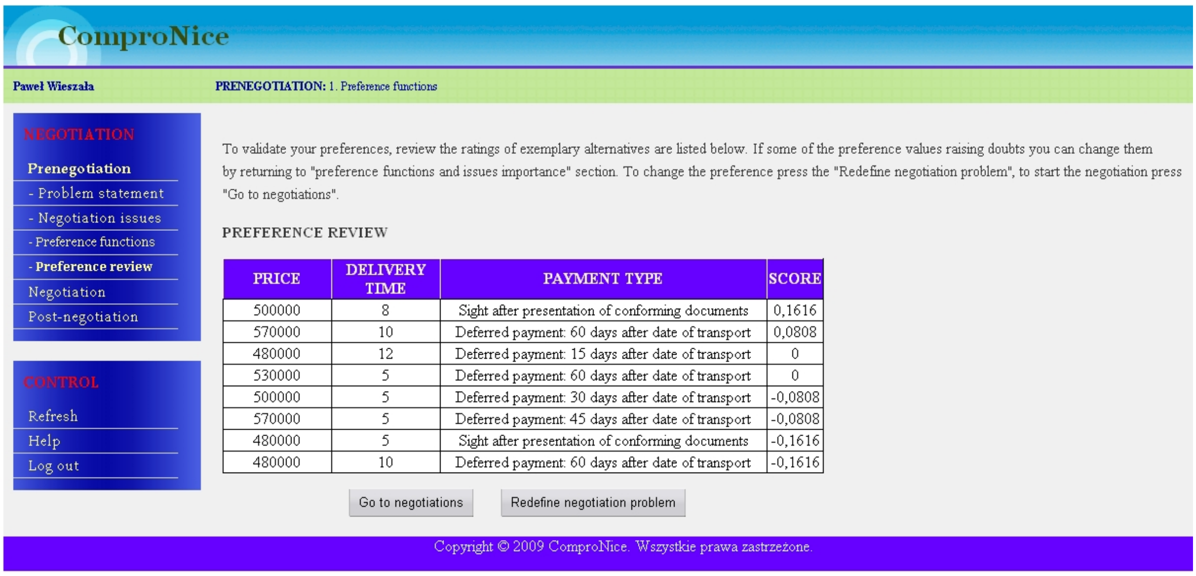


Figure 5: The ranking of the negotiation offers

Analyzing the list of the negotiation offers negotiators can prepare for the subsequent phase of the actual negotiation conduct. She or he realizes the tradeoffs that could be made within different pairs of issues. If her/his counterpart requires making concession in terms of a selected issue, the negotiator can find another issue she/he could compensate the losses in the former one. For instance, the negotiator, who has the ranking of offers as shown in Fig 5., proposes a compromise like {Price: 500000; Delivery time: 5; Payment type: Deferred payment – 30 days after date of transport}, but he counterpart requires the further concessions in the price. Looking at the ranking the negotiator knows she/he can pay even 70000 more for the contract, but will require the 45 days of delay in payment, which guarantee her/him the same level of satisfaction from this contract (these two offers are at the same ranking level constructed according to the negotiator's preferences).

5 SUMMARY

Deriving from PrOACT mechanism we have proposed a pre-negotiation phase protocol as a procedure requiring both mutual and individual activities of both the negotiators. The procedure starts with the mutual discussion over the negotiation issues and after the agreement it requires of negotiators the declaration of their aspirations and reservation levels. Then the analytic engine builds the sets of the options for negotiation issues that are used to build the preliminary set of the negotiation offers. The set is then presented to the negotiators and can be modified according to their individual requirements. Next the offers are evaluated by means of modified PROMETHEE method, which operates on the options instead on the alternatives within the first three steps of the algorithm. The prototype of the electronic negotiation system was also presented to show that the protocol we had proposed can be easily supported by means of a software tool.

The major advantage of the protocol we proposed is that, as oppose to the utility based scoring systems, it allows to avoid the tiring preference elicitation procedures that require the multitude of the utility assignments to the negotiation options and issues. It also does not force negotiators to predefine precisely the negotiation space by themselves. The system derives form their aspiration and reservation levels and the indifference thresholds assigned to each issue to propose the list of salient options. Moreover, it does not involve negotiators in another preference elicitation process if they decided to change the initial set of alternatives proposed by supporting software system. Since they defined the preference and indifference thresholds at the beginning of the pre-negotiation phase, the system automatically determines the new ranking for the extended set of negotiation offers.

There is one problem we found and it concerns evaluating the qualitative issues. Since we are not able to automatically detect the differences between the verbal descriptions of the options the negotiators are asked to find the intermediate levels and declare the order of the options within this issue. The detailed study of the qualitative nature of the issues needs to be conducted to find some alternatives for atomization of the above and we will focus on it in our further research. We will also continue the development of the electronic negotiation system by adding the other features like the communication unit for exchanging offers and tracking the negotiation process or introducing the arbitration and mediation procedures to make the software system act more proactively.

References

[1] Bellucci E., Zeleznikow J. (2005). Managing Negotiation Knowledge: from negotiation support to online dispute resolution, In: John Zeleznikow, Arno R. Lodder (eds.), *Second international ODR Workshop (odrinfo.info)*, Tilburg: Wolf Legal Publishers, pp. 11-22.

[2] Brans, J.P., Vincke, Ph., Mareschal, B., How to select and how to rank projects: the PROMETHEE method. European Journal of Operational Research 1986, (24), s. 228-238.

[3] De Keyser W., Peeters P. (1996). A note on the use of PROMETHEE multicriteria methods. European Journal of Operational Research, 89,457–461 .

[4] Gulliver, P. (1979), Disputes and Negotiations: A Cross-Cultural Perspective. New York, NY: Academic Press.

[5] Hammond J., Keeney R., Raiffa H. (1998). Even Swaps: A Rational Method for Making Trade-offs. Harvard Business Review, March-April 1998.

[6] Kersten, G.E.; Lai, H. (2006). Negotiation Support and E-Negotiation System. InterNeg Research Papers, 013/06.

[7] Kersten G.E., Noronha S.J. (1999) WWW-Based Negotiation Support: Design, Implementation And Use. Decision Support Systems, 25:135–154.

[8] Lewicki R.J., Saunders D.M., Minton J.W. (1999). Negotiation. Irwin/McGraw-Hill.

[9] Raiffa H. (1982). The Art and Science of Negotiation. Harvard University Press, Cambridge.

[10] Schoop M., Jertila A., List T. (2003). Negoisst: a negotiation support system for electronic business-to business negotiations in ecommerce. Data Knowledge Engineering, 47:371–401.

[11] Scott B. (1999). Preparing for Negotiations. In: Lewicki R., Saunders D., Minton J. (eds.) Negotiation. Irwin/McGraw-Hill, pp. 60 – 67.

[12] Wachowicz T. (2008). NegoCalc: Spreadsheet Based Negotiation Support Tool with Even-Swap Analysis. In: J. Climaco, G. Kersten, J.P. Costa (eds.), Group Decision and Negotiation 2008: Proceedings – Full Papers, INESC Coimbra, pp. 323 – 329.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section III:

Production and Inventory

THE VALUE OF SHARING ADVANCE CAPACITY INFORMATION UNDER “ZERO-FULL” SUPPLY CAPACITY AVAILABILITY

Marko Jakšič
University of Ljubljana, Faculty of Economics, Kardeljeva ploščad 17, Ljubljana, Slovenia
School of Industrial Engineering, Eindhoven University of Technology, The Netherlands
marko.jaksic@ef.uni-lj.si

Abstract: The importance of information sharing within the modern supply chains has been established both by practitioners as well as the researches. The accurate and timely information helps firms to effectively reduce the uncertainties of the volatile and uncertain business environment. We model a periodic review, single-item, stochastic demand and stochastic supply, where in a given period supply is either available or completely unavailable. Additionally, a supply chain member has the ability to obtain advance capacity information (ACI) about the future supply capacity availability. We show that the optimal ordering policy is a base stock policy with the optimal base stock level being a function of future capacity availability or unavailability given through ACI. In a numerical experiment we quantify the value of ACI and give the relevant managerial insights.

Keywords: Operations Research, Inventory, Stochastic Models, Value of information, Advance Capacity Information.

1 INTRODUCTION

Particularly in the last two decades, the companies working in global business environment have realized the critical importance of effectively managing the flow of materials across the supply chain. Industry experts estimate not only that the total supply chain costs represent the majority of operating expenses for most organizations but also that, in some industries, these cost approach 75% of the total operating budget [1]. Inventory and hence inventory management, plays a central role in the operational behavior of a production system or a supply chain. The fact is that on average 34% of the current assets and 90% of the working capital of a typical company in the United States are invested in inventories [2]. Due to the complexities of the modern production processes and the extent of global supply chains, inventory appears in different forms at each level of the supply chain. A supply chain member needs to control its inventory levels by applying some sort of inventory control mechanism. An appropriate selection of this mechanism may have a significant impact on the customer service level and member’s inventory cost, as well as supply chain system-wide cost.

Due to the focus on providing quality service to the customer, it is no surprise that the demand uncertainties have attracted the attention first. However, through time the companies have realized that effective management of supply is of equal importance. Looking at the supply chain’s production and supply capacities allocated to produce or deliver a certain product reveals that these are generally far from being stable through time. Quite the opposite, supply capacity may be highly variable due to several reasons, like frequent changes in product mix, particularly in the setting where multiple products share the same capacity, changes in workforce (e.g. holiday leaves), machine breakdown and repair, preventive maintenance etc. To compensate for these uncertainties extra inventory needs to be kept. However, as it turns out in practice, companies have problems in determining the appropriate inventory levels. In [3] and [4] they attribute this to the fact that companies typically do not use any formal analytic approach. This results in an inappropriately too low or too high inventory levels and consequently high inventory related costs.

However, there is another, perhaps even more appealing way to tackle the uncertainties in supply. Foreknowledge of future supply availability can help managers anticipate the possible future supply shortages, and allows them to react in a timely manner, by either building up stock to prevent future stockouts, or reducing the stock in the case of favorable supply conditions. Thus, system costs can be reduced by carrying less safety stock while still achieving the same level of performance. These benefits should encourage the supply chain parties to formalize their cooperation to enable the necessary information exchange either through the implementation of the necessary information sharing concepts like Electronic Data Interchange (EDI) and Enterprise Resource Planning (ERP) or through use of formal supply contracts. We can argue that extra information is always beneficial, but further thought has to be put into investigating in which situations the benefits of information exchange are substantial and when it is only marginally useful.

In this paper, we explore the benefit of using the available advance capacity information (ACI) about future uncertain supply capacity conditions to improve the performance of the inventory control policy and reduce the relevant inventory cost. We study a periodic review, single product, single location inventory model, where both the demand and supply capacity are assumed to be stochastic. We assume that due to his intimate knowledge of the production process the supplier is able to provide the upfront information on his capacity availability to the retailer. This is done for a certain limited number of future periods, where in some of the periods the capacity is completely unavailable, while in other periods there is complete availability of the capacity. Through ACI, the retailer can anticipate near future supply shortages with certainty and prepare for periods of overall product unavailability by building up inventory in advance. Throughout the paper we address to main research questions: (1) What is the optimal way to integrate ACI in the inventory control policy and how will ACI affect the optimal inventory levels? (2) What is the value of ACI and in which settings ACI turns out to be of particular importance?

We proceed with a brief review of the relevant literature, where the focus is on presenting different ways that the researchers have suggested to tackle the problem of supply uncertainty. The supply uncertainty is commonly attributed to one of the two sources: yield randomness and randomness of the available capacity. The problem of either fully available or unavailable supply can actually be related to both of these two categories. More specifically, in [5] authors analyze the random yield case where the creation of good products is a Bernoulli process. Hence, the number of good products depends on the order size and the probability of generating a good product from one unit of output, and follows a binomial distribution. In [6] they analyze an inventory model with stochastic limited supply. In this case, the actual order realization is the minimum of the initial order given and the realized random supply capacity. They show that the optimal policy remains to be a base stock policy as in the case of unlimited supply. The optimal base stock level has to be increased to account for the possible capacity shortfalls in the future periods. This work is extended by [7], where they introduce the notion of *advance capacity information*, and the value of ACI under the non-stationary stochastic demand and limited supply is assessed. Optimal ordering policy in this case is a state-dependent base stock policy, where the base stock level is a function of ACI. In [8] they analyze both uncertainty effects by extending the uncertain capacity setting by incorporating the effect of yield uncertainty also. The work closely related to the topic of this paper is the research presented in [9] and [10]. In a deterministic demand setting, they show that the optimality of the base stock policy applies also in the case of a Bernoulli-type supply process. They obtain a newsboy-like formula to characterize the optimal base stock levels. We extend this work by imposing no restrictions on the characteristics of the demand process characteristics, modeling demand as a non-stationary stochastic process, and additionally, we are primarily interested in the effect of ACI on the optimal performance of the inventory system.

Our contributions in this paper can be summarized as follows: (1) We present a new model that incorporates ACI within the limited supply capacity setting, where supply capacity is modeled as a Bernoulli process. (2) We derive the structure of the optimal policy and show that it fits in the group of the base stock policies, where the base stock level is a function of the revealed capacity realizations in the near future. (3) Finally, we establish the value of ACI, plus recognize and describe the system settings in which ACI brings considerable savings.

The remainder of the paper is organized as follows. We present the detailed model formulation and the structural characteristics of the optimal policy in Section 2. In Section 3, we present the results of a numerical study to assess the value of ACI and provide the relevant managerial insights. Finally, we summarize our findings and suggest the possible extensions in Section 4.

2 MODEL AND THE STRUCTURAL PROPERTIES

In this section, we give the notation and the model description. Additionally, we derive the structure of the optimal policy and characterize the optimal base stock levels. We model the supply uncertainty as a Bernoulli process, where p_t, $0 \le p_t \le 1$, denotes the probability of full capacity availability in period t. We introduce the parameter of supply capacity availability a_t, where $a_t = 0$ denotes the zero availability case and $a_t = 1$ the full availability case. In period t, the manager obtains ACI a_{t+n} on the supply capacity availability in period $t+n$, where the parameter n denotes the length of the ACI horizon. Thus, in period t the supply capacity availability for n future periods is known and we record it in the ACI vector $\vec{a}_t = (a_{t+1}, a_{t+2}, \ldots, a_{t+n})$. Note that the capacity availabilities in periods $t+n+1$ and later are still uncertain.

Presuming that unmet demand is fully backlogged, the goal is to find an optimal policy that would minimize the inventory holding costs and backorder costs over a finite planning horizon T. The demand is in general modeled to be stochastic non-stationary with known distributions in each time period, however, independent from period to period. For simplicity of presentation the zero supply lead time case was chosen, however the model can be easily extended to consider positive supply lead times.

We assume the following sequence of events. (1) At the start of the period t, the manager reviews inventory position before ordering x_t and ACI a_{t+n} on supply capacity availability in period $t+n$ is received, which could potentially limit the order z_{t+n} that will be given in period $t+n$. (2) In the case of $a_t = 1$, the ordering decision z_t is made and correspondingly the inventory position is raised to inventory position after ordering y_t, $y_t = x_t + z_t$. (3) The order placed at the start of the period is received. (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 net inventory.

To follow the evolution of the inventory system through time we need to keep track of the starting inventory position x_t, current supply capacity availability a_t, and the vector of

ACI $\vec{a}_t$. The state space can therefore be described by an $n+2$-dimensional vector $\left(x_t, a_t, \vec{a}_t\right)$ and gets updated in period $t+1$ in the following manner:

$$\begin{aligned} x_{t+1} &= x_t + z_t - d_t, \\ \vec{a}_{t+1} &= \left(a_{t+2}, a_{t+3}, \ldots, a_{t+n+1}\right). \end{aligned} \tag{1}$$

Finally, we can derive the minimal discounted expected cost function that optimizes the cost over a finite planning horizon T from time t onward, starting in the initial state $\left(x_t, a_t, \vec{a}_t\right)$:

$$f_t\left(x_t, a_t, \vec{a}_t\right) = \begin{cases} \min\limits_{y_t \geq x_t} \{C_t(y_t) + \alpha E_{d_t} f_{t+1}(y_t - d_t, a_{t+1}, \vec{a}_{t+1})\}, & \text{if } T-n \leq t \leq T, \\ \min\limits_{y_t \geq x_t} \{C_t(y_t) + \alpha E_{d_t, a_{t+n+1}} f_{t+1}(y_t - d_t, a_{t+1}, \vec{a}_{t+1})\}, & \text{if } 1 \leq t \leq T-n-1, \end{cases} \tag{2}$$

where α is a discount factor, $C_t(y_t) = h\max(0, y_t - d_t) + b\max(0, d_t - y_t)$ represents the single period cost function, and the ending condition is defined as $f_{T+1}(\cdot) \equiv 0$.

We proceed with the characterization of the optimal solution given by the dynamic programming formulation in Eq. (2). The optimality result is based on establishing the convexity of the optimal cost function f_t. Note that the single period cost function $C_t(y)$ is convex in y, since it is the usual Newsboy cost function [11]. Let g_t denote the cost-to-go function of period t, defined as

$$g_t\left(y_t, \vec{a}_t\right) = \begin{cases} C_t(y_t) + \alpha E_{d_t} f_{t+1}(y_t - d_t, a_{t+1}, \vec{a}_{t+1}), & \text{if } T-n \leq t \leq T, \\ C_t(y_t) + \alpha E_{d_t, a_{t+n+1}} f_{t+1}(y_t - d_t, a_{t+1}, \vec{a}_{t+1}), & \text{if } 1 \leq t \leq T-n-1, \end{cases} \tag{3}$$

and we rewrite the minimal expected cost function $f_t\left(x_t, a_t, \vec{a}_t\right)$ as

$$f_t\left(x_t, a_t, \vec{a}_t\right) = \min_{y_t \geq x_t} g_t\left(y_t, \vec{a}_t\right), \quad \text{if } 1 \leq t \leq T. \tag{4}$$

We first show the essential convexity results that will allow us to establish the optimal policy.

Theorem 1. *For any arbitrary value of the information horizon n and value of the ACI vector $\vec{a}_t$, the following holds for all t:*

1. *$g_t\left(y, \vec{a}\right)$ is convex in y,*
2. *$f_t\left(x, \vec{a}\right)$ is convex in x.*

Based on the convexity results, minimizing g_t is a convex optimization problem for arbitrary ACI horizon parameter n. This allows us to establish the structure of the optimal policy in the following theorem.

Theorem 2. *Let $\hat{y}_t(\vec{a}_t)$ be the smallest minimizer of the function $g_t\left(y_t, \vec{a}_t\right)$. For any $\vec{a}_t$, the following holds for all t:*

1. *The optimal ordering policy under ACI is a state-dependent base stock policy with the optimal base stock level $\hat{y}_t(\vec{a}_t)$.*
2. *Under the optimal policy, the inventory position after ordering $y_t(\vec{a}_t)$, in the case of*

supply capacity availability $a_t = 1$, *is given by:*

$$y_t(\vec{a}_t) = \begin{cases} \hat{y}_t(\vec{a}_t), & x_t \le \hat{y}_t(\vec{a}_t), \\ x_t, & x_t > \hat{y}_t(\vec{a}_t). \end{cases} \quad (5)$$

Base stock policy obtained is characterized by a single base stock level $\hat{y}_t(\vec{a}_t)$, which determines the optimal level of inventory position after ordering. The base stock level $\hat{y}_t(\vec{a}_t)$ is a function of future supply availability given by the vector of ACI $\vec{a}_t$. The optimal inventory policy instructs the manager to raise the inventory position up to the base stock level in the case when the inventory position before ordering is below base stock level. However, if the inventory position exceeds the base stock level, the order should not be placed.

3 VALUE OF ACI

In this section we present the results of the numerical analysis, which was carried out to quantify the value of ACI, and to gain insights on how the value of ACI changes with the change in the relevant system parameters. Numerical calculations were done by solving the dynamic programming formulation given in Eq. (2). We use the following set of input parameters: $T = 20,\ n = (0,1,2,3,4,5)$, probability of capacity availability $p = (0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1)$, the cost structure $h = 1,\ b = (5,20,100)$, discount factor $\alpha = 0.99$, uniformly distributed demand with the expected value of $E(D) = 5$ and coefficient of variation $CV_D = (0,0.3,0.6)$. The results are presented in Figures 1 and 2.

To determine the value of ACI, the performance comparison between the no information case, $n = 0$, and the case where ACI is given for a certain number of future periods, $n > 0$. We define the *relative value* of ACI for $n > 0$ as:

$$\%V_{ACI}(n > 0) = \frac{f_{n=0} - f_{n>0}}{f_{n=0}}. \quad (6)$$

We also define the *absolute change* in the value of ACI with which we measure the extra benefit gained by extending the length of the ACI horizon by one time period, from n to $n+1$:

$$\Delta V_{ACI}(n+1) = f_n - f_{n+1}. \quad (7)$$

Let us first observe the effect of changing the system parameters on the total cost. Obviously, decreasing the extent of supply capacity availability through decreasing the value of p will increase the costs. Due to increased probability of multiple consecutive periods of zero capacity, the likelihood of backorders occurring will increase dramatically, and the costs will increase particularly in the case of high b/h cost ratio. Additionally, costs are higher when one has to deal with the effect of increasing demand uncertainty. The demand uncertainty causes the deviations of the actual inventory levels from the desired target levels set by the manager. This results in frequent mismatches between the demand and the available inventory on-hand, and consequently high costs.

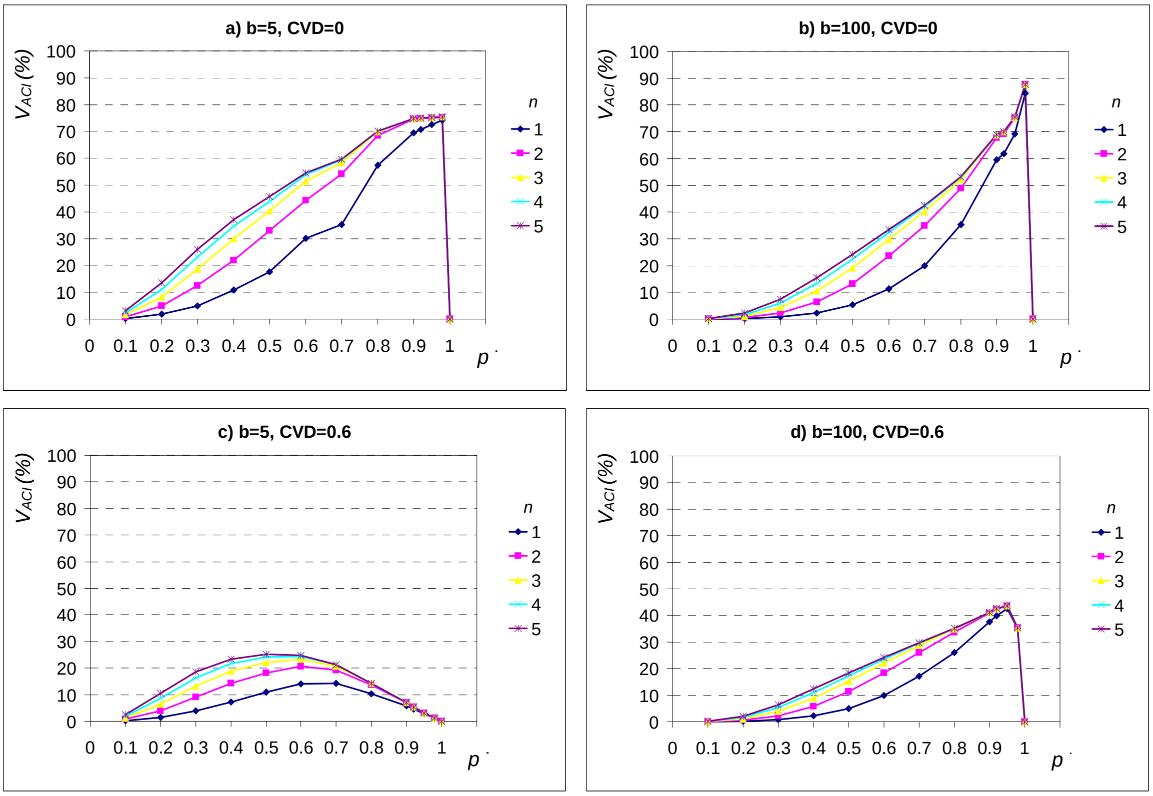


Figure 1: Relative value of ACI.

These costs can be effectively decreased when ACI is available. Extending the ACI horizon obviously also increases the extent of the cost savings. However the marginal gain of increasing n by 1 period varies substantially depending on a particular setting. When we consider the case of low supply capacity unavailability (p close to 1), we observe surprisingly high relative decrease in costs measured through $\%V_{ACI}$. This can be attributed to the fact that ACI enables us to anticipate and prepare for the rare periods of complete capacity unavailability. Thus, we can avoid the backorders and at the same time lower the inventory levels that we would otherwise need to compensate for the event of multiple successive periods with zero capacity. Especially in the case of low demand uncertainty, and also high b/h ratio, $\%V_{ACI}$ can reach levels above 80%, even close to 90% (Figure 1a and 1b). When manager wants to gain the most from the anticipation of future supply capacity unavailability, it is helpful if there are no additional uncertainties present that would prevent him to meet the desired target inventory level. Observe also that these high relative cost savings are already gained with a short ACI horizon. Extending n above 1 only leads to small additional cost reductions. This is an important insight for the practical usage of ACI, when the majority of the gains are already possible with a limited future visibility it is more likely that the manager will be able to obtain ACI (possibly also more accurate) from his supplier. While in the case of p being close to 1 the short ACI horizon is sufficient, we see that longer ACI horizon is needed in a setting with high supply capacity unavailability. Observe that for low p values the relative marginal savings are actually increasing. When multiple periods of zero capacity can occur one following another it is particularly important to anticipate the extent of future capacity unavailability. In such a setting, it is very important if one can have an additional period of future visibility. Several researchers that have studied a conceptually similar problem of sharing advance demand information (ADI) suggest that prolonging the information horizon has diminishing returns [12]. Although we consider a special case of zero or full supply availability in this paper, this result actually shows that this does not hold in general.

While we have observed a large relative decrease in costs in some settings, it may be more important for a particular company to determine the potential decrease in absolute cost figures. Intuitively, we would expect that highest absolute gains would occur in the setting where the uncertainty of supply is high, and the possible shortage anticipation through ACI would be most beneficial. We see that the biggest absolute savings are attained for availability probability between 0.2 and 0.4. Here, the cost decrease is bigger for higher n (Figure 2). In fact, the lower the availability the more we gain by prolonging the ACI horizon. In the case of extremely low levels of capacity availability the system becomes too hard to manage due to an extremely long inventory pre-build phase, and the gains from using ACI are limited.

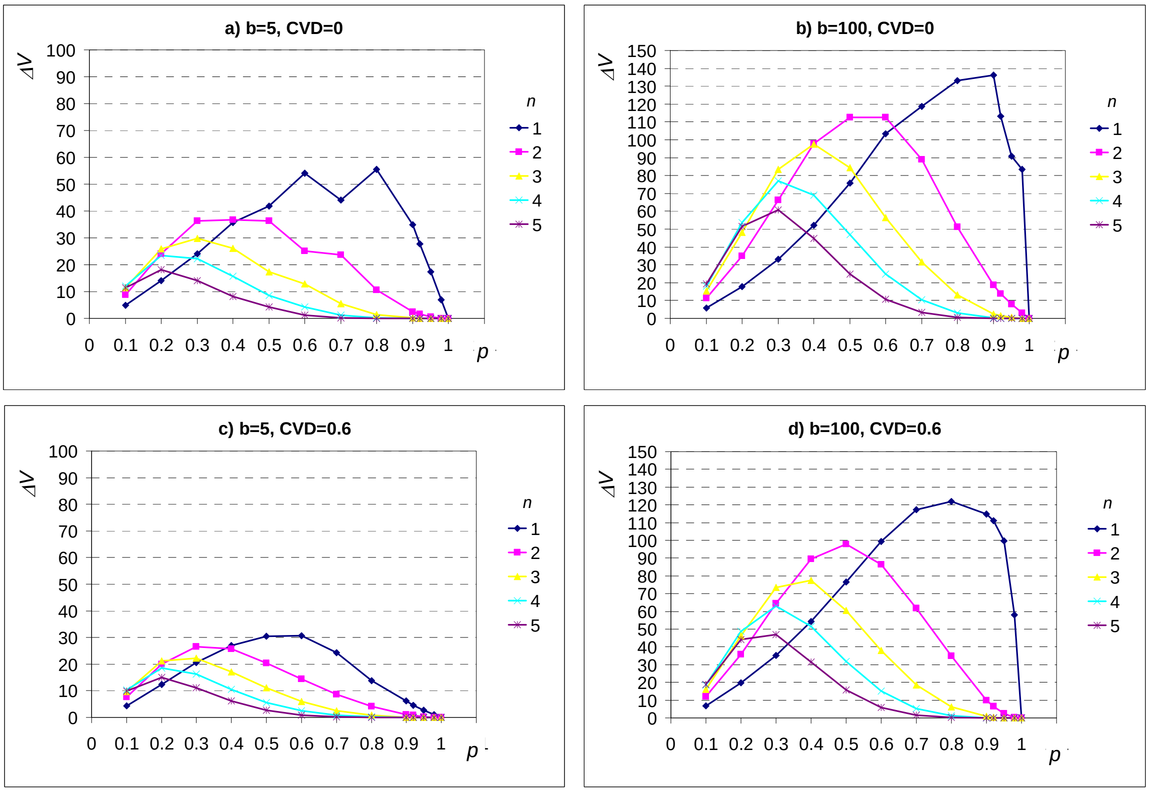


Figure 2: Absolute change in the value of ACI.

The effect of demand uncertainty on absolute decrease in costs is not as obvious as it was in the relative case. This can be attributed to the interaction of two factors. While increased demand uncertainty intensifies the difficulties of managing the inventories as described already above, it also contributes to higher costs and thus provides more potential for savings through ACI.

4 CONCLUSIONS

In this paper, we study a periodic review inventory model under stochastic demand and limited supply availability. The supply capacity is modeled as a Bernoulli processes, meaning that there are randomly interchanging periods of complete capacity unavailability and full availability. We upgrade the base case with no information on future supply capacity availability by considering the possibility that a supply chain member can obtain advance capacity information (ACI) from his upstream partner. We develop an optimal policy and show that it is a base stock policy with a state-dependent base stock level. The optimum base stock level is determined by the currently available ACI, where in the case of information about the unfavorable supply capacity conditions in the future periods, the base stock level is raised sufficiently to avoid the probable stock-outs. By means of numerical analysis, we quantify the benefit of ACI and determine the situations when obtaining ACI can be of particular importance. While the relative cost savings are the highest for the case of close to full availability due to the fact that one can completely avoid backorders with only little extra inventory, the cost reduction in absolute terms is higher for the cases with medium to low supply capacity availability. Further, we show that in most of the cases only little future visibility already offers considerable savings, however when one is faced with a possibility of consecutive periods of supply unavailability it can be very beneficial to extend the ACI horizon. In general, the managers should recognize that the extent of the savings shown clearly indicates that sharing ACI should be encouraged in supply chains with unstable supply conditions. In our experience the current dynamic programming cost formulation is manageable in terms of complexity of calculations and can be used also for larger practical problems. However, a natural extension to this work would be to develop a simpler, preferably also optimal, inventory policy that would capture the effect of sharing ACI.

Appendix

Lemma 1: *If* $g(y,e)$ *is convex then* $f(x,c,e) = \min_{x \le y \le x+c} g(y,e)$ *is also convex for any* $c \ge 0$.

Proof: Let $h(b,e) := \min_{Ay \le b} g(y,e)$ where $A = [1,1]$ and $b = [-x, x+c]$. By minimization on polyhedra property ([11], [13]), we conclude that $h(b,e)$ is convex. Since $h(b,e) = f(x,c,e)$, f is also convex. □

Proof of Theorem 1: The proof is by backward induction starting in time period T.

$t = T$: From Eq. (3), we have $g_T(y_T) = C_T(y_T)$ by taking $f_{T+1}(\cdot) \equiv 0$ into account. Since the reassigned single-period cost function $C_T(y_T)$ is assumed to be convex, function $g_T(y_T)$ is also convex. For $f_T(x_T, a_T) = \min_{y_T \ge x_T} g_T(y_T)$ we apply Lemma 1, and show that function $f_T(x_T, a_T)$ is convex.

$t = T-1$: $f_{T-1}(x_{T-1}, a_{T-1}, \vec{a}_{T-1}) = \min_{y_{T-1} \ge x_{T-1}} \{C_{T-1}(y_{T-1}) + \alpha E_{d_{T-1}} f_T(y_{T-1} - d_{T-1}, a_T)$. We have shown that $f_T(x_T, a_T)$ is convex, thus using affine mapping property [14] we show that function $\tilde{f}_T(y_{T-1}, d_{T-1}, a_T) := f_T(y_{T-1} - d_{T-1}, a_T)$ is also convex (update of the inventory position is linear; thus linear translation with $b = d_{T-1}$). $\alpha E_{d_{T-1}} f_T$ is convex since expectation preserves convexity [15] and by adding cost function $C_{T-1}(y_{T-1})$ and using weighted sum property [14] we show that $g_{T-1}(y_{T-1}, a_T)$ is convex. g_{T-1} is then minimized and through Lemma 1 we conclude that $f_{T-1}(x_{T-1}, a_{T-1}, \vec{a}_{T-1})$ is also convex.

$t = T-2, \ldots, 1$: The proof follows the same line as a previous step using backward induction on t, and thus proving the convexity of $f_t(x_t, a_t, \vec{a}_t)$. □

Proof of Theorem 2: Convexity results of Theorem 1 directly imply the proposed structure of the optimal policy. □

References

[1] Quinn, F. J., 1997. What's the buzz? Supply Chain Management; part 1. Logistics Management, 36, 43.

[2] Silver, A. E., Pyke D.F., Peterson R., 1998. Inventory Management and Production Planning and Scheduling. 3rd Edition. John Wiley & Sons, New York, NY.

[3] Bush, C., Cooper W., 1988. Inventory level decision support. Production and Inventory Management Journal, 29(1), 16–20.

[4] Metters, R., 1997. Production planning with stochastic seasonal demand and capacitated production. IIE Transactions, 29, 1017–1029.

[5] Henig M.., Gerchak Y., 1990. The structure of periodic review policies in the presence of random yield. Operations Research, 38 (4), 634-654.

[6] 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, 40, 320–332.

[7] Jakšič, M., Fransoo J.C., Tan T., de Kok A. G., Rusjan B., 2008. Inventory management with advance capacity information. Beta publicatie, wp 249, Beta Research School for Operations Management and Logistics, Eindhoven University of Technology, The Netherlands.

[8] Wang Y., Gerchak Y., 1996. Periodic review production models with variable capacity, random yield, and uncertain demand. Management Science, 42, 130-137.

[9] Güllü, R., Önol E. and Erkip N., 1997. Analysis of a deterministic demand production / inventory system under nonstationary supply uncertainty. IIE Transactions, 29, 703-709.

[10] Güllü, R., Önol E. and Erkip N., 1999 Analysis of an inventory system under supply uncertainty. Int. J. Production Economics, 59, 377-385.

[11] Porteus E. L., 2002. Foundations of Stochastic Inventory Theory. Stanford University Press, Stanford.

[12] Özer, Ö., Wei W., 2004. Inventory control with limited capacity and advance demand information. Operations Research, 52, 988–1000.

[13] Mincsovics, G., Tan T. and Alp O. 2009. Integrated capacity and inventory management with capacity acquisition lead times. European Journal of Operational Research, 196, 949-958.

[14] Hiriart-Urruty, J.B. and Lemaréchal C. 1996. Convex Analysis and Minimization Alogorithms II. Springer-Verlag, Berlin.

[15] Heyman, D. and Sobel M. 1984. Stochastic Models in Operations Research Vol. II. McGraw-Hill, New York, NY.

MULTISTAGE REVERSE LOGISTICS OF ASSEMBLY SYSTEMS IN EXTENDED MRP THEORY

Danijel Kovačić and **Ludvik Bogataj**
University of Ljubljana, Faculty of Economics
Kardeljeva ploščad 17, SI-1000 Ljubljana, Slovenia
{ludvik.bogataj,danijel.kovacic}@ef.uni-lj.si

Abstract: This paper is addressing the issue of environmental problems in connection with extended MRP Theory. Based on previous work on MRP Theory developed by Grubbström, R.W. and extended by Bogataj, M. and Bogataj, L. to distribution part of supply chain Grubbström, Bogataj and Bogataj later extended the theory including reverse logistics of final and used products. Here we present the input-output model including all possible flows into recycling sub-process, having several stages of recycling. Contribution to the net present value of all activities in a supply chain is expressed which can be compared with environmental damages if the reverse processes would not be introduced. Using this approach the losses and gains between economy and environment could be better evaluated.

Keywords: MRP theory, environment, reverse logistics, input-output analysis, recycling.

1 INTRODUCTION

Environmental issues have been investigated and presented in a wide range of academic papers. However, the decision-making process, regarding optimal waste management in more complex systems demands employment of appropriate tools. Development and implementation of such decision-making tools is in the domain of operational research and production economics. Decision-making models can be classified into two groups: (1) models supporting macro level decisions, and (2) models supporting supply chain management (see Fig. 1), which also contains an environmental component.

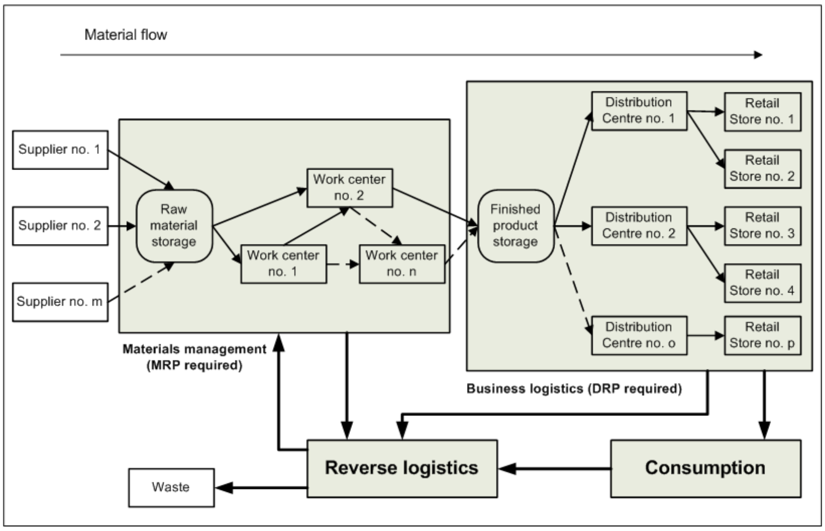


Figure 1: Supply chain extended to reverse logistics part.

Both types of models considered in the literature are based on classical input-output analysis [15], whereby initial investigation of cross-regional flows was expanded to incorporate macro level environmental issues as well. During development of this theory spatial input-output analysis has been developed [14] and input-output analysis has been introduced in production management as well [16]. By distribution of identical items to different locations at different regions, regional aspect could be combined with supply chain approach and further developed in compound theory [9], but this is not the main topic here.

Theory of Leontief and Isard was also used in studies of environmental issues in Slovenia [4], [5]. Additionally, examination of environmental issues with input-output analysis in Central European region emerged within Hungarian academic circles, more specifically with a joint Slovenian-Hungarian project [6]. Input-output analysis was further developed by a number of researchers. [7] provides a review of numerous studies that theoretically expand the input-output system, introduce novel coefficient matrices and present practical applications of the system. In addition, input-output models can be implemented in the analyses of environmental issues [17], which are based on two environmental aspects: (1) natural resources, which represent the system's inputs, and (2) pollution, which is the output. Moreover, input-output analysis represents one of the main components of the MRP theory, as presented in [3], [8], [9], [11] and [12].

Modeling of production processes based on MRP and input-output analysis is investigated in several academic studies. However, Grubbström and Ovrin were the first to additionally include stocks and timing of production activities based on time lags, where they used a *z*-transformation [12]. The authors introduced a generalized input matrix which contains both inputs and time lags. Grubbström and Molinder improved the model in such a way that the discrete scale has been replaced by the continuous one, using the Laplace transformation [11]. These studies also present fundamental equations of the MRP theory that serve as a base for research in this area. The model was improved in terms of capacity limits [10], spatial issues (e.g., [2], [3]), inclusion of game theory (e.g., [1], [13]), and inclusion of reverse logistics into the supply chain [9]. There recycling subsystem of the MRP theory was considered as a subsystem with one input (i.e., used final products) and two outputs (i.e., recycled components and waste). Further, labor was included as an input, thus the net preset value of the recycling subsystem can be specified as follows:

$$\mathrm{NPV}_{\text{recycling}} = \frac{\left(p_2 - p_7\right) A\hat{L}^{\gamma}\hat{P}_6^{\delta} + p_7\hat{P}_6 - c_L\hat{L} - K_6}{1-e^{-\rho T_6}} - \frac{p_6\hat{P}_6}{\rho T_6} \qquad (1)$$

where: (1) p_2 is value of returned products, (2) p_7 is environmental tax, (3) p_6 is price of wasted/used products, (4) $\hat{P}_6$ is batch size of used items, (5) K_6 is setup cost, (6) $c_L\hat{L}$ is labor cost measured as price per unit of labor c_L times quantity of labor used in production $\hat{L}$, (6) T_6 is time period and (7) A, γ, δ are parameters of Cobb-Douglas production function.

The model developed by Grubbström, Bogataj and Bogataj [9] shows how to present complete system consisting of four sub-systems and also gives a simple numerical example. However, the purpose of this paper is to extend the basic MRP theory in such a way that it includes all possible flows of final or unfinished items from any level of the system into recycling subsystem. We will demonstrate that it is possible to specify this model in a generalized form, where lean production and immediate removal of bad items from the main cycle can be modeled and consequently that extended MRP theory is appropriate tool for modeling environmental-economic problems.

2 PRESENTATION OF THE EXTENDED MODEL

Pollution is a real life problem, which is not only a result of household consumption but mainly arises from economic activities. Therefore, studying additional flows in the input-output model of extended MRP theory consisting of four main sub-processes is sensible. Fig. 2 presents all possible flows between sub-processes and is based on multistage assembly production system.

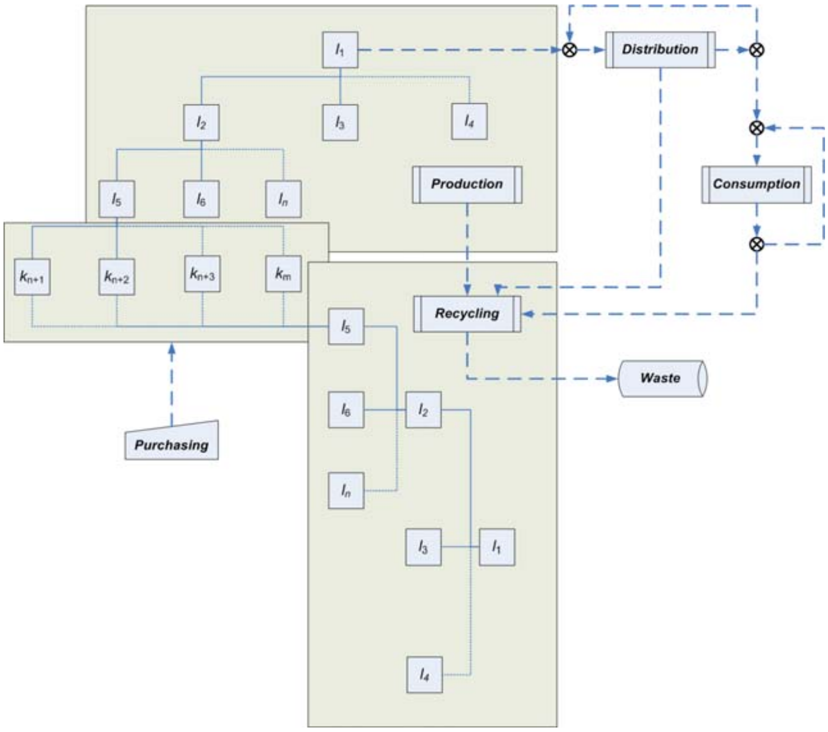


Figure 2: Presentation of all possible material flows in extended MRP theory.

Multistage assembly system can be presented as bill of materials (BOM), where every item on a higher level is assembled from all components or semi-products from a lower level. k_{n+1}, k_{n+2}, …, k_m are items entering the system on the lowest level, from which components are further assembled. They are basic so they cannot be disassembled further. They enter the system as new items from purchasing activities on higher level or exit it to the recycling where the sum of both flows is of the same quantity as the quantity of inputs (no inventory is accumulating on any level for a longer time). Items going to recycling also circulate in the system as a result of recycling activities partly, and are partly disposed forever. Returned items are being extracted from recycled components, entering recycling sub-process from any stage of production, distribution or consumption sub-process. l_2, l_3, …, l_n are assembled components, which can be used for further production, resulting in a new component on higher level or they can proceed directly into recycling sub-process. l_1 is a component on the top level and presents final product, which usually enters distribution process, from there it continues to consumption or directly to recycling sub-process. Every assembled component finishes its cycle in recycling sub-process where it is disassembled in inverse way as assembled. Disassembled items can finish as waste in disposal area or can be returned into the system as entry items of next production cycle.

For proper cyclical presentation it is essential to assume a closed system, which means that no component can be lost inside the system, where disposal is also treated as an element of the system. Every component on the lowest level finishes its way either as recycled component that enters production process again or as waste material, which needs to be disposed. Assumption of the closed system is realistic since every component in reality finishes its lifecycle and cannot stay infinitely in any sub-process.

Due to simplicity, the model presented in this paper only considers one final product l_1 and infinite number of assembly levels and sub-components in the assembly process. The presented general model can relatively easily be upgraded with more final products by simply multiplying rows and columns belonging to l_1 for every additional final product. However, it would seem more sensible to observe a separate model in this case.

2.1 General characterization of the input matrix

Input matrix presents input requirements of the system. It has positive sub-matrices where input requirements appear. Based on presented possible flows in this paper input matrix would be structured like:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & 0 & \mathbf{H}_{14} \\ 0 & 0 & \mathbf{H}_{23} & \mathbf{H}_{24} \\ 0 & 0 & 0 & \mathbf{H}_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Sub-matrix H_{ij} relates to items required by sub-system i running processes in sub-system j where 1 represents production, 2 distribution, 3 consumption and 4 recycling sub-system. We can generalize positive sub-matrices in the following way:

$$\mathbf{H}_{11} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ h_{31}^{11} & 0 & h_{33}^{11} & \cdots & h_{3n}^{11} \\ 0 & 0 & 0 & \cdots & 0 \\ h_{51}^{11} & h_{52}^{11} & 0 & \cdots & h_{5n}^{11} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{(2n-1)1}^{11} & h_{(2n-1)2}^{11} & h_{(2n-1)3}^{11} & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ h_{(2n+1)1}^{11} & h_{(2n+1)2}^{11} & h_{(2n+1)3}^{11} & \cdots & h_{(2n+1)n}^{11} \\ h_{(2n+2)1}^{11} & h_{(2n+2)2}^{11} & h_{(2n+2)3}^{11} & \cdots & h_{(2n+2)n}^{11} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{m1}^{11} & h_{m2}^{11} & h_{m3}^{11} & \cdots & h_{mn}^{11} \end{bmatrix}, \mathbf{H}_{12} = \begin{bmatrix} h_{11}^{12} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{H}_{14} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ h_{21}^{14} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & h_{42}^{14} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & h_{63}^{14} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & h_{(2n)n}^{14} \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Coefficients h_{kl}^{11} in $\mathbf{H}_{11}$ represent assembly coefficients in production sub-process. The condition: $h_{kl}^{11} \geq 0$ must be fulfilled. Upper part of sub-matrix $\mathbf{H}_{11}$ has coefficients running from h_{11}^{11} to $h_{(2n)n}^{11}$ which show how many components k we need for production of one component (or final product) l on higher level. Every element k is presented in two rows where row $2k$ always consists of zero elements and row $2k$-1 of non-zero elements. For k = 1 we only have zero elements since this is final product not being used in further assembly processes. Also when $k = l$ we only have zero elements since produced component can not be used for assembly process of itself. For every row $2k$ there will be positive elements in upper part of sub-matrix $\mathbf{H}_{14}$ presenting input coefficients into recycling from running production activities for every component k where k = 1, 2, …, n. In $\mathbf{H}_{14}$ there will only be positive elements when $k = l$ since only assembled components 1, 2, …, n can enter into recycling sub-process. Similarly, sub-matrix $\mathbf{H}_{12}$ will only have one non-zero element (h_{11}^{12}) presenting input coefficient into distribution from production sub-process. Coefficients in $\mathbf{H}_{12}$ and $\mathbf{H}_{14}$ are actually proportions of total production of every component k entering distribution or recycling sub-process. Let us mark these proportions with β_{ij} where i, j present flow from sub-system i to sub-system j. The following conditions must be fulfilled: $0 \leq \beta_{ij} \leq 1$; $\sum_{i=1}^{4} \beta_{ij} = 1$.

Lower part of sub-matrix $\mathbf{H}_{11}$ has coefficients running from $h_{(2n+1)1}^{11}$ to h_{mn}^{11} which show how many basic elements *k* we need for production of one component *l* on higher level. Here every element *k* is presented in one row since basic elements have no input flow into recycling sub-process in sub-matrix $\mathbf{H}_{14}$.The second row of matrix **H** consists of two sub-matrices with non-zero elements:

$$\mathbf{H}_{23}=\begin{bmatrix} h_{11}^{23} \\ \vdots \\ h_{r1}^{23} \\ 0 \end{bmatrix}, \mathbf{H}_{24}=\begin{bmatrix} 0 \\ \vdots \\ 0 \\ h_{(r+1)1}^{24} \end{bmatrix}$$

Coefficients h_{r1}^{23} show proportion of all produced final components distributed to locations 1, 2, ..., *r* as demand from consumption sub-process because of distribution activities. Similarly, $h_{(r+1)1}^{24}$ shows proportions of all produced final components demanded from recycling sub-process because of distribution activities. We can define h_{r1}^{23} as: $h_{r1}^{23}=\beta_{12}*\beta_{23}*\mu_r$ where μ_r presents proportions of items distributed to location 1, 2, …, *r*. The following conditions must be fulfilled: $0 \leq \mu_r \leq 1$; $\sum_{r=1}^{r}\mu_r = 1$. Further, $h_{(r+1)1}^{24}$ can be defined as: $h_{(r+1)1}^{24}=\beta_{12}*\beta_{24}$. Sub-matrix $\mathbf{H}_{34}$ presents inputs required by recycling sub-system because of running the consumption activities. $\mathbf{H}_{34}$ has only one coefficient h_{11}^{34} which can be defined as: $h_{11}^{34}=\beta_{12}*\beta_{23}*\beta_{34}$.

2.2 General characterization of the output matrix

Output matrix presents output coefficients of the system. It has positive sub-matrices where outputs appear. Based on presented possible flows in this paper output matrix would be structured like:

$$\mathbf{G}=\begin{bmatrix} \mathbf{G}_{11} & 0 & 0 & \mathbf{G}_{14} \\ 0 & \mathbf{G}_{22} & 0 & 0 \\ 0 & 0 & \mathbf{G}_{33} & 0 \\ 0 & 0 & 0 & \mathbf{G}_{44} \end{bmatrix}$$

Sub-matrix $\mathbf{G}_{ij}$ relates to outputs of sub-system *i,* running processes in sub-system *j*. The dimension of output matrix **G** should always coincide with input matrix **H**. We can generalize positive sub-matrices in the following way:

$$\mathbf{G}_{11}=\left[\begin{array}{c|c|c|c|c} g_{11}^{11} & 0 & 0 & \cdots & 0 \\ g_{21}^{11} & 0 & 0 & \cdots & 0 \\ \hline 0 & g_{32}^{11} & 0 & \cdots & 0 \\ 0 & g_{42}^{11} & 0 & \cdots & 0 \\ \hline 0 & 0 & g_{53}^{11} & \cdots & 0 \\ 0 & 0 & g_{63}^{11} & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & 0 & \cdots & g_{(2n-1)n}^{11} \\ 0 & 0 & 0 & \cdots & g_{(2n)n}^{11} \\ \hline 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{array}\right], \mathbf{G}_{22}=\begin{bmatrix} g_{11}^{22} \\ \vdots \\ g_{r1}^{22} \\ g_{(r+1)1}^{22} \end{bmatrix}, \mathbf{G}_{33}=\begin{bmatrix} g_{11}^{33} \end{bmatrix}$$

$\mathbf{G}_{11}$ presents output proportions of production sub-process and coincides with $\mathbf{H}_{11}$ in dimension. Output of every component k is presented in two rows where row $2k$ always presents its output to recycling sub-process and row $2k$-1 presents outputs further in production process. We should point out that for k = 1, output of the first row, presented by g_{11}^{11}, shows proportion of total final production going to distribution sub-process. $g_{(2l-1)l}^{11}$ presents proportions of component l used for further assembly processes and $g_{(2l)l}^{11}$ proportions of these components sent directly to recycling sub-system. The following conditions must be fulfilled: $0 \leq g_{(2l)l}^{11}, g_{(2l-1)l}^{11} \leq 1$; $g_{(2l)l}^{11} + g_{(2l-1)l}^{11} = 1$. These proportions coincide with β_{ij} in input matrices $\mathbf{H}_{ij}$.

$\mathbf{G}_{22}$ presents output proportions of distribution sub-process g_{r1}^{22} to locations 1, 2, …, r and $g_{(r+1)1}^{22}$ to recycling sub-system. We can define g_{r1}^{22} as: $g_{r1}^{22} = \beta_{12} * \beta_{23} * \mu_r$ and $g_{(r+1)1}^{22}$ as: $g_{(r+1)1}^{22} = \beta_{12} * \beta_{24}$. This means that: $\sum_{r=1}^{r+1} g_{r1}^{22} = \beta_{12}$.

$\mathbf{G}_{33}$ presents output proportions of consumption sub-process presented as g_{11}^{33} which can be defined as: $g_{11}^{33} = \beta_{12} * \beta_{23} * \beta_{34}$.

From reverse logistics point of view sub-matrices $\mathbf{G}_{14}$ and $\mathbf{G}_{44}$ are very important. They include return rate coefficients α which define proportion of recycled item as a result of running recycling activities on any component in any state of the system.

$$\mathbf{G}_{14} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \\ \hline g_{(2n+1)1}^{14} & g_{(2n+1)2}^{14} & \cdots & g_{(2n+1)n}^{14} \\ g_{(2n+2)1}^{14} & g_{(2n+2)2}^{14} & \cdots & g_{(2n+2)n}^{14} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1}^{14} & g_{m2}^{14} & \cdots & g_{mn}^{14} \end{bmatrix}, \mathbf{G}_{44} = \begin{bmatrix} g_{11}^{44} & g_{12}^{44} & \cdots & g_{1n}^{44} \\ g_{21}^{44} & g_{22}^{44} & \cdots & g_{2n}^{44} \\ \vdots & \vdots & \ddots & \vdots \\ g_{(m-2n)1}^{44} & g_{(m-2n)2}^{44} & \cdots & g_{(m-2n)n}^{44} \end{bmatrix}$$

$\mathbf{G}_{14}$ presents output coefficients of recycling activities which can be reused in next cycle of production activities and $\mathbf{G}_{44}$ outputs of recycling activities which have to be disposed (waste elements). Lower part of $\mathbf{G}_{14}$ coincides with $\mathbf{G}_{44}$. Since we suppose closed system where none of the elements stay at any state for a longer period, meaning that every component comes to recycling sub-process, we can define that: $g_{kl}^{14} + g_{(k-2l)l}^{44} = h_{kl}^{11}$ for every k, l. This means that every assembled element on the lowest level will end as recycled element, appropriate for reuse, or waste element which will be disposed off. g_{kl}^{14} and $g_{(k-2l)l}^{44}$ can now be expressed as: $g_{kl}^{14} = h_{kl}^{14} * \alpha_{kl}$ and $g_{(k-2l)l}^{44} = h_{kl}^{11} * (1-\alpha_{kl})$.

Coefficients α_{kl} shows return rate of elementary item k being recycled from component l. They will have value greater than 0 only in case when recycled component l will be exactly one level above item k in the grid. Final return rate α_{kl} depends on all coefficients α and β defining further flows until recycling and can be formulated in the following way:

$$\alpha_{kl} = \sum_{n'=l}^{3} \left(\prod_{i=k,l,l-1,\ldots n',4}^{n',4} \beta_i \alpha_i \right)_{n'} \tag{2}$$

Index i shows sequence of activity cells from lowest to highest level and can take any feasible combination of the following values from assembly system: any basic element, component or final product from k or l consisted in $\mathbf{H}_{11}$ or any of sub-systems 1, 2, 3 or 4. The first parameter in this sequence is always component from the lowest level which is being recycled. Index n' defines activity cell from where we enter into recycling sub-system. Last element of the sequence is always 4, since every flow must end in recycling subsystem.

If we take for example i = k5, l4, l1, 2, 4 this would present extraction process of element k5 which was further assembled to l4 and at the end to the final product l1. This final product entered the distribution sub-process but never came to consumption. Instead, it entered directly into recycling sub-process. We should multiply all coefficients α_i and β_i which occurred during all transformations to higher level: (1) k5 from l4, (2) l4 from l1, (3) l1 from 2 and (4) 2 from 4. Obviously n' here is distribution sub-system (2).

Differentiating return rates α_i is sensible. If we recycle l from its forerunner l-1 we can expect higher return rates when extracting from lower level of system state from where transformation from activity cell n' to recycling process happened. The higher in grid n' stands, the lower return rates between lower levels are to be expected.

2.3 General characterization of the net production

Activity vector **P** represents cumulative quantities of every assembled component l. Since we assumed closed system, net production **z** can be generalized in the following way:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \\ \mathbf{P}_3 \\ \mathbf{P}_4 \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \\ \hline Q_n \\ \hline Q_n \\ \hline Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{bmatrix}, \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \mathbf{z}_3 \\ \mathbf{z}_4 \end{bmatrix} = (\mathbf{G} - \mathbf{H})\mathbf{P} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ Q_{2n+1}^{14} \\ Q_{2n+2}^{14} \\ \vdots \\ Q_m^{14} \\ \hline 0 \\ \vdots \\ 0 \\ \hline 0 \\ \hline Q_{2n+1}^{44} \\ Q_{2n+2}^{44} \\ \vdots \\ Q_m^{44} \end{bmatrix}$$

Only final product l1 enters distribution and consumption sub-process, consequently vectors $\mathbf{P}_2$ and $\mathbf{P}_3$ only contains cumulative produced quantity of the final product. In net production vector $\mathbf{z}_1$ presents net production of items used in manufacturing, $\mathbf{z}_2$ of distributed items, $\mathbf{z}_3$ of consumed items and $\mathbf{z}_4$ of recycled items. Components l are being assembled and after use completely disassembled, which means that their net production in vector $\mathbf{z}_1$ is zero. Input elements on the lowest level are being recycled. Consequently Q_m^{14} in vector $\mathbf{z}_1$ represents recycled quantity of item m which will be reused in the next production cycle. Q_m^{44} represents quantity of items m which will be disposed as waste. This quantity of items has to be purchased for next cycle of production process. Vectors $\mathbf{z}_2$ and $\mathbf{z}_3$ both have only zero elements, which means that all final production that enters distribution or consumption sub-process also continues its flow into any possible next sub-process.

3 NUMERICAL EXAMPLE

Fig. 3 shows BOM for assembly system of final product A extended with all possible flows between sub-systems. Final product A is produced from 2 units of subcomponent B and 2 units of elementary subcomponent E. Subcomponent B is further assembled from 4 units of subcomponent C and 2 units of elementary subcomponent E, while C is assembled from 2 components of elementary subcomponent D.

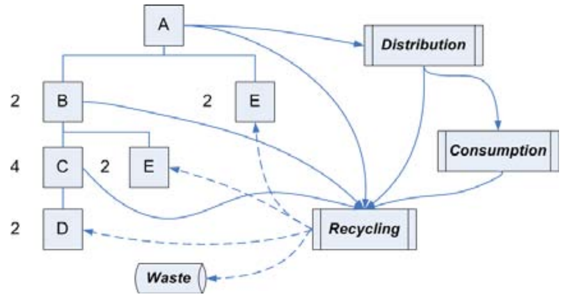


Figure 3: Presentation of all flows in the system.

Elements D and E are basic and consequently recycled from components B and C, and final product A. Considering final product A they can be recycled from three different paths: A entering recycling process directly, A entering recycling process through distribution or A entering recycling process through distribution and consumption. At any possible flow into recycling, 20% of total incoming quantity of the element is being transferred directly into recycling process and 80% of total quantity moves one level higher in the process. When the total amount of 1000 units of final product A is produced, 800 units (80% of 1000) continue their way into distribution sub-process and remaining 200 units (20% of 1000) into recycling sub-system. We need to produce 2500 units of B ($Q_{l=\mathrm{B}} = Q_{l=\mathrm{A}} * h_{31}^{11} \div (\beta_{11})_{l=\mathrm{B}} = 1000 * 2 \div 0.8$) and further 12500 units of C ($Q_{l=\mathrm{C}} = Q_{l=\mathrm{B}} * h_{52}^{11} \div (\beta_{11})_{l=C} = 2500 * 4 \div 0.8$) to cover demand of all processes in the system. From 2500 units of B, 2000 (80% of 2500) are used for further assembly process and remaining 500 (20% of 2500) are sent to recycling. From 12500 units of C, 10000 units (80% of 12500) are used for further assembly process of component B and remaining 2500 units (20% of 12500) are sent to recycling. In the process of distribution, 75% of total quantity of final product A is distributed to location 1 and 25% to location 2.

According to this, we can define coefficients $\beta_{n'j}$ describing inflows into recycling sub-system: β_{34} = 1 and $\beta_{n'4}$ = 0.2 for every other exiting state *n'*. All other coefficients of flows from lower to higher level for this example are β_{ij} = 0.8. Coefficients describing distribution flows of final product A to locations 1 and 2 are: $\mu_1 = 0,75$; $\mu_2 = 0,25$.

Partial return rates α_i for elementary items D and E are showed in Tables 1 and 2. Every α_i in these tables represents recycling rates for all components entering recycling process from production, distribution or consumption process.

Table 1: Partial return rates α_i for elementary item D

Sequence		from 1	from 2	from 3
from C	D,C	0.5		
from B	D,C,B	0.4		
from B	C,B	0.5		
from A	D,C,B,A	0.2	0.2	0.1
from A	C,B,A	0.4	0.3	0.2
from A	B,A	0.5	0.4	0.3

Table 2: Partial return rates α_i for elementary item E

Sequence		from P	from D	from C
from B	E,B	0.5		
from A	E,A	0.7	0.6	0.5
from A	E,B,A	0.4	0.35	0.3
from A	B,A	0.5	0.4	0.3

Next, final return rates have to be calculated. Since D is always being recycled from C and E can be recycled from A or B, we operate with three final return rates, which can be calculated using (2): $\alpha_{k4l3} = 0.1588$, $\alpha_{k5l2} = 0.556$ and $\alpha_{k5l1} = 0.2578$.

Table 3 shows detailed calculations for every feasible *n'*. Empty fields in this table show unfeasible flows of material, i.e. E can never be recycled from component C.

Table 3: Calculations of final return rates $\alpha_{k4l3}, \alpha_{k5l2}, \alpha_{k5l1}$

n'	*l3*	*l2*	*l1*	2	3	Sum
α_{k4l3}	0.2*0.5	0.8*0.2*0.5*0.4	(2*0.8)*0.2*0.5*0.4*0.2	(3*0.8)*0.2*0.4*0.3*0.2	(4*0.8)*0.3*0.2*0.1	0.1588
α_{k5l2}		0.2*0.5	0.8*0.2*0.5*0.4	0.8*0.8*0.2*0.4*0.35	(3*0.8)*0.3*0.3*0.5	0.2579
α_{k5l1}			0.2*0.7	0.8*0.2*0.6	(2*0.8)*0.5	0.556

According to the general characterization we can now define input matrix H and activity vector P as follows:

$$
\mathbf{H} = \left[\begin{array}{ccc|c|c|ccc}
 & & & 0,8 & & & & \\
 & & & & & 0.2 & & \\
2 & & & & & & & \\
 & & & & & & 0.2 & \\
 & 4 & & & & & & \\
 & & & & & & & 0.2 \\
 & & 2 & & & & & \\
2 & 2 & & & & & & \\
\hline
 & & & & 0.8*0.8*0.75 & & & \\
 & & & & 0.8*0.8*0.25 & & & \\
 & & & & & 0.8*0.2 & & \\
\hline
 & & & & & 0.8*0.8 & & \\
\hline
 & & & & & & & \\
 & & & & & & &
\end{array}\right], \quad
\mathbf{P} = \left[\begin{array}{c}
1000 \\ 2500 \\ 12500 \\ \hline 1000 \\ \hline 1000 \\ \hline 1000 \\ 2500 \\ 12500
\end{array}\right]
$$

Output matrix G for this example is as follows:

$$
\mathbf{G}=\left[\begin{array}{ccc|c|c|ccc}
0.8 & & & & & & & \\
0.2 & & & & & & & \\
 & 0.8 & & & & & & \\
 & 0.2 & & & & & & \\
 & & 0.8 & & & & & \\
 & & 0,2 & & & & & \\
 & & & & & & & 2*0.1588 \\
 & & & & & 2*0.556 & 2*0.2579 & \\
\hline
 & & & 0.8*0.8*0.75 & & & & \\
 & & & 0.8*0.8*0.25 & & & & \\
 & & & 0.8*0.2 & & & & \\
\hline
 & & & & 0.8*0.8 & & & \\
\hline
 & & & & & & & 2*(1-0.1588) \\
 & & & & & 2*(1-0.556) & 2*(1-0.2579) &
\end{array}\right]
$$

Next, we can calculate input requirements $\mathbf{x}=\mathbf{HP}$, outputs $\mathbf{y}=\mathbf{GP}$ and net production $\mathbf{z}=\mathbf{y}-\mathbf{x}=\left(\mathbf{G}-\mathbf{H}\right)\mathbf{P}$ as follows:

$$
\mathbf{x}=\mathbf{HP}=\left[\begin{array}{c}
800\\ 200\\ 2000\\ 500\\ 10000\\ 2500\\ 25000\\ 7000\\ \hline 480\\ 160\\ 160\\ \hline 640\\ \hline 0\\ 0
\end{array}\right],
\mathbf{y}=\mathbf{GP}=\left[\begin{array}{c}
800\\ 200\\ 2000\\ 500\\ 10000\\ 2500\\ 3969\\ 2402\\ \hline 480\\ 160\\ 160\\ \hline 640\\ \hline 21031\\ 4598
\end{array}\right],
\mathbf{z}=\mathbf{y}-\mathbf{x}=\left(\mathbf{G}-\mathbf{H}\right)\mathbf{P}=\left[\begin{array}{c}
0\\ 0\\ 0\\ 0\\ 0\\ 0\\ -21031\\ -4598\\ \hline 0\\ 0\\ 0\\ \hline 0\\ \hline 21031\\ 4598
\end{array}\right]
$$

As shown, from 1000 units of final product A produced, 800 enter into distribution sub-system and 200 go directly to recycling. From 800 units, which enter production sub-system, 480 units are distributed to location 1 and 160 units to location 2, and another 160 units are going directly into recycling sub-system. Product that entered consumption sub-process is used and afterwards completely sent to recycling sub-system. From 2500 units of B, 2000 units are used for production of A and 500 units go directly to recycling sub-system. 10000 units of C are used for assembling component B and 2500 units go directly to disposal. For every cycle we need 25000 units of elementary component D and we get 3969 units from recycling. We also need 7000 units of E, whereby we get 2402 units from recycling and we have to buy additional 4598. In sub-process of distribution and consumption all components which entered also left the sub-system and finished in recycling system.

4 NET PRESENT VALUE OF RECYCLING ACTIVITIES IN TOTAL

In such an extended system we would like to evaluate total NPV of all recycling activities. It is equal to sum of individual NPVs at each stage of supply chain. Following the calculations in (1) we can write:

$$\mathrm{NPV}_{\mathrm{recycling}} = \sum_{n'=l}^{3} \frac{\left(p_{2n'} - p_{7n'}\right) A\hat{L}_{n'}^{\gamma}\hat{P}_{6n'}^{\delta} + p_{7n'}\hat{P}_{6n'} - c_L\hat{L}_{n'} - K_{6n'}}{1 - e^{-\rho T_{6n'}}} - \frac{p_{6n'}\hat{P}_{6n'}}{\rho T_{6n'}} \tag{3}$$

Index *n'* defines all activity cells from where we can enter into recycling sub-system.

Following parameters are dependent on entry flow *n'* to recycling sub-system: (1) p_2, (2) p_7, (3) p_6, (4) $\hat{P}_6$, (5) K_6, (6) $\hat{L}$ and (6) T_6. All other parameters of the system are constant for all product flows.

4.1 Calculation of NPV using numerical example

We can now make a short simulation of NPV calculation for numerical example presented in section 3. As presented, 2500 units of component C, 500 units of component B and 200 units of final product A exit production sub-system directly to recycling sub-process. Additional 160 units of final product A enter into recycling sub-system from distribution and remaining 640 from consumption sub-process. Entry quantities are presented with parameter $\hat{P}_6$ and are showed in table 4, together with all other parameters, dependent on *n'*.

Table 4: Presentation of parameters dependent on *n'*

n'	p_2	p_7	p_6	$\hat{P}_6$	K_6	$\hat{L}$	T_6
l3	6	-0.3	0.5	2500	1.6	110	10
l2	7	-0.35	0.8	500	1.8	130	13
l1	9	-0.5	1.2	200	2	140	15
2	8.5	-0.5	1.1	160	2	140	17
3	8	-0.5	1	640	2	140	19

We set values for other *n'* independent parameters as: (1) A = 1.1, (2) γ = 0.4, (3) δ = 0.4, (4) ρ = 0.65 and (5) c_L = 2. Using (3) we can now calculate NPV for every entry flow into recycling sub-system: $\mathrm{NPV}_{l3} = \text{-}125.16$, $\mathrm{NPV}_{l2} = 196.44$, $\mathrm{NPV}_{l1} = 221.43$, $\mathrm{NPV}_2 = 166{,}25$ and $\mathrm{NPV}_3 = 240.00$. NPV of recycling for complete system is sum of all partial NPVs and has value of $\mathrm{NPV}_{\mathrm{recycling}} = 698.96$.

NPV of recycling component C with direct flow into recycling system is not economical since it is negative. In optimal case component C would be completely disposed of when exiting production sub-process with direct flow into recycling sub-system which would increase $\mathrm{NPV}_{\mathrm{recycling}}$ to 824.12.

5 CONCLUSION

In this paper we present a generalized form of the input-output model based on multistage reverse logistics of assembly systems in MRP theory. Multiple flows from any possible stage of the system have been considered using variable coefficients. Let us point out that there are other potential variations of modeling this system. For example, it would be possible to write every outflow of the unit which is being recycled into separate row of input-output matrices, which would result in more detailed information. As a result we would get flow of recycled components for every possible flow in the system. However, getting cumulative data as a result of the model presented in this paper simplifies the system and makes it easier for further studies.

This paper shows how to calculate net present value of complete system using extended model of all possible flows into recycling sub-system. Every single flow can be evaluated. Contribution of every flow to cumulative NPV can be positive or negative which would influence optimization of the system as a whole.

It has been shown that using extended MRP theory is a good basis for researching the wide area of reverse logistics problems. Moreover, taking in consideration many EU directives covering recycling activities such research is necessary and sensible.

References

[1] Bogataj, L., & Horvat, L. (1996). Stochastic considerations of Grubbström-Molinder model of MRP, input-output and multi-echelon inventory systems. International Journal of Production Economics, 45, 329-336.

[2] Bogataj, M., & Bogataj, L. (2001). Supply chain coordination in spatial games. International Journal of Production Economics, 71, 277-285.

[3] Bogataj, M., & Bogataj, L. (2004). On the compact presentation of the lead times perturbations in distribution networks. International Journal of Production Economics, 88, 145–155.

[4] Bogataj, M., & Drobne, S. (1990). Vloga geodezije pri reševanju ekoloških problemov. Geodetski vestnik, 34(1), 41-49.

[5] Bogataj, M., & Drobne, S. (1993). Vidiki internalizacije eksternih stroškov in politike varstva okolja v input-output modelih. V V. Rupnik (ed.), & L. Bogataj (ed.), Zbornik del SOR '93 (pp. 179-192). Ljubljana: Slovensko društvo Informatika.

[6] Bogataj, M., Drobne, S., & Bogataj, L. (1994). Market-oriented real estate policy as a challenge to sustainable development. V Z. Hajdú (ed.), & G. Horváth (ed.), European challenges and Hungarian responses in regional policy (pp. 436-484). Pécs: Centre for Regional Studies, Hungarian Academy of Sciences.

[7] Ciaschini, M. (1988). Input-output analysis: current developments. London, New York: Chapman and Hall.

[8] Grubbström, R.W. (1999). A net present value approach to safety stocks in a multi-level MRP system. International Journal of Production Economics, 59, 361-375.

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

[10] Grubbström, R.W., & Huynh, T.T.T (2006). Multi-level, multi-stage capacity-constrained production–inventory systems in discrete time with non-zero lead times using MRP theory. International Journal of Production Economics, 101, 53–62.

[11] Grubbström, R.W., & Molinder, A. (1994). Further theoretical considerations on the relationship between MRP, input-output analysis and multi-echelon inventory systems. International Journal of Production Economics, 35(1-3), 299-311.

[12] Grubbström, R.W., & Ovrin, P. (1992). Intertemporal generalization of the relationship between material requirements planning and input-output analysis. International Journal of Production Economics, 26(1-3), 311-318.

[13] Horvat, L., & Bogataj, L. (1999). A market game with the characteristic function according to the MRP and input-utput analysis model. International Journal of Production Economics, 59, 281-288.

[14] Isard, W. (1971). Regional Input-output Study: Recollections, Reflections, and Diverse Notes on the Philadelphia Experience. Cambridge, Mass: MIT Press.

[15] Leontief, W. (1966). Input-output economics. New York, Oxford: Oxford University Press.

[16] Orlicky, J.A. (1975). Material Requirements Planning. New York: McGraw-Hill.

[17] Ten Raa, T. (2005). The Economics of Input-Output Analysis. Cambridge, New York: Cambridge University Press.

HOW TO IMPLEMENT KEY PERFORMANCE INDICATORS IN A COMPANY

Dubravko Mojsinović
Consule d.o.o, Dr. Franje Tuđmana 8, 10 434 Strmec Samoborski, Croatia
dmojsinovic@consule.hr

Abstract: Key Performance Indicator (KPI) can be used as a good tool for improving business results. It has a quantitative nature and shows business performance. Example from a retail company illustrates how a KPI can be implemented. Example shows that in a retail company it seems reasonable to monitor Pieces sold per customer and Number of customers as KPIs. Calculation shows the threshold values for these two indicators.

Keywords: KPI, Key Performance Indicator, Retail Industry

1 INTRODUCTION

A key performance indicator (KPI) can be defined as a specific measure of an organization's performance in an area of its business [1]. Companies usually use more KPIs since one KPI can't show all aspects of business. There are attempts, such as Balanced Scorecard model, trying to integrate more KPIs [2]. Usually management chooses a KPI, adds a desired value to it and then monitors actual values against the goal. If values of a KPI are changing unfavorably, managers try to make business decisions to improve performance. This work addresses some questions regarding KPIs using an example from a retail company. Some of these questions are as follows: How to choose KPI? How to measure it and how often? How to extract the necessary data [3]? How to fine tune the methodology? Is one value of a KPI enough?

2 CHOOSING KPI

Managers frequently meet their business partners and attend various conferences. Therefore they have a clear picture of indicators used in different companies. It is important that the chosen figures are related with the company goals and easily understood by employees [4]. In our example the company monitors the following figures: Number of pieces sold, Number of visitors who entered the stores, Number of customers who entered the stores and Number of entrants (as a sum of Number of visitors and Number of customers).

Let us assume that there is enough stock in stores, that there are no assortment and pricing problems and that the capacity limitations do not exist. The favorable direction of these figures is very straightforward:

- Number of pieces sold-INCREASE IT
- Number of visitors who entered the stores-DECREASE IT
- Number of customers who entered the stores-INCREASE IT
- Number of entrants-INCREASE IT

Store employees can significantly influence the first three figures with their selling skills. As for the last figure, they can influence it through the word of mouth of good customers. Word of mouth is a reference to the passing of information from person to person [5]. So, more they sell to the specific customer, more the customer is satisfied and more of the positive word of mouth will happen. Word of mouth is advertising "free of charge". Satisfied customers will talk to their friends and spread the news that the company is excellent and has good assortment, service level and so on. In this way new customers will come. The last figure is significantly influenced by the advertising activities of a company which are usually centralized. The consequence of change in the last figure is the change in the second and third figure. Positive word of mouth also means that existing customers are buying more and more pieces of goods. Looking solely at Pieces sold and Number of customers is not enough. This means that we have to introduce the ratio of the first and the third figure: Number of pieces sold per customer. It seems that the best strategy is:

- Number of pieces sold-INCREASE IT THROUGH VERY GOOD LEVEL OF SERVICE (1)
- Number of visitors who entered the stores-DECREASE IT THROUGH INCREASING NUMBER OF CUSTOMERS
- Number of customers who entered the stores-INCREASE IT THROUGH VERY GOOD LEVEL OF SERVICE (2)
- Number of entrants-INCREASE IT THROUGH INCREASING NUMBER OF CUSTOMERS
- Number of pieces sold per customer-INTRODUCE THIS RATIO AND THEN INCREASE IT THROUGH EXCELLENT LEVEL OF SERVICE (3)

"Very good level of service" is different from "Excellent level of service". The second one is related to the fifth indicator. It means that each customer, on average, is buying more and more pieces of goods. The fifth indicator is Number of pieces sold per customer which represents the ratio of Number of pieces sold to Number of customers. This is definitely one step further from monitoring only pieces sold or number of customers. This means that the bonds between sales staff and customers are becoming increasingly stronger and this figure is pointing directly to the quality of relationship which is more and more important to the companies. It is important to understand that sold pieces can increase slowly and Number of customers fast. In this case quality of relationship is deteriorating. Therefore the fifth indicator is essential. This indicator is used in other companies as well [6]. This leaves us with 3 indicators:

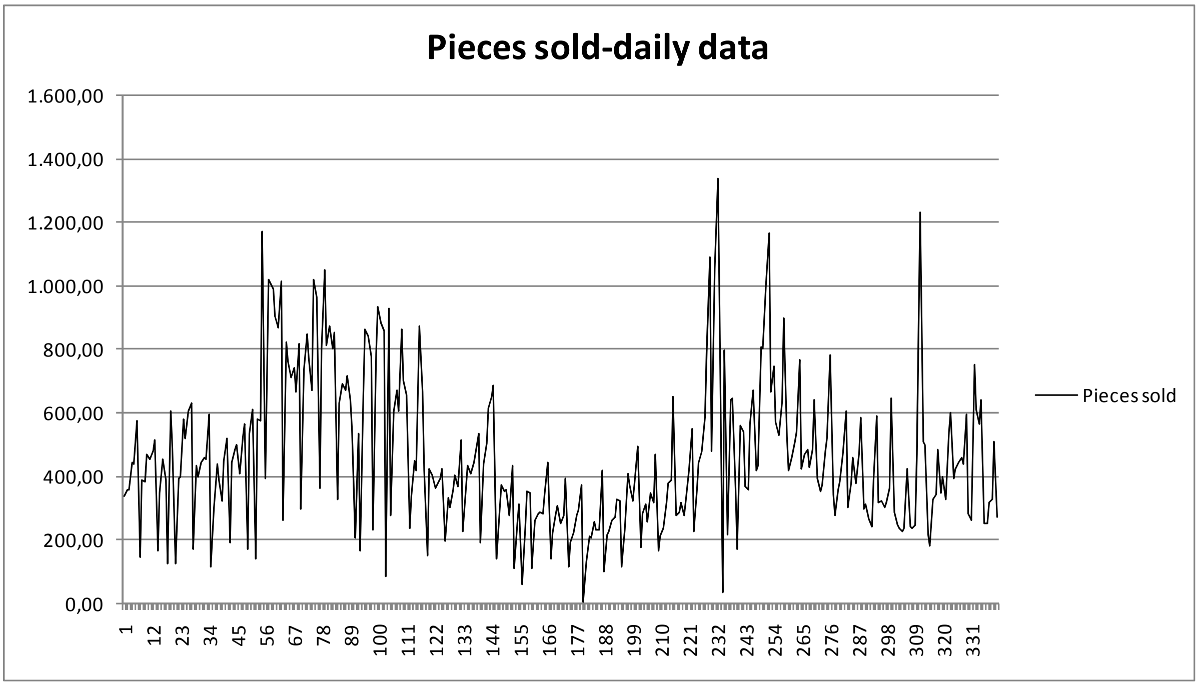
- Number of pieces sold
- Number of customers who entered the stores
- Number of pieces sold per customer

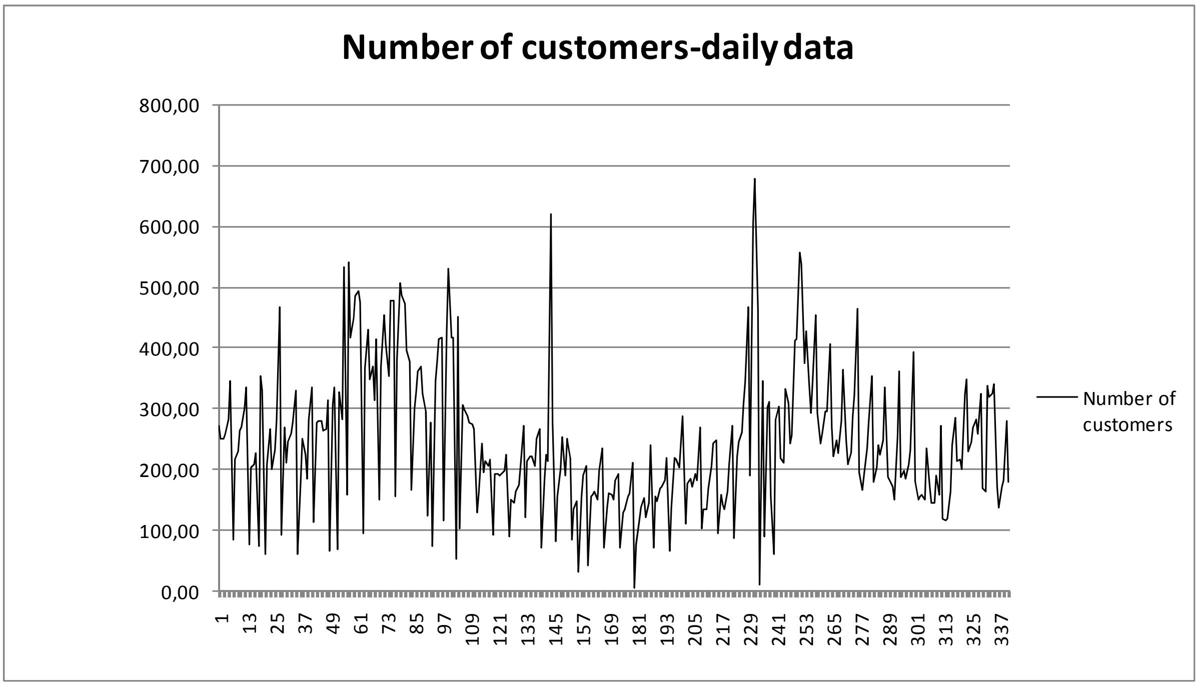
3 MONITORING FREQUENCY

Usual accounting cycle is one year. Different people have different buying habits for different products. As for the fashion industry, habits of people are very heterogeneous depending on richness, climate, culture, age, gender and so on. Some goods can be worn for many years and people keep stocks of it. Fashion industry traditionally has two seasons in one year: spring-summer and autumn-winter. They are connected to the climate, but also to some other factors such as production possibilities. It makes sense, from the point of view of company, to follow the figures seasonally. Six months is a long time during which a lot of good business decisions can be made and the bad ones avoided. The business in fact is done all the time. So what can we do? We can do it yearly, by season or even daily. Management should know daily weather the company is heading in the good direction. This reason seems logical, so we can conclude that we need the data on a daily basis. On the other hand, it seems that the logical cycle in the fashion industry is one season. So what we need are daily figures for the three variables. Since the cycle is 6 months, we can use moving average which clearly shows seasonal fluctuations. Moving average is a standard method used in time series analysis in statistics. Instead of looking at the average figure of all available data, we are looking at the local average figure. In this way it is possible to keep “a little bit” of fluctuations and simultaneously disregard too much emphasis on some unusual events. More on this can be found in [7] and [8]. For a 31 days moving average on a particular day we take 15 measurements before, 15 measurements after and the middle one on that day. We divide their sum by 31 and get the moving average, which in fact represents an estimate for the middle one. We do the same for each day except for the first and last 15 days.

4 DATA

It is usually possible to derive good reports on Pieces sold per day from the information systems. There can be some methodological issues because some pieces are not sold but given to customers as presents. There can also be a significant number of goods returned due to the existence of faulty products or the fact that the customer bought the wrong size, color or even model and wants to exchange it for something else. Sometimes goods can be lent and then returned. There are also problems due to stolen goods. Counting number of customers is subject to human error. Some customers, such as loyal customers who are recognized by company staff can be counted fairly well, but the majority of others cannot be easily counted. Some companies do not even count the customers, but instead count the number of invoices that their systems produce. This can be used as an estimate of number of customers. Some companies have technological support in terms of counters which count number of people who entered the stores. The reason for mentioning all this is that if we want to have daily data from the system, data collection process must be automated or otherwise the amount of manual work would be huge with the results lagging behind. So, once it is possible to extract the data needed from the system, we can move on. We extracted the daily data for 341 days from the information system, that is to say, for two seasonal cycles. It is not 365 due to eliminating 2 days for which the data were wrong (number of pieces per customer less than 1) and there were some nonworking days during the year. The figure shows the development of variables in time.





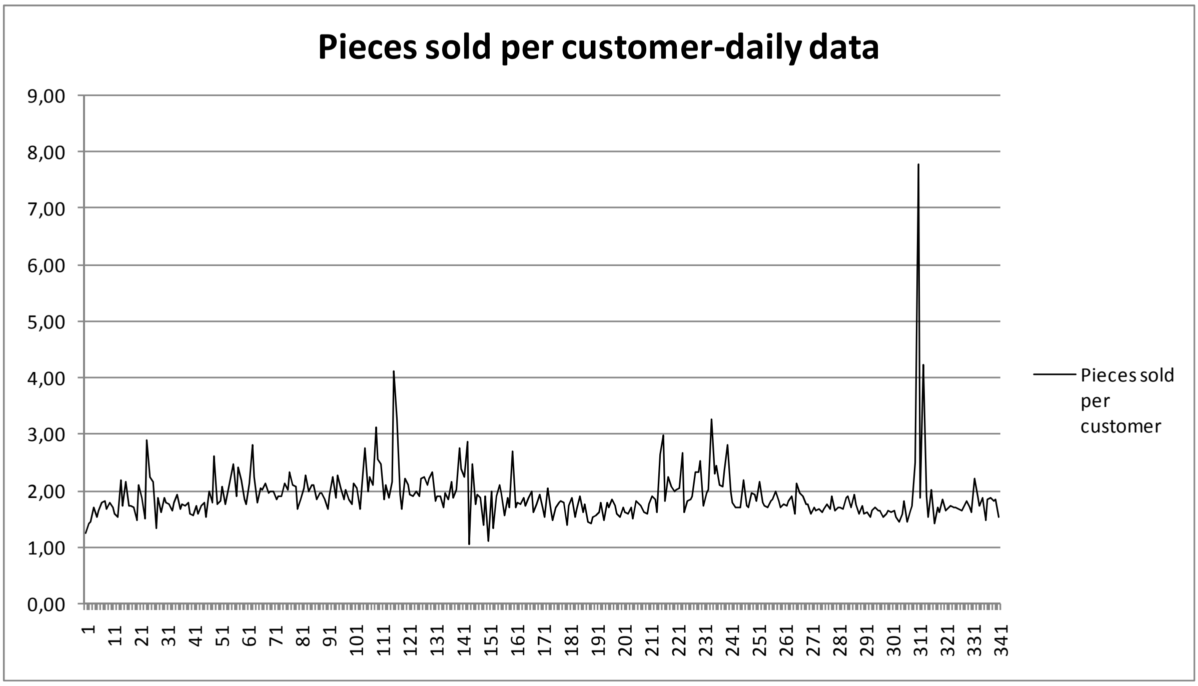


Figure 1: Development of 3 series in time

It can be seen that there exist seasonal variations in the first two variables. The third variable seems fairly stationary with no such seasonal cycles. So for the third variable we can conclude that no matter in which period of the year we are in, it is favorable to go above the year average. As far as the first two are concerned, it seems it is not the same in which part of the year we are in and therefore it is favorable to avoid daily and weekly oscillations and be above, for example, 31 days moving average.

5 ADDITIONAL SIMPLIFICATION

If we look at the correlation coefficients between the 3 variables in question, we can come to an additional simplification. If we look at the daily data for one year interval of the 3 variables mentioned, we get the following correlation coefficients.

Table 1: Correlation coefficients

	1	2	3
1	1	0,91321	0,29474
2	0,91321	1	-0,07833
3	0,29474	-0,07833	1

First and second variable are positively correlated. There is also positive correlation between Number of pieces sold and Pieces sold per customer, but it is not too strong. What can be concluded is that if we follow the first or the second variable it is rather the same, but we must follow the third one since it is not enough correlated with the first two. Since second one is less correlated with the third one than the first one with the third one, it seems reasonable to eliminate the first one. This leaves us with monitoring variables two and three. So from the initial four variables, we introduced the fifth one and we ended with two variables to follow:

- Number of customers and
- Number of pieces sold per customer.

The Figure shows the variable selection process, with the selected variables shown in bold.

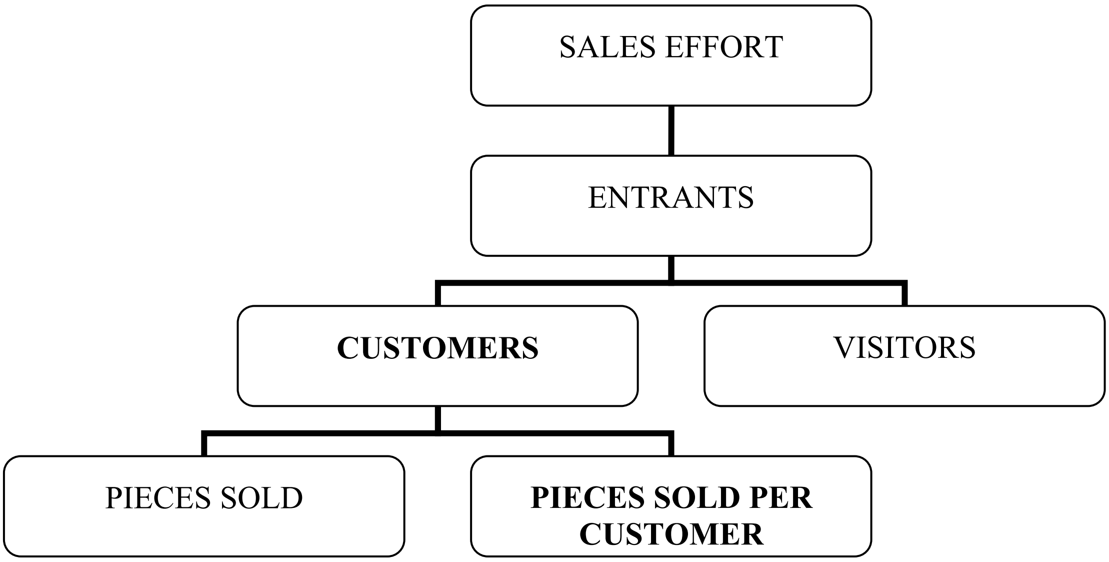


Figure 2: Selection of KPI-s

6 SETTING GOALS

The average value for Pieces sold per customer for a whole year is 1.89. From the previous discussion we can come to the conclusion that every day when this ratio is calculated it is favorable that it is above this threshold value. This would signify an improvement in the quality of service in stores.

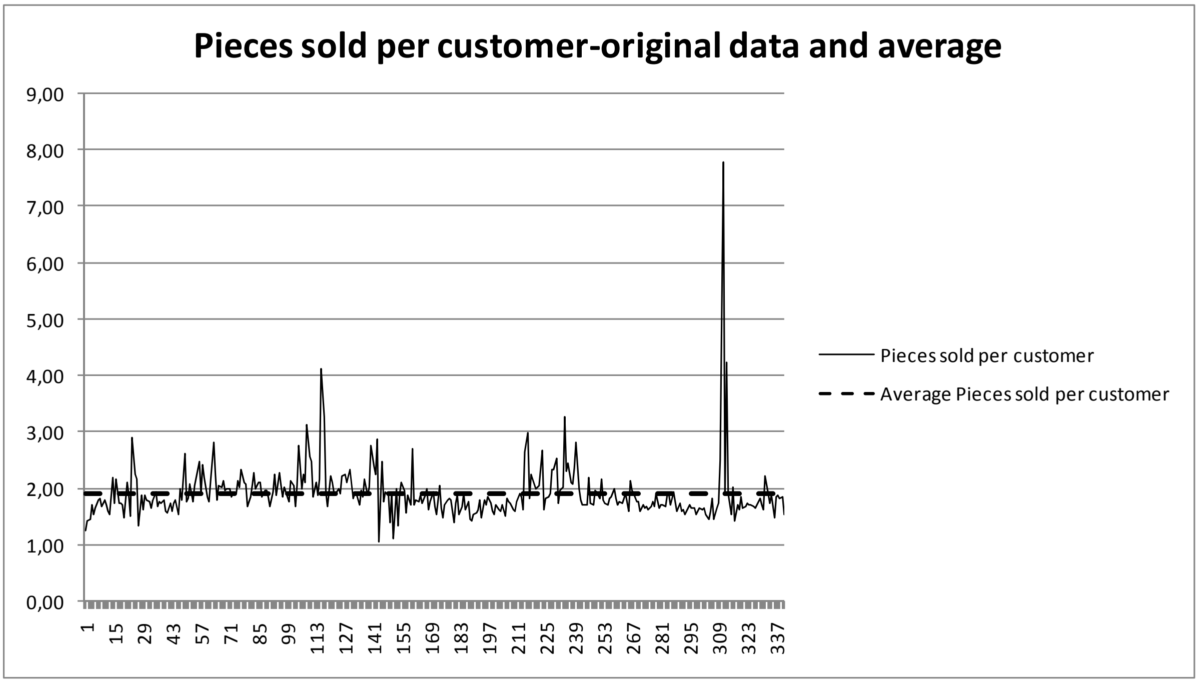


Figure 3: Pieces sold per customer and average

Regarding Number of customers things are not so simple since the seasonal influence is obvious. In average it should be above 243.25 customers daily.

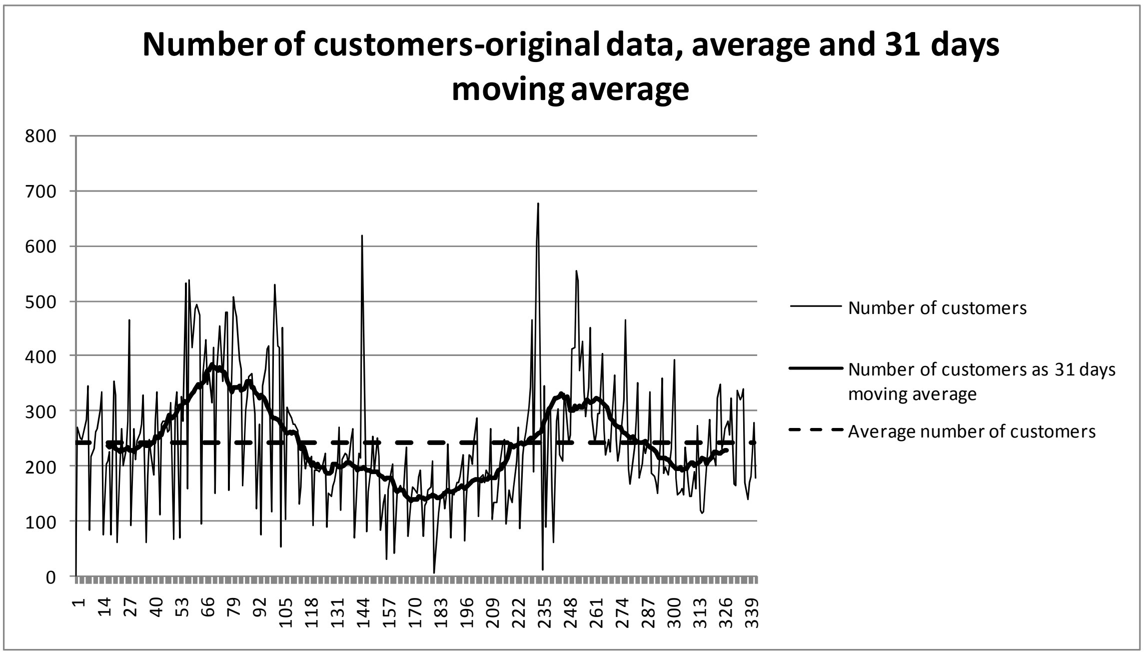


Figure 4: Number of customers and average

If we take average value as a threshold for daily performance, we could make a huge mistake. For example, if we achieve value of 300 around day 66, it could mean that our performance was very good, but in fact it was poor. In this case it is better to benchmark against the moving average figure corresponding to this period of the year. In the case of the day 66 above average performance would be achieved with the value of 400 customers. So, when examining the daily figures, managers can look at the graph and see, if they are above the seasonally adjusted threshold value [9]. It is advisable to analyze the historic figures of the KPIs during the budget process and set them up then. If there is some knowledge why KPI values could change, this can be built in the prediction model. Putting high value for the KPIs does not make sense if there is no action plan or some external forces which can bring that change.

Let us analyze what happens if the stores increase their quality of service for 10%. In that case customers would perceive it and increase the Number of pieces sold for 10%. If we are selling 100 pieces to 50 customers then we will sell additional 10 pieces to 50 customers. These 50 customers would by word of mouth bring 10% additional customers. So we would have 55 customers. These additional 5 customers would buy additional (110/50)X5=11 pieces. In total we have an increase of 21 pieces sold from 100 to 121 or 21%. What we see in this example is that the managers can monitor KPIs and try to increase their value. If they do that, the result can be very beneficial to the company.

7 CONCLUSION

Key Performance Indicator (KPI) can be used as a good tool in order to improve performance of a company.

It is a quantitative figure which shows how some aspect of business is functioning.

Prerequisite for implementation of KPIs in a company is a good information system.

Sometimes it is not enough to fix the KPI value, but instead it should be changeable depending on the time of the year.

In a retail company used as an example, it seems reasonable to monitor Pieces sold per customer and Number of customers as KPIs.

If on a particular day Pieces sold per customer are above 1.89, then the company is doing a good job.

Number of customers should be compared according to the graph shown in the text. In average it should be above 243.25 customers daily.

There are possibilities to integrate different KPIs by means of Balanced Scorecard methodology.

References

[1] Australian Government Information Management Office Department of Finance and Administration, 2006. Performance Indicator Resource Catalogue, http://www.finance.gov.au/budget/ICT-investment-framework/docs/AGIMO_PerfIndicatorReport_v1_2.pdf 2 p.

[2] Ravlić, S., 2002. Primjena koncepta Balanced Scorecard u Ericssonu Nikoli Tesli, http://kvaliteta .inet.hr/e-quality/prethodni/8/Ravlic_Sinisa.pdf 8 p.

[3] Kaplan, R. S., Norton, D. P., 2005. Using the Balanced Scorecard as a Strategic Management System, Harvard Business Review On Point, http://portal.sfusd.edu/data/strategicplan/Harvard%20Business%20Review%20article%20BSC.pdf 6 p.

[4] Nimax, G., Shuler, P., 2008. Management Methodologies in Practice at UVa, http://www.virginia.edu/processsimplification/Management%20Methodologies.pdf 5 p.

[5] Wikipedia, the free encyclopedia, 2009. Word of Mouth. http://en.wikipedia.org/wiki/Word_of_mouth.

[6] Kaplan, R. S., Norton, D. P., 2005. Using the Balanced Scorecard as a Strategic Management System, Harvard Business Review On Point, http://portal.sfusd.edu/data/strategicplan/Harvard%20Business%20Review%20article%20BSC.pdf 9 p.

[7] Hillier, F. S., Lieberman, G. J., 1995. Introduction to Operations Research. McGraw-Hill, Inc, Singapure, 6th ed., pp. 807-827.

[8] Šošić, I., Serdar, V., 1994. Uvod u statistiku. Školska knjiga-Zagreb, Zagreb, 8. izmijenjeno izdanje, pp. 193-211.

[9] Kaplan, R. S., Norton, D. P., 2005. Using the Balanced Scorecard as a Strategic Management System, Harvard Business Review On Point, http://portal.sfusd.edu/data/strategicplan/Harvard%20Business%20Review%20article%20BSC.pdf 7 p.

SCHEDULING THE PRODUCTION OF TURNED PARTS WITH GENETIC ALGORITHMS

Aleš Slak and **Jože Tavčar**
Iskra ISD-Strugarstvo d.o.o., Savska loka 4, 4000 Kranj
And Iska Mehanizmi, Lipnica 8, SI 4245 Kropa
email: ales.slak@iskra-isd.si, joze.tavcar@iskra-mehanizmi.si

Abstract: The article presents optimisation of distributing products across machines in the production of turned parts. Genetic algorithms (GA) were used as a method to find a good distribution of orders. On the basis of a model and GA, we carried out a simulation. It showed that appropriately selected genetic operators provide a satisfactory distribution of orders across machines, which in turn reduces the throughput time of products and their total costs.

Keywords: genetic algorithms, scheduling, production planning, production optimisation, production of turned parts, tool setting, batch production optimization, elitist selection

1 INTRODUCTION

Manufacturing enterprises are constantly in need of reengineering due to technical development and never-ending market requirements. In order to recognise critical points and provide improvements, we need a clear and easily understandable model of enterprise functioning [1]. We surveyed the situation according to the ARIS methodology and found hidden potential for better management of processes and cost reductions.

Because the optimisation model was just one fraction of the whole picture and it is too extensive to be presented in this article, the focus will be on the problem of optimising, for which the model was executed and its applicability in the manufacturing of turned products was proved.

1.1 Problem presentation

In the presented example, we optimised scheduling of orders by means of genetic algorithms. We took the order type of production with limited resources. Several products are manufactured at the same time but only one product can be manufactured in a single place, which makes for a scheduling problem. Re-scheduling is as important task as the basic scheduling. Once the basic scheduling has been finished, it is taken as a working framework that can be executed, subject to no changes during the execution. In case of urgent orders arriving after the basic scheduling has been completed, certain tasks are given higher priority. If orders are cancelled or in case of an equipment malfunction, it is necessary to re-schedule things. Due to the above-mentioned reasons, it again serves as a working framework. Scheduling is a complex tasks due to limited resources, time constraints, setting times and priorities.

The task was optimum distribution of 30 batches on 18 machines, which means 30! or $2*10^{32}$ possible combinations. However, the number of possible arrangements is greatly reduced due to restrictions. Production was followed for 2 weeks or 10 working days, which amounts to 240 hours (3 shifts). Completion of each batch requires four working operations. It is possible to complete the first operation of each order even before the previous batch has been finished. This operation is termed tool presetting. Other operations are: tool preparation, manufacturing process adoption and batch production.

For the purpose of the optimisation task, we measured machine setting times and batch production time for a specific product. These times were systematically arranged, which resulted in:

- tool presetting time,
- time of tool setting inside a machine
- total time of machine setting according to client's drawing
- manufacturing process adoption time (PPA) and
- manufacturing time

Restrictions that affect scheduling of orders first of all prevent the latter from being able to be executed on all machines. This is not possible due to manufacturing technology, sophistication of the product, processing costs and machine capacities (degrees of freedom).

Mathematical model:
Product's throughput time ($t_{p,i}$) [8] was used as a basis. In order to calculate the time, we should know the technological time of manufacturing one piece ($t_{t,i}$), unit (min/piece), volume of order (Q_i) and preparation/finishing times ($t_{pz,i}$). The manufacturing time of one piece ends when the last operation of a product within a schedule is completed.

$$t_{p,i} = t_{t,i} \times Q_i + t_{pz,i} \tag{1}$$

Because products come to machines one by one, the busiest machine is marked at the end of the schedule, and this is the throughput time. By means of the target function we shall try to minimise it up to the level where we try to find the most equal distribution of machines occupancy and to achieve minimum throughput time.

j-machine's throughput time is the sum of all product's throughput times that are performed on a single machine.

$$t_{p,j} = \sum_i t_{p,i} \tag{2}$$

The maximum time of all machines T_p.is marked as

$$T_p = \max t_{p,j} \tag{3}$$

In our example, the target function (f) tends to find the minimum time of the busiest machine in the population.

$$f = \min T_p \tag{4}$$

Together with time, costs are also often a subject of interest. Therefore, we chose total production costs as the target function. Order execution costs ($s_{s,i}$) include amortisation costs (s_a) that run all the time, irrespectively of machine occupancy, labour costs (s_d) and preparation/finishing times ($t_{pz,i}$). This applies to tool presetting time, tool setting, manufacturing process adoption and clearing away tools, after the end of the batch production. At the end there is also the effect of standard cost price ($s_{ssc,i}$), covering tool costs, material costs and operation costs (turning, drilling, milling,...) together with the quantity (Q_i).

$$s_{s,i} = t_{p,i} \times s_a + t_{pz,i} \times s_d + Q_i \times s_{ssc,i} \tag{5}$$

Manufacturing costs on the j-machine include manufacturing costs of all products, manufactured on this machine:

$$s_{s,j} = \sum_i s_{s,i} \tag{6}$$

Maximum costs on the machine (S_s) shall be marked as:

$$S_s = \max s_{s,j} \tag{7}$$

The population capability function (f) is the minimum value of maximum costs on the machine:

$$f = \min S_s \tag{8}$$

2 PRODUCTION SCHEDULING

Due to increased sophistication of industrial production, requirements for better efficiency, better adaptability, higher quality and lower costs, the production in general has changed dramatically in recent years. Due to the complexity of today's systems, deterministic mathematical methods have become insufficient to achieve optimum production. The last decades have brought forward a number of new methods, based on Darwinistic principles, for researching, planning and predicting complex systems. The genetic algorithm (GA) is also one of such methods. This is an evolutionary optimisation method, based on the principles of biologic evolution, such as genetic crossover and selection [6].

Research studies on the application of genetic algorithms have been carried out in different fields. The most famous one is the problem of a salesman who is looking for the shortest way through towns that he has to visit, without going to the same town twice. Janez Abram used this problem in his work, replacing towns with product orders. He proved that it is possible to schedule orders in the production of chairs by means of Heller-Logemann algorithm and genetic algorithms, which results in time savings [9].

It's vital to know the effect of population size, type of selection and genetic operators. Comparisons between different types of selection have showed that the elitist selection is the most suitable one for scheduling orders into the manufacturing process. With this type of selection, the linear crossover operator (LOX) and mutation by replacement yield the best results. The population size of 60 chromosomes is suitable for a scheduling case study [4]. The combination between evolutive calculation and schedule optimisation has proved effective and successful.

The application of genetic algorithms for production planning is very useful. Comparison between time consumption of manual planning and planning with the use of genetic algorithms is very much in favour of the latter. Kljajić and Breskvar proved it with a simulation model and use of GA for production planning. GA is of much assistance in creating a good quality schedule. As a bench mark of a good schedule, they used the common idle time; i.e. time when working facilities are not engaged. For their simulation, they used the order type of a manufacturing enterprise [3].

When scheduling involves urgent orders or their delivery time is expiring, they require immediate execution. In this case, scheduling requires decomposition. It means that urgent orders are scheduled apart from those that are not urgent. The schedule is divided into logical units, according to delivery dates. The number of possible schedules of orders decreases significantly, which also simplifies the simulation model [21]. The schedule plan is performed individually for each unit. Implementation of GA for production planning was shown on the example of four working days whereby a new schedule was executed each day (Breskvar,2004).

Optimising production planning, we are therefore looking for an optimum schedule of clients' orders at minimum costs and minimum throughput time. Besides, it has to take account of all restrictions that the production system imposes on the planner, such as production capacity, quantity of material in stock, delivery periods,… [2].

A good production plan means minimum total manufacturing time and minimum total idle time in the production as time also incurs production costs.

Planning is basically about predicting how a task is supposed to be done. However, it is necessary to impose time restriction on the predicted work, i.e. specify the beginning and end, divide individual working phases within this time interval, and specify resources, tools and materials that it requires. Under the condition, of course, that the work will be done at minimum costs.

Preparing and implementing plans is not the same for all types of production. In our company, we deal with temporary, repeated and serial production with many changes of tools due to small batches or short manufacturing times of large numbers of products. For this reason, plans should be made for short periods. They are called operational plans.

For operational or time planning it is necessary to be familiar with data on the production process, workplaces and work tasks. It is necessary to know processing times, preparation/finishing time and workplaces where an operation can take place. These data have been collected in tables.

In our case, planning and scheduling of products across machines into the production is carried out in two parts. First, the planner makes a weekly schedule for each machine, followed by foremen's decisions on the concrete distribution of work. It results in shiftings of working plan quality, depending on the experiences of planners and foremen.

The so-called drop-in orders that need urgent attention and delivery to the customer are another problem of the order-type of production.

A good working plan means good exploitation of working facilities, taking account of deadlines and production restrictions. Because the actual working plan should adjust to current needs and is subject to changes, it can yield poor results in terms of working facilities exploitation.

3 SCHEDULING WITH GENETIC ALGORITHMS

It has been mentioned at the beginning that this is the order-type of production and that 30 batches should be distributed onto machines. One product's batch is represented as a gene and chromosomes represent the distribution of products. A group of chromosomes makes up a population. A single chromosome or distribution within the population is also termed an individuum. A population in any time is referred to as a generation. Although a population can vary from a generation to generation, the size of the population and the structure of chromosomes remain unchanged.

The database contains information on products that are being considered for order scheduling. The genes list (chromosome) presents a possible plan in the production. From the data on all orders, we made a basic schedule. With GA, on the other hand, we are looking for a schedule that is good enough and comes close to the optimum. Using GA, we have to be aware that there might exist a better distribution than our current result, however, finding the optimum solution would take too much time. The GA principle is similar to natural evolution. Only the strongest survive, providing a healthier population of descendants through genes mutations or mutual crossover.

Genetic algorithms represent a random search technique, whereby little knowledge about a problem is used, however the problem should be optimised. The algorithms use operators, representing parameters. The first problem with the use of genetic algorithms therefore appears with the presentation of problem parameters or the so-called chromosome syntax.

Binary presentation is the most widespread application of the chromosome syntax. In this method, each order is given a binary number. The resulting chromosome is a cluster of binary numbers, assigned to orders.

A simpler approach to orders scheduling optimisation is to shape up the chromosomes in the form of permutations of the list of tasks that should be done. The programme is so built that it develops the allowed task schedules; i.e. by executing each task only once. Such chromosome syntax was used also on the example of orders scheduling in the production of chairs (Abram, 2005). We opted for permutation of orders list because our case also involves scheduling of orders into the production.

The quality of each possible schedule was assessed by the capability function. In our case it is the minimum time of the busiest machine in the population and the lowest total costs.

Through the selection process, chromosomes are chosen and they are changed into new schedules by means of genetic operators. The selected chromosomes of the preceding generation and new, altered chromosomes, designed by a genetic operator, make up the population of the next generation. Chromosomes of the next generation are again subject to evaluation, selection and genetic operators as well as another crossover.

The choice of suitable genetic operators for generating possible schedules is the experimental part that should be performed when the algorithm is created. The initial distribution of products into the production reflects the sequence of order arrivals that were executed during the observation period. With the genetic algorithm, we found the optimum schedule.

For the time-demanding capability function, the elitist selection is used. It has also been used on our model of scheduling of orders into the production. The basic principle of the elitist selection is that it keeps one or more best individuums throughout the generations. It helps improve the convergence of solution and is more suitable than other types of selection (random) from the viewpoints of both results and time complexity.

In the application, in order to create new populations, we used the mutation by replacement principle and one point crossover (1PX). For the mutation, a randomly selected gene within a chromosome was replaced by a random value within the limits. The probability of mutation is generally very low, between 0,5 and 5% while the crossover probability is very high, between 80 and 95%. In both methods, operators take care that none of the genes within a chromosome are repeated.

The size of the population should be adjusted to each single problem with regard to the speed of solution converging.

We should not forget that a larger population probably yields better results, however, GA functions more slowly. In our case, we performed the effect of the size of a population, arriving at the result that the ideal size is 60 chromosomes (Bernik, 2003).

Functioning of a genetic algorithm in seeking the best solutions can be presented in six steps (Bernik, 1999):

1. Initialisation of the chromosome population
2. Evaluation of each chromosome in the population
3. Creating a new chromosome by means of crossover and mutation operator
4. Removing population members in order to make way for new chromosomes
5. Evaluation of new chromosomes and their inclusion in the population, if they exceed the threshold.
6. If the difference is small enough, stop the algorithm and return the best chromosomes; otherwise, go to item 3.

4 PROGRAMME ENVIRONMENT

The developed application should enable better and automated scheduling of orders. For the purpose of the application, we used data such as preparation/finishing time, manufacturing time, quantity, standard cost price, labour costs, amortisation costs and possible number of machines per individual order.

The application runs in the MS Excel environment and control is carried out in Visual Basic. The data that we used were taken from tables in MS Excel. The speed of the calculation was not recognized as a problem for our optimisation problem. An average calculation time was between 40 and 450 seconds.

The main part of the software is universal and could be used also for other purposes (e.g. optimisation of cutting parameters on CNC machines). Part of the programme has been adjusted to our specifications. The adjustments refer to the choice of the target function and genetic operators. Within the programme, the calculation is followed by sorting by size and a little less than half of the chromosomes were removed. From the better ones, we created offsprings and included them into the population.

Because the MS Excel software is widely spread, the application could be used by all companies with the same type of production and facing similar production planning problems. It would only be necessary to adjust the simulation model and choose good GA parameters for a specific problem.

5 RESULTS

Because the genetic algorithm uses a randomly generated population all the presented results are chromosomes' mean values. The population capability was measured with the maximum time of machine occupancy in the production of turned parts.

The conclusion was that one point crossover (1PX) operator yields better results in the initial generations, compared to mutation by replacement. In later generations, mutation by replacement yields better results, which is a proof that mutation provides for diversity of the population. With optimum schedule in the shortest possible time being the target, we merged both operators. It would mean that both, crossover and mutation, would be used for searching new schedules.

Both operators yield better capability assessments results for populations of between 50 and 60 chromosomes. The size of the population in terms of its capability is in agreement of Kljajić's and Bernik's results, showing that the size of 60 chromosomes is suitable for optimising scheduling of orders into the production. Too small populations (20 chromosomes) yield unsatisfactory results.

Figure 1 shows the impact of the population size on convergence. The curves denote population sizes of 20, 40, 60 and 80 chromosomes. New schedules were searched by means of both operators, one point crossover, mutation by replacement and elitist selection. The diagram shows that increasing population sizes yield better results, however, excessive enlargement does not yield better results but only extends the evolution process. The bigger the population the more time is needed to arrive at a good solution. What follows is solutions with the use of the population size of 60 chromosomes.

The results showed that the genetic algorithm yields better results than both planned and actual schedules. The simulation also revealed which machine was engaged throughout the production monitoring time and which one was idle all the time. There are several reasons for the idleness. One of them is that there has been no planned order for a particular machine because machining on it is very expensive. Other reasons may include expensive tools, delivery time of the necessary tool, small quantities.

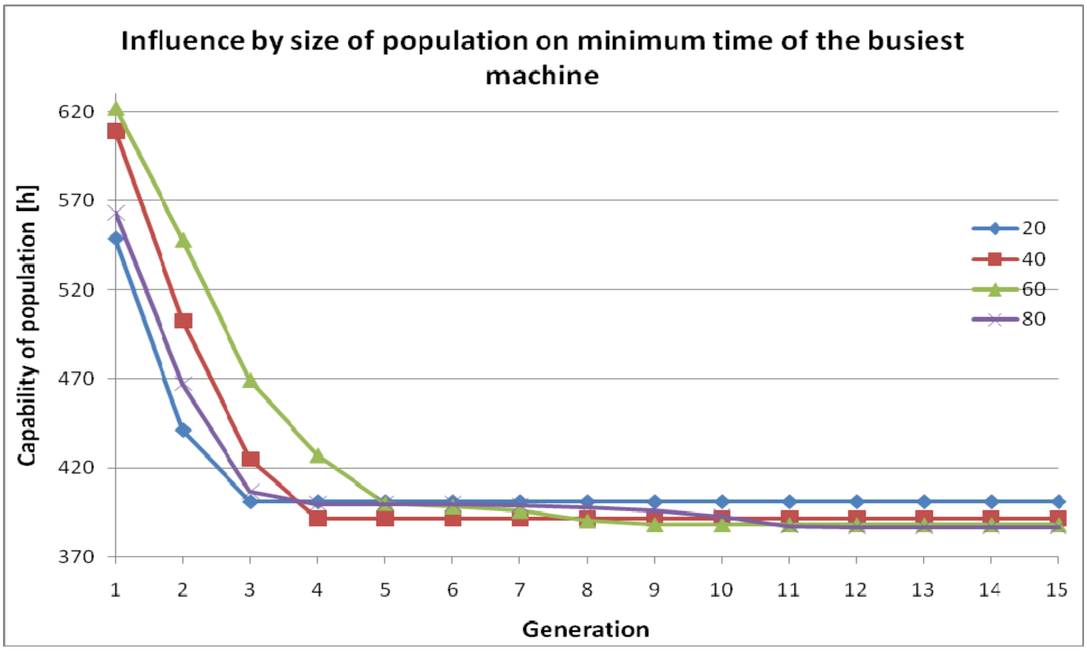


Figure 1: Influence of population on convergence; both operators used

By means of genetic algorithms and target function for measuring the minimum time of the busiest machine, the throughput time was reduced by 4.6%, compared to the planned one. The actual throughput time on the machine and the production time for all products was longer than planned. One of the reasons is without doubt the worker's consistency to make corrections in time and thus preventing production of defective pieces. Due to worker's inconsistency, the machine has to work longer than planned. Other reasons include tool replacement due to wear or breakage, shortage of spare parts in case of machine failure, wrong prediction (number of pieces per shift), problems with rod suppliers.

The times served as a basis to calculate the use share of a machine in relation to the production monitoring time. The production was monitored for 240 hours. The total time for 18 machines is therefore 4320 hours. Table 1 shows average use of all machines, taking account of their initial use. The result 0.891 represent the manually planed one. The last two lines show the results, obtained by genetic algorithms. The only difference between them is the target function. The result of 0.811 was obtained by the target function, used to find the minimum time of the busiest machine. A little better result was achieved with the target function, used to find the minimum value of the common costs of the plan.

Table 1: Average use of the 18 machines

	average
Planned:	0,891
Actual:	1,025
GA (time):	0,811
GA (cost):	0,808

With genetic algorithms, the average below 1 is achieved. It means that the production time of all products is shorter than the production monitoring time. In our case, value 1 represents 10 days or 4320 hours for all machines. Actual value is longer then production monitoring time. Table 1 shows that the execution of all orders takes 8.08 days or 3492.6 hours.

Figure 2 reveals the influence of the size of the population, with the function of maximum common costs of the plan used as the target function. It is interesting that we reach the same conclusion as in the first target function – the maximum time on the machine. Increasing the number of chromosomes in the population yields better results. The size of 80 chromosomes does not yield significantly better results that would make this population size optimum. Population sizes of 60 chromosomes yield fairly good results, so we can used it for next results presentations.

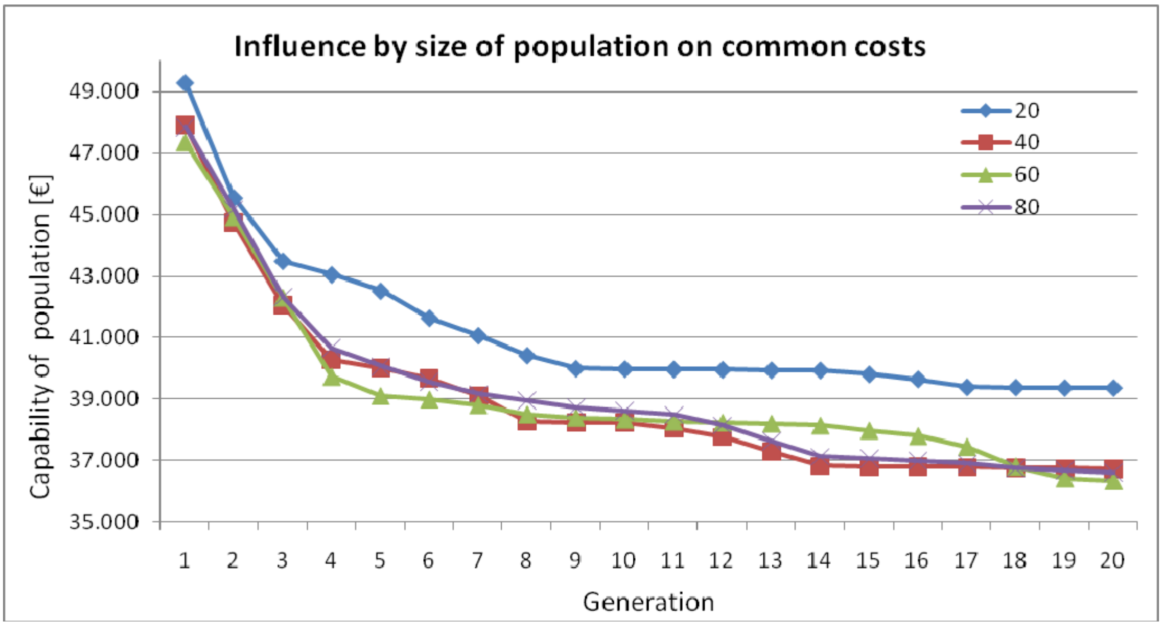


Figure 2: Influence of population on convergence; both operators used

Figure 3 reveals the comparison of the costs calculated by both target functions by 60 chromosomes. The first function is common costs of all products and the second one is the time of the busiest machine. With the first function an insignificantly better result was acchived; the exactly value of the common costs is 35.730,5 €. The result of the minimum time as a target function the busiest machine is 36.129,3 €. The use of genetic algorithms reduced common costs for 16.23 percent with the first target function and for 15.29 percent with the second.

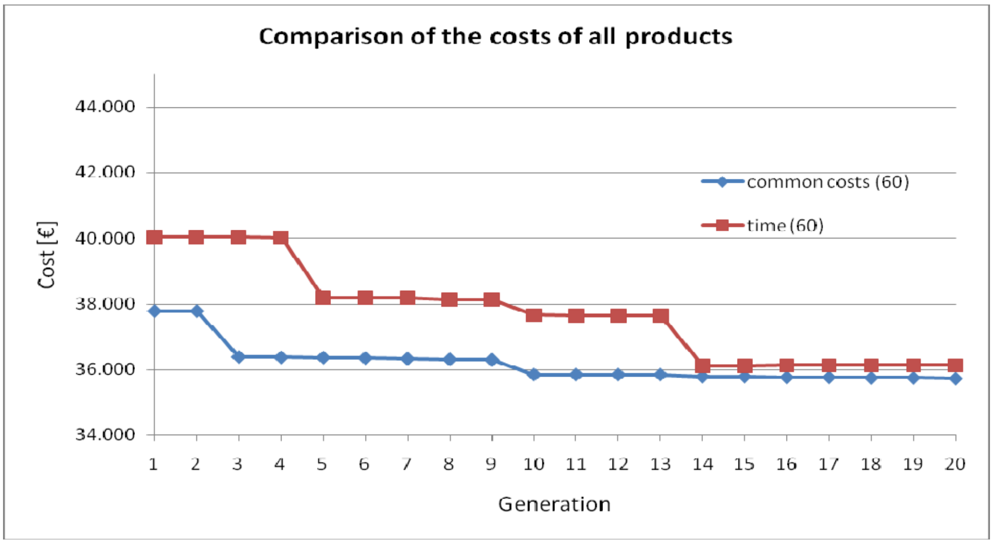


Figure 3: Comparison of the costs with population size of 60 chromosomes in relation to the target function

Throughput times of all products manufacturing times were also calculated and changes of the value at various target functions were surveyed. For the calculations, machines occupancy before the simulation was taken account of. The curves in Figure 4 illustrate the total product throughput time, caused by a schedule. The minimum values of each generation are shown.

Things have shown that the target function, used to find the minimum common costs of the plan, yields better results.

With the maximum time on the machine as the target function, we have got the minimum throughput time 3502.5 hours, with another target function we have got 3492.9 hours. Both throughput times are less the planning, which is 3848,58 hours. The reduction of throughput times is 9.00 and 9.24 percent.

Taking account of machines occupancy increases the throughput time. By how much it increases depends on correct distribution of products across machines. The achieved results show that the minimum throughput time on a machine is reached in the same generations as the minimum throughput time of all products. At this point, machines occupancy is the most evenly distributed. The minimum time of the busiest machine, no. 11 in our case, was 368.23 hours. The obtained result did not take account of the initial machines occupancy. Taking account of machines occupancy yielded the time of 381.9 hours on the same machine.

It means that taking account of machines occupancy increased the throughput time on the machine by only 3.7 percent. It shows that using genetic algorithms and the choice of parameters provided a good solution and good distribution of products across machines.

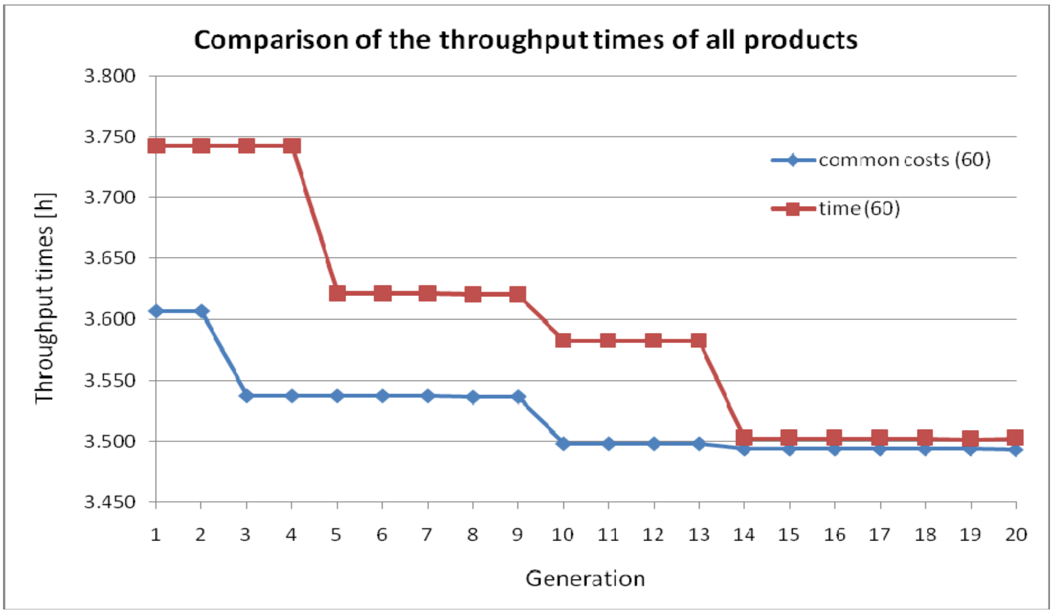


Figure 4: Throughput time in relation to the target function

6 CONCLUSION

Comparison between machine's planned times, actual ones and those achieved by GA has shown that orders can be scheduled much better by means of genetic algorithms than manually. The throughput time of all products was reduced by approximately 400 hours, compared to the manually planned one, which improved machines' productivity and efficiency. Reducing machine costs, total costs were reduced, too. Because of similarities between machines, restrictions and similar products, total costs were reduced by 16 percent. In the current form, the recalculation speed is not a bottleneck.

Both target functions yield good results. However, each producer has to decide whether to give priority to costs or timely delivery.

GA results are good and justify the integration into the planning process. It is necessary to assure the best possible data transfer from the information system to the optimising algorithm. In the next phase, the focus will be creating a user friendly interface and application in the production for every day work.

References

[1] Jože Duhovnik, Jože Tavčar- Elektronsko poslovanje in tehnični informacijski sistemi, LECAD,Univerza v Ljubljani, Fakulteta za strojništvo, Ljubljana 2000.

[2] Tone Ljubič: Planiranje in vodenje proizvodnje, modeli-metode-podatki, založba moderna, Kranj, 2000.

[3] Jože Balič: Inteligentni obdelovalni sistemi, Univerza v Mariboru, fakulteta za strojništvo, Maribor, 2004.

[4] Janez Abram: Optimizacija terminskega in kapacitetnega načrtovanja pri proizvodnji stolov, magistrska naloga, oddelek za lesarstvo, Biotehnična fakulteta, 2005

[5] Bernik I., M. Kljajić: Selekcija parametrov genetskih algoritmov pri optimizaciji proizvodnje, conference proceedings, 2003.

[6] Breskvar U.,M. Kljajić: Prednosti planiranja proizvodnje s simulacijo in genetskimi algoritmi, Organizacija , revija za management, informatiko in kadre, založba moderna organizacija, volume 34, no. 10, December 2001.

[7] U. Breskvar: Scheduling with Genetic Algorithms and Visual Event Simulation Model, 26th Int. Conf. Information Technology Intergfaces ITI 2004, Cavtat, Croatia.

[8] Bernik I., M. Kljajić: Petri mreže in dogodkovna simulacija pri optimizaciji proizvodnje z genetskimi algoritmi, Evropska skupnost in management: Zbornik posvetovanja z mednarodno udeležbo, prispevek na konferenci, 1999.

RESEARCH ON STOCK RETURNS IN CENTRAL AND SOUTH-EAST EUROPEAN TRANSITIONAL ECONOMIES

Jelena Vidović
University Centre for Professional Studies
Livanjska 5/III, Split, Croatia;
E-mail: jvidovic@oss.unist.hr

Zdravka Aljinović
University of Split, Faculty of Economics
Matice hrvatske 31
E-mail: zdravka.aljinovic@efst.hr

Abstract: In this paper normality tests are performed using the data on daily stock returns from 11 Central and South Eastern European countries. The article focuses on the skewness and kurtosis of stock returns providing information about extreme returns that can be expected on these stock markets. Seen from the EU integration perspective results indicate that there is no difference between stock returns of countries that already accessed EU and others. Stock markets regardless their belonging to the EU integration show different stages of stock market development.

Keywords: stock returns, skewness, kurtosis, normality tests, transitional stock markets

1 INTRODUCTION

Transition economies in Central and South Eastern Europe represent very attractive investment area for foreign investors. In the past few years these stock markets witnessed tremendous growth both in number of listed securities as well as in market capitalization (Figure 1). In 2007 stock market indices grew tremendously. Yet there are many problems that foreign investor must be concerned about when investing in these countries. The problem of insufficient corporative informing seems to be improving in the recent years while the problems of liquidity and risk steel remain undefined. These last two issues become quite important in the recent period when the global economy was struck by financial crisis. This paper investigates stock returns of Central and South East European Counties and questions the stock returns in the last years.

Paper is organized as follows: section 2 presents a brief review of previous researches. The research methods employed when examining stock returns in transitional economies are outlined in section 3. Section 4 resumes main results of conducted normality tests of stock returns in selected countries while the main conclusions are outlined in section 5.

2 PREVIOUS RESEARCHES

The origin of discussion about the risk of stock returns started in 1952 when Markowitz introduced his Portfolio Theory. In his theory he supposed that stock returns are normally distributed but also indicated that this assumption should be taken into consideration. In the years that followed many authors investigated stock returns, mainly for developed countries.

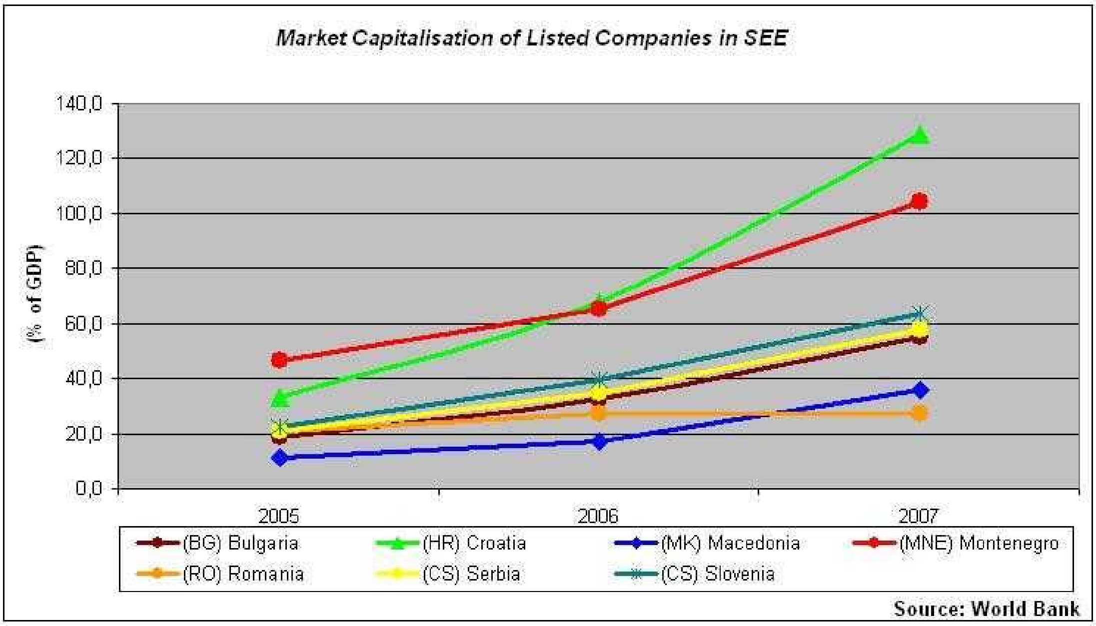


Figure 1: Market Capitalization of Listed Companies in South Eastern European Countries. Source: World Bank

In the recent years transitional stock markets witnessed tremendous growth. In 2007 value of stock market index on Zagreb Stock Exchange rose 71%, Slovene Stock market index rose 71,0%, while Deutsche stock market index rose 22,3% [10].

Conclusions of many studies reject the normal distribution for many emerging countries as well as for efficient markets in the world. The conditions on transitional stock markets are not consistent with the assumption of theoretical models. Bekaert et.al. concluded that emerging equity markets do not behave as developed markets and standard distributional models are rejected for observed stock returns in many countries [2]. Previous researches were focused on the many emerging markets worldwide but there is no paper which analyses stock returns of Central European countries as in this paper.

Similar research was conducted for Latin American emerging stock markets. Results showed that these markets have significantly fatter tails than industrial markets, especially the lower tail of the distribution. On the other hand most industrial markets pass the JB normality test [9].

There are no many researches on investment perspectives in Central and South Eastern Europe, especially there are no researches considering stock market returns as it was made in this paper. Most researches concentrate on the stock market indices and do not deal with stock returns. From this point of view there are only two recent researches that refer to the similar problem.

Middelton [8] et.al. studied facts that institutional investors take into account when investing in these countries. Authors of this paper noticed that there is no substantive literature on the CEE investment processes. In order to gain some information about the investment in these countries they interviewed professional investors and private equity fund managers. All the investors had chosen to invest in these markets attracted by the economic growth in the region. In the same time they pointed out that there are many problems when investing in these markets: the absence of reliable data from audited financial statements, small number of companies listed and problem of overpriced shares.

Chuhan´s[4] study discovered that investor's perception regarding the riskness of emerging markets act as a barrier for investors allocating funds to emerging economies. Investors are concerned about the illiquidity of securities traded on the emerging market stock exchanges as well as the small number equities listed on these exchanges. In many emerging markets market capitalization was dominated by a handful of companies which tended to increase market volatility. For example in Croatia in 2007 only eight stocks held 50% of total market capitalization and only 13 shares held 50% of the overall trading volume in the same year. Poland stock market, which is the most successful stock market in region, has a problem of excess demand for shares. Main problem is that Polish pension funds primary invest in Poland producing excess demand for shares.

Overall the different factors addressed in Middelton`s study were not seen as barriers for investing in CEE region, the interviews highlighted EU accession as the key contributor to the reduction in the risks which they faced. Such a finding is not surprising since all of the interviews had chosen to invest in these markets.

Bley[3] investigated dynamics and interactions of Euro stock markets. His sample consists of Austria, Belgium, Germany, Spain, Finland, France, Greece, Ireland, Italy, Netherlands and Portugal. Results imply that the highest risk measured in standard deviation is presented by Finland, followed by Greece and Germany. The stock return distribution appears to be leptokurtic for all markets. Over time, the number of euro markets that display various degrees of negative skewness has increased from four in the first period (1998-2000) to ten in third period (2004-2006). Investors are more exposed to the risk since the distribution of returns has a greater exposure to outlier events and bias to the downside.

According to the described situation on these markets we form the hypothesis that transitional stock markets in selected countries offer high returns which are associated with high level of risk. Under this assumption we expect that stock return in these countries do not follow the normal distribution and have fat tails indicating existence of extreme returns. Another hypothesis is based on the assumption that stock returns of countries that already assessed EU, which means a higher stage in economic growth and development, should show differences in behavior of extreme returns. At the moment EU member states are Poland, Czech Republic, Hungary, Bulgaria and Romania which accessed EU in 2007 while Croatia and Turkey have status of candidate countries.

3 DATA AND METHODOLOGY

In this paper stock returns of twelve countries from the region of Central and South East Europe are tested for normality. Selected countries and their corresponding stock indices are: Chez Republic (PX), Hungary (BUX), Poland (WIG20, WIG, SWIG80, COMP, BUDW, SPOZ, BNKI, MWIG40, TELE,), Slovak Republic (SAX), Slovenia (SE SBI I), Bosnia and Herzegovina (SASX 10), Romania (BETC, BETI), Serbia (BELEXLINE, BELEX 15), FYR Macedonia (MBI10), Croatia (CROBEX), Bulgaria (SOFIX) and Turkey (XU30, XU50, XU100). For every country 10 stocks were selected according to their presence in calculation of national stock indices. Selected stocks are first 10 stocks according to their weight in constitution of national stock indices.

Data in this paper consist of 250 daily closing prices of selected stocks as well as the daily stock index values obtained from the national stock exchanges in period from April 2008 until the end of March 2009. For each stock and for the each index, series of daily logarithmic returns were calculated. Daily stock returns and returns on national stock exchange indices are tested for normality.

For few stock markets was not possible to obtain 250 daily observations for 10 stocks. Slovakia stock exchange index SAX in the past period had only four constituents. Only one stock that participates in calculation of SAX index had more than 200 observations in selected period and as a result analysis of stock returns of 10 stocks from the national Stock Exchange index could not be performed for Slovakia. Similar, Bosnia and Herzegovina stock index SASX-10 had 9 constituents while only two of them had more than 200 observations of 250 possible in the selected time period while others had from 147 to 177 observations. Stocks from Macedonia stock index are also quite illiquid. Only six stocks which are constituents of MBI10 Index have between 180 and 243 daily observations while others have less than 120 daily observations in the same period. From the Macedonia stock exchange only nine stocks were taken into the analysis and for Bosnia and Herzegovina 7 stocks.

In order to determine whether the stock returns follow the normal distribution in this paper are presented both the descriptive statistics (mean, standard deviation, skewness and kurtosis) and tests for normality. Normality tests conducted in this paper are Shapiro Wilk (W test) and Kolmogorov – Smirnov (K-S) D test (Lilliefors test). In general all tests work well against skewed alternative distributions while strong differences appear when alternative distribution is symmetrical. Shapiro Wilk W test is usually recognized as the best overall test for normality against the skewed alternatives[6]. KS and W test compare empirical data with a theoretical distribution.

Shapiro-Wilk statistics was constructed by considering the regression of ordered sample values on corresponding expected normal order statistics which for sample of normally distributed population is linear. The statistics is positive and equal or less than one, being close to one indicates normality.

Kolmogorov – Smirnov test statistics is based on different distances between theoretical distribution function P given by the null hypothesis, and the Q distribution function of the sample. KS distance (d) is the largest absolute difference between cumulative theoretical distribution function and cumulative empirical distribution function. Under the 95% confidence level critical value of KS statistics is 0,2647. If the value of KS statistics is above the critical value the normality assumption can not be accepted.

Measures of kurtosis and skewness also show the deviation of a variable from a normal distribution. Skewness is third standardized moment that measures the degree of symmetry of probability distribution. If skewness is greater than zero the distribution has long right tail. Kurtosis is fourth central moment that measures thinness of tails or peakedness of probability distribution.

4 THE RESULTS

Figures 2 and 3 show the skewness and excess kurtosis of returns on stock market indices in observed period. There is considerable variation in the skewness of the returns on individual country indices. The excess kurtosis is almost always greater than 0 indicating fatter tails than the normal distribution. Macedonia, Slovakia, Serbia, Croatia, Chez Republic and Hungary report kurtosis 4 points bigger than normal distribution kurtosis. Only seven stock market indices have kurtosis smaller than two, six of those indices belong to Warsaw Stock Exchange and the last belongs to Istanbul Stock Exchange.

When analyzing skewness of stock market return distributions Slovakia, Serbia, Bosnia and Herzegovina, Croatia and Turkey stock indices have positive skewness indicating long right tail and possibility of extreme positive returns. Other stock market indices report negative skewness which means long left tail and extreme negative returns.

Slovakia, Serbia and Bosnia and Herzegovina stock exchange indices report severe kurtosis and large positive skewness. The possible reason for these results is that stocks which are constituents of these stock market indices are quite illiquid. Liquidity problems appear among stocks from Serbia, Slovakia, Bosnia and Herzegovina and Macedonia stock exchanges. On average 3 of 10 stocks from these markets which were included in the analysis had less than 150 daily observations of 250 possible.

Bulgaria and Macedonia stock exchange indices have large negative skewness indicating long left tail and extreme negative returns.

In the next step stock market indices were tested for normality. Results indicate that all stock market indices do not pass W normality test, while only two stock market indices pass the KS normality test, one of them is BUDW (Warsaw Stock Exchange Construction Index) and another is XU030 (Istanbul Stock Exchange National 30 Index).

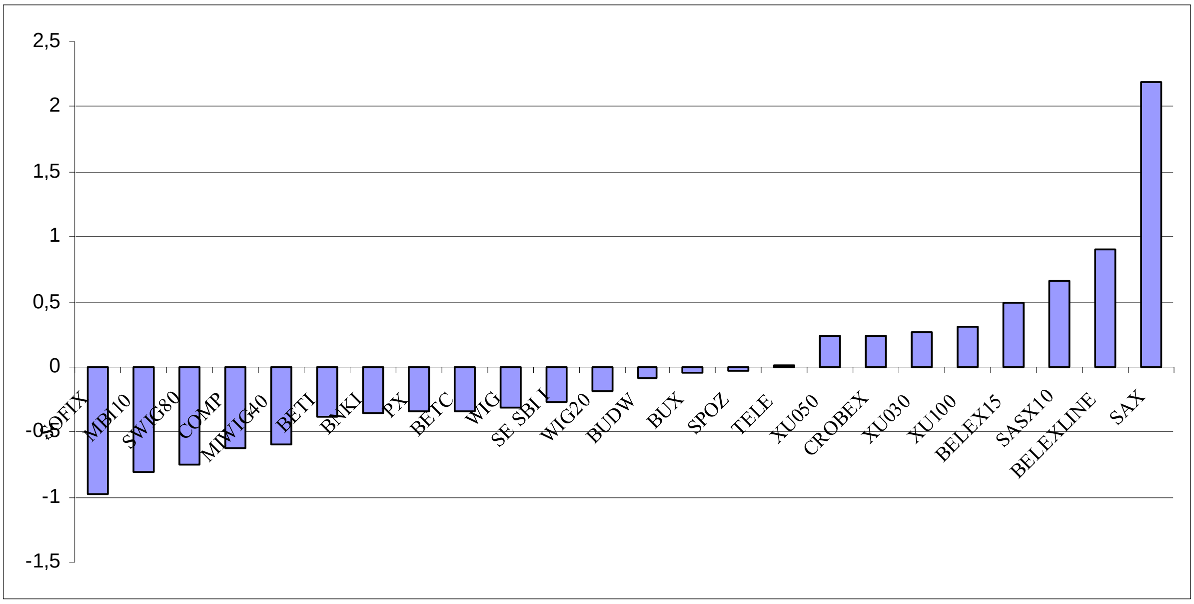


Figure 2: Skewness of stock market indices returns in period from April 2008 until the end of March 2009. Source: own calculation.

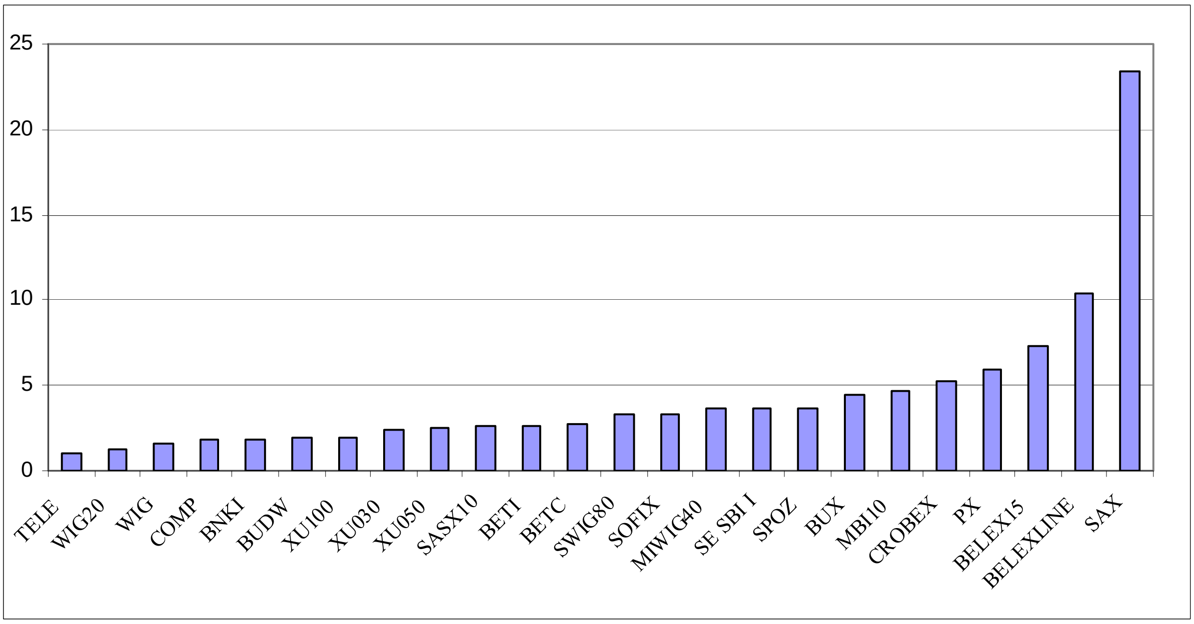


Figure 3: Kurtosis of stock market indices returns in period from April 2008 until the end of March 2009. Source: own calculation.

In the following tables and figures are presented results of descriptive statistics for stocks from Central and South East European stock markets. Mean, minimum, maximum, skewness and kurtosis were calculated for 10 stocks from every observed stock exchange.

All stocks in the past period had negative mean return which can be explained as an impact of the financial crises that struck the world economy. Bulgaria, Bosnia and Herzegovina and Macedonia have the highest negative mean returns for the daily rates of return, followed by Hungary, Croatia and Romania. Respectively, Turkey and Poland provide the lowest negative mean returns in observed one year time period.

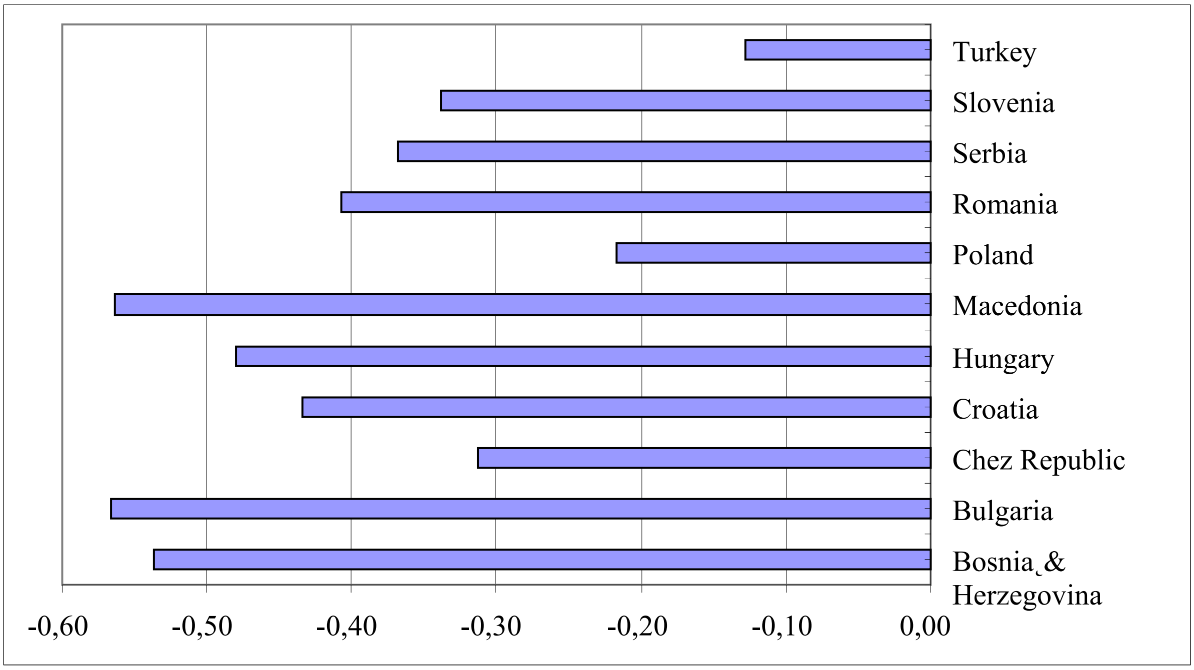


Figure 4: Average stock returns in period from April 2008 until the end of March 2009. Source: own calculation.

Stocks from Croatia, Hungary, Bulgaria, Czech Republic, Macedonia and Bosnia and Herzegovina stock exchanges exhibit the biggest kurtosis. Along with the excess kurtosis stocks on these markets have biggest extreme returns. Extreme negative returns in these countries vary from 16.20% in Hungary to 86.78% in Romania. In tables 3 and 4 are presented results on the number of extreme returns and their distance from expected returns in terms of standard deviations (z value). Main characteristic of extreme returns on these stock markets is that they appear far behind two standard deviations and only minority of extreme returns appears inside 2,5 standard deviations.

Table1: Kurtosis of stock returns in period from April 2008 until the end of March 2009. Source: own calculation.

	Kurtosis of stock returns				
Country	*0,2-1,2*	*1,2-2,2*	*2,2-3,2*	*3,2-4,2*	*>4,2*
Bosnia & Herzegovina	2	1			4
Bulgaria				3	7
Chez Republic		2	1	1	6
Croatia		1		1	8
Hungary			2		8
Macedonia		1	3		5
Poland	5	1	3	1	
Romania	1	1	4	1	3
Serbia	3	3	2		2
Slovenia		6	3	1	
Turkey	1	7	1		1
Total	12	23	19	8	44

Table2: Skewness of stock returns in period from April 2008 until the end of March 2009.

	Skewness of stock returns				
Country	*(-11)-(-10)*	*(-2)-(-1)*	*(-1)-0*	*0-1*	*1-2*
Bosnia & Herzegovina			3	3	1
Bulgaria		3	7		
Chez Republic			6	3	1
Croatia			3	5	2
Hungary			6	3	1
Macedonia		3	3	2	1
Poland			6	4	
Romania	1		5	4	
Serbia			4	6	
Slovenia			4	6	
Turkey			4	6	
Total	1	6	51	42	6

Source: own calculation.

Extreme positive returns vary from 22.31% in Hungary to 45.44% in Bosnia and Herzegovina. When examining skewness, stocks do not follow any specific pattern. All stocks from Bulgaria stock exchange exhibit negative skewness and have extreme negative returns which are in absolute value always bigger than extreme positive returns. Overall, 48 stocks exhibit positive skewness while 58 stocks had negative skewness.

Table 3: Number of stocks according to extreme negative returns shown in terms of standard deviation in period from April 2008 until the end of March 2009.

Country	z value						
	(-13.5)-(-12.5)	*(-7.5)-(-6.5)*	*(-6.5)-(-5.5)*	*(-5.5)-(-4.5)*	*(-4.5)-(-3.5)*	*(-3.5)-(-2.5)*	*>(-2.5)*
Bosnia & Herzegovina			1	1	2	3	
Bulgary		1	2	4	3		
Chez Republic			1	4	5		
Croatia					8	2	
Hungary				2	6	2	
Macedonia			1	2	2	4	
Poland				3	2	5	
Romania	1				6	2	1
Serbia					2	4	4
Slovenia					4	6	
Turkey				1	3	6	
Total	1	1	5	17	43	34	5

Source: own calculation.

Stocks which belong to Poland and Slovenia stock exchanges have the smallest kurtosis. Extreme negative returns on these two stock exchanges vary from 23.62 % to 12.74%, while extreme positive returns vary from 15.35% to 17.69%. Stocks on these two stock exchanges exhibit lowest standard deviations followed by Hungary, Turkey and Macedonia.

Table 4: Number of stocks according to extreme positive returns shown in terms of standard deviation in period from April 2008 until the end of March 2009.

Country	z value					
	1.5-2.5	*2.5-3.5*	*3.5-4.5*	*4.5-5.5*	*5.5-6.5*	*6.5-7.5*
Bosnia & Herzegovina	1	3	2		1	
Bulgary		4	5	1		
Chez Republic		4		5	1	
Croatia			2	4	2	2
Hungary		1	4	3	2	
Macedonia	1	4	2		1	1
Poland		6	4			
Romania	2	5	3			
Serbia	1	6	2	1		
Slovenia		1	8	1		
Turkey		2	6	2		
Total	5	36	38	17	7	3

Source: own calculation.

By far the highest level of absolute risk measured in standard deviation is in Bosnia and Herzegovina followed by Romania, Serbia and Bulgaria.

Table 5: Number of stocks according to their standard deviation

Country	Std. Deviation of stock returns						
	1-2	*2-3*	*3-4*	*4-5*	*5-6*	*6-7*	*7-8*
Bosnia & Herzegovina			2	3	1		1
Bulgary		1	2	4	3		
Chez Republic	1	2	1	3	1	1	1
Croatia	1	1	1	5	2		
Hungary		4	4	2			
Macedonia		3	6				
Poland		4	3	3			
Romania			3	5		2	
Serbia			2	3	5		
Slovenia		4	6				
Turkey		1	8	1			
Grand Total	2	20	38	29	12	3	2

Source: own calculation.

All stock returns were tested for normality. The results of normality test indicate that stock returns are far away from normal distribution. Shapiro-Wilk test which is the most powerful test against skewed alternatives rejects normality assumption for almost all observed stocks under 1% significance level[6]. Only two stocks from Poland stock exchange pass the W test. Four stocks from Warsaw stock exchange and one stock from Istanbul stock exchange pass the KS normality test.

In addition normality test were also conducted for 250 weekly returns and the results are similar. Only four stocks from Warsaw stock exchange, three stocks from Istanbul stock exchange and one stock from Prague stock exchange pass the W test. KS does not reject normality assumption for 6 stocks from Istanbul and Warsaw stock exchange and one stock from Prague and Budapest stock Exchange.

5 CONCLUSION

These transitional equity markets do not behave like developed markets. This paper specifies two main characteristics of these stock markets, the level of risk which could be met on these stock markets and liquidity of shares. All these stock markets share the same common feature; stock price returns are not normally distributed.

We find that central and south-eastern European equity markets have fat tails. From all stock exchanges Bulgaria stock exchange shows the worst performances, all stocks have negative mean returns, negative skewness and kurtosis that significantly exceeds normal distribution. More than half stocks from Serbia, Bosnia and Herzegovina, Macedonia and Slovakia stock exchange had less than 150 daily observations of 250 in selected time period. Stocks on these four stock markets along with stocks from Macedonia, Hungary and Romania have kurtosis much bigger then normal distribution indicating many extreme observations.

Only stocks from Slovenia, Poland, and Turkey stock markets share similar characteristics. They have kurtosis around 2, standard deviation around 3 and negative mean return from -0,12 in Turkey to -0,34 in Slovenia.

In 2004 Czech Republic, Hungary, Poland and Slovenia assessed European Union, followed by Bulgaria and Romania in 2007 while Croatia and Turkey have status of candidate countries.

As we can see from these results EU accession can not be driven in relation with characteristic of Central and South European stock markets while they share different characteristics. The EU accession key can not be observed as indicator reducing risk on transitional stock markets.

References

[1] Aljinović Z., Marasović B., Pivac S., 2007: Transition capital markets comparison by efficient frontiers, in: Seventh International Conference on Enterprise in Transition, Proceedings, Faculty of Economics Split, Split-Bol, May 24-26: 227-229 (extended abstract, full-text on CD-ROM)

[2] Bekaert G., Campbell R. H., 2002: Research in emerging markets finance: looking to the future, Emerging Markets Review, Vol. 3, pp. 429-448.

[3] Bley, J., 2009: European Stock Market Integration, Fact or Fiction?, Journal of International Financial Markets, Institutions & Money, Accepted manuscript, doi:10.1016/j.intfin.2009.02.002

[4] Chuhan, P., 1994: Are institutional investors an important source of portfolio investment in emerging markets?, Policy Research Paper No 1243, The World Bank, International Economics Department, Debt and International Finance Division, January, 1-35.

[5] Chunhachinda P., Dandapani K., Hamid S., Parakash A. J., 1997: Portfolio selection and skewness: Evidence from international stock markets, Journal of Banking&Finance, Vol. 2, pp. 143-167.

[6] Coin D., 2008: A goodness-of-fit test for normality based on polynomial regression, Computational statistics & data analysis, Vol. 52, pp. 2185-2198

[7] Fama, E., 1965: The behavior of stock market prices, Journal of Business, Vol. 38, pp. 34-105.

[8] Middleton C. A. J., Fifield S. G. M., Power D.M., 2007: Investment in Central Eastern European equities – An investigation of the practices and viewpoints of practitioners, Studies in Economics and Finance, 1: 13-30.

[9] Susmel R., 2001: Extreme observations and diversification in Latin America emerging equity markets, Journal of International Money and Finance, 20: 971-986.

[10] Stock market indicators http://www.world-exchanges.org/statistics/annual/2007/indicators, http://www.zse.hr/default.aspx?id=179

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section IV:

Scheduling and Control

THE PERTURBED ADJOINT OPERATOR DEVELOPED FOR FILTERING THE SUPPLY SYSTEMS WITH STOCHASTIC DELAY IN CONTROL

Ludvik Bogataj and **Marija Bogataj**
University of Ljubljana, EF_KMOR, Kardeljeva ploscad 17, 1000 Ljubljana, SI
Mediterranean Institute for Advanced Studies, Sempeter pri Gorici, SI
Marija.Bogataj@guest.arnes.si; Ludvik.Bogataj@ef.uni-lj.si

Abstract: The perturbed adjoint operator for linear hereditary differential systems considering finite magnitude perturbations of the delay in the control, is characterized. The system is described by linear differential- delay equation with delays in the state of the system and in the control, which are the case often encountered in real world problems. Such characterization of the perturbed adjoint operator is important for sensitivity studies of linear – quadratic problems with control that is near to optimal, by filtering with delay in observation, especially important in evaluations of perturbations which appear in global supply chains. MRP Theory approach is needed for presentation of such perturbed global supply chains. Using obtained straightforward robust perturbation results changes of criterion function could be predicted.

Keywords: Hereditary systems, delay, perturbations, finite magnitude, adjoint operator, supply chain, MRP Theory.

1 INTRODUCTION

In the real world problems, like in supply chain where lead times are substantial, we encounter on systems that can be modeled by differential-delay equations with time delays in the state of the system and in the control:

$$\dot{\mathbf{x}}(t) = \mathbf{A}_0\mathbf{x}(t) + \mathbf{A}_1\mathbf{x}(t-a) + \mathbf{B}_0\mathbf{u}(t) + \mathbf{B}_1\mathbf{u}(t-b), \qquad t > 0 \tag{1}$$

In such a system we often do not know the value of time delays with certainty. Therefore we consider them as a kind of distributed parameters, for which only the maximal degree of variability can be predicted. So in this case uncertain distributed parameters can be included in the mathematical model as finite magnitude perturbations described in Bogataj (1989, 1989 a, 1990) and Bogataj, Pritchard (1979). The applications of this models to inventory control are presented in Bogataj and Bogataj (1991, 1994), Bogataj and Čibej (1994) and Čibej and Bogataj (1994) and specially elaborated for MRP Theory, which was previously developed by Grubbström (1996, 1998, 1980, 2007) and in many other papers also together with some researchers of his Linköping Institute of Productions Economics, in Bogataj, Bogataj, and Vodopivec (2005) and Bogataj and Bogataj (2007). In mentioned models some sensitivity results are given, describing influence of perturbations of the system matrices and of certain delay in the control (or both simultaneously) on the quadratic objective function. L Bogataj (1989) has also characterized the adjoint operator of the infinitesimal generator of a strongly continuous semigroup for linear differential – difference equation with delays in the state and in the control, generalizing the results of Vinter (1978) to the case including delays also in the control. Importance of mentioned approach for comparative static in inventory management , global sensitivity analysis in inventory control and especially for evaluation of risk in a supply chain is described in Borgonovo (2008) and especially in Borgonovo, Peccati (2007, 2009). The characterization of the adjoint operator is important in the case of solving the dual problem of a quadratic cost optimal control problem, the so called filtering with delay in observation. The important question that naturally arises is: Can we split the adjoint operator in the case of finite magnitude perturbations to the original part, corresponding to the unperturbed model plus the part, which is a consequence of perturbations? In the case of linear hereditary systems including delay in the state of the system and in the control, the answer is affirmative and is given in this paper. To the author's knowledge, this kind of problem has been unsolved properly up to now. This approach could substantially improve the known sensitivity results for linear quadratic optimal control problems, especially those which appear in MRP Theory when parallel study of behavior of supply system is consider in time and frequency domain, like presented in the paper of Grubbström, Bogataj and Bogataj (2009).

The usefulness of the results can be also find when study Optimal tracking control (OTC) for time-delay systems affected by persistent disturbances with quadratic performance index. By introducing a sensitivity parameter, the original OTC problem could be transformed into a series of two-point boundary value (TPBV) problems without time-advance or time-delay terms. Recently obtained OTC law by Tang, Sun, and Pang (2008) consists of analytic feedforward and feedback terms and a compensation term which is the sum of an infinite series of adjoint vectors.

2 THE NEW APPROACH TO MRP THEORY

The line of research, now designated *MRP Theory*, has attempted at developing a theoretical background for multi-level production-inventory systems, Material Requirements Planning (MRP) in a wide sense. Grubbström developed MRP Theory on the basic methodologies of "Input-Output Analysis", (Leontief, 1928) and *Laplace transform*. *Laplace transform* is a mathematical methodology dating back to the latter part of the 18th century and used for solving differential equations, for studying stability properties of dynamic systems, especially useful for evaluating the Net Present Value (NPV).

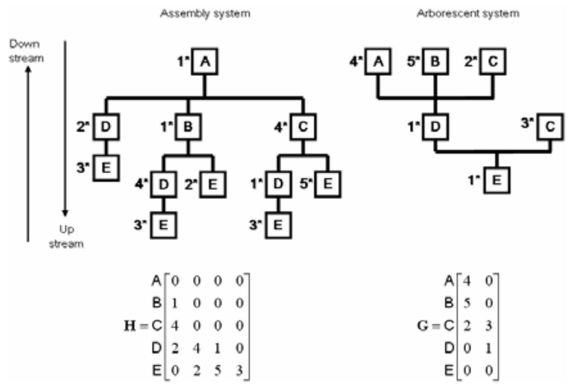


Figure 1: Examples of a pure assembly system and a pure arborescent system, in the form of product structures and their input and output matrices **H** and **G**, respectively (Grubbström, Bogataj, Bogataj 2007).

Basic in MRP theory are the rectangular input and output matrices **H** and **G**, respectively, having the same dimension. Different rows correspond to different items (products) appearing in the system and different columns to different activities (processes). We let m denote the number of processes (columns) and n the number of item types (rows). If the jth process is run on activity level P_j, the volume of required inputs of item i is $h_{ij}P_j$ and the volume of produced (transformed) 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 of realized intensity of flows through the activity cells in a supply chain. In case when the plan of this intensity $\mathbf{P}_0$ is not equal to $\mathbf{P}$, we have to write the total of inputs by $\mathbf{HP}_0$, from which the net production is determined by $\mathbf{GP}$-$\mathbf{HP}_0$. Here $\mathbf{P}_0 = \mathbf{P}_0(t)$ is assumed to be known vector which could change in time. In MRP systems, lead times are essential ingredients. The lead time of a process is the time in advance of completion that the requirements are requested. If $P_j(t)$ is the volume (or rate) of item j planned to be completed at time t, then $h_{ij}P_j(t)$ of item i needs to be available for production (assembly) the lead time τ_j in advance of 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 $(t-\tau_j)$ is reserved for the specific production $P_j(t)$ and then moved into *work-in-process* (*allocated component stock*, *allocations*). At time *t*, when this production is completed, the identity of the items type *i* disappear, and instead the newly produced items $g_{kj}P_j(t)$ appear. This approach has been developed for production systems, when transportation time did not influence lead time substantially. In case of transportation and production lead time τ_j should be split on two parts: production part of lead time τ_j^{pr} and transportation part τ_{ij}^{tr}. Therefore $h_{ij}P_j(t)$ of item *i* need to be available for production (assembly) the lead time $\tau_{ij} = \tau_j^{pr} + \tau_{ij}^{tr}$ in advance of *t*, i.e. at time $(t-\tau_j^{pr}-\tau_{ij}^{tr})$.

In order to incorporate the lead times for assembly and arborescent processes in MRP systems without transportation time lags, Grubbström (1967, 1980,1996,1998) suggested to transform the relevant time functions into Laplace transforms in the frequency domain.

It was clearly presented in Bogataj, Grubbström, and Bogataj (2008) that transportation lead time and transportation costs can be simply embedded in input – output matrices as a special matrices which multiplied previous two in frequency domain so that previous matrices including delays of production, written in frequency domain like $\tilde{\mathbf{G}}(s)$ and $\tilde{\mathbf{H}}(s)$ have been transformed in $\tilde{\mathbf{G}}^{pt}(s)$ and $\tilde{\mathbf{H}}^{pt}(s)$.

In such a system inventories in supply chain flows were described, using frequency domain as

$$\tilde{\mathbf{R}}(s) = \frac{\mathbf{R}(0) + \tilde{\mathbf{A}}(s,\tau)\tilde{\mathbf{R}}(s) + \tilde{\mathbf{G}}^{pt}(s)\tilde{\mathbf{P}}(s) - \tilde{\mathbf{H}}^{pt}(s)\tilde{\mathbf{P}}_0(s) - \tilde{\mathbf{F}}(s)}{s}, \text{ under condition } £^{-1}\left\{\tilde{\mathbf{R}}(s)\right\} \geq \mathbf{0} \tag{2}$$

where all kind of time delays have been hidden in $\tilde{\mathbf{G}}^{pt}(s)$ and $\tilde{\mathbf{H}}^{pt}(s)$ and $\mathbf{A}(s,\tau)\mathbf{R}(s)$ describe the perishability or obsolescence of the items in a supply chain.

Here we shall stay in time domain where lead times are expressed in control vector of production. Therefore the expression (2) gets the form:

$$\dot{\mathbf{R}}(t) = \mathbf{A}_0\mathbf{R}(t) + \mathbf{A}_1\mathbf{R}(t-a) + \mathbf{GP}(t,\tau) - \mathbf{HP}_0(t,\tau) - \mathbf{F}(t) \tag{3}$$

We assume that Net Present Value (NPV) has been optimized for unperturbed system in advanced and optimal safety stock function $\mathbf{R}_0(t)$ and optimal production plan $\mathbf{P}_0(t,\tau)$ have been found, using MRP Theory results. We wish to control the perturbed system with perturbed production function for perturbation: $\mathbf{P}(t,\tau-k) - \mathbf{P}_0(t,\tau)$ which influence differences in planned inventory level: $\mathbf{R}(t) - \mathbf{R}_0(t)$.

3 TIME DELAY IN THE STATE OF THE SYSTEM AND IN CONTROL

We now consider the following linear differential- delay equation with time delay in the state of the system and in the control:

$\dot{\mathbf{R}}(t) = \mathbf{A}_0\mathbf{R}(t) + \mathbf{A}_1\mathbf{R}(t-a) + \mathbf{B}_0\mathbf{P}(t) + \mathbf{B}_1\mathbf{P}(t-b),\ 0 < t < t_1$

With initial data

$$\mathbf{R}(0) = \boldsymbol{\varphi}^0, \quad \mathbf{R}(\tau) = \boldsymbol{\varphi}^1(\tau) \quad -a \le \tau < 0$$

$$\mathbf{P}(\tau) = \boldsymbol{\varphi}^2(\tau) \quad -b \le \tau \le 0; \quad \mathbf{P}(t) = \begin{Bmatrix} \boldsymbol{\varphi}^2(t) \in L_2\left(\left[-b, t_1\right], R^m\right) \\ 0 \Leftarrow t > t_1 \end{Bmatrix} \tag{4}$$

In (4) $\mathbf{A}_0, \mathbf{A}_1$ are nxn matrices, $\mathbf{B}_0, \mathbf{B}$ are nxm matrices, $\mathbf{R} \in R^n$, $\mathbf{P} \in L_2\left(\left[-b, t_1\right], R^m\right)$

and

$$\begin{aligned} \boldsymbol{\varphi} = \left(\boldsymbol{\varphi}^0, \boldsymbol{\varphi}^1, \boldsymbol{\varphi}^2\right) \in M^2\left(\left[-a, 0\right], R^n\right) \mathrm{x} L_2\left(\left[-b, t_1\right], R^m\right) \equiv \\ \equiv R^n \mathrm{x} L_2\left(\left[-a, 0\right], R^n\right) \mathrm{x} L_2\left(\left[-b, t_1\right], R^m\right); \quad a, b \in R, \quad a \ge b \ge 0. \end{aligned} \tag{5}$$

With (4) we associate cost function:

$$J(\mathbf{P}) = \int_0^T \left[\left(\left(\mathbf{R}(t) - \mathbf{R}_0(t)\right)^T, \boldsymbol{\Phi}(t)(\mathbf{R}(t) - \mathbf{R}_0(t)) \right)_{R^n} + \left((\mathbf{P}(t) - \mathbf{P}(t))^T, \boldsymbol{\Psi}(t)(\mathbf{P}(t) - \mathbf{P}(t)) \right)_{R^m} \right] dt \tag{6}$$

where $\boldsymbol{\Phi}$ is positive definite nxn matrix and $\boldsymbol{\Psi}$ is nxm matrix. First matrix $\boldsymbol{\Phi}(t)$ is describing costs of changing inventory level described by (2) and generalized in (4) and $\boldsymbol{\Psi}$ depends on costs of changing production intensity under robust perturbations.

Theorem 1: The infinitesimal generator $\tilde{\mathbf{A}}$ of a strongly continuous semigroup of operators for the model (4) has the following form:

$$\tilde{\mathbf{A}} \begin{bmatrix} \boldsymbol{\varphi}^0 \\ \boldsymbol{\varphi}^1 \\ \boldsymbol{\varphi}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_0\boldsymbol{\varphi}^1(0) + \mathbf{A}_1\boldsymbol{\varphi}^1(-a) + \mathbf{B}_0\boldsymbol{\varphi}^2(0) + \mathbf{B}_1\boldsymbol{\varphi}^2(-b) \\ \dfrac{\partial \boldsymbol{\varphi}^1}{\partial \theta} \\ \dfrac{\partial \boldsymbol{\varphi}^2}{\partial \eta} \end{bmatrix} \tag{7}$$

with the domain

$$D(\tilde{\mathbf{A}}) = \begin{Bmatrix} \boldsymbol{\varphi} = \left(\boldsymbol{\varphi}^0, \boldsymbol{\varphi}^1, \boldsymbol{\varphi}^2\right) \in Z, \boldsymbol{\varphi}^1 \in W^{1,2}\left(\left[-b, t_1\right], R^n\right), \boldsymbol{\varphi}^1(0) = \boldsymbol{\varphi}^0, \\ \dfrac{\partial \boldsymbol{\varphi}^2}{\partial \eta} \in L_2\left(\left[-b, t_1\right], R^m\right), \boldsymbol{\varphi}^2(t) \in L_2\left(\left[-b, t_1\right], R^m\right), \boldsymbol{\varphi}^2(t_1) = 0, \end{Bmatrix}$$

,and $\boldsymbol{\varphi}^2$ is absolutely continuous. (7 a)

For proof see Čibej and Bogataj (1994), Bogataj and Čibej (1994) and Ichikawa (1982). Using the results of Theorem 1 the equation 4 can be written in Hilbert space as the following abstract evolution equation using the results of Bogataj (1989, 1989a, 1990) and Ichikawa (1982).

$\dot{\mathbf{z}}(t) = \tilde{\mathbf{A}}\ \mathbf{z}(t)$ with initial data $\mathbf{z}(0) = \mathbf{z}_0$ and

$$\mathbf{P}(t) = \begin{cases} \boldsymbol{\varphi}^2(t) \in L_2\left(\left[-b, t_1\right], R^m\right) \\ 0 \Leftarrow t > t_1 \end{cases}, \text{where} \quad \mathbf{z}(t) = \begin{bmatrix} \mathbf{R}(t) \\ \mathbf{R}(t+\phi) \\ \mathbf{P}(t+\eta) \end{bmatrix}, \quad -a \le \phi < 0, \quad -b \le \eta \le 0$$

$$\mathbf{z}(t) \in Z, \qquad Z = R^n \, \text{x} L_2\left(\left[-a,0\right], R^n\right) \text{x} L_2\left(\left[-b,t_1\right], R^m\right) \tag{8}$$

The cost function (6) can be written in abstract form as

$$J(\mathbf{P}) = \int_0^{t_1} \left[\left(\mathbf{z}(t)^T, \tilde{\mathbf{\Phi}}^+(t)\mathbf{z}(t) \right)_Z + \int_{-b}^{0} \left(\mathbf{P}(t+\eta)^T, \mathbf{\Psi}^+(t)\mathbf{P}(t+\eta) \right)_{R^m} d\eta \right] dt$$

where $\tilde{\mathbf{\Phi}}^+ = \begin{bmatrix} \mathbf{\Phi}^+ & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, and transformations $\tilde{\mathbf{\Phi}}$ to $\tilde{\mathbf{\Phi}}^+$ and $\mathbf{\Psi}^+$ to $\mathbf{\Psi}^+$

appear when servomechanism gets regulator's form. (9)

In supply chain models described by (4) the perturbations of delays in control take place. So we can write in stead of the model (4) the following perturbed equation with finite amplitude perturbation k of the mentioned type:

$$\dot{\mathbf{R}}(t) = \mathbf{A}_0\mathbf{R}(t) + \mathbf{A}_1\mathbf{R}(t-a) + \mathbf{B}_0\mathbf{P}(t) + \mathbf{B}_1\mathbf{P}(t-k), \qquad 0 < t < t_1$$

With initial data

$$\mathbf{R}(0) = \boldsymbol{\varphi}^0, \qquad \mathbf{R}(\tau) = \boldsymbol{\varphi}^1(\tau) \qquad -a \le \tau < 0$$

$$\mathbf{P}(\tau) = \boldsymbol{\varphi}^2(\tau) \qquad -b \le k \le \tau \le 0; \qquad \mathbf{P}(t) = \begin{Bmatrix} \boldsymbol{\varphi}^2(t) \in L_2\left(\left[-b,t_1\right], R^m\right) \\ 0 \Leftarrow t > t_1 \end{Bmatrix} \tag{10}$$

We note that in perturbed system (10) the perturbed delay k in the control appears.
The equation (10) can be written as the following abstract evolution equation in the space

$$Z : \dot{\mathbf{z}}(t) = \tilde{\mathbf{A}}\mathbf{z}(t) + \Delta\tilde{\mathbf{A}}\mathbf{z}(t) \qquad \mathbf{z}(0) = \mathbf{z}_0, \qquad \mathbf{P}(t) \in L_2\left(\left[-b,t_1\right], R^m\right), \mathbf{P}(t) \equiv 0 \Leftarrow t > t_1 \tag{11}$$

The abstract differential equation (11) can be written in condensed form as :
$\dot{\mathbf{z}}(t) = \tilde{\mathbf{A}}_p\mathbf{z}(t)$, where $\tilde{\mathbf{A}}_p$ is the perturbed infinitesimal generator of a strongly continuous semigroup corresponding to the perturbed system (10). In (11) the operator $\tilde{\mathbf{A}}$ is defined by (7) and domain (7 a). Here $\Delta\tilde{\mathbf{A}}$ can be expressed by:

$$\Delta\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{B}_0\mathbf{P}(t) + \mathbf{B}_1\mathbf{P}(t-k) \\ 0 \\ 0 \end{bmatrix} \tag{12}$$

We note that all perturbations can be of a finite magnitude.

4 THE PERTURBED ADJOINT OPERATOR FOR PERTURBATIONS OF FINITE MAGNITUDE

For the original unperturbed model described by linear diff with delay in the state and in the control differential – difference equations, the known characterization of adjoint operator of the infinitesimal generator can be expressed by the following theorem:

Theorem 2: The adjoint $\tilde{\mathbf{A}}^*$ of the infinitesimal generator $\tilde{\mathbf{A}}$ for the model (4) is given by

$$D\left\{\tilde{\mathbf{A}}^*\right\} = \begin{Bmatrix} \left(\mathbf{h}^0, \mathbf{h}^1, \mathbf{h}^2\right) \in M^2\left(\left[-a,0\right], R^n\right) \text{x} L_2\left(\left[-b,t_1\right], R^m\right), \\ \mathbf{z} \in W^{1,2}\left(\left[-a,0\right], R^n\right) \text{x} W^{1,2}\left(\left[-b,t_1\right], R^m\right), \mathbf{z}^1(-a) = 0, \\ \mathbf{z}^2(-b) = 0, \qquad \mathbf{h}^1(\alpha) = \mathbf{z}^1(\alpha) + A_1^{T_h 0}, \mathbf{h}^2(\alpha) = \mathbf{z}^2(\beta) + B_1^{T_h 0} \end{Bmatrix}$$

$$\int_{-b}^{t_1} \mathbf{D}z^2(\eta)d\eta = -\left(\mathbf{B}_0^T + \mathbf{B}_1^T\right)\mathbf{h}^0 \tag{13}$$

and the action is:

$$\left[\tilde{\mathbf{A}}^*\mathbf{h}\right]^0 = \left(\mathbf{A}_0{}^T + \mathbf{A}_1{}^T\right)\mathbf{h}^0 + \int_{-a}^{0} \mathbf{D}\mathbf{z}^1(\alpha)d\alpha \quad \left[\tilde{\mathbf{A}}^*\mathbf{h}\right]^1(\phi) = -\mathbf{D}\mathbf{z}^1(\phi) \quad \left[\tilde{\mathbf{A}}^*\mathbf{h}\right]^2(\eta) = -\mathbf{D}\mathbf{z}^2(\eta) \tag{14}$$

For the perturbed system (10), where the finite magnitude perturbations of the delay in the control take place, the following Theorem characterizes the perturbed adjoint operator $\mathbf{A}^*$:

Theorem 3: The adjoint of perturbed operator $\tilde{\mathbf{A}}$ is given by:

$$D\left\{\tilde{\mathbf{A}}_p{}^*\right\} = \left\{\begin{array}{l} \left(\mathbf{h}^0,\mathbf{h}^1,\mathbf{h}^2\right) \in M^2\left(\left[-a,0\right],R^n\right) \mathrm{x} L_2\left(\left[-b,t_1\right],R^m\right), \\ \mathbf{z} \in W^{1,2}\left(\left[-a,0\right],R^n\right) \mathrm{x} W^{1,2}\left(\left[-b,t_1\right],R^m\right), \mathbf{z}^1(-a) = 0, \\ \mathbf{z}^2(-b) = 0, \quad \mathbf{h}^1(\alpha) = \mathbf{z}^1(\alpha) + \mathbf{A}_1{}^{T_h 0}, \mathbf{h}^2(\beta) = \mathbf{z}^2(\beta) + \boldsymbol{\chi}(\beta)\mathbf{B}_1{}^{T_h 0} \end{array}\right\} \tag{15}$$

$\int_{-b}^{t_1} \mathbf{D}z^2(\eta)d\eta = -\left(\mathbf{B}_0^T + \mathbf{B}_1^T\right)\mathbf{h}^0$ and the action is

$$\left[\tilde{\mathbf{A}}_p{}^*\mathbf{h}\right]^0 = \left(\mathbf{A}_0{}^T + \mathbf{A}_1{}^T\right)\mathbf{h}^0 + \int_{-a}^{0} \mathbf{D}\mathbf{z}^1(\alpha)d\alpha, \quad \left[\tilde{\mathbf{A}}_p{}^*\mathbf{h}\right]^1(\phi) = -\mathbf{D}\mathbf{z}^1(\phi), \quad \left[\tilde{\mathbf{A}}_p{}^*\mathbf{h}\right]^2(\eta) = -\mathbf{D}\mathbf{z}^2(\eta) \tag{16}$$

In expression (16) we have $\boldsymbol{\chi}(\beta) = \begin{cases} 1, & -k \le \beta \le 0 \\ 0, & elsewhere \end{cases}$

Proof: Chosen on arbitrary $\boldsymbol{\mu} \in D(\tilde{\mathbf{A}}_p{}^*)$ and let it be $\mathbf{v} = \tilde{\mathbf{A}}_p^*\boldsymbol{\mu}, \quad \Delta\mathbf{v} = \Delta\tilde{\mathbf{A}}^*\boldsymbol{\mu}$, where $\tilde{\mathbf{A}}^*$ is characterized by **Theorem 2** and $\Delta\tilde{\mathbf{A}}^*$ is the perturbed part of the adjoint. Using the definition of the adjoint operator we have

$$\left(\tilde{\mathbf{A}}_p\mathbf{h},\boldsymbol{\mu}\right) = \left(\mathbf{h},\tilde{\mathbf{A}}_p{}^*\boldsymbol{\mu}\right) = \left(\mathbf{h},\mathbf{v}\right) = \left(\mathbf{h},\tilde{\mathbf{A}}^*\boldsymbol{\mu}\right) + \left(\mathbf{h},\Delta\tilde{\mathbf{A}}^*\boldsymbol{\mu}\right) \tag{17}$$

For all $\mathbf{h} \in W^{1,2}\left(\left[-a,0\right],R^n\right) \mathrm{x} W^{1,2}\left(\left[-b,t_1\right],R^m\right)$. Using the definition of the infinitesimal generator (7) for $\tilde{\mathbf{A}}$ and perturbation part of generator $\tilde{\mathbf{A}}_p$ denoted by $\Delta\tilde{\mathbf{A}}$ given by (12) it follows that:

$$(\tilde{\mathbf{A}}\mathbf{h},\boldsymbol{\mu}) = \mathbf{A}_0\mathbf{h}^0 + \mathbf{A}_1\mathbf{h}^1(-a) + \mathbf{B}_0\mathbf{h}^2(0) + \mathbf{B}_1\mathbf{h}^2(-b) + \mathbf{B}_1(\mathbf{h}^2(-k) - \mathbf{h}^2(-b)), \boldsymbol{\mu}(0) +$$

$$+\int_{-a}^{0} \mathbf{D}\mathbf{h}^1(\mathbf{v}),\boldsymbol{\mu}^1(\mathbf{v}))d\mathbf{v} + \int_{-b}^{t_1} \mathbf{D}\mathbf{h}^2(\eta),\boldsymbol{\mu}^2(\eta))d\eta = (\mathbf{h}^0,(\mathbf{A}_0{}^T + \mathbf{A}_1{}^T)\boldsymbol{\mu}(0)) +$$

$$+(\mathbf{h}^2,(\mathbf{B}_0{}^T + \mathbf{B}_1{}^T)\boldsymbol{\mu}(0)) - (\mathbf{A}_1\int_{-a}^{0} \mathbf{D}\mathbf{h}^1(\alpha)d\alpha,\boldsymbol{\mu}\ (0)) + \int_{-a}^{0} (\mathbf{D}\mathbf{h}^1(\alpha),\boldsymbol{\mu}^1(\alpha))d\alpha -$$

$$-\mathbf{B}_1\int_{-b}^{t_1} \mathbf{D}\mathbf{h}^2(\eta)d\eta,\boldsymbol{\mu}(0)) + \int_{-b}^{t_1} (\mathbf{D}\mathbf{h}^2(\eta),\boldsymbol{\mu}^2(\eta))d\eta + \mathbf{B}_1\int_{-b}^{t_1} \mathbf{D}\mathbf{h}^2(\eta)d\eta,\boldsymbol{\mu}(0)) -$$

$$-\mathbf{B}_1 \int_{-k}^{t_1} \mathbf{Dh}^2(\eta) d\eta, \boldsymbol{\mu}(0)) = (\mathbf{h}^0, (\mathbf{A}_0{}^T + \mathbf{A}_1{}^T)\boldsymbol{\mu}(0)) +$$

$$+(\mathbf{h}^2(0), (\mathbf{B}_0{}^T + \mathbf{B}_1{}^T)\boldsymbol{\mu}(0)) + \int_{-a}^{0} \mathbf{Dh}^1(\alpha), \boldsymbol{\mu}^1(\alpha) - \mathbf{A}_1{}^T \boldsymbol{\mu}(0)) d\alpha + \qquad (18)$$

$$\int_{-b}^{t_1} (\mathbf{Dh}^2(\eta), \boldsymbol{\mu}^2(\eta) - \mathbf{B}_1{}^T \boldsymbol{\mu}(0)) d\eta + \int_{-b}^{t_1} (\mathbf{Dh}^2(\eta), \mathbf{B}_1{}^T \boldsymbol{\mu}(0)) d\eta - \int_{-b}^{t_1} (\mathbf{Dh}^2(\eta), \chi(\eta)\mathbf{B}_1{}^T \boldsymbol{\mu}(0)) d\eta$$

Similarly we have

$$(\mathbf{h}, \mathbf{v}) = (\mathbf{h}^0, \mathbf{v}^0) + \int_{-a}^{0} (\mathbf{h}^1(\phi), \mathbf{v}^1(\phi)) d\phi + \int_{-b}^{t_1} (\mathbf{h}^2(\eta), \mathbf{v}^2(\eta)) d\eta = (\mathbf{h}^0, \mathbf{v}^0 + \int_{-a}^{0} \mathbf{v}^1(\phi) d\phi) +$$

$$+(\mathbf{h}^2(0), \int_{-b}^{t_1} \mathbf{v}^2(\eta)) d\eta - \int_{-a}^{0} (\mathbf{Dh}^1(\alpha), \int_{-a}^{\alpha} \mathbf{v}^1(\phi)) d\phi) d\alpha - \int_{-b}^{t_1} (\mathbf{Dh}^2(\eta), \int_{-b}^{\eta} \mathbf{v}^2(\varepsilon)) d\varepsilon) d\alpha \qquad (19)$$

To obtain (19) we applied twice Fubini's Theorem which is justified as the mapping $(\alpha, \phi) \to (\mathbf{Dh}^1(\alpha), \mathbf{v}^1(\phi))$ is in $L_1\left(\left[-a,0\right] \mathrm{x} \left[-a,0\right], (R^n)^T \mathrm{x} R^n\right)$ and the mapping $(\left(\mu, \varepsilon\right) \to (\mathbf{Dh}^2(\eta), \mathbf{v}^2(\varepsilon))$ is in $L_1\left(\left[-b,t_1\right] \mathrm{x} \left[-b,t_1\right], (R^m)^T \mathrm{x} R^m\right)$. For details see Bogataj (1989). Comparing (18) and (19) we see that the following equalities have to be fulfilled:

$$\mathbf{v}^0 + \int_{-a}^{0} \mathbf{v}^1(\phi) d\phi = \left(\mathbf{A}_0{}^T + \mathbf{A}_1{}^T\right)\boldsymbol{\mu}(0), \qquad \int_{-b}^{t_1} (\mathbf{v}^2(\eta) d\eta = \left(\mathbf{B}_0{}^T + \mathbf{B}_1{}^T\right)\boldsymbol{\mu}(0)$$

$$-\int_{-a}^{\alpha} \mathbf{v}^1(\phi) d\phi = \boldsymbol{\mu}^1(\alpha) - \mathbf{A}_1{}^T \boldsymbol{\mu}(0), \quad -\int_{-b}^{\eta} \mathbf{v}^2(\varepsilon) d\varepsilon = \boldsymbol{\mu}^2(\eta) - \mathbf{B}_1{}^T \boldsymbol{\mu}(0) + \mathbf{B}_1{}^T \boldsymbol{\mu}(0) - \boldsymbol{\chi}(\eta)\mathbf{B}_1{}^T \boldsymbol{\mu}(0) \qquad (20)$$

This means that there exist a $\mathbf{z} \in W^{1,2}\left(\left[-a,0\right], R^n\right) \mathrm{x} W^{1,2}\left(\left[-b,t_1\right], R^m\right), \mathbf{z}^1(-a) = 0, \mathbf{z}^2(-b) = 0$ such that $\boldsymbol{\mu}^1(\alpha) = \mathbf{z}^1(\alpha) + \mathbf{A}_1{}^T \boldsymbol{\mu}(0)$ for all $\alpha \in \left[-a,0\right]$, $\mathbf{Dz}^1\left(\alpha\right) = -\mathbf{v}^1\left(\alpha\right)$ in $L_2\left(\left[-a,0\right], R^n\right)$.
$\mathbf{v}^2(\eta) = \mathbf{z}^2(\eta) - \mathbf{B}_1{}^T \boldsymbol{\mu}(0) + \mathbf{B}_1{}^T \boldsymbol{\mu}(0) + \boldsymbol{\chi}(\eta)\mathbf{B}_1{}^T \boldsymbol{\mu}(0)$ for all $\eta \in \left[-b,0\right]$,
$\mathbf{Dz}^2\left(\eta\right) = -\mathbf{v}^2\left(\eta\right)$ in $L_2\left(\left[-b,t_1\right], R^m\right)$ and $\mathbf{v}^0 = \left(\mathbf{A}_0{}^T + \mathbf{A}_1{}^T\right)\boldsymbol{\mu}(0) + \int_{-a}^{0} Dz^1(\alpha) d\alpha$

$$0 = (\mathbf{B}_0{}^T + \mathbf{B}_1{}^T)\boldsymbol{\mu}(0) + \int_{-b}^{t_1} \mathbf{Dz}^2(\eta) d\eta, \quad v^1(\alpha) = -\mathbf{Dz}^1(\alpha), \qquad v^2(\eta) = -\mathbf{Dz}^2(\eta), \qquad (21)$$

Consequently, $D(\mathbf{A}^*)$ is contained in the set (22):

$$D^* \equiv \left\{ \left(\mathbf{r}^0, \mathbf{r}^1, \mathbf{r}^2\right) \in Z \middle| \; \mathbf{z} \in W^{1,2}\left(\left[-a,0\right], R^n\right) \mathrm{x} W^{1,2}\left(\left[-b,t_1\right], R^m\right), \mathbf{z}^1(-a) = 0, \mathbf{z}^2(-b) = 0, \right.$$
$$\left. \mathbf{r}^1(\alpha) = \mathbf{z}^1(\alpha) + \mathbf{A}_1{}^T \mathbf{r}^0; \;\; \mathbf{r}^2(\eta) = \mathbf{z}^2(\eta) + \mathbf{B}_1{}^T \boldsymbol{\mu}(0) - \mathbf{B}_1{}^T \boldsymbol{\mu}(0) + \boldsymbol{\chi}(\eta)\mathbf{B}_1{}^T \boldsymbol{\mu}(0) \right\}$$

The easy verification that $D^* \subset D(\mathbf{A}^*)$ by applying again Fubini's Theorem is left to the reader. **Q.E.D.**

Corollary 1: The adjoint of the perturbed part of the operator $\mathbf{A}:\Delta\mathbf{A}^*$ is given by

$$D\left\{\Delta\mathbf{A}^*\right\} \equiv \left\{ \left(\mathbf{h}^0,\mathbf{h}^1,\mathbf{h}^2\right) \in M_2\left(\left[-a,0\right],R^n\right) \mathrm{x} L_2\left(\left[-b,t_1\right],R^m\right) \right|$$

$$\mathbf{z} \in W^{1,2}\left(\left[-a,0\right],R^n\right) \mathrm{x} W^{1,2}\left(\left[-b,t_1\right],R^m\right); \quad \mathbf{z}^1(-a)=0, \mathbf{z}^2(-b)=0.$$

and $\mathbf{h}^1(\alpha)=z^1(\alpha),\ \ \mathbf{h}^2(\beta)=z^2(\beta)-\mathbf{B}_1{}^T\mathbf{h}^0+\boldsymbol{\chi}(\beta)\mathbf{B}_1{}^T\mathbf{h}^0,\ \ \int_{-b}^{t_1}\mathbf{D}\mathbf{z}^2(\eta)d\eta=-(\mathbf{B}_0{}^T+\mathbf{B}_1{}^T)\mathbf{h}^0 \Big\}$

and the action is

$$\left[\Delta\mathbf{A}^*\mathbf{h}\right]^0=\int_{-a}^{0}\mathbf{D}\mathbf{z}^1(\alpha)d\alpha,\ \left[\Delta\mathbf{A}^*\mathbf{h}\right]^1=-\mathbf{D}\mathbf{z}^1(\alpha),\ \left[\Delta\mathbf{A}^*\mathbf{h}\right]^2=-\mathbf{D}\mathbf{z}^2(\eta)$$

Proof can be easily derived by splitting $\mathbf{z}_p=\mathbf{z}+\Delta\mathbf{z}$, Here $\mathbf{z}$ correspond to $\mathbf{v}=\mathbf{A}^*\boldsymbol{\mu}$ and $\Delta\mathbf{z}$ correspond to $\Delta\mathbf{v}=\Delta\mathbf{A}^*\boldsymbol{\mu}$. Then we consider the characterization of $\mathbf{A}^*$ given in **Theorem 2,** in the proof of **Theorem 3** and finally the residuum terms give the result of **Corollary 1.**

Q.E.D.

Remark: For easy verification of Corollary 1 we left in the proof of Theorem 3 also terms that obviously cancel.

5 CONCLUSION

In the paper we have shown how to derive a perturbed adjoint operator for the infinitesimal generator of a strongly continuous semigroup of linear operators for linear hereditary systems which can describe perturbed supply networks, with delays in the state of the system and in the control. The importance of the results for supply networks lies in the fact that considered perturbations in the delay of control can be of finite magnitude, which is the case of today's actions against recession which is affecting smoothness of flows in a supply chain. The achieved results are particularly useful by the sensitivity study because we have succeeded to split the perturbed adjoint operator into the part which is equal to the adjoint operator of unperturbed model plus the part which is the result of perturbations . This enables us to get the straightforward sensitivity results for filtering with delays in observations of supply chains. The results can improve the findings of Basin et al (2004, 2007) on optimal and robust control for linear state-delay systems in general, and could extend the possibilities for using MRP Theory of Grubbström also in case of robust perturbations in control of global supply chain.

The usefulness of the results can be also find when study Optimal tracking control (OTC) for time-delay systems affected by persistent finite disturbances with quadratic performance index. By introducing a sensitivity parameter, the original OTC problem could be transformed into a series of two-point boundary value (TPBV) problems without time-advance or time-delay terms. Recently obtained OTC law consists of analytic feed-forward and feedback terms and a compensation term which is the sum of an infinite series of adjoint vectors. A simulation example shows that the approximate approach is effective in tracking the reference input and robust with respect to exogenous persistent disturbances. The mentioned procedure can be improved substantially by presented results for finite perturbations in delay, which could be one of directions for further research too.

References

[1] Basin, M., Rodriguez-Gonzalez, J., Martinez-Zuniga, R. (2004) Optimal control for linear systems with time delay in control input. *Journal of the Franklin Institute*, 341 (3), p.267-278.

[2] Basin, M., Rodriguez-Gonzalez, J. and Fridman, L. , 2007 Optimal and robust control for linear state-delay systems. *Journal of the Franklin Institute*, 344 (6), p.830-845

[3] Bogataj L (1989) Sensitivity of linear-quadratic systems with delay in the state and in the control for perturbation of the systems matrices. *Glas. mat.*, 24, 2-3, str. 355-360.

[4] Bogataj L (1989a) The Adjoint Operator of the Infinitesimal Generator for the Linear Differential – Difference Equation With Delays in the State and in the Control. *Glas. mat.*, 24, pp. 349-354.

[5] Bogataj L (1990) The estimation of the influence of uncertain delay in the control on the objective function. *Glas. mat.*, 1990, 25, 1, str. 49-55.

[6] Bogataj M and Bogataj L (1991) Perturbed inventory systems with delays. *Int. j. prod. econ.*. [Print ed.], vol. 26, str. 277-281.

[7] Bogataj L. and Čibej, J. A. (1994) Perturbations in living stock and similar biological inventory models. *Int. j. prod. econ.*. let. 35, št. , str. 233-239.

[8] Čibej J A. and Bogataj L (1994) Sensitivity of quadratic cost functionals under stochastically perturbed controls in inventory systems with delays. *Int. j. prod. econ.*, 35, pp. 265-270.

[9] Bogataj M and Bogataj L (1994) Optimal control of hereditary inventory system with short-time conservation effects. *Int. j. prod. econ.*. [Print ed.], vol. 35, str. 241-244.

[10] Bogataj M and Bogataj L and Vodopivec, R (2005). Stability of perishable goods in cold logistic chains. *Int. j. prod. Econ.* , vol. 93/94, str. 345-356.

[11] Bogataj L, Pritchard, A.J. *1978.*Sensitivity analysis for control problems described by delay equations. *Advances in measurement and control, MECO'78,*Zurich: Acta Press, pp.741-746.

[12] Bogataj, D., Bogataj, M., 2007. Measuring the supply chain risk and vulnerability in frequency space. *Int. j. prod. econ.*. [Print ed.], vol. 108, Issues 1-2, pp. 291-301.

[13] Bogataj, M., Grubbström, R.W., Bogataj, L (2008), *Location of activity cells in distribution and reverse logistics,* ISIR Symposium 2008, Budapest, to appear: *Int.j. Prod. Econ.*

[14] Borgonovo E, Peccati L. (2009) Financial management in inventory problems: Risk averse vs risk neutral policies, , *International Journal of Production Economics*, 118 (1), p.233-242,

[15] Borgonovo E, Peccati L. (2007) Global sensitivity analysis in inventory management, *International Journal of Production Economics*, 108 (1), p.302-313,

[16] Borgonovo, E. (2008) Differential importance and comparative statics: An application to inventory management *International Journal of Production Economics*, 111 (1), p.170-179.

[17] Grubbström, R.W., 1998. A Net Present Value Approach to Safety Stocks in Planned Production , *International Journal of Production Economics*, Vol. 56-57, pp. 213-229.

[18] Grubbström, R.W., 1996. Material Requirements Planning and Manufacturing Resources Planning, in Warner, M., (Ed.), *International Encyclopedia of Business and Management,* Routladge, London,

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

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

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

[22] Grubbström, R.W., Bogataj, M., Bogataj, L (2007), *A compact representation of distribution and reverse logistics in the value chain*, Ljubljana: Faculty of Economics, KMOR.

[23] Ichikawa,A. (1982) Quadratic Control of Evolution equations with delays in control, *SIAM* J. Control and Optimization, pp. 645-668.

[24] Tang, G. , Sun, H. and Pang, H., 2008 Approximately optimal tracking control for discrete time-delay systems with disturbances, Progress in Natural Science, 18 (2), p.225-231.

[25] Vinter R. (1978), On the evaluation of the state of linear differential –delay equations in M; properties of generator, *J. Inst.Math.Appl*. 21 , pp.13-23.

SCHEDULING JOBS WITH FORBIDDEN SETUPS WITH METAHEURISTICS AND PENALTIES – A CASE STUDY AT A CONTINUOUS CASTING PLANT

Karsten Hentsch and **Peter Köchel**
Chemnitz University of Technology, Department of Computer Science
Straße der Nationen 62, 09107 Chemnitz, Germany
{karsten.hentsch, peter.koechel}@informatik.tu-chemnitz.de

Abstract: In real world scheduling problems often constraints arise, which are not covered by traditional scheduling problems. In this paper we tackle a forbidden setups constraint which was found at a continuous casting plant. Because of alloy issues it is not possible to get from one job to every other. When dealing with such constraints, often penalties are suggested to resolve this problem, so we tried to apply this solution. Some tests with a black box SPEA2 algorithm were made to show the difficulties of this approach.

Keywords: scheduling, constraints, forbidden setups, penalties, genetic algorithm.

1 INTRODUCTION

In almost all service and manufacturing systems one of the most important tasks is how to schedule arriving jobs such that some criteria will be satisfied. In the consequence up to now there have been defined a great variety of scheduling problems as well as corresponding models and solution approaches (see e.g. [1, 15]). However, most models suffer from such more or less restrictive assumptions like single machine, unique processing times, zero set-up times or a single criterion. On the other hand some classical approaches like linear or dynamic programming are practicable for small-size problems only. Therefore over the past years we can observe an increased application of heuristic search methods as for instance Genetic Algorithms, Tabu Search, Simulated Annealing and many more (cp. [16]). If well designed these approaches find good solutions, if not even optimal ones. In the literature one can find several successful applications. But this does not mean that there are no more open problems with scheduling. Practice is richer than theory. For instance, scheduling problems with forbidden setups are scarcely considered. Only a few papers examine such problems, whereas mixed-integer linear programming is the applied approach (see [2, 8, 10]). Furthermore, the comprehension of various constraints is an important topic. Constraints are an important part of most real world optimization problems. In our paper we will report on some experiences and results with such a problem.

In the following chapter we describe the use case and our research approach. Chapter 3 contains some general considerations on constraint-handling techniques, which are necessary to realize our decision to concentrate on penalty functions. Our test conditions and test results are described in Chapter 4. Chapter 5 concludes the paper.

2 CASE DESCRIPTION AND RESEARCH APPROACH

The case history is connected with a project we are realizing in cooperation with a commercial software firm located in Chemnitz and a Saxon metal-working company. The project is supported by the Development Bank of Saxony (SAB) under contract 4349-257050-73. We have the following situation. A finite number of continuous casting lines manufacture a great variety of strip stocks. The strip stocks differ by their profiles, more than 100, and used alloys, about 45 to 50. The lines can operate day and night. Setups cause idle times and costs that are sequence dependent. In distinction to most cases considered in literature we have a lot of forbidden item sequences caused by the fact that not all alloys can follow some set of other alloys. Additionally, not all lines can manufacture all items. And finally, we have two criteria – total tardiness of all jobs and a cost criterion. The cost criterion comprises setup cost and cost for finishing a job before maturity. The actual situation at the company is that a dispatcher schedules new orders continuously when arrived (dynamic scheduling). Scheduling a new order means that the dispatcher has to answer three questions:

1. At which of the available continuous casting lines to manufacture?
2. At which place in the already scheduled job sequence to insert the new order?
3. To manufacture more than ordered or not and to store the surplus production?

The main goal of the dispatcher is to get a good schedule with respect to the two criterions. Our task is not only to aid and to replace the dispatcher by corresponding software but also to find better solutions for the above mentioned questions. In principle we had to distinguish between two situations: First, the *optimal scheduling* situation in which we have to find as good as possible answers on the three questions, and second, the *admissible scheduling* situation. In the admissible scheduling situation a client wants to get a quick answer if his order can be fulfilled within special time constraints. Whereas the response time in the first situation can amount up to one or two days an answer must be given after a few minutes in the second situation.

At the actual stage of the project we considered the optimal scheduling situation only. The first observation was that dynamic scheduling in our case leads to a complete rescheduling of all not yet fixed jobs. Therefore we started our research with the goal to find optimal solutions for the first two questions from above for a given set of orders or jobs. The NP-hardness of such scheduling problems is well-known (cp. [1]). In the consequence we oriented ourselves on heuristic search methods. Because we applied Genetic Algorithms (GA) successful to several optimization problems (see [6, 7]) and we have also developed corresponding software products, where various heuristic search methods are available ([5, 9]), we concentrated on GAs.

As constraints are an important part of real world optimization problems, numerous constraint-handling techniques were proposed by several authors. To name only a few there are constraint programming, penalty functions, repair procedures and transformations to unconstrained multi-objective optimization problems. However, for GAs only a subset of the mentioned techniques is available. In the literature repair procedures and penalty functions are favorable for use with evolutionary approaches. Although a repair procedure seems to be a good choice it has a big disadvantage. As we face this NP-hard problem of generating a feasible schedule from given item sequences, we are running quickly into intractable problem sizes. To circumvent the dimensional explosion when handling problem sizes of 40 jobs, we decided to tackle the feasibility problem by penalties. A comprehensive review of such techniques is made by [3]. In the following chapter we discuss some of these techniques with respect to their applicability for our problem.

3 SOME CONSIDERATIONS ON PENALTIES FOR GA-OPTIMISATION

3.1 Constraint impact

The job sequencing problem, which is part of the scheduling problem beside the resource assignment problem, can be seen as a variation of the Hamiltonian path problem which decision variant is known to be NP-complete. This problem can be modeled as a graph with one node for every job and edges between two nodes if and only if the setup between the two corresponding jobs is allowed. Unconstrained scheduling problems are represented by a complete graph, where every possible job sequence corresponds to a Hamiltonian path. Thus the problem is trivial and every job sequence is feasible. But if some edges are missing the problem becomes hard. In our case 30% of the edges are missing, so the probability of a randomly generated job sequence of 40 jobs to be feasible, which is a practical problem size, is about 10^{-6} respectively on average it takes about one million trials to get one feasible solution. Thus it is obvious that to work only with feasible solutions is not possible. As stated in Chapter 2 we decided to use penalty functions for non-feasible solutions.

3.2 General notes on penalties

Although it sounds quite easy to define some penalty factors this is a tricky task and can be an optimization problem itself [12]. Our problem is a bi-objective optimization problem which makes this task even harder as we have to define penalty factors for both objectives. As [4] stated in reference to [14], the penalty should be kept as low as possible, just above the limit below which infeasible solutions are optimal. This is called the "minimum penalty rule". If the penalty is too high or too low, then the problem might become very difficult for a GA to solve. This difficulty is due to several properties of GAs and the solution space structure. If the penalty is too high, the search might converge quickly towards a feasible region in the solution space. But if there are several disjoint feasible regions it will be hard to get to the others [12, 13]. Besides optimal solutions often seem to lie on the border of feasible and unfeasible regions [4], but once inside a feasible region it's hard to move towards the unfeasible region to reach the border. On the other hand if the penalty is too low, the search might explore the unfeasible regions too long hence being unable to find a good feasible solution in reasonable time [13]. To aid researches in designing penalty functions, [3] derived in reference to [11] some rules, which are quoted below and discussed afterwards in relation to our problem.

1. Penalties which are functions of the distance from feasibility are better performers than those which are only functions of the number of violated constraints.
2. For a problem having few constraints, and few feasible solutions, penalties which are solely functions of the number of violated constraints are not likely to produce any solutions.
3. Good penalty functions can be constructed from two quantities: the *maximum completion cost* and the *expected completion cost*. *Completion cost* refers to the distance to feasibility.
4. Penalties should be close to the *expected completion cost*, but should not frequently fall below it. The more accurate the penalty, the better will be the solution found. When a penalty often underestimates the *completion cost*, then the search may fail to find a solution.

Applying these rules to our problem induces some difficulties. The distance to feasibility, mentioned in the 3rd rule, refers to the solution space. First of all estimating the distance to feasibility requires knowledge about the feasible region and, to be more precisely, knowledge of the nearest feasible region, due to the 4th rule. In addition to that an appropriate distance metric is required to measure the distance.

As we lack knowledge about the feasible regions, due to the hardness of the Hamiltonian path problem, these rules are difficult to realize. A search over possible job displacements, a simple distance metric, to find the nearest feasible solution is not applicable. Even the number of violated constraints, though it is easy to compute, has some properties lowering their suitability to penalize an infeasible solution.

The examples from ***Figure 1*** reveal some problems with penalizing a solution by the number of violated constraints. We use the number of job displacements as distance metric to measure the distance to feasibility and J_i denotes job *i*. Although there are two violations in permutation a), marked with a dashed arrow and the letter V, there is only a single displacement needed to make it feasible - scheduling J_3 after J_5. From distance to feasibility it is like permutation b), which can be made feasible by scheduling J_5 between J_2 and J_3. So in permutation a) the distance to feasibility is overestimated. But permutation c) shows an underestimation. Though there is only a single violation, two displacements are needed to make it feasible, scheduling J_4 after J_5 and J_3 after J_4.

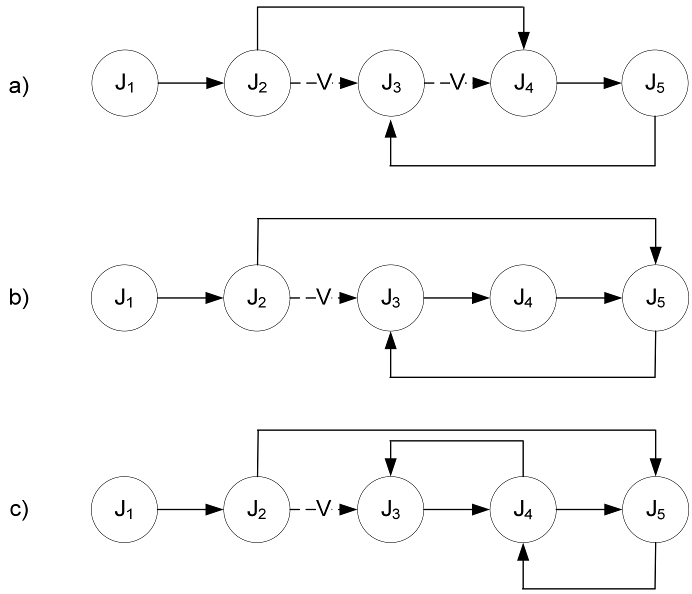


Figure 1: job sequences with forbidden setups

3.3 Local Penalties / Penalties through setup times and costs

A simple way to apply penalties for our problems is to transform the constraint problem into an unconstraint one by inserting the missing edges and penalizing them by assigning high weights. For some simple objective functions like cycle time or sum of completion times this approach might be suitable. But as scheduling problems often come along with due date related objectives, for example tardiness, lateness, earliness or other customer satisfaction related objectives, penalties incorporated through high setup times and costs have a major drawback. Considering ***Figure 2*** one can see that a high setup time after J_2 will delay all jobs after it and therefore altering the objective function value with higher magnitude than a delay after J_6. Keeping in mind that such a permutation cannot be scheduled the effects of those delays are some kind of virtual.

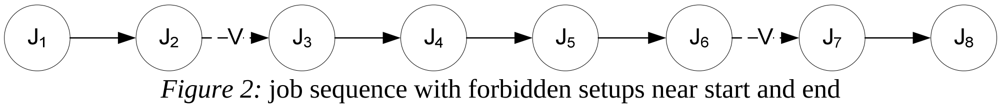


Figure 2: job sequence with forbidden setups near start and end

So in addition of being unable to determine the distance to feasibility in a sensible manner and hence being unable to define a suitable penalty factor, the effects of applying this penalty through setup times and costs depends on the position of the violation.

Unfortunately there is just another drawback associated with certain objective functions. As we pointed out results an incorporation of penalties through setup times and costs in a delay of the following jobs. For certain circumstances this could even improve the objective function value if earliness related measures are part of it. According to ***Figure 2*** we assume that job J_7 and J_8 are finished too early resulting in inventory costs. By penalizing the constraint violation from J_6 to J_7, we delay J_7 and J_8 thus compensating the penalty with lower inventory costs. Depending on penalty factor and objective function, the result of this effect ranges from lowering the penalty over even it out to improving the objective function value. Although this is bound to a limited number of objective functions and scenarios, it increases the present variability of penalty impact even more.

3.4 Global penalties

Another possibility of penalizing an infeasible solution is to schedule as much as possible of the given sequence. For all jobs that can't be scheduled a penalty is added to the objective function value. Again consider the example from ***Figure 2*** and assume that a setup from job J_2 to J_4 and from job J_6 to J_8 is allowed. This approach would result in a sequence outlined in ***Figure 3***.

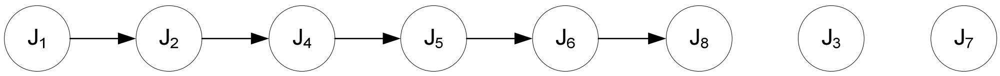


Figure 3: valid job sequence and two unscheduled jobs

Despite the advantage of being able to apply a constant penalty for all unscheduled jobs, this approach has similar drawbacks as the local penalty approach. The problem was an additional virtual delay through high setup times and costs which is replaced with a missing real delay, which the unscheduled jobs would trigger in a fully scheduled sequence. So when trying to define an appropriate penalty for an unscheduled job, one not only has to estimate the effects on the objective function value caused by the job itself, but also its delaying effect to his successors which effects the objective function value too.

In addition this approach could suffer even more from the problem of overestimating the distance to feasibility. Assuming two feasible partial sequences bound together through a violation. Furthermore assume that every setup from the last job of the first partial sequence to every other job from the second partial sequence is forbidden. This problem is depicted by ***Figure 4***. The scheduler can only schedule J_1 to J_3, afterwards all setups from J_3 are forbidden resulting in 5 jobs which cannot be scheduled. Owing to this overestimation we do not expect this approach being suitable for our problem, but since these are only some theoretical considerations, we will examine this in practice.

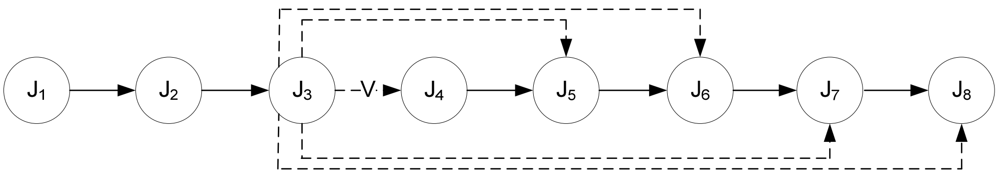


Figure 4: job sequence with a forbidden setup

4 TESTS AND RESULTS

4.1 Test setup

The test is based on the schedule of one month made by a human planner at the company. This schedule contains 125 jobs and was developed by an iterative planning process. As optimizer we used a black box version of the well known SPEA2, which is part of a framework (see [5]). A solution consists of two parts, one representing the resource assignment, the other representing the job sequence. The population size was set to 50, probability for crossover was 0.7, and mutation probability was set to 0.05. Various operators for both crossover and mutation were used. We set the optimization duration to 6 hours and took snapshots every 10 minutes to see, how the solutions evolve.

As we wanted to test both the local and global penalty approaches, we conducted tests with four different penalty value levels. Because of the *minimum penalty rule*, we took duration and cost of the worst setup as level 1. Though this setup is very rare, we made three other levels with 75%, 50% and 25% of this setup (see ***Table 1***). We will refer to those levels by the penalty factors, e.g. 24/3000 for penalty level 1.

Table 1: penalty levels with assigned penalty values for both objective functions

penalty level	time penalty in hours	cost penalty in Euro
1	24	3000
2	18	2250
3	12	1500
4	6	750

To rate the optimization results we used two solutions, to which we compare all approaches. The first solution is a schedule made by a human planner. The objective function value of this schedule highlights the border which the optimizer should beat at minimum. The optimal solution would be the most suitable to determine the quality for all approaches but as we lack that knowledge, we tried to find the best solution we could. With tremendous computational effort we were able to determine some Pareto optimal solutions, which dominate the human planner's solution. These solutions were reviewed by the human planner to validate them. We refer to them as *best known Pareto front*. As we have all necessary job data apart from release dates, our optimizer might assign a job to a time slot, for which this job was not yet known to the human planner. This advantage should be kept in mind when comparing optimized solutions against the human planner solution.

To validate the results, we started every test case 10 times, but only with 1 hour for every run to limit the computational effort. In a single objective optimization problem we would present mean and variance of the objective function values, but for multi objective optimization problems this is not applicable, since we cannot simply average multiple Pareto fronts. So we made scatter diagrams containing the Pareto optimal solutions from these 10 runs and try to validate the results with these.

4.2 Results

At first ***Figure 5*** presents the results with the global penalty approach. There is only a single feasible solution, the leftmost point near (0, 63000). Though this solution is not dominated by the human planner solution, it is unacceptable due to its high cost. Sure there is no hard upper bound for this criterion, but the costs of the optimized solutions shouldn't be much worse than the human planner solution. Among all solutions we can also see that the solution variety for both objective functions is poor. With all penalty levels the solutions converge to a Pareto front or a single point containing the feasible solution, which the optimizer isn't able to escape. For the penalty level 24/3000 one can see that after 120 minutes the optimizer gets stuck in this single solution, which he cannot improve any further. With the lower penalty levels the optimizer converges even quicker, but seems to find more solutions. But the appearance of the Pareto fronts suggests that all approaches found similar or even equal solutions, but due to the higher penalty level these solutions have worse objective function values (level 12/1500) and with higher penalty levels they are dominated by the feasible solution (level 18/2250 and level 24/3000), so they do not appear.

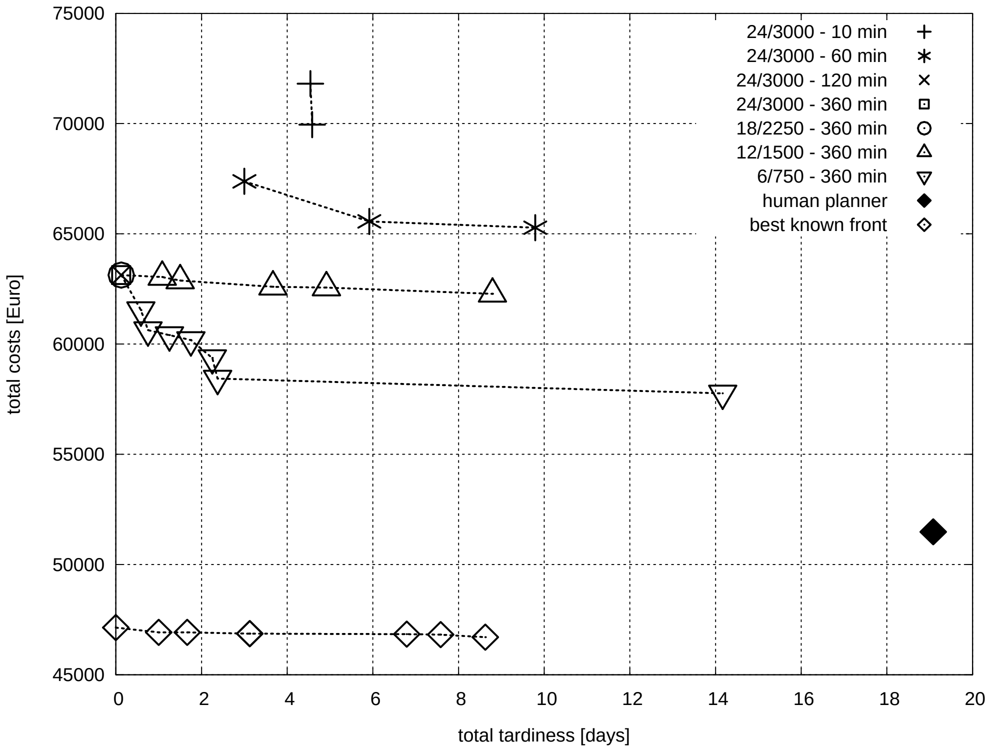


Figure 5: Comparison of different penalty values with global penalty approach

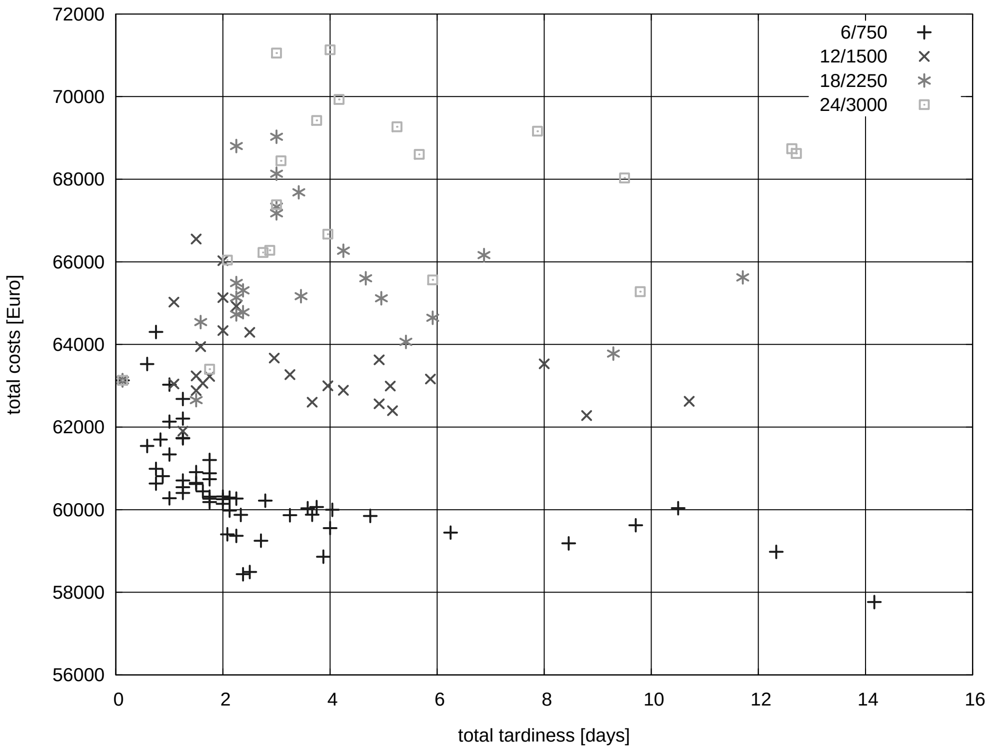


Figure 6: Solutions of 10 different runs for every penalty level with global penalty approach

All results could be confirmed by a scatter diagram from the multiple runs (see ***Figure 6***). The solution near (0, 63000) is the only feasible solution just as seen in ***Figure 5***. All other solutions are infeasible and form a layer structure apart from some outliers. Though the layers are not clearly separated from each other, some points seem to represent similar or equal solutions which are shifted to greater objective function values due to higher penalty levels. Therefore it is not possible to mark a better performing penalty level with this approach.

The results from the local penalty seemed to be more encouraging at first glance. This can easily been verified by looking at Figure 7. At first one can see that there is a greater variety of solutions, the Pareto fronts are broader and more populated as with the global penalty approach. All penalty levels except of level 6/750 produced feasible solutions and in addition level 12/1500 and 18/2250 produced solutions dominating the human planner solution. But there is still a gap about 5000 Euro to the best known solutions. Apart from that level 12/1500 seems to be the best approach to our problem, so we suggest an optimal penalty level around 12/1500.

But a look at the scatter diagrams revealed some disadvantages. ***Figure 8*** shows the important part of this diagram. We cut at a cost level of 60.000 Euro. Above this level, the solutions behave similar to ***Figure 6***, but even if they were feasible, they are unacceptable due to their high costs. We also cut at a tardiness value of 100 where only two outliers with total tardiness greater than 140 days lie beyond. These are unacceptable too. Both actions were necessary to maintain a reasonable resolution.

The solutions from the different runs have a great variety forming a big chaotic cloud. Comparing every run we saw one run, a penalty level performs best, another run it performs worst. With such a great variance in solution quality, this approach needs further attention to reduce this variance considerably. The only result which could be confirmed is that level 6/750 seems to be unsuitable, because only 3 of 10 runs found feasible solution.

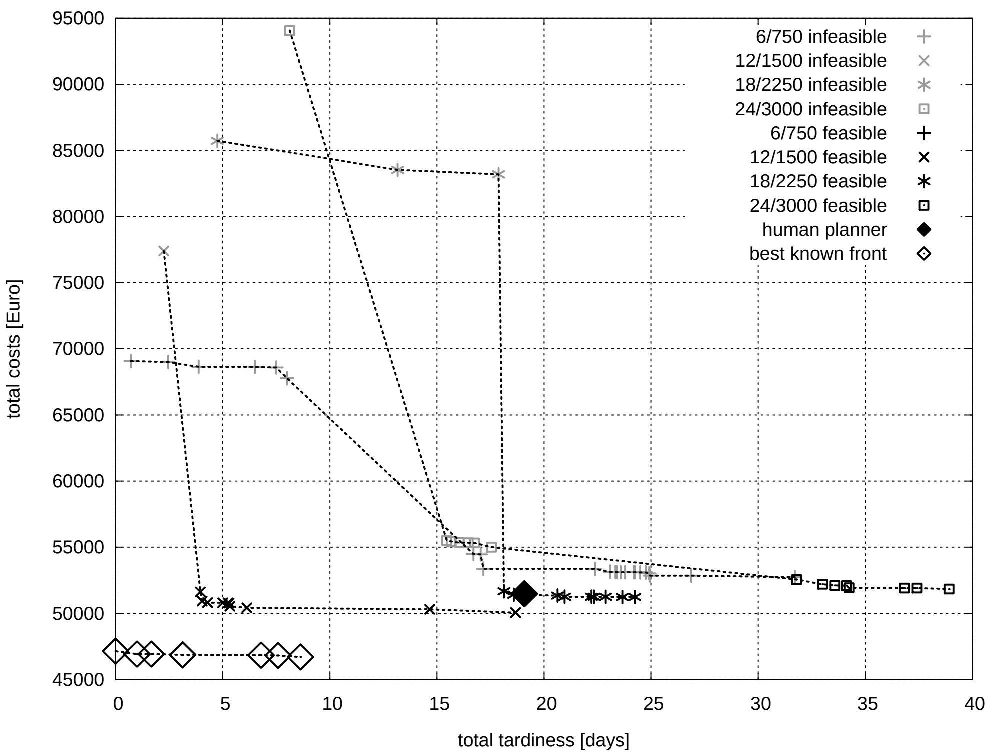


Figure 7: Comparison of different penalty values after 360 minutes with local penalty approach

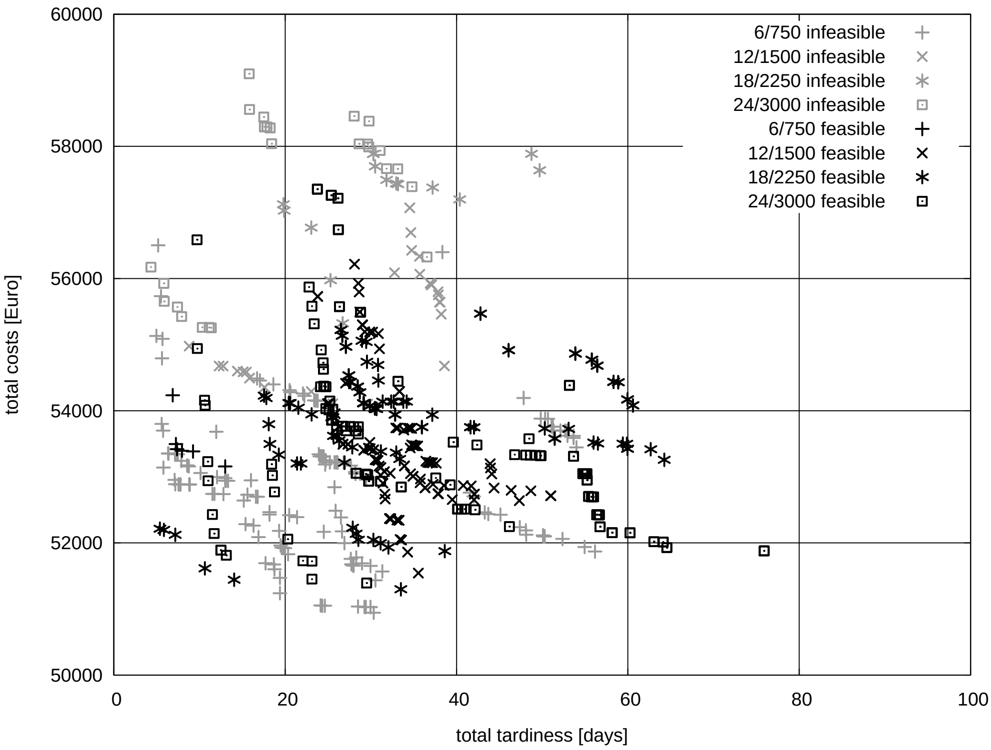


Figure 8: Solutions of 10 different runs for every penalty level with local penalty approach

5 CONCLUSION AND REMARKS

In this paper we tried to tackle a constraint scheduling problem with forbidden setups by applying genetic algorithms and penalties. Therefore we applied two different penalty types, a local penalty approach which incorporates the penalty through setup times and costs and a global penalty approach which leaves jobs after violated constraints unscheduled and adds a penalty for unscheduled jobs.

Both approaches were tested with four different penalty levels and the tests were verified with 10 independent runs. The results were compared against a human planner solution and the best known solutions found so far.

The global penalty approach produced only a single feasible solution with poor quality and some infeasible others. Moreover the solutions converged quickly towards a poor Pareto front, which could not be improved any further. The local penalty approach performed clearly better than the global one but has some other disadvantages. Although this approach produced many and wide-ranging solutions, which were distributed about fifty-fifty between feasible and infeasible ones, the solution variance between different optimization runs is very high. And even though the best feasible solutions beat the human planner solution, they were unable to get within close proximity to the best known solutions.

In future we will concentrate on schedule generation schemes which can construct or repair invalid schedules. In addition we will try to improve the local penalty approach to decrease solution variance. The global penalty approach won't be followed up. Moreover other constraint handling techniques could be incorporated to solve our problem reasonably.

6 ACKNOWLEDGEMENT

Support provided by the Development Bank of Saxony (SAB) under contract 4349-257050-73 is gratefully acknowledged.

References

[1] Brucker, P., 2007. Scheduling Algorithms. 5th edition, Springer Berlin Heidelberg.

[2] Chen, C. L., Xiao, C. L., Feng, D., Shao, H. H., 2002. Optimal shortterm scheduling of multiproduct single-stage batch plants with parallel lines. Industrial Engineering and Chemistry Research, 41(5), pp. 1249–1260.

[3] Coello Coello, C. A., 1999. A Survey of Constraint Handling Techniques used with Evolutionary Algorithms. Technical Report, Lania-RI-99-04, Laboratorio Nacional de Informática Avanzada, Xalapa, Veracruz, México.

[4] Coello Coello, C. A., Lamont, G. B., Van Veldhuizen, D. A., 2007. Evolutionary Algorithms for Solving Multi-Objective Problems. 2nd edition, Springer US.

[5] Kämpf, M., 2009. Software-Framework zur Simulationsbasierten Optimierung mit Anwendung auf Produktions- und Lagerhaltungssysteme. Cuvillier Verlag.

[6] Köchel, P., Nieländer, N. U.,2002. Kanban Optimization by Simulation and Evolution. Production Planning & Control, 13(8), pp. 725–734.

[7] Köchel, P., Nieländer, N. U.,2005. Simulation-Based Optimisation of Multi-Echelon Inventory Systems. International Journal of Production Economics, 93-94, pp. 505–513.

[8] Kopanos, G. M., Lanez, J. M., Puigjaner, L., 2009. An efficient mixedinteger linear programming scheduling framework for addressing sequence-dependent setup issues in batch plants. Industrial Engineering and Chemistry Research, Webpublication, Article ASAP.

[9] Nieländer, N. U.,2009. CHEOPS: Das Chemnitzer hybrid-evolutionäre Optimierungssystem. http://archiv.tu-chemnitz.de/pub/2009/0100.

[10] Relvas, S., Matos, H. A., Barbosa-Póvoa, A. P. F. D., Fialho, J., 2007. Reactive scheduling framework for a multiproduct pipeline with inventory management. Industrial Engineering and Chemistry Research, 46(17), pp. 5659–5672.

[11] Richardson, J. T., Palmer, M. R., Liepins, G. and Hilliard, M., 1989. Some Guidelines for Genetic Algorithms with Penalty Functions. In J. D. Schaffer, editor, Proceedings of the Third International Conference on Genetic Algorithms, pp. 191–197, San Mateo, California.

[12] Runarsson, T.P., Xin Yao, 2000. Stochastic ranking for constrained evolutionary optimization. IEEE Transactions on Evolutionary Computation, 4(3), pp. 284–294.

[13] Smith, A. E., Coit, D. W., 1997. Constraint Handling Techniques - Penalty functions. In: Bäck, T., Fogel, D. B., Michalewicz, Z. (eds.), Handbook of Evolutionary Computation, chapter C5.2, Oxford University Press and Institute of Physics Publishing.

[14] Smith, A. E., Tate, D. M., 1993. Genetic Optimization Using A Penalty Function. In: Forrest, S., editor, Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 499–503. San Mateo, California. Morgan Kaufmann Publishers.

[15] T‘kindt, V., Billaut, J.-C., 2006. Multicriteria Scheduling. 2nd ed., Springer Berlin Heidelberg.

[16] Xhafa, F., Abraham, A., 2008. Metaheuristics for Scheduling in Industrial and Manufacturing Applications, Springer Berlin Heidelberg.

COMBINED APPLICATION OF OPERATIONS RESEARCH AND OPTIMAL CONTROL: A NEW VIEWPOINT ON THE PROBLEM SEMANTIC IN SUPPLY CHAIN MANAGEMENT

Dmitry Ivanov[1], Boris Sokolov[2], and Joachim Kaeschel[1]
Chemnitz University of Technology, Faculty of Economics and Business Administration
09107 Chemnitz, Germany
Dmitry.ivanov@wirtschaft.tu-chemnitz.de, www.ivanov-scm.com

Russian Academy of Science, St. Petersburg Institute of Informatics and Automation
39, 14 Linia, VO, St.Petersburg, 199178, Russia
sokol@iias.spb.su

Abstract: In operations research, remarkable advancements have been achieved in the domain of supply chain management (SCM) for the last two decades. On the other hand, in recent years, the works on SCM have been broadened to cover the whole supply chain dynamics with feedback loop control. Based on a combination of fundamental results of the modern optimal control theory with the optimization methods of operations research, an original approach to supply chain modelling is presented in this paper to answer the challenges of dynamics, uncertainty, and adaptivity in SCM. We illustrate the general framework with the help of a modelling complex that has been developed for supply chain planning and scheduling. In this paper, we show explicitly how to interconnect a static optimization linear programming model with an optimal control dynamic model. The proposed approach allows establishing a new viewpoint on SCM problem semantic. Conventional SCM problems may be considered from a different viewpoint and new problems may be revealed and formulated due to the mutual enriching of operations research and optimal control techniques.

Keywords: supply chain management, optimal control, mathematical programming, planning.

1 INTRODUCTION

The term "*supply chain management*" (SCM) was coined in the 1980-90s. Presently, SCM is considered as the most popular strategy for improving organizational competitiveness along the entire value chain in the twenty-first century [2], [4], [5], [18].

A *supply chain* (SC) is a network of organizations, flows, and processes wherein a number of various enterprises (suppliers, manufacturers, distributors, and retailers) collaborate (cooperate and coordinate) along the entire value chain to acquire raw materials, to convert these raw materials into specified final products, and to deliver these final products to customers [5].

SCM studies human decisions in relation to cross-enterprise collaboration processes to transform and use the supply chain resources in the most rational way along the entire value chain, from raw material suppliers up to customers, based on functional and structural integration, cooperation, and coordination throughout [5].

SCM's impact on the business performance can be estimated as up to 30%. From decisions on the supply chain configuration arise up to 80% of the total supply chain costs and up to 75% of the operational costs in supply chains [5]. Regarding the merit and performance of SCM, the following figures can be shown as examples. The increase in sales and reduction in costs in the value-adding chain due to supply chain management amounts to 15 to 30%. Partial effects such as inventory reductions, an increase in service level, supply chain reliability and flexibility, a reduction in transaction costs etc. amount to 10 to 60%. These effects occur due to balancing supplies along the entire value-adding chain to ensure mutual iterative balances of production and logistics processes subject to full customer satisfaction. As customer orientation, globalization, and IT advancements are still the ongoing trends in markets, the importance of SCM will become ever greater. Currently, responsiveness, agility, and flexibility shape enterprise competitiveness and demand a value chain management throughout [2], [4], [5], [18].

In operations research (OR), remarkable advancements have been achieved in the domain of supply chain management (SCM) for the last two decades. Studies [2], [5], [18] provide a systematic summary of operations research on quantitative models of the SCM, especially for inventory management, tactical planning decisions, and supply contracts.

In OR, improvements in supply chain planning and scheduling are usually algorithmic. For the last decade, considerable advancements have been achieved in this area [13]. However, a number of limitations can be observed in operations research techniques with regard to supply chain planning from the dynamics point of view. First, problems of high dimensionality are whether reduced to a simple dimensionality or heuristics are applied. Secondly, complex dynamics of real supply chain execution can not be reflected in single-class models. Thirdly, models of planning and control are not explicitly interconnected in terms of uncertainty.

In recent years, the works on supply chain management have been broadened to cover the whole supply chain dynamics. In these settings, control theory is becoming of even greater interest to researchers and practitioners. In the studies [3], [5], [7], [9], [10], [14], a number of valuable control frameworks, models, and algorithms have been developed to the SCM domain.

In the planning and scheduling of supply chains, a number of particular features should be taken into account. The processes of supply chain execution are non-stationary and non-linear. It is difficult to formalize various aspects of supply chain functioning. Besides, the supply chain models have high dimensionality. There are no strict criteria of decision-making and no *a priori* information about many supply chain parameters. The supply chain execution is always accompanied by perturbation impacts. Unlike the automatic systems, adjustment control decisions in supply chains are taken by managers and not by automats. This results in individual interests, risk perceptions, and subjective multi-criteria decision-making. Besides, the constellation of the discreteness of decision-making and continuity of flows should be taken into account.

In these settings, planning and scheduling in the supply chain environment should be considered not as a static jobs appointment to behaviourally passive machines but as dynamic procedures in accordance with an actual execution environment and active behaviour of supply chain elements [4], [5]. Hence, supply chain modelling problems are more complex than they appear in some studies.

In order to conduct research on such complex problems, a combination of different methods becomes necessary. While describing a complex system mathematically, this is almost impossible to consider only one system model [1]. As emphasized in [5], [6], [7], [18], supply chain problems are tightly interlinked with each other and have multi-dimensional characteristics that require the application of different integrated frameworks of decision-making support. Isolated application of only one solution method leads to a narrowing in problem formulation, overdue constraints and sometimes unrealistic or impracticable goals. A combined application of operations research, control theory, systems analysis, and agent-based modelling may potentially provide new insights into supply chain domain.

2 OPTIMAL CONTROL FOR SCM: CONTRIBUTIONS AND SHORTCOMINGS

The basics of the supply chain multi-disciplinary treatment were developed according to the DIMA (Decentralized Integrated Modelling Approach) methodology [6] to contribute to comprehensive supply chain modelling and to establish foundations for network theory as called for by an increasing number of researchers

The first strong contribution of the control theory to the SCM domain is the interpretation of planning and scheduling not as discrete operations but as a continuous adaptive process. Secondly, the possibility of covering the supply chain dynamics at the process level (unlike the systems dynamics) and the permanent changes in supply chain processes and environment is also a strong contribution of the control theory in the supply chain planning and scheduling domain. Finally, the control theory allows the consideration of goal-oriented formation of supply chain structures and the solution of problems in this system as a whole.

However, the control theory application also has its challenges and limitations. Conventionally, the conceptual basics of control theory lead to the fields of automatic control, signal identification, and automatic regulation. Unquestionably, the optimal control and systems analysis provide a number of advanced insights into the dynamics, stability, adaptability, non-linearity, and high-dimensionality of complex systems. However, these techniques have not been widely applied to complex business systems. There are some reasons for this.

First, the decision-making in business systems is of a discrete nature. In technical control systems, it is assumed that the control u can be selected continuously in time. The problem becomes even more complex as, though the decision-making in business systems is discrete, the processes and flows remain nevertheless continuous. Hence, the mathematical models of classical optimal control need do-main-specific modifications.

Secondly, the mathematics of optimal control also has its limitations. In the 1960–70s, significant advances were made in optimal control. However, many problems were caused by the non-linearity of operation execution processes at the algorithmic level. Usually, the right parts of the differential equations have reflected the non-linearity. To solve these problems algorithmically, step functions and the arising sectionally continuous functions have been usually applied. The fact that the derivative of the step function does not always exist has negatively influenced the further development of optimal control techniques for planning and scheduling. Thirdly, the narrow understanding of control as a regulation function has negatively influenced researchers in the application of control theory to management domains.

Summarizing, our investigations have shown that the supply chain planning and scheduling is challenged by the following problems:

- the problem of high dimensionality and non-linearity of models of SC structures and operations,
- the problem of describing uncertainty factors, and
- the problem of multi-criteria subjective decision-making.

The main motivation for this research approach is to combine the possibilities of different decision-making techniques, such as OR, control theory, and systems analysis, to achieve new quality in the decision support systems for SCM, e.g., in applying the proved fundamentals of control theory to the SCM domain, the conventional OR-based modelling techniques for SCM can be enriched by new viewpoints on dynamics, stability, adaptability, consistency, non-linearity, and high-dimensionality of the complex system.

The mathematics of the optimal control can help in revealing new conformities to natural laws that remain unrevealed within the operations research field. Hence, the conventional SCM problems may be considered from a different viewpoint and new problems that are near to the real-world settings may be revealed and formulated.

3 MULTIPLE-MODEL COMPLEXES, QUALIMETRY AND FUNCTORS

As discussed above, modelling adequacy cannot be ensured within a single model, thus multi-model complexes should be used. Each class of models can be applied to the objects for analysis of their particular aspects. The integration and combined application of various models is implemented by means of multiple model complexes [5], [6], [10], [17], [19], which are based on the application of functors [17], [19]. The coordinated application of different models improves modelling quality, as disadvantages of one model class are compensated for by advantages of other classes, which enable the examination of the determined classes of modelling tasks.

The multiple-model complexes allow problem examination and solution in different classes of models, and result representation in the desired class of models (the concept of "virtual" modelling). This becomes possible under the terms of collective application of structural–mathematical and categorical–functorial conceptions of the models' architecture. The study [19] demonstrated the capabilities of the categorical–functorial approach to qualimetry of models by an example of a polymodel description. In the work [5], we provided an example of a multiple model complex for the SCM domain.

The qualimetry is rooted in the quality science and reflects the concept of comprehensive quality, which, according to the ISO 8402-2000 international standard, means a totality of characteristics of an object that determine its capability to satisfy the established or supposed requirements. The study [19] extensively discussed the concept of qualimetry of models.

A category is a mathematical construction that consists of a class of objects and a class of morphisms (arrows). Categories provide one of the most convenient methods for describing objects within the framework of the developed qualimetry of models since, first, the conceptual level of the representation of objects in the given theory has to be invariant relative to the method of description of their internal structure, and, second, an object in a category is integral, since its internal structure is not considered. Finally, objects acquire properties only in a certain category and these properties reveal themselves only in comparison with other objects. Functors establish relationships between different categories. A one-place covariant functor can be characterized as a mapping from one category to another [15].

4 INTEGRATION OF STATIC AND DYNAMIC MODELS: CONCEPT

Let us discuss an example of how to apply the multiple model complexes. The interconnection between different models is ensured by means of *functors*. The problem of the supply chain analysis and synthesis is mostly formalized using either graph (network) models or models of *linear and integral programming*. As a rule, the problem of analysing and synthesizing programmes for supply chain execution is formalized with the help of *dynamic models*. However, the problems of coordination and consistency of the results remain open. To obtain a constructive solution to these problems, we propose to use a functorial transition from the category of digraphs that specifies the models of execution of operations to the category of dynamic models, which describes the processes of supply chain execution. The mathematical model of the above-described transition is presented in [5]

The investigations have shown that, in the framework of the considered poly-model description, not only are the functionality conditions held, but also the conditions of the general position of the adjacency mapping [17], [19]. The proposed general scheme of inter-model coordination needs detailed elaboration of the main classes of dynamic models, such as system dynamics, logical dynamics models, Petri nets, and dynamic models of supply chain operations. In the next paragraph, we will consider how explicitly to interconnect a static optimization linear programming model with an optimal control dynamic model within the supply chain planning and scheduling domain.

5 INTEGRATING STATIC AND DYNAMIC MODELS: FORMAL STATEMENT

Let A be the set of nodes of supply chain (SC). It includes the set A_0 of suppliers, the set $A_р$ of transit nodes, and the set $A_п$.of customers. We have $A_0 \cup A_р \cup A_п = A$; $A_0 \cap A_п = \varnothing$; $A_0 \cap A_р \neq \varnothing$; $A_р \cap A_п \neq \varnothing$. The structure and the parameters of SC undergo changes at discrete time points $(t_0, t_1, ..., t_k)$. These points divide the planning interval $(t_0, t_k]$ into subintervals. The SC structure does not vary at each subinterval $(t_{l-1}, t_l]$.

We assume that each network element is characterized by the following characteristics: the inventory V_j at a node A_j; the capacity φ_{jl}; the capacity ψ_{ijl} of the inter-node channel for the nodes A_i and A_j, where $A_i, A_j \in \{A_1, A_2, ..., A_n\}$. The following variables are needed for problem formalization: the amount x_{ijl} of products transmitted from A_i to A_j at time interval number l; product amounts y_{jl}, g_{jl}, z_{jl} relating to the node A_j and respectively to be stored (according to the plan of supply chain operation), to be processed, and to be delayed (as caused by limited capacity of SC). Let q_{jl}^{+} be the set of node numbers for the nodes transmitting product to A_j at time interval l and q_{jl}^{-} be the set of node numbers for the nodes receiving products from A_j at time interval l.

Now the model of SC planning can be stated as follows (1-3):

$$\alpha_1 \sum_{j=1}^{m} \sum_{l=1}^{k} g_{ij} - \alpha_2 \sum_{j=1}^{m} \sum_{l=1}^{k} z_{ij} = I \to \max , \qquad (1)$$

$$x_{jl} = \sum_{i \in q_{jl}^{(-)}} x_{ijl} - \sum_{i \in q_{jl}^{(+)}} x_{ijl} + (y_{jl} - y_{j(l-1)}) + z_{jl} + g_{jl} , \qquad (2)$$

$$0 \le x_{ijl} \le \psi_{ijl} ;\ 0 \le y_{jl} \le V_j ;\ 0 \le g_{jl} \le \varphi_{jl} ;\ z_{jl} \ge 0 . \qquad (3)$$

Constraint (2) reflects that the amount of products x_{jl} delivered to the node A_j at the time interval l can be transmitted to other nodes, or processed and storaged at this node, or delayed. The required program of SC operation should provide the maximal value for the generalized quality index I. Here the first component $I_1 = \alpha_1 \sum_{j=1}^{m} \sum_{l=1}^{k} g_{ij}$ of index (1) characterizes the total amount of customer orders processed in SC and the second component $I_2 = \alpha_2 \sum_{j=1}^{m} \sum_{l=1}^{k} z_{ij}$ characterizes the amount of not fulfiled customer orders. Both I_1 and I_2 refer to supply chain service level.

The planning problem (1)–(3) is a multi-criteria linear-programming (LP) problem of high dimensionality. It includes $m \times k$ equations (2) and $\left(\sum_{l=1}^{k} f_l + 3 \times m \times k \right)$ unknown variables, where m is the number of nodes in SC; k is the number of time intervals; f_l is the number of variables characterizing amounts of received (transmitted) products at the l-interval. An additional peculiarity of the problem is related to two-sided constraints for the variables x_{ijk} , y_{jl} , g_{jl} . A linear convolution of two indices is used to overcome the problem of goal uncertainty. Thus, we assume that $\alpha_1 + \alpha_2 = 1$, $\alpha_1, \alpha_2 \geq 0$ in expression (1).

Analysis of problem (1)-(3) showed that in spite of its high dimensionality it could be efficiently solved via special decomposition procedures and the method of successive improvement of plans for LP problems with two-sided constraints [16].

With regard to SC settings, a special consideration must be given to the fact that real SC operate heterogeneous products of different importance and priority of customer orders. We propose to concretize the obtained plan of SC operation (in the terms of model (1)-(3)) via the following deterministic dynamic model (DM) (4):

$$\dot{x}_l^{(o,1)} = u_l^{(o,1)}; \dot{x}_{ij\rho l}^{(п,1)} = u_{ij\rho l}^{(п,1)}; \dot{x}_{\tilde{j}\rho l}^{(п,2)} = u_{\tilde{j}\rho l}^{(п,2)}; \dot{x}_{ij\rho l}^{(п,3)} = u_{ij\rho l}^{(п,1)};$$
$$\dot{x}_{\tilde{j}\rho l}^{(п,4)} = u_{\tilde{j}\rho l}^{(п,4)};\ i, j, \tilde{j} \in \{1,...,m\}, \quad (4)$$

where $x_l^{(o,1)}$ is a duration of the time interval l; $x_{ij\rho l}^{(п,1)}$ is an amount of product transmitted from A_i to A_j on the time interval l and attributed to the type ρ; $x_{\tilde{j}\rho l}^{(п,2)}$ is an amount of ρ-type product processed at $A_{\tilde{j}}$ on the time interval l; $u_{ij\rho l}^{(п,1)}(t)$ is a transmission rate in the channel $A_i \to A_j$ for the product of the type ρ; $u_{\tilde{j}\rho l}^{(п,2)}(t)$ is a rate of product processing at the node $A_{\tilde{j}}$ on the time interval l for the product of the type ρ; $x_{ij\rho l}^{(п,3)}$ and $x_{\tilde{j}\rho l}^{(п,4)}$ are auxiliary variables denoting correspondingly the time passed from the beginning of product transmission (for the channel $A_i \to A_j$, the interval l, and the product of the type ρ) and the time passed from the beginning of product processing or product storage (for the node $A_{\tilde{j}}$, the interval l, and the information of the type ρ), $u_{ij\rho l}^{(п,3)}(t)$ and $u_{\tilde{j}\rho l}^{(п,4)}(t)$ are auxiliary control inputs assuming values from the set $\{0, 1\}$. The control and state variables should meet the following constraints:

$$u_l^{(o,1)}[a_{(l-1)}^{(o,1)} - x_{(l-1)}^{(o,1)}] = 0; \quad (5)$$

$$0 \leq u_{ij\rho l}^{(п,1)} \leq c_{ij\rho}^{(п,1)} u_{ij\rho l}^{(п,3)};\ 0 \leq u_{\tilde{j}\rho l}^{(п,2)} \leq c_{\tilde{j}\rho l}^{(п,2)} u_{\tilde{j}\rho l}^{(п,4)}; \quad (6)$$

$$0 \leq u_{ij\rho l}^{(п,3)} \leq u_l^{(o,1)};\ 0 \leq u_{\tilde{j}\rho l}^{(п,4)} \leq u_l^{(0,1)}; \quad (7)$$

$$\sum_{i \in q_{jl}^{(+)}} u_{ij\rho l}^{(п,3)} + u_{j\rho l}^{(п,4)} \leq 1;\ \sum_{i \in q_{jl}^{(-)}} \sum_{\rho=1}^{P} x_{ij\rho l}^{(п,1)} u_{j\rho l}^{(п,4)} \leq \sum_{l=1}^{k} y_{jl} u_l^{(o,1)}; \quad (8)$$

$$\sum_{\rho=1}^{P} u_{ij\rho l}^{(п,1)}(t) \leq d_{ijl}^{(1)}(t);\ \sum_{\rho=1}^{P} u_{\tilde{j}\rho l}^{(п,2)}(t) \leq d_{\tilde{j}l}^{(2)}(t); \quad (9)$$

$$\psi_{ijl} = \int_{t_{l-1}}^{t_l} d_{ijl}^{(1)}(\tau) d\tau;\ \varphi_{jl} = \int_{t_{l-1}}^{t_l} d_{\tilde{j}l}(\tau) d\tau; \quad (10)$$

$$x_l^{(o,1)}(t_0) = x_{ij\rho l}^{(п,1)}(t_0) = x_{\tilde{j}\rho l}^{(п,2)}(t_0) = x_{ij\rho l}^{(п,3)}(t_0) = x_{\tilde{j}\rho l}^{(п,4)}(t_0) = 0; \quad (11)$$

$$x_l^{(o,1)}(t_f) = a_l^{(o,1)},\ x_{ij\rho l}^{(п,1)}(t_f), x_{\tilde{j}\rho l}^{(п,2)}(t_f), x_{ij\rho l}^{(п,3)}(t_f), x_{\tilde{j}\rho l}^{(п,4)}(t_f) \in \mathbf{R}^1. \quad (12)$$

Constraints (5) specify an ordered sequence of time intervals of SC structure constancy, $a_{(l-1)}^{(o,1)}$ is a given duration of the $(l-1)$ interval of structure constancy. Equation (6) specifies the limits for the intensity of product transmission and processing ($c_{ij\rho l}^{(п,1)}$, $c_{\tilde{j}\rho l}^{(п,2)}$ are given constants). The first inequality in (8) specifies the rules for the operation of the node A_j, that is the ρ-type product either being transmitted to other nodes ($u_{ij\rho l}^{(\text{п},3)}(t)=1$), or being processed or stored in the same node ($u_{\tilde{j}\rho l}^{(\text{п},4)}(t)=1$). The second inequality in (8) states the limit y_{jl} for the amount of product to be stored (processed) at the node A_j on the interval l. The constraints (9) and (10) specify capacities of channels. The constraints (11) and (12) specify end conditions. The quality indices for the plans of SC operation are expressed as follows:

$$I_3=\frac{1}{2}\sum_{l=1}^{k}\left[\left(a_l^{(\text{п},1)}-\sum_{i,j,\rho}\gamma_\rho^{(1)}x_{ij\rho l}^{(\text{п},1)}(t_f)\right)^2+\left(a_l^{(\text{п},2)}-\sum_{\tilde{j},\rho}\gamma_\rho^{(2)}x_{\tilde{j}\rho l}^{(\text{п},2)}(t_f)\right)^2\right], \quad (13)$$

$$I_4=\sum_{l=1}^{k}\sum_{\rho=1}^{P}\sum_{j=1}^{m}\gamma_\rho^{(2)}x_{j\rho l}^{(\text{п},4)}(t_f), \quad (14)$$

$$I_5=\frac{1}{2}\sum_{l=1}^{k}\sum_{\tilde{j}=1}^{m}\sum_{p=\tilde{j}+1}^{m-1}\left\{\int_{t_{l-1}}^{t_l}(x_{\tilde{j}}^{(\text{п},4)}(\tau)-x_p^{(\text{п},4)}(\tau))^2d\tau+[x_{\tilde{j}\rho l}^{(\text{п},4)}(t_f)-x_{\tilde{j}\rho l}^{(\text{п},4)}(t_f)]^2\right\}, \quad (15)$$

where $a_l^{(\text{п},1)}=\sum_{i=1}^{m}\sum_{j=1 j=1}^{m}x_{ijl}$; $a_l^{(\text{п},2)}=\sum_{j=1}^{m}g_{jl}$. The index (13) expresses the completeness of heterogeneous product processing taking into account priorities of different customer orders. This performance indicator corresponds to the index I_1 of the static model. The index (14) expresses total supply delay caused by product processing, storage, and transmission in transit nodes. The given coefficients $\gamma_\rho^{(1)}$ and $\gamma_\rho^{(2)}$ establish priorities for transmission and processing of product of the type ρ. This performance indicator corresponds to the index I_2 of the static model. Indexes (13) and (14) represent the dynamic interpretation of SC service level. The index (15) lets evaluate uniformity (irregularity) of resources allocation in SC at each time (integral component of functional (15)) and at the time $t=t_f$ (terminal part of (15)).

6 IMPLEMENTATION AND EXPERIMENTAL ENVIRONMENT

Based on the above-described principles, an adaptation framework and a dynamic supply chain scheduling model have been elaborated [5], [7]. In these studies, the model presented in this paper is enriched by taking into account perturbation impacts and uncertainty. The proposed approach has the following particular features [5], [11], [12], [17]. *Firstly*, we consider planning and scheduling as *an integrated function* within an adaptive framework. *Secondly*, we formulate the planning and scheduling models as optimal control problems, taking into account the *discreteness of decision making* and the *continuity of flows*. In the model, a multi-step procedure for solving a multiple-criteria task of adaptive planning and scheduling is implemented. In doing so, at each instant of time while calculating solutions in the dynamic model with the help of the Pontryagin's maximum principle, the linear programming problems to allocate jobs to resources and integer programming problems for (re)distributing material and time resources are solved. The process control model is presented as a dynamic linear system while the non-linearity and non-stationarity is transferred to the model constraints. This allows us to ensure convexity and to use the interval constraints. As such, the constructive possibility of discrete problem solving in a continuous manner occurs.

Thirdly, the modelling procedure is based on an essential reduction of a *problem dimensionality* that is under solution at each instant of time due to connectivity decreases. As such, the problem under solution can be presented with a polynomial complexity rather than with an exponential one. In contrast, traditional exact scheduling techniques work almost with the compete list of all the operations and constraints in supply chains. In doing so, the dimensionality of problems can be reduced and discrete optimization methods of linear programming can be applied for solution within the general dynamic non-linear model.

Fourthly, for solving the problem, *Pontryagin's maximum principle* is applied. The algorithm of optimal control is based on a transformation of the optimal control problem to the boundary problem. Besides, the *Lagrange multipliers* will be presented as dynamic parameters of the conjunctive vector.

Fifthly, the *multiple objective optimization* for taking into account the individual preferences of decision makers is applied. The construction and narrowing of Pareto's sets is performed in the interaction mode providing decision maker participation. The model basis in such a case is represented by discrete models of mathematical programming, queuing model, simulation models, and development control [17].

For the experiments, we elaborated a software prototype "Supply network dynamics control" [5]. The developed prototype provides a wide range of the analysis possibilities from such points of view as supply chain structure content (initial data for scheduling) and the approachability of the tactical planning goals, taking into account individual managers' risk perception, the execution dynamics of customers' orders, and operations within these orders (including key customer orders and bottleneck operations).

The scheduling model is very flexible. More than 15 parameters can be changed to investigate different interrelations of schedule parameters and supply chain tactical goals (e.g., service level) achievement. E.g., there is an explicit possibility to change:

- The amount of resources, their intensity, and capacities;
- The amount and volumes of customers' orders and operations within these orders (including key customer orders and bottleneck operations);
- The priorities of orders, operations, resources, and goal criteria;
- The lead times, supply cycles, and penalties for breaking delivery terms;
- The perturbation influences on resources and flows in the supply chain (e.g., demand fluctuations, technological failures, purposeful threats like thefts or terrorism);

7 DISCUSSION OF FINDINGS

Analysis of (5)-(12) shows that the coordination of the planning results of the static and the dynamic model can be carried out through the variables x_{ijk}, y_{jl}, and g_{jl}. The realization of the system modeling for the considered problem showed the following advantages of the joint use of the static and dynamic planning models:

- Static models of SC operation let take into account the factors (losses, limitation of capacities) which define the state constraints in dynamic models;
- Static models provide input data for the dynamic models while the direct enumeration of variants is impracticable;
- Static models take into account, to a first approximation, the factors of distributed computation and of structure dynamics in SC and provide quantitative estimations for the amounts of supplied, processed, and delayed products.

However, the detailed time-referenced description of product distribution and processing is difficult within the static model. That is why we proposed to use dynamic model of SC operation. Dynamic description of the considered processes gives the following advantages.

- Dynamic models let establish and optimize SC performance indicators that are difficult to express within a static model (for example the indices of uniformity (irregularity) of a resources allocation for the product processing in a supply chain as a requirement on supply chain collaboration strategy; see equation (15);
- Dynamic models let use the advanced methods of the optimal control theory for a synthesis of control programs as applied to SC and its subsystems (enterprises).

The experiments with the proposed models showed that, as compared with well-known simulation-oriented approaches, they improve responsiveness and performance of SC processes. The modern optimal control theory, in combination with systems analysis, operations research and agent-oriented modelling, is a powerful technique to handle dynamics and uncertainty in supply chains.

8 CONCLUSIONS

Based on a combination of fundamental results of the modern optimal control theory with the optimization methods of operations research, an original approach to supply chain modelling has been presented in this paper to answer the challenges of dynamics and uncertainty in supply chains. We illustrated the general framework with the help of a modelling complex that has been developed for supply chain planning and scheduling.

In this paper, we showed explicitly how to interconnect a static optimization linear programming model with an optimal control dynamic model. The partial SCM problems that have been previously treated in isolation from each other are considered as an integrated domain. Such a problem formulation is near to the real-world settings. Hence, the proposed approach allows establishing a new viewpoint on SCM problem semantic. Conventional SCM problems may be considered from a new viewpoint and new problems may be revealed and formulated due to the mutual enriching of operations research and optimal control techniques.

The proposed modelling complex allows the achievement of better results in many cases in comparison with many heuristics algorithms. However, the most important point is that this approach allows the interlinking of static and dynamic models within an integrated framework. Hence, the proposed modelling complex does not exist as a “thing in itself” but works in the integrated decision-support system and guides the supply chain planning and scheduling decisions in dynamics on the principles of dynamic optimization and adaptation.

Finally, let us discuss the limitations of the proposed approach. For concrete application cases, it is nevertheless almost impossible to take into account the whole variety of complex interrelated constraints and parameters. For each concrete problem, we should actually build a new complex of constraints. This is a very time-consuming process as we should achieve the strong monosemantic problem formulation to ensure only one solution is gained. In some cases, certain model simplifications will be needed. Constructive ways to achieve calculation precision with regard to the interrelations of measured and calculated (control) parameters are to be developed further. Finally, the mathematical formalization of uncertainty factors is complicated by a high complexity of stochastic dynamic models. Hence, further developments on the modelling with interval data (e.g., on the basis of attainable sets [5]) are needed.

9 ACKNOWLEDGEMENTS

The research described in this paper is partially supported by grants from the Russian Foundation for Basic Research (grants 07-07-00169, 08-08-00403, 09-07-00066), the Saint Petersburg Scientific Center RAS, and the Department of nanotechnologies and information technologies of the Institute of Systems Analysis RAS (2.3), and the Deutsche Forschungsgemeinschaft (German Research Foundation) PAK 196 "Competence Cell-based Production Networks", subproject P6 "Automated Running of Production Networks".

References

[1] Casti, J.L. (1979). *Connectivity, complexity and catastrophe in large-scale systems*. Wiley-Interscience, New York and London
[2] Chandra, C., Grabis, J. (2007). *Supply chain configuration*. Springer, New York.
[3] Disney, S.M, Towill, D.R, Warburton, R.D.H. (2006). On the equivalence of control theoretic, differential, and difference equation approaches to modeling supply chains. *International Journal of Production Economics,* 101, 194-208.
[4] Dolgui, A., Proth, J.M. (2009). *Supply Chains Engineering: useful methods and techniques*. Springer, Berlin.
[5] Ivanov, D, Sokolov, B. (2009). *Adaptive Supply Chain Management*. Springer, UK, in press.
[6] Ivanov, D. (2009a). DIMA – A Research Methodology for Comprehensive Multi Disciplinary Modelling of Production and Logistics Networks. *International Journal of Production Research,* 47(5), 1133-1155.
[7] Ivanov, D., Sokolov, B. (2009b). Dynamic Supply Chain Scheduling. *Journal of Scheduli*ng, under review.
[8] Ivanov, D. (2009c). Adaptive aligning of planning decisions on supply chain strategy, design, tactics, and operations. *International Journal of Production Research*, DOI: 10.1080/002075409028935417, in press.
[9] Ivanov, D., Arkhipov, A., Sokolov, B. (2007). Intelligent planning and control of manufacturing supply chains in virtual enterprises. *International Journal of Manufacturing Technology and Management*, 11(2), 209-227.
[10] Ivanov, D., Sokolov, B., Kaeschel, J. (2009). A multi-structural framework for adaptive supply chain planning and operations with structure dynamics considerations. *European Journal of Operational Research*, doi:10.1016/j.ejor.2009.01.002, in press.
[11] Kalinin, V.N., Sokolov, B.V. (1985). Optimal planning of the process of interaction of moving operating objects. *International Journal of Difference Equations*, 21(5), 502-506.
[12] Kalinin, V.N., Sokolov, B.V. (1996). Multi-model description of control processes for airspace crafts. *Journal of Computer and System Sciences International*, 6, 192-199.
[13] Kreipl, S., Pinedo, M. (2004). Planning and Scheduling in Supply Chains: An Overview of Issues in Practice. *Production and Operations Management*, 13(1), 77-92.
[14] Lalwani, C.S., Disney, S., Towill, D.R. (2006).Controllable, observable and stable state space representations of a generalized order-up-to policy. *International Journal of Production Economics*, 101, 172-184.
[15] Mesarovic, M.D., Takahara, Y. (1975). *General systems theory: mathematical foundations*. Academic Press, New York, Can Francisco, London.
[16] Moskvin, BV. (1987) Optimization of information transmission in packet network // Trans. of All-Union conf. "Compac 87", Riga, pp 168-171.
[17] Okhtilev, M., Sokolov, B., Yusupov, R. (2006). *Intelligent technologies of complex systems monitoring and structure dynamics control*. Nauka, Moskau (in Russian).
[18] Simchi-Levi, D., Wu, S.D. and Zuo-Yun, S. (2004). *Handbook of quantitative supply chain analysis.* Springer, New York.
[19] Sokolov, B.V, Yusupov. R.M. (2004). Conceptual foundations of quality estimation and analysis for models and multiple-model systems. Journal of Computer and Systems Sciences International, 6, 5–16.

HANDLING OF A NON-LINEAR MODEL OF THE EVACUATION PLAN DESIGN BY IP-SOLVER

Jaroslav Janáček
University of Žilina, Faculty of Management and Informatics,
Univerzitná 1, 010 26 Žilina, Slovak Republic,
jaroslav.janacek@fri.uniza.sk

Abstract: When an evacuation plan is designed, vehicles are assigned to evacuated municipalities so that the entire population is evacuated in minimal time. This evokes an idea to design a model of the problem and then to solve it by an IP-solver. Nevertheless an attempt for completing of a model ends by obtaining of a non-linear one and a succeeding attempt for its linearization causes an extreme increase of the number of variables. This paper deals with an approach to the linearization with small number of variables. Furthermore, a special iterative process is presented here to find an optimal solution in a short time.

Keywords: evacuation plan, vehicle assignment, min-max problem, integer programming, IP-solver

1 INTRODUCTION

Let population of a given set of municipalities is endangered by some threat. The casualties can be avoided by evacuation of the endangered population to some refuges, which are predestinated for each evacuated place in advance. To perform the evacuation, some available vehicles are disposable at several places located in the neighborhoods of the endangered municipalities. It is necessary to determine a route for each used vehicle so that the population is evacuated from its original places to the predetermined refuges. The evacuation should be performed so that the time of evacuation is as short as possible. The end of evacuation is given by the time, when the last inhabitant reaches its predestinated refuge.

If an arbitrary vehicle is used to provide an evacuated community with this service, the route of vehicle may take a prescribed form. The route starts at an original vehicle location, continues to the served municipality, picks up a portion of the evacuated population and takes them to the predetermined refuge [5]. If necessary, the vehicle may return back to the evacuated place and save another portion of population by taking them to the same refuge. This cycle can be repeated several times.

Even if this unique form of route is assumed, the determination of a time optimal assignment of the vehicles to the evacuated places represents a hard combinatorial problem, the solution of which must be usually found in a short time of several minutes. In this paper, we present two different approaches to this problem. Each of these approaches enables to employ a commercial IP-solver, to obtain a final concrete set of decisions on the vehicle assignment. These approaches consist of a linear programming model formulation and a solving process performed by commercial software with usage of its particular characteristics.

To formulate the following mathematical models for the individual approaches, we shall use a general notation, where symbol I denotes a set of all considered homogenous fleets of vehicles. Each homogenous fleet $i \in I$ is characterized by a number N_i of vehicles and by a vehicle capacity K_i. We shall assume that the fleet i is located at a node $u(i)$ of a road network covering the serviced area. The endangered municipalities form a set J and each municipality $j \in J$ is described by a number b_j of its population and by a road network node $v(j)$, where the municipality j is located. We assume for the purpose of evacuation that a refuge $w(j)$ is assigned to each municipality $j \in J$. Furthermore, let t_{ij} denote the time, which is necessary for a vehicle from the fleet i to traverse the distance between the nodes $u(i)$ and $v(j)$. In addition, let s_j denote the time necessary for traversing the distance between the nodes $v(j)$ and $w(j)$. Using this notation, we can introduce two approaches to the evacuation plan design problem.

The first approach consists of a non-linear problem formulation and a subsequent problem linearization so that the number of auxiliary newly introduced variables is as small as possible. The linearized problem is then solved on the whole by a commercial IP-solver, which makes use of branch-and-bound method.

The second approach avoids the problem linearization by a special iterative process, which successively solves linear problems searching for the first feasible solution in each step.

Advantages and disadvantages of both approaches were studied and some results of the numerical experiments are presented in the concluding part of this paper to demonstrate efficiency of both approaches, in the case, when a commercial software tool is used for obtaining final decisions on the evacuation plan.

2 EVACUATION PROBLEM

The objective is to determine a route of each vehicle so that all population is evacuated and the longest route is minimal [4]. Coming out of the assumption on the possible form of a vehicle, we introduce the variable $z_{ij} \in \{0,1\}$ defined for each pair of fleet i and dwelling place j. This variable takes the value of 1 if and only if the positive number of vehicles of fleet i is assigned to the municipality j.

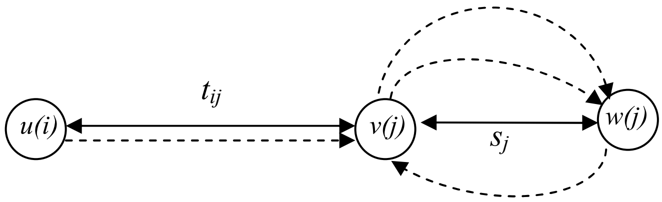


*Figure 1 :*A route with two visits at *v(j)*

Taking into account that the given part of fleet i can visit an evacuated place several times, we denote the number of visits of this part at place j by variable $x_{ij} \in Z^+$. Now, when the part of fleet i is assigned to the municipality j, its travelling time (see Fig. 1) is equal to:

$$t_{ij} + s_j + 2s_j(x_{ij} - 1) \tag{1}$$

The variable $T \geq 0$ denotes an upper bound of all route times.

We can take into account that a maximal sensible time T^{max} of the evacuation can be given. In this case, the maximal possible number of visits of fleet i vehicles at the place j can be evaluated to comply with this limit. The inequality (2) must hold for the x_{ij}.

$$t_{ij} + s_j + 2s_j(x_{ij} - 1) \leq T^{\max} \tag{2}$$

From the associated inequality (2) it follows that:

$$x_{ij} \leq \left\lfloor \frac{T^{\max} - t_{ij} + s_j}{2s_j} \right\rfloor \tag{3}$$

and then the value $P_{ij}(T^{max})$ can be defined as the value of right-hand-side of inequality (3). If $P_{ij}(T^{max}) \leq 0$, then the fleet i is not able to service the place j. To minimize the size of a built model, we introduce a set $J(i)$ of all the places from J, for which inequality $P_{ij}(T^{max}) > 0$ holds for the fleet i. Similarly, we define a set $I(j)$ of all the fleets from I, for which

$P_{ij}(T^{max}) > 0$ holds for the place j. In addition to these variables, we introduce variables $q_{ij} \in Z^+$, which denote number of vehicles of the fleet i assigned to the municipality j. Taking into account expressions (1) and (3), a model of evacuation plan design can be written as follows:

$$Minimize\ T \tag{4}$$

$$Subject\ to\quad (t_{ij} - s_j)z_{ij} + 2s_j x_{ij} \le T \quad for\ i \in I,\ j \in J(i) \tag{5}$$

$$x_{ij} \le P_{ij}(T^{\max})z_{ij} \quad for\ i \in I,\ j \in J(i) \tag{6}$$

$$q_{ij} \le N_i z_{ij} \quad for\ i \in I,\ j \in J(i) \tag{7}$$

$$\sum_{j \in J(i)} q_{ij} \le N_i \quad for\ i \in I \tag{8}$$

$$\sum_{i \in I(j)} K_i q_{ij} x_{ij} \ge b_j \quad for\ j \in J \tag{9}$$

$$z_{ij} \in \{0,1\} \quad for\ i \in I,\ j \in J(i) \tag{10}$$

$$x_{ij} \in Z^+ \quad for\ i \in I,\ j \in J(i) \tag{11}$$

$$T \ge 0 \tag{12}$$

$$q_{ij} \in Z^+ \quad for\ i \in I,\ j \in J(i) \tag{13}$$

The constraints (5) assure that the travelling time of each part of the fleet is less or equal to the upper bound T. The constraints (6) represent binding constraints between the variables x_{ij} and z_{ij}. These constraints cause that if the variable z_{ij} is equal to zero, then the variable x_{ij} is also zero, which means that if the fleet i is not assigned to the place j, then the fleet cannot visit this place at all. The constraints (7) represent binding constraints between the variables q_{ij} and z_{ij}. These constraints cause that if the variable z_{ij} is equal to zero, then the variable q_{ij} is also zero, which means that if the fleet i is not assigned to the place j, then no vehicle of the fleet i can visit this place. The constraints (8) assure that the total number of designated vehicles of the fleet i does not exceed the number N_i.

The constraints (9) ensure that each dwelling place j is provided with a sufficient capacity, which enables to evacuate all the population of size b_j. These constraints are nonlinear and thus this model cannot be input to an IP-solver. An approach to the model linearization follows.

To rewrite the model into a linear form, we make use of the fact that the variable x_{ij} may take only one of several few values from the range of 0, 1, …, P_{ij}. We introduce auxiliary variable m_{ij}, which serves as a lower bound of the number of visits performed by the vehicles of fleet i at the dwelling place j. Then the following constrains must hold.

$$q_{ij} x_{ij} \ge m_{ij} \quad for\ i \in I,\ j \in J(i) \tag{14}$$

Now the series (9) of constraints can be replaced by the linear constraints (15).

$$\sum_{i \in I(j)} K_i m_{ij} \ge b_j \quad for\ j \in J \tag{15}$$

The constraint $q_{ij}x_{ij} \ge m_{ij}$ can be replaced by the following system of logical constraints: If $x_{ij}=k$ then $kq_{ij} \ge m_{ij}$ for $k = 0, 1, \ldots, (P_{ij}\text{-}1)$.

It holds that if some of the logical constraints is fulfilled for $k=p$, then the constraints are fulfilled for each $k>p$ and, vice versa, if a constraint is not fulfilled for $k=p$, then they cannot be fulfilled for any $k<p$.

Now we introduce variables $y^p_{ij} \in \{0, 1\}$ for $i \in I$ and $j \in J(i)$ and $p=0, 1, \ldots, P_{ij}-1$ and the system of logical constraints can be replaced by the series (24) of constraints.

$$N_i P_{ij} y^p_{ij} + pq_{ij} \geq m_{ij} \qquad \text{for } i \in I,\ j \in J(i),\ p = 0, \ldots, P_{ij} - 1 \tag{16}$$

If the system (16) holds for given values of the variables q_{ij} and k is the minimal value of subscript p for which $y^k_{ij}=0$, then the system must also hold for such a setting of variables y^p_{ij}, where $y^p_{ij}=1$ for $p=0, \ldots, k-1$ and $y^p_{ij}=0$ for $p=k, \ldots, P_{ij}-1$, i.e. the setting of variables y^p_{ij} fulfils the equation (17).

$$\sum_{p=0}^{P_{ij}-1} y^p_{ij} = k \qquad \text{for } i \in I,\ j \in J(i) \tag{17}$$

Then the following linear model describes the evacuation plan design problem:

$$\text{Minimize } T \tag{18}$$

$$\text{Subject to} \quad (t_{ij} - s_j) z_{ij} + 2 s_j \sum_{p=0}^{P_{ij}(T^{max})-1} y^p_{ij} \leq T \quad \text{for } i \in I,\ j \in J(i) \tag{19}$$

$$\sum_{p=0}^{P_{ij}(T^{max})-1} y^p_{ij} \leq P_{ij}(T^{max}) z_{ij} \qquad \text{for } i \in I,\ j \in J(i) \tag{20}$$

$$q_{ij} \leq N_i z_{ij} \qquad \text{for } i \in I,\ j \in J(i) \tag{21}$$

$$\sum_{j \in J(i)} q_{ij} \leq N_i \qquad \text{for } i \in I \tag{22}$$

$$\sum_{i \in I(j)} K_i m_{ij} \geq b_j \qquad \text{for } j \in J \tag{23}$$

$$N_i P_{ij}(T^{max}) \geq m_{ij} \qquad \text{for } i \in I,\ j \in J(i) \tag{24}$$

$$N_i P_{ij}(T^{\max}) y^p_{ij} + pq_{ij} \geq m_{ij} \qquad \text{for } i \in I,\ j \in J(i),\ p = 0, \ldots, P_{ij}(T^{\max}) - 1 \tag{25}$$

$$z_{ij} \in \{0,1\} \qquad \text{for } i \in I,\ j \in J(i) \tag{26}$$

$$T \geq 0 \tag{27}$$

$$m_{ij} \geq 0 \qquad \text{for } i \in I,\ j \in J(i) \tag{28}$$

$$y^p_{ij} \in \{0,1\} \qquad \text{for } i \in I,\ j \in J(i),\ p = 0, \ldots, P_{ij}(T^{\max}) - 1 \tag{29}$$

The constraints (19) mean the same as the constraints (5). These constraints were derived from the constraints (5) by substitution of the left-hand-side of the equality (17) for x_{ij}. The constraints (20) are the binding constraints, which assure relations between the variables z_{ij} and the sum of y^p_{ij}, which corresponds with the number x_{ij} of visits of the part of fleet i at the municipality j. The constraints (19) are equivalent to the constraints (7), which assure that if no part of the fleet i is designated to the place j ($z_{ij}=0$), then the number q_{ij} of vehicles is equal to zero. The constraints (22) assure that the total number of designated vehicles of the fleet i does not exceed the number N_i.

As m_{ij} represents a lower bound of the product $x_{ij}q_{ij}$, which is the number of visits of vehicles of the fleet i at the place j, then the constraints (23) ensure that population of the place j can be evacuated. The constraints (24) ensure that the estimation m_{ij} does not exceed the upper bound $N_iP_{ij}(T^{max})$ of visits and the constraints (25) ensure that the sum of y^p_{ij} over p corresponds with the number of necessary visits of group of q_{ij} vehicles of the fleet i at j.

Even if the above model is linear, the number of its variables is extremely high. Only the number of variables y^p_{ij} is $|I| * |J| * P$, where P denotes the average of $P_{ij}(T^{max})$. That is the reason, for which we try to decrease the number of these variables by a better construction of the constraints (25). We realize that the sum of the variables y^p_{ij} over the subscript p substitutes an integer number x_{ij}. This substitution can be done by a more thrifty way using the binary code of a nonnegative integer number. If a new series of binary variables $\underline{y}_{ijr}$ is introduced, then an integer number x_{ij} from 0 to $P_{ij}(T^{max})$ can be written as:

$$x_{ij} = \sum_{r=0}^{R_{ij}-1} 2^r \underline{y}_{ijr}\,, \tag{30}$$

where

$$R_{ij} = \left\lceil \log_2 (P_{ij}(T^{\max}) + 1) \right\rceil. \tag{31}$$

This way, if $P_{ij}(T^{max})$ is 63, then the original linearized model (18)-(29) would involve 64 variables y^p_{ij} to substitute the number x_{ij}. In the later case, only 8 variables $\underline{y}_{ijr}$ are necessary.

Nevertheless, it is possible to use this substitution only in those constraints, where x_{ij} is used. This is not the case of the constraints (25), where the variables y^p_{ij} are used individually. The construction of the constraints (25) is based on the fact that the lowest subscript p, for which the associated variable y^p_{ij} is not forced to take the value of 1, corresponds with the lowest value of x_{ij}, for which the constraint (14) holds. We try to find a series of linear expressions Y^p_{ij} of the variables y^p_{ij} for p=0, ..., $P_{ij}(T^{max})$-1 so that the same effect is achieved. It means that if p is the lowest subscript for which Y^p_{ij} is not forced to take the value greater than zero, then the constraint (14) holds for corresponding x_{ij}. Furthermore, the equality $Y^p_{ij}=0$ must not prevent the expressions Y^k_{ij} for $k<p$ from taking a value greater than 0.

Let us denote the linear expressions as:

$$Y^p_{ij} = \sum_{r=0}^{R_{ij}-1} a_{pr} \underline{y}_{ijr}\,. \tag{32}$$

Assuming that the sufficient coefficients a_{pr} are found, the resulting model can take the following linear form:

$$\mathit{Minimize}\ T \tag{33}$$

$$\mathit{Subject\ to}\quad (t_{ij} - s_j) z_{ij} + 2 s_j \sum_{r=0}^{R_{ij}-1} 2^r \underline{y}_{ijr} \le T \quad \mathit{for}\ i \in I,\ j \in J(i) \tag{34}$$

$$\sum_{r=0}^{R_{ij}-1} 2^r \underline{y}_{ijr} \le P_{ij}(T^{\max}) z_{ij} \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{35}$$

$$q_{ij} \le N_i z_{ij} \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{36}$$

$$\sum_{j \in J(i)} q_{ij} \le N_i \qquad \mathit{for}\ i \in I \tag{37}$$

$$\sum_{i \in I(j)} K_i m_{ij} \ge b_j \qquad \mathit{for}\ j \in J \tag{38}$$

$$N_i P_{ij}(T^{max}) \ge m_{ij} \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{39}$$

$$N_i P_{ij}(T^{\max}) \sum_{r=0}^{R_{ij}-1} a_{pr} \underline{y}_{ijr} + p q_{ij} \ge m_{ij} \qquad \mathit{for}\ i \in I,\ j \in J(i),\ \ p = 0, ..., P_{ij}(T^{\max}) - 1 \tag{40}$$

$$z_{ij} \in \{0,1\} \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{41}$$

$$T \ge 0 \tag{42}$$

$$q_{ij} \in Z^+ \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{43}$$

$$m_{ij} \ge 0 \qquad \mathit{for}\ i \in I,\ j \in J(i) \tag{43}$$

$$\underline{y}_{ijr} \in \{0,1\} \qquad \mathit{for}\ i \in I,\ j \in J(i),\ r = 0, ...,\ R_{ij} - 1 \tag{44}$$

A question arises, how to construct the linear forms (32)? Let us consider the integers p=0, ..., $P_{ij}(T^{max})$-1. Each such integer can be written uniquely in the binary code in accordance to

$$p = \sum_{r=0}^{R_{ij}-1} 2^r b_{pr} , \tag{45}$$

where $b_{pr} \in \{0, 1\}$ and R_{ij} is given by (31). Let us define a_{pr}=1-b_{pr} for p=0, ..., $P_{ij}(T^{max})$-1 and r=0, ..., R_{ij} -1.

It can be proved that Y^p_{ij} satisfy our demands. i.e. if p is the minimal integer for which $pq_{ij} \geq m_{ij}$ holds, then for such y_{ijr}, for which the right-hand-side of (30) is equal to p, must hold $Y^p_{ij} = 0$ and $Y^k_{ij} \geq 1$, for each $k \leq p$-1. The first equality follows from the definition of a_{pr} and from the fact that the variables y_{ijr} take the values of b_{pr}. As concerned $k \leq p$-1, there must exist a subscript s of the binary codes of p and k that b_{ps}=1 and b_{ks}=0. Then $a_{ks}.y_{ijs} =(1-b_{ks}).b_{ps}$=1 and it follows that $Y^k_{ij} \geq 1$.

3 AN ITERATIVE APPROACH

Even if the non-linearity of the original model was successfully mastered, the preliminary numerical experiments showed that a common branch-and-bound method needs a very long computational time to find an optimal or near-to-optimal solution of some problem instances. That is why, we have suggested an iterative approach, which resembles the Tanako-Asai's method used in the fuzzy optimization [6].

The approach is based on a procedure, which searches only for a feasible solution of the evacuation problem formulated for a fixed value of T^{max}. If the feasible solution is found, then the value of T^{max} is decreased, the associated model of the problem is reformulated and the searching process is repeated. In the opposite case, there is found that no feasible solution exists and this way a lower bound of the optimal time of evacuation is obtained. By a subsequent searching for feasible solutions for the increased or decreased values of T^{max} the optimal time of evacuation can be estimated with an arbitrary precision.

For the purpose of the above-mentioned searching process, the model (4)-(13) of the evacuation problem was reformulated this way:

$P_{ij}(T^{max})$ obtained from the right-hand-side of (3) was substituted for x_{ij} in the model. Due to this substitution, the constraints (5), (6) and (7) become redundant. Furthermore, it can be noticed that after this substitution, the objective function takes the value of T^{max} and that is why it can be replaced by some other expression, e.g. the total time spent by all vehicles on their routes.

After this rearrangement the model (4)-(13) can be reformulated as:

$$Minimize \quad \sum_{i \in I} \sum_{j \in J(i)} (t_{ij} - s_j + 2 s_j P_{ij}(T^{max}))q_{ij} \tag{46}$$

$$Subject \;\; to \quad \sum_{j \in J(i)} q_{ij} \leq N_i \qquad for\; i \in I \tag{47}$$

$$\sum_{i \in I(j)} K_i P_{ij}(T^{max}) q_{ij} \geq b_j \quad for\; j \in J \tag{48}$$

$$q_{ij} \in Z^+ \quad for\; i \in I,\; j \in J(i) \tag{49}$$

The constraints (47) assure that the total number of the designated vehicles of the fleet i does not exceed the number N_i and the constraints (48) ensure that demands of all evacuated municipalities are covered by the assigned vehicles.

4 NUMERICAL EXPERIMENTS

To be able to perform the computation, we used the general optimisation software environment XPRESS-IVE [7], [8] for our study. This software system includes a branch-and-cut method and it also enables exploitation of the premature stopping rules. The software is equipped with the programming language *Mosel*, which can be used for both input of model and writing of input and output procedures. To avoid the situation, when demand of the computational process on free memory exceeds a disposable limit and the searching process fails, the searching process can be equipped with premature stopping procedures, which terminate the searching process whenever some stopping rule is met and output the best-found solution. The stopping rules may terminate the searching process either if a time limit is exceeded or if the best-found solution quality differs from each relevant lower bound by less than a given percentage of the lowest bound.

To verify the suggested approaches employing of the IP-solver, we formulated ten different instances of the problem. One of the instances denoted as "Hradza" comes out of a possible emergency situation, which can occur if the dam Liptovska Mara breaks. Then, under given assumptions, 26 communities would have to be evacuated to 26 predestined places.

For this evacuation, 411 vehicles of different capacities located at three bigger towns of the area are available. This set of vehicles was partitioned into fleets so that all vehicles of one fleet have the same capacity and location. The other instances were formulated in nine areas of the Slovak Republic in the similar way. These instances are denoted by names of the biggest towns of the areas. The numbers and capacities of available vehicles were generated similarly to the first instance for the vehicles to be able to satisfy the demand on evacuation. These benchmarks were used to verify the suggested method, which consists in particular model building and employing the general IP-solver for obtaining of a good solution of the problem.

We used the possibility of the solver, which enables premature termination of the searching process whenever a fixed time limit is exceeded. The experiments were performed on a personal computer equipped with Intel Core 2 6700 with parameters: 2.66 GHz and 3 GB RAM.

To find characteristics of the first approach involving the linearized model (33)-(44) of evacuation by the divisible fleets, we solved each of these instances. The best results obtained in computational time of 40 minutes are presented in Table 1, where "No. of F" denotes the considered number of fleets, "No. of EP" denotes the number of evacuated dwelling places, "Rows" denotes the number of structural constraints of the model and "Columns" denotes the number of used variables including the auxiliary ones, which are automatically introduced by the solver. In the row denoted as "Tmax", there are reported the predetermined values of T^{max} in minutes. The symbol "Tbest" denotes the row, where the best-found time of evacuation is plotted. The label "Ctime" denotes the computation time, which was consummated by the respective computational process to obtain the associated solution.

To compare the both approaches, the same instances were solved by the iterative approach and the results are reported in Table 2.

Table 1: Results of numerical experiments for divisible fleets obtained in 40 minutes of computational time

Instances:	Brati-slava	Dubnica nad Váhom	Hrádza	Košice	Liptovský Mikuláš	Leo-poldov	Micha-lovce	Nitra	Nové Zámky	Púchov
No. of F.	14	10	16	13	9	11	10	10	9	12
No. of EP.	25	25	26	25	25	25	25	25	25	25
Rows	2083	3449	5000	1508	3555	1135	4185	1997	3476	2323
Columns	1457	1517	2564	1243	1412	965	1581	1289	1435	1543
Tmax [min]	162	938	202	184	668	140	1420	338	1003	278
Ctime [s]	1.1	2449.3	1232.6	0.9	2439.9	1212	2433	2425.3	2442.1	3
Tbest [min]	**81**	**368**	**81**	**92**	**245**	**93**	**486**	**168**	**613**	**139**

Table 2: Results of numerical experiments for divisible fleets obtained by the iterative approach

Instances:	Brati-slava	Dubnica nad Váhom	Hrádza	Košice	Liptovský Mikuláš	Leo-poldov	Micha-lovce	Nitra	Nové Zámky	Púchov
No. of F.	14	10	16	13	9	11	10	10	9	12
No. of EP.	25	25	26	25	25	25	25	25	25	25
Rows	39	35	42	38	34	36	35	35	34	37
Columns	273	250	215	240	223	275	249	250	225	246
Tmax [min]	100	370	200	100	250	100	500	200	630	150
Ctime [s]	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.1	1.8	0.0
Tbest [min]	**81**	**360**	**81**	**92**	**236**	**93**	**469**	**167**	**568**	**139**

5 COMPARISON AND CONCLUSIONS

We suggested two approaches to the evacuation plan design problem. These approaches come out from a nonlinear model of the evacuation problem. The first approach is based on linearization of the problem, when a special construction was used to minimize the number of involved auxiliary variables. The second approach comes out from an iterative concept, when each step of the iterative process is based on searching of a feasible solution of the problem for a fixed value of the objective function. It can be found from the above-presented results that the second approach is much more robust and reaches better solutions in a much shorter computational time than the first approach. This phenomenon follows from a much smaller size of the problems, which are solved in the individual steps of the second approach. The relative timesavings are so big that they enable to perform whole sequence of the steps in the time, which is much shorter than the time necessary for completion of the branch-and-bound search of the first approach. This way we have derived a convenient approach, which enables to use a common IP-solver for solving of the evacuation plan design.

References

[1] Janáček, J., 2004. Service System Design in the Public and Private Sectors. In: Proceedings of the international conference Quantitative Methods in Economics (Multiple Criteria Decision Making XII), , Virt, pp. 101-108.

[2] Janáček, J., 2006. Time Accessibility and Public Service Systems, Proc. of conference Quantitative Methods in Economics, Bratislava, pp. 57-63.

[3] Janáček, J., Sibila, M., 2009. Optimal Evacuation Plan Design with IP-Solver. Communications –Scientific Letters of the University of Žilina, Vol. 11, No 3, pp 29-35

[4] Sibila, M., 2009. Management and Transport of Inhabitant Evacuation (in Slovak), [Dissertation thesis], Zilinská univerzita v Ziline, Fakulta specialneho inzinierstva, Katedra technickych vied a informatiky, , p. 136.

[5] Teichmann, D., 2008. Contribution to the Problems of Inhabitant Evacuation and Usage of Mathematical Programming for the Evacuation Planning (in Czech), Proc. Krízový management, Vol. 7, 2/2008.

[6] Teodorovič, D., Vukadinonič, K., 1998. Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach. Boston: Kluwer Academic Publishers, 387 p

[7] XPRESS-MP Manual "Getting Started", 2005. Dash Associates, Blisworth, UK, p. 105.

[8] XPRESS-Mosel "User guide", 2005. Dash Associates, Blisworth, UK, p. 99.

Acknowledgment: This work was supported by project "Centre of excellence for systems and services of intelligent transport" ITMS 26220120028

Project Part-Financed by the European Union

European Regional Development Fund

Operačný program VÝSKUM a VÝVOJ

DIRECT USE OF FUNCTIONAL GRADIENTS AND LINEAR PROGRAMMING IN OPTIMAL CONTROL

Lado Lenart[1] **, Jan Babič**[1] and **Janez Kušar**[2]
[1] 'Jožef Stefan' Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
[2] Faculty of Mechanical Eng.,Aškrčeva 2. SI-1000 Ljubljana, Slovenia
{lado.lenart@ijs.si}

Abstract: In the paper techniques for writing simple, transparent and effective programs for solving some problems in optimal control are presented. The calculation of functional gradients is the background for the fixed-time problems. The terminal function can be constant or else it has to be minimized. The direct algorithm is based on Euler discretization of a dynamic system and object functional. The discretized system is linearized. Then the best solution is searched using linear programming iteratively. The advantage of this algorithm is that the constraints in the control and state variables reduce the computational complexity of a non constraint problem. Typical examples illustrate the use of the described formalism. All solutions are validated with other methods.

Keywords: optimal control, functional gradients, dynamic systems, linearization, linear programming

1 INTRODUCTION

When describing methods for solving optimal control problems (OCP), a technique is often classified as either direct or indirect [1]. A direct method constructs a sequence of control trajectories $u_1, u_2, \ldots, u^*$ such that the objective function V is minimized and typically $V(x_1) > V(x_2) > \ldots > V(x^*)$. An indirect method attempts to find a root of the necessary condition $\nabla V(x) = 0$. Thus, for an indirect method, it is necessary to explicitly derive the adjoint equations, the control equations and all the transversality conditions. For the practitioner there are three main reasons to avoid indirect methods. Firstly, base knowledge of the control theory is required. Secondly, if the problem includes path constraints, it is necessary to make an a priori estimate of the constrained arc sequence. And thirdly, when the basic method is not robust, the user must guess the initial values for the adjoint variables, which are not physical quantities and therefore the intuition is of no use. Even when the optimal solution is known, derivation is needed to get the adjoint variables. The positive aspects of indirect methods are the possibility to theoretically prove the convergence of a discrete problem to the underlying continuous model and to cope with the control and parameter perturbation.

The most elementary techniques for the direct methods that are treated here allow the reader to write the optimal control programs by himself, using only limited number of external functions if any, with preference those from MATLAB environment. Some typical examples are elaborated.

The layout of the paper is as follows. In the second section the functional gradients $\partial V/\partial u$ are presented as the most important term in a descent optimization. Relatively simple algorithms with empty target set are combined with fixed time, fixed endpoint system.

In the section three the linearization [2, 3, 4] of nonlinear control systems is shown to be the queue for relatively straightforward processing of optimal control tasks. All algorithms are tested within the examples. The most illustrative is the last example with the linearization of a double link manipulator. The dynamic model is collocated using Euler- type discretization. The obtained system is linearized and herewith the problem is reduced to the linear programming (LP) task. Afterwards the steepest descent algorithm is used to solve

OCP iteratively. All test solutions are compared with the reference solutions, obtained by the analytical methods or using MATLAB solvers.

2 FUNCTIONAL GRADIENTS IN OPTIMAL CONTROL PROBLEMS

OCP are mostly formulated in the Bolza form :

$$V = \min_u \{ \ \Phi\big(x(T)\big) + \int_0^T F\big(x(t),u(t)\big)dt \ \} \ , \quad \text{s.t.:} \quad \dot{x} = f\big(x(t),u(t)\big); \quad x(0)=0 \tag{2.1}$$

In (2.1) V is performance index, x is the state variable, T is the terminal time, $\Phi\big(x(T)\big)$ is the terminal profit (in economics) and *u(t)* is the control function.. F is the utility function. The integral term in (2.1) can be omitted without the reduction of the generality as the problem (2.1) can be transformed into (2.2) while introducing a new variable $\dot{z} = F\big(x(t),u(t)\big)$. Then (2.1) is written as:

$$\begin{aligned} &V = \min_u \{\Phi(x(T)) + z(T)\}; \text{ s.t.} \\ &\dot{x} = f\big(x(t),u(t)\big); \quad \dot{z} = F\big(x(t),u(t)\big); \quad x(0) = x_0; z(0) = 0; \end{aligned} \tag{2.2}$$

OCP then is solved as problem (2.11). We continue here with problem (2.1). In the next, integral limits $0,T$ shall be omitted in the formulas if they are not necessary. Now the leading idea will be, that the control u is disturbed with the variation $\delta u(.)$, inducing herewith changes in $x(.)$ and $V(.)$ Let $\Phi\big(x(T)\big)$ in (2.1) be zero, as this does not affect the generality in gradient calculation. The resulting expression for variation in V then is:

$$\delta V\big[\delta u(.)\big] = \int F_x \delta x \, dt + \int F_u \delta_u \, dt \tag{2.3}$$

Let the boundary conditions for ODE in (2.1) be written as $G(x)=0$. Then the variation in the dynamic system ODE in (2.1) results in:

$$\frac{d\delta x}{dt} = f_x(t)\delta x + f_u(t)\delta u; \quad \text{RW: } G_x \delta x = 0 \tag{2.4}$$

To get the gradient $\partial V/\partial u$ one has to take δx in (2.4) and put it into (2.3). The way to make this correctly is to apply a well known theorem for linear operators [5] on the arbitrary adjoint function λ, the known function ∂x and known operator f_x with his adjoint f_x^*, in (2.5).

$$\big(\lambda, f_x \delta x\big) - \big(f_x^* \lambda, \delta x\big) = 0 \tag{2.5}$$

The next is to use the Lagrangian identity in the form:

$$\int \left\{ \lambda\left(\frac{d\delta x}{dt} - f_x \delta x \right) + \delta x \left(\frac{d\lambda}{dt} + f_x^* \lambda \right) \right\} dt = \lambda \, \delta x \, |_0^T \tag{2.6}$$

(2.6) is easy to verify. On the left side, by (2.5) the terms with f_x, f_x^* vanish because scalar product is commutative with real factors. The left side in (2.6) is then :

$$\int \Big(\ \lambda \frac{d\delta x}{dt} + \delta x \frac{d\lambda}{dt} \ \Big) dt = \int \frac{d}{dt}\big(\lambda \, \delta x\big) dt = \lambda \, \delta x \, |_0^T \, , \tag{2.7}$$

as it was to prove. Now the arbitrary function λ is concretized to be the solution of (2.8):

$$\dot{\lambda} + f_x^* \lambda = -F_x; \quad \Rightarrow \dot{\lambda} + \lambda f_x = -F_x \tag{2.8}$$

If the second parenthesis in (2.6) is replaced with (2.8), the term $d\,\delta x/dt$ in the first parenthesis is calculated from (2.4) and the hypothesis is made that the right side in (2.6) equals zero, then it follows:

$$\int \delta x\, F_x dt = \int \lambda\, f_u \delta u\, dt \tag{2.9}$$

Combining (2.9) and (2.3) after omitting the integration operator one gets the functional gradient:

$$\frac{\partial V}{\partial u(t)} = F_u + f_u \lambda \tag{2.10}$$

In the construction above the problem remains open, how to bring the term $\lambda\delta x|_0^T$ in (2.7) to zero: under the assumption, that $\Phi(x(T))$ is zero too. Now in the term $\lambda(0)\delta x(0)$ the variation $\delta x(0)=0$ as $x(0)=x_0$. But $\delta x(T)$ is free, therefore $\lambda(T)=0$ and this is a terminal boundary condition. (2.8) can now be integrated backwards.

For the case, that $\Phi(x(T))\#0$, the problem (2.1) shall to be transformed to (2.2) and calculate the functional gradient as presented below: the functional gradient shall be calculated for problem (2.2) , rewritten to (2.11):

$$V = F(x(T)) \tag{2.11}$$

The variation of (2.11) delivers:

$$\delta V[\delta u(.)] = F_x[x(T)]\delta x(T) \tag{2.12}$$

Now $F_x[x(T)]$ is arbitrary in interval $t\in[0,T)$, so in (2.8) F_x can be zeroed. Then the second summand in (2.6) vanishes, remains.

$$\lambda(T)\delta x(T)|_0^T = \int \lambda(t) f_u(t)\delta u(t)\,dt \tag{2.13}$$

Let in (2.13) be $\lambda(T) = F_x[x(T)]$. Then the left side in (2.13) evaluates to $F_x[x(T)]\delta x(T)$ - as $\delta x(0)=0$. It follows:

$$F_x[x(T)]\delta x(T) = \int \lambda(t)\, f_u(t)\delta u(t)\,dt \tag{2.14}$$

Combining (2.14) with (2.12) and omitting the integration sign one gets the gradient:

$$\frac{\partial V}{\partial u(t)} = f_u(t)\lambda(t) \tag{2.15}$$

With (2.10) and (2.15) it is possible to calculate the functional gradients. These expressions were developed under the next restrictions: the operators f_x, f_x^* are linear operators and the target set $x(T)$ is free, it means that for state vector *x* only initial values are given.

Next the calculation of functional gradients shall be given for the fixed time, fixed endpoint task. These problems are typical in mechanics and robotics, where the practice requires to bring the system from an initial state $x(0)$ into a well defined terminal state $x(T)$. Gradient algorithm is based on influence functions $p(t)$ and $R(t)$, which predict, how changes in the control histories $\delta u(t)$ will change the performance index and the terminal conditions. But these are fixed and therefore the gradient cannot be calculated directly as in (2.8), (2.10). Let then the OCP be a modified variant of (2.1) :

$$V = \min_u \left\{ \int F\left(x(t),u(t)\right)dt \right\}, \quad \text{s.t.:} \quad \dot{x} = f\left(x(t),u(t)\right);$$
$$x(0)=0; \quad \psi\left(x(T)\right)=0 \tag{2.16}$$

In (2.16) the term $\Phi\left(x(T)\right)$ from (2.1) is constant, as all the terminal values are fixed and can be cancelled. Then [6] the change of V is:

$$\delta V = \int\left(p^T \frac{\partial f}{\partial u} + \frac{\partial F}{\partial u}\right)\delta u(t)\,dt; \; \partial\psi \,\square\, \left[\delta x_1(T),\dots,\delta x_n(T)\right]^T = \int R^T \frac{\partial f}{\partial u}(t)\,\delta u\,dt \tag{2.17}$$

Influence function p is given e.g. in [7] – comparison has to be made with (2.8) :

$$\dot{p} = -\left(\frac{\partial f}{\partial x}\right)^T p - \left(\frac{\partial F}{\partial x}\right)^T; \quad p_i(T)=0; \quad i=1,\dots,n \tag{2.18}$$

Influence matrix equation for R is:

$$\dot{R} = -\left(\frac{\partial f}{\partial x}\right)R; \quad R(T)=I \tag{2.19}$$

In (2.19) I is $n\times n$ identity matrix. As (2.17), (2.18) are the result of linearization of (2.16), there is no bound for δu , what is unfeasible. To master the situation, the penalty matrix W is introduced and combined with (2.16):

$$\delta V_1 = \delta V + \frac{1}{2}\int (\delta u)^T W \delta u\,dt \tag{2.20}$$

(2.20) must be combined with the second condition in (2.17), one gets:

$$\delta\overline{V} = \delta V_1 + \nu^T\left[\int R^T \frac{\partial f}{\partial u}(t)\,dt - \delta\psi\right] \tag{2.21}$$

One is now searching for δu , which makes (2.21) to minimum. Then the first variation of (2.21) must be zero:

$$\delta\left(\delta\overline{V}\right) = \int\left[\frac{\partial F}{\partial u} + (p+R\nu)^T \frac{\partial f}{\partial u} + (\delta u)^T W\right]\delta(\delta u)\,dt = 0 \tag{2.22}$$

In construction of (2.22) the change in constants ν was neglected, as they must be relatively great values to keep the expression in square brackets in (2.21) small. Further more from (2.22) it follows, after the term in square brackets has been zeroed:

$$\delta u = -W^{-1}\left[\frac{\partial F}{\partial u} + (p+R\nu)^T \frac{\partial f}{\partial u}\right]^T \tag{2.23}$$

δu in (2.23) can be inserted into the right hand term in the second equation in (2.17).

$$\delta\psi = -\int R^T \frac{\partial f}{\partial u} W^{-1}\left[\frac{\partial F}{\partial u} + (p+R\nu)^T \frac{\partial f}{\partial u}\right]^T dt = -A_\psi - B_\psi \nu$$
$$A_\psi = \int\left(p^T \frac{\partial f}{\partial u} + \frac{\partial F}{\partial u}\right) W^{-1}\left(\frac{\partial f}{\partial u}\right)^T R\,dt; \quad B_\psi = \int R^T \frac{\partial f}{\partial u} W^{-1}\left(\frac{\partial f}{\partial u}\right)^T R\,dt \tag{2.24}$$

From (2.24) one gets ν:

$$\nu = B_\psi^{-1}\left(-\delta\psi - A_\psi\right) \tag{2.25}$$

ν from (2.25) is inserted into (2.23), so δu is known function of $\delta\psi$ and can be put into the first equation in (2.17). Let function C_ψ be introduced.

$$C_\psi = \int\left(p^T \frac{\partial f}{\partial u} + \frac{\partial F}{\partial u}\right) W^{-1}\left[\left(\frac{\partial f}{\partial u}\right)^T p + \left(\frac{\partial F}{\partial u}\right)^T\right]dt \tag{2.26}$$

Then δV in (2.17) can be written as:

$$\delta V = -\left(C_{\psi} - A_{\psi} B_{\psi}^{-1} A_{\psi}\right) + A_{\psi} B_{\psi}^{-1} \delta\psi \qquad (2.27)$$

Now gradient algorithm can be based exclusively on $\delta\psi$. The algorithm stops, if all terms in (2.28) are approaching to zero.

$$C_{\psi} - A_{\psi} B_{\psi}^{-1} A_{\psi} \to 0; \quad \nu + B_{\psi}^{-1} A_{\psi} \to 0; \quad \frac{\partial F}{\partial u} + \left(p + R\nu\right)^T \frac{\partial f}{\partial u} \to 0; \quad 0 \le t \le T \qquad (2.28)$$

Executing the algorithm iteratively changing $\delta\psi = -\varepsilon\psi; \quad 0 \le \varepsilon \le 1$, one gets δu from (2.23) providing that ν has been calculated in (2.25). The performance change δV is calculated in (2.27) and herewith the gradient $\partial V/\partial u$ is known already.

The direct use of functional gradients is demonstrated in examples 1,2

Example 1:

$$V = \min_u \int_0^2 \left(u^2 + x^2\right) dt; \quad \dot{x} = u + x; \quad x(0) = 1 \qquad (2.29)$$

In the trivial case the analytic solution will be used as the reference. The canonical equation system then is a dual point problem (2.30):

$$\dot{x} = x - 0.5p; \quad \dot{p} = -p - 2x; \quad x(0) = 1; \; p(2) = 0; \qquad (2.30)$$

Let $x(t, x_0, p_{01}), p(t, x_0, p_{01}), x(t, x_0, p_{02}), p(t, x_0, p_{02})$ be the solutions of (2.30) without the condition $p_2(0) = 0$. Initial values p_{01}, p_{02} must be found so that $p(T, x_0, p_{01})$ and $p(T, x_0, p_{02})$ are different in sign. Interval $\left[p_{01}, p_{02}\right]$ is then halved etc.. The result of this convergent algorithm is $p(0)$. Therefore solution of the canonical equations (2.30) is an initial problem.

The other reference can be obtained with quadratic programming. The process (2.29) is transcribed into the discrete form:

$$x_{k+1} - x_k = h\left(x_k + u_k\right); \quad h = T\left(N - 1\right)^{-1}; \quad k = 1, 2, \ldots, n-1 \qquad (2.31)$$

The discrete status vector is defined:

$$s_{ux} = \left(u_1, u_2, \ldots, u_{n-1}, x_1, x_2, \ldots, x_n\right); \quad b_{eq} = \left(x_0, 0, \ldots\right); \quad \dim b_{eq} = n \qquad (2.32)$$

Matrix A^{eq}, $\dim A^{eq} = n \times \left(2n - 1\right)$, is formed with entries:

$$A_{11}^{eq} = 1; \; A_{i,i-1}^{eq} = h; \; A_{i,n-i+2}^{eq} = 1 + h; \; A_{i,n+i-1}^{eq} = -1; \quad i \in \left[2, n\right] \qquad (2.33)$$

Quadratic problem is defined as:

$$\min_{u_1, \ldots, u_n} \left(\sum u_i^2\right); \; \text{ s.t. } A^{eq} s_{ux} = b_{eq} \qquad (2.34)$$

References can be compared with the result of the steepest descent gradient method, based on (2.10). In (2.10) $F_u = 2u; \quad f_u^* = 1;$. The function λ is the solution of (2.35).

$$\dot{\lambda} + \lambda = -2x; \quad \lambda(T) = 0 \qquad (2.35)$$

The boundary value in (2.35) is calculated from the right hand expression in (2.7), There $\delta x(0) = 0$ and the product $\lambda \delta x \,|_0^T$ must be zero. As $\delta x(T) \# 0$ (free end point), then the boundary condition is $\lambda(T) = 0$. For a given $x(u)$ (2.35) is integrated backwards numerically and herewith all terms in (2.10) are known.

Example 2:

The process model is given in (2.36). The constant control vector was guessed, $u_1(t) = -4.2; u_2(T) = -0.5;$ If *I* is unity matrix, set the matrix $W = 10 * I$.

$$V = \min_u \int_0^T F(x,u)\,dt; \quad F = x_1^2 + x_2^2 + u_1^2 + u_2^2;$$
$$\dot{x} = f(x,u); \quad \dot{x}_1 = x_1 + x_2 + u_1; \quad \dot{x}_2 = x_1 - x_2 + u_2 \tag{2.36}$$
$$T = 2; \quad x_{10} = 4; \quad x_{20} = -2; \quad x_{1T} = 7; \quad x_{2T} = 3; \quad \psi(T) = x_T - \int_0^T f(x,u)\,dt$$

The program proceeds in the following steps:

STEP a) Take the current control $u(t)$ and integrate the forward function $f(x,u)$. Record $x(t)$ and terminal the functions $\psi(T)$.

STEP b) Calculate the influence function $p(t)$ with (2.18).

STEP c) Calculate the integrals in (2.24) to get A_ψ and B_ψ. For each value of integrating variable *t* the influence matrix equation (2.19) must be solved to get the solution $R(t)$.

STEP d) As A_ψ and B_ψ are known, one chooses the iteration step $\delta\psi = -\varepsilon\psi(T)$. In this example $\varepsilon = 0.1$. Then, vector ν can be calculated by (2.25).

STEP e) with ν obtained – this is now a two dimensional constant vector – one calculates δu by (2.23). δu is used to change the current control vector *u*. If the desired convergence tolerance in object function *F* and terminal functions ψ has not been reached yet, one proceeds with step a).

The convergence behaviour is given in fig.1. for object function *F* and terminal functions ψ .

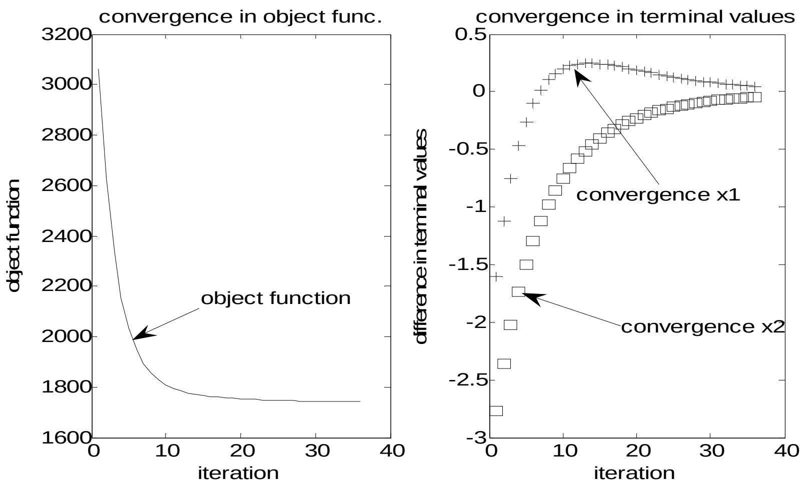


Fig. 1: Iterative convergence for fixed time, fixed endpoint OCP example 2

3 LINEARIZATION IN OPTIMAL CONTROL PROBLEMS

Let the nonlinear dynamic system be given with:

$$\dot{x}(t) = f(t, x(t), u(t)) \text{ a.e. on } [0,1] = D;\ x(0) = x_0 \tag{3.1}$$

With any given control function $u(t)$ the solution of (3.1) is an absolutely continuous function which shall be denoted x^u. The corresponding linearized system is [2]:

$$\delta\dot{x}(t) = f_x(t, x^u(t), u(t))\,\delta x(t) + f_u(t, x^u(t))\,\delta u(t); \quad \delta x(0) = 0 \tag{3.2}$$

The solution which depends of u and δu is written $y^{u,d}$. If then the norm $\| f_x \|$ is limited then (3.1) and (3.2) have unique solutions and there exist finite constant c_3 so that

$$\| y^{u,d} \| \le c_3 \| \delta u \| \tag{3.3}$$

From (3.3) it follows, that in the OCP with linearization $\| y^{u,d} \|$ can be made small and then (3.2) can be so close to the original system as necessary and this linear system is also unique. For the linearization the statement is important, that δx is arbitrary and one can choose

$$\delta x(T) = 0, \tag{3.4}$$

whether the terminal values are fixed or not. They exists several ways to transcribe (3.1) into the discrete form, e.g. the automatic [8], here the Euler approximation was used [9], similar to the algorithm in example 1.

$$x_{k+1} - x_k = f(x_k, u_k, t_k); \quad t_k = (k-1)\frac{T}{n-1}; \quad k = 1,2,\ldots,n; \quad x_1 = x_{10} \tag{3.5}$$

If in (3.5) control vector u_k is disturbed with Δu_k, this induces the change Δx_k. The changes must be compatible with (3.5):

$$\begin{aligned}&(x_{k+1} + \Delta x_{k+1}) - (x_k + \Delta x_k) = h\, f(x_k + \Delta x_k, u_k + \Delta u_k, t_k) = \\ &h\big(f_x(x_k, u_k, t_k)\Delta x_k + f_u(x_k, u_k, t_k)\Delta u_k\big)\end{aligned} \tag{3.6}$$

(3.5) can be subtracted from (3.6). Then the linearized form of (3.1) is:

$$\Delta x_{k+1} - \Delta x_k = h\big(f_x(x_k, u_k, t_k) + f_u(x_k, u_k, t_k)\big) \tag{3.7}$$

If the object function is written in the integral from F in (2.1), then the integral can be replaced with the sum: the linearized form of object function appears to be:

$$\Delta V = h\sum_{k=1}^{n}\big[F_x(x_k, y_k, t_k)\Delta x_k + F_u(x_k, y_k, t_k)\Delta u_k\big] \tag{3.8}$$

The problem of minimizing (3.8) subject to (3.7) is a linear programming problem LP [10,11,12]. The solution of LP can be straightforward. The control and status variables are packed into an expanded status vector, using the next principle::

$$\begin{aligned}&\Delta U_j = (\Delta u_1^j, \Delta u_2^j, \ldots, \Delta u_n^j); \quad \Delta X_j = (\Delta x_1^j, \Delta x_2^j, \ldots, \Delta x_n^j); \\ &\Delta S_{LP} = (\Delta U_1, \Delta U_2, \ldots, \Delta X_1, \Delta X_2, \ldots)\end{aligned} \tag{3.9}$$

(3.6) is then completed with boundary conditions, in an analogous way the Jacoby-matrices in (3.8) can be put in another vector, say ΔS_F. LP problem then is:

$$\min(\Delta S_F, \Delta S_{LP}); \quad \text{s.t.:} \quad A_{eq}\Delta S_{LP} = b_{eq} \tag{3.10}$$

The simple details about construction of A_{eq}, b_{eq} are explained in the examples below. The resulting ΔS_{LP} is then used to correct ΔU_j for the next iteration. To force the convergence of the consecutive LP calls in iteration algorithm, the absolute values of components in vector ΔU_j must be limited, moreover they must be kept relatively small. Then to get feasible initial solution in LP, it is obligatory, that any initial values for ΔU_j which fulfil the terminal conditions are found.

Example 3:

A dynamic system can e.g. be interpreted as mass *M*, supported in parallel with a shock absorber with dumping factor *L* and a spring with stiffness factor *K*. The factor of gravity is excluded.

$$V=\min_{U}\int_0^T u^2dt;\quad u=M\,\ddot{x}+L\,\dot{x}+K\,x;\quad x(0)=x_0;\dot{x}(0)=v_0;x(T)=x_T;\dot{x}(T)=v_T;\quad (3.11)$$

The LP vector includes discrete vectors for $u,x,\dot{x}=v$;

$$\Delta X_L=(\Delta u_1,\ldots,\Delta u_n;\Delta x_1,\ldots,\Delta x_n;\Delta v_1,\ldots,\Delta v_n)\quad (3.12)$$

The linearized object function and linearized dynamic system can be written in the form:

$$\min_{\Delta u_1,\ldots\Delta u_n}\sum_1^n u_i\Delta u_i;\quad A_{eq}=b_{eq}\quad (3.13)$$

The entries for matrix A_{eq} and vector b_{eq} are:

$$\begin{aligned}&\dim(A_{eq}=2(n-1)+4,3n);\dim b_{eq}=3n;b_{eq}=(0);h=T(n-1)^{-1}\\&A_{eq}(i-1,i+n)=1;A_{eq}(i-1,i-1+n)=-1;A_{eq}(i-1,i-1+2n)=-h;i=2,\ldots n\\&A_{eq}(n+i-2,i+2n)=1;\ A_{eq}(n+i-2,i+2n-1)=-1+{h}/{M}\,L;\\&A_{eq}(n+i-2,i-1)=-{h}/{M};\ A_{eq}(n+i-2,i-1+n)=-{h}/{M}\,K;i=2,\ldots n\end{aligned}\quad (3.14)$$

The lines 2,3 in (3.14) stay for iterative discretization in variable Δx and the line 4 stays for discretization in variable Δv. The boundary conditions complete A_{eq} in (3.15). The elements, not assigned as entries to A_{eq} in (3.14), (3.15), are set to be zero.

$$\begin{aligned}&A_{eq}(2(n-1)+1,n+1)=1;\ A_{eq}(2(n-1)+2,2n+1)=1;\\&A_{eq}(2(n-1)+3,2n)=1;\ A_{eq}(2(n-1)+4,3n+1)=1;\end{aligned}\quad (3.15)$$

The algorithm to solve the example runs in next two steps:
STEP 1: guess control vector (u_i) in (3.11), which satisfies the boundary values
STEP 2: Calculate A_{eq},b_{eq} in (3.14), (3.15). Run LP (3.13), obtaining as result vector (Δu_i). Form the new control vector with components: $u_i=u_i+\Delta u_i$. Repeat step 2 until the algorithm converges.

The convergence of the algorithm for the vector x is shown in fig. 2 , with the next setting of parameters: $T=2;M=1;L=3;K=1;n=100;x_0=0.0;x_T=0.3;v_0=1.4;v_T=1.4;$

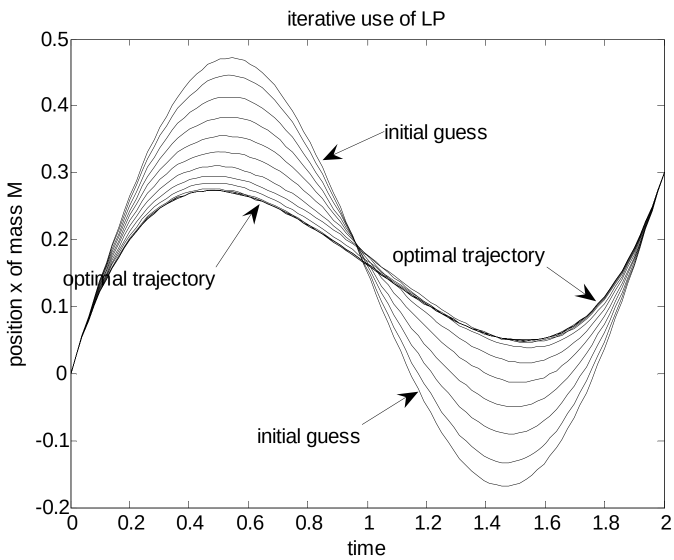


Fig. 2: Optimal trajectory for control system (3.11)

As the system (3.11) is linear and quadratic, it can be solved with quadratic programming to get the reference. The matrix A_{eq} in (3.13) doesn't change. The entries for vector b_{eq} are the boundary values as components $b_{eq}(2n-2+i); i=1,\ldots,4$, consistent with (3.15).

Example 4:
As the dynamic system dual link manipulator was used, original model is reduced to (3.16):

$$\dot{\vartheta}_1=\omega_1;\ \dot{\omega}_1=\frac{1}{I_1}(u_1-S_1-S_2);\ \dot{\vartheta}_2=\omega_2;\ \dot{\omega}_2=\frac{1}{I_2}(u_2-S_2);$$
$$\text{boundary conditions: } \vartheta_{10},\vartheta_{20},\omega_{10},\omega_{20},\vartheta_{1T},\vartheta_{2T},\omega_{1T},\omega_{2T} \qquad (3.16)$$
$$\text{object function: } V=\min_{u}\int_0^T u^2 dt$$

In (3.16) ϑ_1,ϑ_2 are the absolute angles of the first rsp. second link with the vertical, I_1,I_2 are the inertial moments, S_1,S_2 are the static moments, u_1,u_2 are motor torques. The vector of LP variables in (3.10) is:

$$X=\left(\Delta\vartheta_1^1,\ldots,\Delta\vartheta_n^1,\Delta\vartheta_1^2,\ldots,\Delta\vartheta_n^2,\Delta\omega_1^1,\ldots,\Delta\omega_n^1,\Delta\omega_1^2,\ldots,\Delta\omega_n^2,u_1^1,\ldots,u_n^1,\Delta u_1^2,\ldots,\Delta u_n^2\right) \quad (3.17)$$

The upper indices in (3.17) are the link indices. The matrix A_{eq} is formed compatible with (3.17): The conceptual difference when comparing with (3.14) is the numeric calculation of Jacoby matrix , e.g. for variable ω_k^1 :

$$\Delta\omega_k^1=\Delta\omega_{k-1}^1+h\left(\frac{\partial\dot{\omega}_1}{\partial\vartheta_1}\Delta\vartheta_{k-1}^1+\frac{\partial\dot{\omega}_1}{\partial\vartheta_2}\Delta\vartheta_{k-1}^2+\frac{\partial\dot{\omega}_1}{\partial u_1}\Delta u_{k-1}^1+\frac{\partial\dot{\omega}_1}{\partial u_2}\Delta u_{k-1}^2\right) \qquad (3.18)$$

The result is presented on fig. 3.

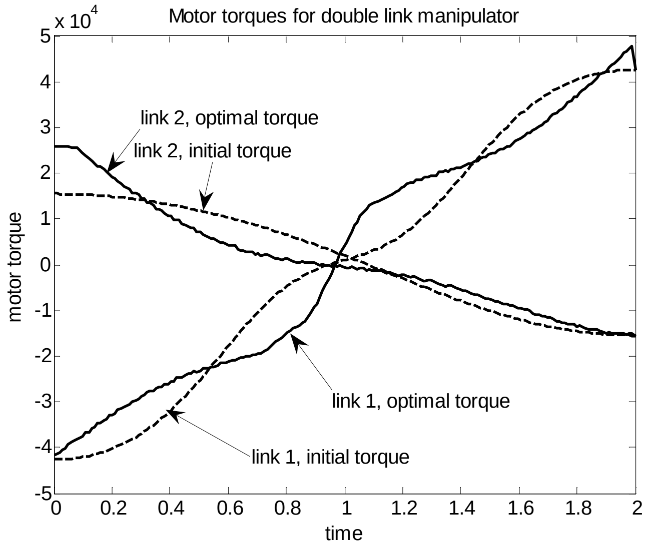


Fig. 3: Optimal and initial torques to move the dual link manipulator from the 'low' $(\vartheta_1=-\pi/2;\ \vartheta_1=-\pi/2;)$ to the high position

In particular the result from (3.16) depends strongly on the guessed start trajectory. If for the start guess links are rotating in the same direction or in the opposite direction, this behaviour repeats for the optimal trajectory. The results could be confirmed with the MATLAB solver fmincon.

4 CONCLUSION

On the fundamental level it is shown, that in OPC functional gradients exist and can be calculated at least numerically. Therefore the perturbation in the control at any instant cannot be annulled later along the trajectory and cannot be made inactive because of the other perturbations, as the system of variations is linear. In particular, this statement about functional gradients remains valid even if the target values are fixed. This theory is based on linear operators. Based on the theory of functional gradients it is quite a formality to linearize the dynamic system together with the object function. Then the OPC is reduced to linear programming. In the paper, only the first order gradient methods were used in the examples. On behalf of a series of LP solvers it is possible to deal even with numerically complex problems. However, nonlinear problems with their immanent non-convexity require pre-processors. In principle they must be built on combinatorial optimization. Here the multi-decision problem is generated as it occurs in the combination of fixed terminal conditions and the requirement to master the object function simultaneously.

The used method of LP shows, that even in OPC field this is a quite robust method. For more complicated cases the simple discretization of Euler – type should be replaced with other collocation methods, e.g. Runge- Kutta methods. The fixed terminal time problem should be replaced with variable time problem introducing the terminal time as an additional parameter in LP.

References

[1] T.J.Betts, 2003. Practical Methods for optimal control using nonlinear programming, SIAM, Philadelphia.

[2] R.Pytlak, 1999. Numerical Methods for optimal control problems with state constraints, Springer-Verlag, Berlin-Heidelberg-New York

[3] A.V.Fiacco, 1983. Introduction to sensitivity and stability analysis in nonlinear programming, Mathematics in Science and Engineering 165, Academic Press, New York.

[4] R.W.H.Sargent.2000. Optimal control. Journal of computational and applied mathematics, Vol.124. Issues 1,2. pp.361-371

[5] L.Collatz, 1964. Funktionalanalysis und numerische Mathematik, Springer-Verlag, Berlin-Goettingen-Heidelberg

[6] A.E.Bryson, Yu-Chi Ho. 1975. Applied Optimal Control, John Wiley & Sons, New York

[7] F.L.Lewis,V.L.Syrmos. 1995. Optimal Control, John Wiley & Sons, New York

[8] A.Walther. 2006. Automatic differentiation of explicit Runge-Kutta methods for optimal control. J. Computer Optimization and Applications, Vol 36: 83-108

[9] L.Dontchev,W.W.Hager. 2000, The Euler approximation in state constrained optimal control. Mathematics of Computation Vol 70, No.233, pp. 173-203

[10] V.G.Antonik,V.A.Srochko.1992. The solution of optimal control problems using linearization methods. Journal of Computing mathematics and mathematical physics, Vol.32, No.7, pp.859-

[11] J.E.Rubio. 1980. Solution of nonlinear optimal control problems in Hilbert spaces by means of linear programming techniques. Journal of optimization theory and applications. Vol. 30, No.4 pp. 643-660.

[12] P.De Farias,B.Van Roy. 2003. The linear programming approach to approximate dynamic programming. Operations research, Vol. 51, No. 6, pp. 850-865

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section V:

Finance and Investments

CROSS-COUNTRIES CAPITAL MARKETS DEVELOPMENT ANALYSIS IN FRAMEWORK OF DYNAMIC PANEL MODELS

Josip Arnerić, **Blanka Škrabić,**
University of Split, Faculty of Economics
Matice hrvatske 31, 21000 Split, Croatia
{jarneric, bskrabic}@efst.hr,

Nataša Erjavec
University of Zagreb, Faculty of Economics
Trg J. F. Kennedya 6, 10000 Zagreb, Croatia
nerjavec@efzg.hr

Abstract: Financial development has the major role in economic growth of Central and Eastern European emerging countries, which is emphasized in the period after their political and structural changes. Together with the processes of Europeanization and the accession driven reforms, foreign direct inflows and portfolio investments resulted in transforming the banking sector and in creating and developing capital markets, particularly in emerging countries.

Therefore, this paper deals with capital markets development analysis of the CEE emerging countries, which are still more bank-based. Empirical results will show which macroeconomic variables affects capital market development in these countries. Moreover, capital market development from emerging countries and developed countries will be compared. Dynamic panel models will be specified as the most appropriate ones to fit the observed data. Parameters of the panel data models will be estimated using GMM method proposed by Arellano and Bond with one lag of dependent variable. For instruments in the panel data model a maximum of two lags of the dependent variables will be used.

Keywords: capital market development, dynamic panel model, Arellano and Bond estimators.

1 INTRODUCTION

The opening up of capital accounts to international capital flows became an essential aspect of economic changes following macroeconomic stabilization in the countries of Central and Eastern Europe during 1990s. These changes required a reform of financial sectors, such as capital market development, banking privatization, trade openness and institutional changes. Emerging countries firstly have become trade opened. Accelerated trade openness has resulted in a small businesses bankruptcy because of increased import of goods and services. Later on financial openness of these countries was necessary for their economic and financial development. Namely, financial liberalization has resulted in changes of international capital flows structure. In the second half of 90s international capital flows to developing countries marked the tendency to replace official lending with a greater share of private capital flows.

Hence, this paper deals with capital market development analysis across emerging and developed countries in the period from 1994 to 2007. Before applying panel data technique hierarchal cluster analysis from 25 European countries will be performed. Hierarchical cluster analysis will be used to determine which subsets of countries can be classified into emerging or developed markets. It can be expected that cluster analysis will confirm the classification of selected CEE countries into emerging markets and selected Western European countries into developed markets. Estimated dynamic panel models between obtained clusters will be compared.

The structure of the paper is organized as follows. In the second section the review of recent theoretical and empirical findings is given. It the third section data and methodology are presented, while the fourth section contains the results of the research and their comparative analysis. The last section summarizes empirical findings and provides concluding remarks.

2 LITERATURE REVIEW

In recent years, with the accumulation of evidence that development of capital markets is positively correlated with economic growth, there has been increasing interest in understanding the determinants of capital market development [14].

Rajan and Zingales (2003) proposed an "interest group" theory of financial development where incumbents oppose financial development because it produces fewer benefits for them than for potential competitors. The incumbents will shape policies and institutions to their own advantage when they have the power. Incumbents can finance investment opportunities mainly with retained earnings, whereas potential competitors need external capital to start up. When a country is open to trade and capital flows, it is more likely to develop its financial system. This is because openness to both trade and finance breeds competition and threatens the rents of incumbents. The Rajan and Zingales hypothesis has very important policy implications, calling for simultaneous trade and financial liberalisation. Its implications, therefore, run contrary to the sequencing literature, which advocates that trade liberalisation should precede financial liberalisation and that capital account opening should be the last stage in the liberalisation process [17].

Law and Demetriades (2006) provided the evidence which suggests that trade openness and institutions are important determinants of financial development, using dynamic panel data technique and data from 43 developing countries during 1980-2001. Moreover, they also suggest that trade openness affects developing countries' financial development differentially. In middle-income countries, trade promotes financial development; and the effect is smaller in low-income economies. On the other hand, capital inflows have a positive effect in determining financial development, especially capital market development in middle-income countries [13].

Huang (2006) has examined the effect of financial openness on the development of financial systems in a panel of 35 emerging markets during the period of 1976 to 2003. He concluded that financial openness is the key determinant of cross-country differences in the development of financial systems. When testing financial openness against the development of the banking sector and capital market separately, he found strong and robust evidence that this link between openness and development exist in capital markets[11].

Bekaert and Harvey (2000) defined financial liberalization as allowing inward and outward foreign equity investment. Henry (2003) argues that strictly speaking, equity market liberalization is a specific type of capital account liberalization, which is the decision to allow capital in all forms to move freely in and out of the domestic market. There are other forms of financial openness relating to bond market, banking sector and foreign exchange reforms [7].

Chinn and Ito (2005) have focused on the links between capital account liberalization, legal and institutional development and financial development, especially in equity markets. In a panel data analysis encompassing 108 countries and twenty years ranging from 1980 to 2000, they explored several dimensions of the financial sector. The empirical results suggest that a higher level of financial openness contributes to the development of equity markets only if a threshold level of general legal systems and institutions is attained [8].

Garcia and Liu (1999) used pooled data from fifteen industrial and developing countries from 1980 to 1995, to examine the macroeconomic determinants of stock market development, particularly market capitalization. They find out that: (1) real income, saving rate, financial intermediary development, and stock market liquidity are important determinants of stock market capitalization; (2) macroeconomic volatility does not prove significant; and (3) stock market development and financial intermediary development are complements instead of substitutes [9].

3 DATA AND METHODOLOGY

In this paper yearly data covers period from 1994 to 2007 for 25 European countries. For endogenous variables two indicators of capital market development are chosen: market capitalization to GDP (MC) and stock trade total value to GDP (STTV). For exogenous variables determinants of capital market development are used: GDP per capita (GDPPC in US $), trade openness to GDP (TO), foreign direct investment to GDP (FDI), and private capital flows to GDP (PCF). All series were ln-transformed.

Initially, hierarchical cluster analysis was performed in order to determine which subsets of countries can be classified into developed or emerging markets according to indicators and determinants of capital market development. All variables were standardized before clustering procedure and Ward's method with Euclidean distance metric used.

Cluster analysis began with each observation (country) in defining a separate group. After recomposing the distance between the groups, the two closest groups were merged. This process was repeated until only one group remained. The dendrogram shows how each of the clusters was formed.

Many economic relationships are dynamic in nature. The current behaviour depends upon past behaviour. Therefore in many cases of describing economic relationships, dynamic panel model is estimated. These dynamic relationships are characterized by the presence of lagged dependent variable among the regressors:

$$y_{it} = \mu + \gamma y_{i,t-1} + \beta_1 x_{it1} + \beta_2 x_{it2} + ... \beta_K x_{itK} + \alpha_i + \varepsilon_{it};\ i = 1,...N, t = 1,...,T\ , \quad (1)$$

where i denoting individual and time t , μ is an intercept, γ is a parameter of lagged dependent variable and $\beta_1, \beta_2, ..., \beta_K$ are the parameters of exogenous variables. It is assumed that ε_{it} are $IID(0,\sigma_\varepsilon^2)$; identically and independently distributed error terms. Unobservable individual-specific effect α_i is time invariant and it accounts for any individual-specific effect that is not included in the regression. If lagged dependent variable $y_{i,t-1}$ is included in model the variable is correlated with individual-specific effect α_i. That renders the OLS estimator biased and inconsistent even if ε_{it} are not correlated. As a result, a new method for estimation was required. Arellano and Bond (1991) proposed new estimator for the dynamic panel model. They argued that additional instrument must be included in dynamic panel data model [2].

Arellano and Bond first considered panel autoregression model without exogenous variables:

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it};\ i = 1,...N, t = 1,...,T, \quad (2)$$

where ε_{it} are $IID(0,\sigma_\varepsilon^2)$. In model (2), lagged dependent variable $y_{i,t-1}$ and individual-specific effect α_i are correlated. To overcome the correlation, problem the first differences of the equation (3) are needed:

$$y_{it} - y_{i,t-1} = \gamma\left(y_{i,t-1} - y_{i,t-2}\right) + \left(\varepsilon_{it} - \varepsilon_{i,t-1}\right);\ i = 1,...N, t = 1,...,T\ . \quad (3)$$

OLS estimator of γ in equation (3) is inconsistent, even when $T \to \infty$, and $\left(\varepsilon_{it} - \varepsilon_{i,t-1}\right)$ follows MA(1) process. This inconsistency is a result of correlation between $y_{i,t-1}$ and $\varepsilon_{i,t-1}$. Equation (3) for time period $t = 3$ is defined:

$$y_{i3} - y_{i2} = \gamma\left(y_{i2} - y_{i1}\right) + \left(\varepsilon_{i3} - \varepsilon_{i2}\right);\ i = 1,...N\,. \tag{4}$$

For $t = 3$, y_{i1} is a valid instrument for $\left(y_{i2} - y_{i1}\right)$. Moreover, y_{i1} is highly correlated with $\left(y_{i2} - y_{i1}\right)$ and it is not correlated with $\left(\varepsilon_{i3} - \varepsilon_{i2}\right)$, as long as ε_{it} are not correlated. If this procedure is continued for $t = 4,5,...,T$, for period $t = T$, $(y_{i1}, y_{i2,}...y_{i,T-2})$ are valid instruments for $y_{i,T}$. Arellano and Bond suggested that the list of instruments can be extended by exploiting additional moment conditions. It is well known that imposing more moment conditions increases the efficiency of estimators [18].

Number of moment conditions varies with *T*. For $t = T$, there are $T - 2$ moment conditions and $T - 2$ valid instruments:

$$\begin{aligned} &E\left\{\left(\varepsilon_{iT} - \varepsilon_{i,T-1}\right) y_{i1}\right\} = 0 \\ &E\left\{\left(\varepsilon_{iT} - \varepsilon_{i,T-1}\right) y_{i,2}\right\} = 0 \\ &\qquad\vdots \\ &E\left\{\left(\varepsilon_{iT} - \varepsilon_{i,T-1}\right) y_{i,T-2}\right\} = 0. \end{aligned} \tag{5}$$

All these moment conditions can be exploited in GMM framework. To introduce the GMM estimator, for general sample size *T*, a vector $\Delta\varepsilon_i$ is defined as a vector of transformed error terms

$$\Delta\varepsilon_i = \begin{pmatrix} \varepsilon_{i4} - \varepsilon_{i3} \\ \varepsilon_{i5} - \varepsilon_{i4} \\ \vdots \\ \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{pmatrix} \tag{6}$$

and

$$Z_i = \begin{pmatrix} y_{i1} & 0 & 0 & ..... & 0 & 0.... & 0 \\ 0 & y_{i1}\, y_{i2} & & & 0 & 0... & 0 \\ & & & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0... & & 0 & y_{i1}y_{i2}...y_{iT-2} & \end{pmatrix} \tag{7}$$

as a matrix of instruments.

The set of all moment conditions (7) can be written as:

$$E\left\{Z_i{}'\Delta\varepsilon_i\right\} = 0\,. \tag{8}$$

To derive GMM estimator, the equation (8) is written as:

$$E\left\{Z_i^{'}(\Delta y_i - \gamma\Delta y_{i,-1})\right\} = 0\,. \tag{9}$$

To obtain the estimates of γ, a quadratic expression

$$\min_{\gamma}\left[\sum_{i=1}^{N} Z_i' \ (\Delta y-\gamma\Delta y_{i,-1})\right]' W_N\left[\sum_{i=1}^{N} Z_i'(\Delta y-\gamma\Delta y_{i,-1})'\right], \tag{10}$$

is minimized using the corresponding sample moments. In (10), W_N is a symmetric positive definite weighting matrix.

Differentiating (10) with respect to γ and solving for γ gives:

$$\hat{\gamma}_{GMM}=\left(\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)W_N\left(\sum_{i=1}^{N} Z_i'\Delta y_{i,-1}\right)\right)^{-1}\times\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)W_N\left(\sum_{i=1}^{N} Z_i'\Delta y_i\right). \tag{11}$$

The general GMM approach does not impose that ε_{it} are $IID(0,\sigma_\varepsilon^2)$ over individuals and time [3]. Therefore, optimal weighting matrix is estimated without imposing these restrictions. It is also possible to impose that ε_{it} are not autocorrelated:

$$E\left\{\Delta\varepsilon_i\Delta\varepsilon_i'\right\}=\sigma_\varepsilon^2 G=\sigma_\varepsilon^2\begin{pmatrix} 2 & -1 & 0 & & \dots \\ -1 & 2 & & \dots & 0 \\ & & \dots & & \\ 0 & & & 2 & -1 \\ 0\dots & & & -1 & 2 \end{pmatrix}. \tag{12}$$

Optimal weighting matrix is then defined as:

$$W_N^{opt}=\left(\sum_{i=1}^{N} Z_i' G Z_i\right)^{-1}. \tag{13}$$

Combining optimal weighting matrix (13) into (11) one step Arellano and Bond GMM estimator can be obtained as:

$$\hat{\gamma}_{GMM}=\left(\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)W_N^{opt}\left(\sum_{i=1}^{N} Z_i'\Delta y_{i,-1}\right)\right)^{-1}\cdot\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)W_N^{opt}\left(\sum_{i=1}^{N} Z_i'\Delta y_i\right). \tag{14}$$

Replacing differenced error terms $\Delta\varepsilon_i$ in (12) by differenced residuals $\Delta\hat{\varepsilon}_i$, obtained from the preliminary consistent estimator $\hat{\gamma}_{GMM}$, two step Arellano Bond estimator is estimated as:

$$\hat{\gamma}_{GMM}=\left(\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)\hat{W}_N^{opt}\left(\sum_{i=1}^{N} Z_i'\Delta y_{i,-1}\right)\right)^{-1}\cdot\left(\sum_{i=1}^{N}\Delta y_{i,-1}' Z_i\right)\hat{W}_N^{opt}\left(\sum_{i=1}^{N} Z_i'\Delta y_i\right). \tag{15}$$

Estimators in (14) and (15) are asymptotically equivalent if the ε_{it} are $IID(0,\sigma_\varepsilon^2)$.
Dynamic panel model which includes K exogenous variables can be written in matrix form:

$$y_{it}=\mu+\gamma y_{i,t-1}+x_{it}'\beta+\alpha_i+\varepsilon_{it}=1,\dots N, t=1,\dots,T, \tag{16}$$

where β is a vector of parameters $\beta_1,\dots,\beta_K$ and x_{it}' is a matrix of exogenous variables $x_{it1},x_{it2},\dots,x_{itK}$. Assuming that there is no correlations between exogenous variables and any ε_{is} in (16), valid instruments are given in the matrix:

$$Z_i = \begin{pmatrix} y_{i1}\ x'_{i1}\ x'_{i2} & 0\ \ 0\ \ 0\ \ 0\ 0 & \cdots & 0\ 0 \cdots 0 & 0\ \ 0 \cdots 0 \\ 0\ \ \ 0\ \ 0 & y_{i1}\ y_{i2} x'_{i1}\ \ x'_{i2}, x'_{i3} & \ddots & 0\ 0 \cdots 0 & 0\ \ 0 \cdots 0 \\ \vdots\ \ \ \vdots\ \ \vdots & \vdots\ \ \vdots\ \ \vdots\ \ \vdots\ \ \vdots & \vdots & \vdots\ \vdots\ \ \ \vdots & \vdots\ \ \vdots\ \ \vdots \\ 0\ \ \ 0\ \ 0 & 0\ \ 0\ \ 0\ \ 0\ \ 0 & \cdots & \multicolumn{2}{c}{y_{i1,} y_{i2} \ldots y_{i,T-2} x'_{i1}\ x'_{i2} \ldots x'_{i,T-1}} \end{pmatrix} \tag{17}$$

Validity of chosen instruments for parameters estimation can be tested using Sargan test [6]. Therefore, Sargan test is used for testing overidentification of the restrictions. If a null hypothesis is accepted by Sargan test it means that all chosen instruments are valid, i.e. dynamic panel model is adequately specified. Namely, optimal number of instruments must be chosen for acceptance of the null hypothesis which reduces estimation bias.

Two key tests for serial correlation are derived by Arellano and Bond. Test for the first-order serial correlation (usually labelled m_1) and test for the second-order serial correlation in differenced residuals (usually labelled m_2). The first-order autocorrelation in the differenced residuals does not imply that the estimates are inconsistent [1]. However, the second-order autocorrelation would imply that the estimates are inconsistent.

4 EMPIRICAL RESULTS

According to the cluster analysis, a dendrogram shows how each of the clusters was formed. A horizontal line connecting two groups shows that the groups were combined at the distance shown on the vertical axis. From dendrogram (Figure 1) it can be seen that the optimal number of clusters is 2 since the greater distance is in combining the two clusters into one cluster.

```
                         Rescaled Distance Cluster Combine

    C A S E              0         5        10        15        20        25
  Label             Num  +---------+---------+---------+---------+---------+

  Romania            21   -+
  Serbia             23   -+-+
  Poland             20   -+ |
  Croatia            13   -+ +---+
  Lithuania          18   -+ |   |
  Latvia             17   -+ |   |
  Ukraine            25   -+-+   |
  Macedonia, FYR     19   -+     +-----------------+
  Slovenia           24   -+     |                 |
  Czech Republic     14   -+     |                 |
  Hungary            16   -+-+   |                 +-----------------------+
  Bulgaria           12   -+ +---+                 |                       |
  Russian Fed        11   ---+                     |                       |
  Estonia            10   -------+-----------------+                       |
  Belgium            15   -------+                                         |
  Germany             1   -+-+                                             |
  Austria             6   -+ +---+                                         |
  Denmark             7   ---+   |                                         |
  Italy               2   -+     +---+                                     |
  Greece              8   -+---+ |   |                                     |
  Portugal            9   -+   +-+   |                                     |
  France              3   -+-+ |     +-------------------------------------+
  Spain               4   -+ +-+     |
  Switzerland        22   ---+       |
  United Kingdom      5   -----------+
```

Figure 1: Dendrogram computed according to the cluster analysis using Ward's method with Euclidean metric.

The Figure 1 confirms the finding that the first cluster consists of emerging markets belonging to CEE countries (fifteen of them), except Belgium, while the second cluster consists of developed markets belonging to Western European countries (ten of them). Therefore 96% of countries are correctly classified.

Variance decomposition of each variable into two components: a between-group component and a within-group component is performed to obtain F-statistic (Table 1). The F-statistic from ANOVA – *ANalysis Of VAriance* is used for testing the differences in means between clusters obtained from hierarchical cluster analysis. Since the p-values of the F-ratios are less than 0.05 for all variables (except FDI and PCF) the conclusion is that there is a statistically significant difference between the means for all variables (except FDI and PCF).

Table 1: Differences testing in means between clusters obtained from hierarchical cluster analysis.

Variables	Clusters		ANOVA results of differences testing in means between clusters	
	Mean value for CEE emerging markets	Mean value for WE developed markets	F-ratio	p-value
GDPPC	23522.74	7165.15	25.245	0.000
FDI	7.68	8.28	0.046	0.833
TO	64.57	114.54	12.038	0.002
MC	77.05	31.41	17.034	0.000
STTV	96.91	11.38	41.780	0.000
PCF	48.10	39.67	0.355	0.557

Taking into consideration the first order autoregressive behaviour of the entire set of capital market indicators dynamic panel model is specified for capital market development analysis.

Dynamic panel model is given by the following equation:

$$\ln CMD_{it} = \alpha + \gamma \ln CMD_{i,t-1} + \beta_1 \ln GDPPC_{it} + \beta_2 \ln FO_{it} + \beta_3 \ln TO_{it} + \varepsilon_{it}, \qquad (18)$$

where dependent variable CMD_{it} is an indicator of the capital market development and $CMD_{i,t-1}$ is a lagged dependent variable. $GDPPC_{it}$ is the measure of economic growth (GDP per capita). FO_{it} is financial openness (gross private capital flow to GDP and/or foreign direct investment to GDP) and TO_{it} is trade openness (the sum of export and import to GDP). Two step Arellano and Bond GMM estimator is used for model estimation. Dependent variables with maximum of 2 time lags are used for the valid instruments. For each selected capital market development indicator, dynamic panel models, for emerging and developed markets, are estimated. Moreover, in the specified panel models different measure of financial openness are used. Empirical results of estimated panel models are given in Table 2 and Table 3.

According to the estimated parameters of dynamic panel models for emerging markets (CEE countries) in Table 2, only the coefficient on lagged dependent variable is statistically significant and positive. In comparison to developed markets, besides lagged dependent variable, the indicators of financial openness as well as trade openness have statistically significant and positive influence on capital market development (market capitalization to GDP) across Western European countries.

Table 2: Estimated panel model parameters with market capitalization to GDP as dependent variable.

Parameters of dynamic panel model [a]	**EMERGING MARKETS**		**DEVELOPED MARKETS**	
	Model 1	**Model 2**	**Model 1**	**Model 2**
α	0.164 (0.127)	0.019** (0.122)	0.030 (0.041)	0.023 (0.054)
lnMC$_{t-1}$	0.450*** (0.064)	0.357*** (0.0624)	0.484*** (0.126)	0.2609** (0.121)
lnGDPPC	-1.700 (2.040)	1.853 (2.277)	-2.778 (2.197)	-1.132 (3.432)
lnPCF	0.337 (0.233)	-	0.266*** (0.085)	-
lnFDI	-	0.0037 (0.046)	-	0.0741* (0.414)
lnTO	-0.788 (0.586)	-0.222 (0.557)	0.764 (0.661)	1.289*** (0.467)
Number of observations	119	132	109	117
Number of groups	13	14	11	11
Sargan test (p-value)	0.944	0.949	0.969	0.955
m_1 test (p-value)	0.008*	0.191	0.221	0.568
m_2 test (p-value)	0.3417	0.520	0.131	0.2737

,,*** indicates significance at 10%, 5%, and 1% level.*
[a] *the numbers in the brackets are standard errors.*

Diagnostic tests (Sargan test, m_1 and m_2 statistics) for all estimated models in Table 2 are satisfying.

Table 3: Estimated panel model parameters with stock trade total value to GDP as dependent variable.

Parameters of dynamic panel model [a]	**EMERGING MARKETS**		**DEVELOPED MARKETS**	
	Model 1	**Model 2**	**Model 1**	**Model 2**
α	-0.1496 (0.105)	0.019** (0.122)		0.201** (0.080)
lnSTTV$_{t-1}$	0.478*** (0.050)	0.357*** (0.0624)	0.176** (0.087)	0.202** (0.087)
lnGDPPC	4.222* (2.539)	8.399* (4.287)	2.919* (3.019)	4.502 (4.697)
lnPCF	0.859*** (0.189)	-	-0.035 (0.199)	-
lnFDI	-	0.358*** (0.0247)	-	-0.031 (0.045)
lnTO	-0.395 (0.971)	-1.467 (0.992)	5.599*** (0.857)	5.093*** (0.926)
Number of observations	113	130	109	117
Number of groups	13	14	11	11
Sargan test (p-value)	0.991	0.969	0.9723	0.972
m_1 test (p-value)	0.0713*	0.150	0.3655	0.356
m_2 test (p-value)	0.895	0.659	0.6848	0.755

,,*** indicates significance at 10%, 5%, and 1% level.*
[a] *the numbers in the brackets are standard errors.*

When stock trade total value to GDP is selected as dependent variable, empirical results differ significantly. Namely, for emerging markets (except lagged dependent variable) financial openness and GDP per capita are statistically significant in dynamic panel model.

For developed markets, lagged dependent variable and trade openness are statistically significant, while the GDP per capita is significant only in the first panel model at 10% significance level. Diagnostic tests (Sargan test, m_1 and m_2 statistics) for all estimated models in Table 3 are also satisfying.

5 CONCLUDING REMARKS

In this paper the influences of financial openness, trade openness and income on capital market development across countries were examined. From empirical results it can be concluded that financial openness has positive and statistically significant effect on capital market development of emerging countries, unlikely trade openness. In developed markets, both financial and trade openness significantly effect capital market development.

Moreover, financial openness has gather influence on capital market development in emerging markets. Therefore, trade openness has high and positive influence on capital market development only in developed markets. Even trade openness in emerging markets is almost double in comparison to developed markets, it has no effect on capital market development because fast trade openness of these countries after the 90s has resulted in small businesses bankruptcy. This has happened due to an increase of import of goods and services without improving their export.

References

[1] Anderson, T. W., Hsiao. C., 1981. Estimation of dynamic models with error components, Journal of the American Statistical Association, 76. pp. 598-606.

[2] Arellano, M., Bond. S., 1991. Some test of specification for Panel data, Monte Carlo Evidence and Application to Employment Equations, Review of Economic Studies, 58, pp. 277-297.

[3] Baltagi, B. H., 2005. Econometric Analysis of Panel Data, Third Edition. John Wiley & Sons, New York, 147 p.

[4] Baltagi, B., Demitriades, P., Law, S. H., 2007. Financial Development, Openness and Institutions: Evidence from Panel Data, Conference on New Perspectives on Financial Globalization. Research Department, Washington, pp. 1-31.

[5] Baltagi, B. H., Demetriades, P. O., Law. S. H., 2008. Financial development and openness: Evidence from panel data. Journal of Development Economics, article in press, 12 p.

[6] Beck. T., Levine, R., 2004. Stock markets, banks, and growth: Panel evidence, Journal of Banking & Finance, 28, pp. 423-442.

[7] Bekaert, G., Harvey, C., 2000. Foreign speculators and emerging equity markets, The Journal of Finance, Vol. LV, No. 2, pp. 1-55.

[8] Chinn, M. C., Ito, H., 2006. What maters for financial development? Capital controls institutions and interactions, Journal of Development Economics, 81, pp. 163-192.

[9] Garcia, V. F., Liu, L., 1999. Macroeconomic Determinants of Stock Market Development, Journal of Applied Economics, Vol. 2, No. 1, pp. 29-59.

[10] Hill, Griffiths, Lim, 2008. Principles of Econometrics, Third edition. John Wiley & Sons, New York, 220 p.

[11] Huang, W., 2006. Emerging Markets Financial Openness and Financial Development, Working paper No. 06/588, Department of Accounting and Finance. University of Bristol, pp. 1-43.

[12] Kivet, J. F., 1995. On bias, inconsistency and efficiency of various estimators in dynamic panel data models, Journal of Econometrics, 68, pp. 53-78.

[13] Law, H. S., Demetriades, P., 2006. Openness, Institutions and Financial Development, Working paper WEF 0012. World Economy & Finance Research Programme, University of London, pp. 1-29.

[14] Levine, R., Zervos, S., 1998. Stock Markets, Banks, and Economic Growth, American Economic Review, 88, pp. 537-558.

[15] Levine, R., 2004. Finance and Growth: Theory and Evidence, NBER Working Paper No. 10766. Brown University-Department of Economics, National Bureau of Economic Research, pp. 1-118.

[16] Levine, R., 1997. Financial development and economic growth: views and agenda, Journal of Economic Literature, 35, pp. 668-726.

[17] Rajan, R. G., Zingales, L., 2003. The great reversals: the politics of financial development in the twenty century, Journal of Financial Economics, 69, pp. 5-50.

[18] Verbeek, M., 2005. A Guide to Modern Econometrics, Second Edition. John Wiley & Sons, London, 341 p.

[19] Wooldrige, J. M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, London, 412.

MARKET VALUE VALUATION WITH OHLSON MODEL: AN EMPIRICAL ANALYSIS OF ATX INDEXED COMPANIES IN VIENNA STOCK EXCHANGE

Ali Osman Kusakci
Faculty of Engineering and Natural Sciences
International University of Sarajevo, 71000 Sarajevo, Bosnia and Herzegovina
akusakci@ius.edu.ba

Abstract:
Scholars of finance and accounting propose several models on equity valuation. One of these models, namely Ohlson Model, relates market value with accounting numbers and thereby, offers an opportunity of equity valuation based on accounting data. This paper aims to weight the explanatory power of the Ohlson Model on determination of market value of ATX (Austrian Traded Index) companies. For this purpose, the model parameters are estimated on the basis of past data and its precision is weighted with dataset of validation period. Although the differences between estimated market value and value calculated based on stock price are not statistically significant, the model is limited mainly with the abnormal consequences of the recent financial crisis.

Keywords: Ohlson Model, ATX, linear information dynamic, market value, equity valuation.

1. Introduction

A company's balance sheet, income statement and statement of cash flow are still most used sources to evaluate its performance despite the fact their reliability has been questioned because of very big scaling accounting scandals, such as Enron (2001) and Worldcom (2002), at the beginning of the new millennium. In addition to these illegal falsifications of the book accounts, there is still enough space to adjust the accounting numbers for a better image in the public within the framework drawn by the legal entities.

On the other hand in order to evaluate a company's market value the accounting numbers provide the most available, clear, and solid basis. Over the decades several performance ratios and equity valuation models are developed based on the accounting data which utilizes of the advantages mentioned above. Other approaches used the market data to compute the market value based on the assumption of having an efficient market with perfect information (i.e. all information is reflected fully in the company's market share.) (Spivey, McMillan, 2003). Although this assumption is not entirely true, it provides a crucial fundament for many evaluation models, since collecting the data prevailing on the market requires no great effort.

Evaluating a firm's market value is of paramount importance not only for shareholders but also for financial institutions and researchers in that area. Cupertino and Lustosa (2004) refer to different literature to highlight the importance of the ability of evaluating assets with precision and states that this is at the heart of theory of finance because many personal and business decisions are to be made by the selection of alternatives that maximize economic value. The Ohlson Model offers clear and very easily applicable approach in measuring the performance, while it builds a descriptive relation between a company's book value and its market value and makes it possible to estimate residual income for the next year based on the data of current year. This ability of the model makes it a powerful and widely accepted valuation tool among researchers. Its popularity is stated in Lo and Lys (2000) as follows: "not surprisingly, this enthusiasm is also evident in the impact of the model on contemporary accounting literature. For example, to date (May 12, 1999) we found an average of 9 annual citations in the Social Sciences Citation Index (SSCI) for Ohlson (1995). If this citation rate continues, Ohlson's work is not just influential, but will become a classic."

Since its introduction in 1995, it has been implemented on different markets in several countries, as in Japan (Koji, 2001), the United Kingdom (Gregor, Saleh, Tucker, 2005), and the United States (Lee, Myers and Swaminathan, 1999) etc. Beyond the empirical analysis the literature on the model has been enriched with critics and proposed improvements in the model.

This paper will deliver first a direct insight into the Ohlson Model based mostly on the original work of Ohlson (1995) and Spremann's (2004) book, while contributions and criticisms to the model will be highlighted in the light of recent studies on financial accounting. In the second part, parameters in the model will be estimated with the help of historical data on the Vienna Stock Exchange. This calibrated model will be used to evaluate the theoretical market values of ATX-indexed companies. As the final stage, the results of the empirical study will be presented and some drawbacks of the model as well as improvements will be discussed.

2. Ohlson Model

2.1. Clean Surplus Relation

Ohlson Model relies on two basic approaches used in accounting and finance, namely Clean Surplus Relation (CSR) and Discounted Dividends Model (DDM). The CSR is creating a relation between the book value of a firm's current year equity and previous year equity. It can be formulated as follows:

$$b_t = b_{t-1} + x_t - d_t \tag{1}$$

where

- b_t book value of firm's equity in year *t*
- b_{t-1} book value of firm's equity in year *t-1*
- x_t earnings for the period *(t-1,t)*
- d_t net dividends paid at date *t*.

The CSR proposes that the current year's book value of equity depends only on the previous year's book value of equity, the earnings and the dividends of the current period. Although this assumption covers most of the cases, there could be different factors affecting book value including capital increases and reductions. CSR also implies that dividends reduce book value but leave current earnings unaffected. To express this statement in the language of mathematics following formulation can be adopted (Ohlson, 1995).

$$\begin{aligned} \partial b_t / \partial d_t &= -1 \\ \partial x_t / \partial d_t &= 0 \end{aligned} \tag{2}$$

This additional feature implies that dividends displace the market value on a dollar-for-dollar, and dividends paid today influence future expected earnings negatively (Ohlson, 1995).

2.2. Discounted Dividends Model

As mentioned in section 2.1, the second fundament of the model is the Discounted Dividends Model (DDM). This feature is named in Ohlson's (1995) original work as PVED, present value of expected dividends.

$$P_t = \sum_{\tau=1}^{\infty} \frac{E_t[\tilde{d}_{t+\tau}]}{(1+r)^{\tau}} \quad \textbf{(3)}$$

Where

- P_t market Value of the firm at date *t*
- $d_{t+\tau}$ net dividends paid at date $t+\tau$
- r risk free interest rate
- $E_t[.]$ the expected value operator conditioned on the date *t* information

One can see that equation (3) indicates the same idea as it is described in the famous Gordon Model (1959) if we assume that paid dividends increase with a constant rate. DDM evaluates firm's equity from the view point of an investor, who purchases shares of that firm and expects dividend payments during the time he keeps the shares of equity. So, the present value of the total expected dividends paid determines the market value of the firm (Lo, Lys, 2000).

Regarding the interest rate r, at which the dividends will be discounted, Ohlson (1995) prefers to keep the problem simple and assumes a case of risk neutrality and so a risk free interest rate.

Before we combine these two approaches, we must define abnormal earnings (AE) or as it is stated in the literature "residual income". It can be formulated as follows:

$$x_t^a \equiv x_t - rb_{t-1} \quad \textbf{(4)}$$

where

- x_t^a abnormal earnings between *t-1* and *t.*

Thus, the abnormal earnings are earnings over the "normal" earnings, which can be computed by the product of a "proper" discount rate and the book value at date *t-1*. So, the "normal" earnings are related with the "normal" return on the capital invested at the beginning of the period. As described by Ohlson (1995), we can interpret x_t^a as earnings minus a charge for the use of capital, so positive abnormal earnings mean a "profitable" period from *t-1* to *t* since the book rate of return x_t^a / b_{t-1} exceeds the firm's cost of capital r.

We denote that the interest rate used in equation (4) is the same rate introduced in DDM-equation. At this point we should state that there are several critics to the assumption of using risk free interest rate as a firm's cost of capital. Cupertino and Lustosa (2004) mention that the discount rate in various works on Ohlson Model is assumed to be cost of equity rather than a risk free rate. Determination of a proper discount rate could be seen a debatable point in the Ohlson Model.

Equation (4) combined with clean surplus relation leads to

$$d_t = x_t^a - b_t + (1+r)b_{t-1} \tag{5}$$

Given that $E_t[b_{t+\tau}]/(1+r)^\tau \to 0$ as $\tau \to \infty$, a firm's equity's market value can be expressed with the following equation when we bring together CSR, DDM, and equation (5)(Spremann, 2004, my translation).

$$P_t = b_t + \sum_{\tau=1}^{\infty} \frac{E_t[x_{t+\tau}^a]}{(1+r)^\tau} \tag{6}$$

This relation implies that a firm's market value at any time *t* equals the sum of its book value and the present value of the expected value of future abnormal earnings. Furthermore, compared with the DDM relation it proposes to estimate future's abnormal earnings instead of future dividends. Feltham and Ohlson (1995) concludes that the difference between a firm's market and book value, which is named "firm's goodwill", can be estimated as the expected value of future abnormal earnings discounted with risk free interest rate. Thus, a firm's good will g_t can be written as follows:

$$g_t \equiv P_t - b_t = \sum_{\tau=1}^{\infty} \frac{E_t[x_{t+\tau}^a]}{(1+r)^\tau} \tag{7}$$

2.3. Linear Information Dynamics

The abnormal earnings in the relation above are uncertain future payments which are subject to a stochastic process and must be estimated. However, estimation of residual incomes is most challenging part of market value relation in (6) since numbers for the book value b_t can be easily gathered from accounting data. Exactly at that point Ohlson (1995) helps us with his proposal that abnormal earnings follow linear information dynamics (LID). This assumption represents a very helpful contribution to equity evaluation models and inspired many different ideas. A similar but more sophisticated evaluation model is developed by Feltham and Ohlson (1995) with the same linear information dynamics idea mentioned as "Markovian structure". The autoregressive structure of abnormal earnings is described in Ohlson (1995) as follows:

$$\tilde{x}_{t+1}^a = \omega x_t^a + v_t + \tilde{\varepsilon}_{1t+1} \tag{8}$$

$$\tilde{v}_{t+1} = \gamma v_t + \tilde{\varepsilon}_{2t+1} \tag{9}$$

Where

- v_t other information term
- $\tilde{\varepsilon}_{1t+1}$ normal distributed disturbance terms
- $\tilde{\varepsilon}_{2t+1}$ normal distributed disturbance terms
- ω parameter of autoregressive process
- γ parameter of autoregressive process

According to LID a firm's abnormal earning at *t+1* can be shown as a function of realized abnormal earning of previous year *t*, a term called "other information", and a random disturbance variable with zero-mean following normal distribution. Thus, for the next year's abnormal earning, $\tilde{x}^a_{t+1}$, the term ωx^a_t provides quiet good range of estimation.

For the purpose of empirical implementation, the other information term in the formulation v_t can be replaced with analysts' forecasts. It covers every possible event that has not yet an effect on the financial statements and not captured within abnormal earnings of current year. Possible factors which can be taken into account by other information may be

- the news sourced from accounting but have not affected the current accounting numbers yet (i.e., revenue of the current period)
- or the information not related with accounting and cannot be captured by financial statements, such as legal environment, image, innovation position in the market, a new product, and management capacity of the firm. (Spremann, 2004, my translation)

Moreover, Ohlson (1995) assumes with the equation (9) that v_t is an independent variable of current and past abnormal earnings, but it influences the future abnormal earnings. Another feature of that equation describing the linear dynamics of $\tilde{v}_{t+1}$ is that it includes a normal distributed random variable $\tilde{\varepsilon}_{2t+1}$ with zero mean which captures possible deviations from the estimated value of $\tilde{v}_{t+1}$.

Although the model offers a revolutionary framework by associating the accounting numbers with the market value, a crucial point is how to estimate the parameters ω and γ in the model. Ohlson (1995) implies in his paper that a firm's economic settings and its accounting principles determine these exogenous parameters, which are fixed and known. In addition he restricts the parameters to be non-negative and less than one. Furthermore, he refers to his previous work Ohlson (1991) for a detailed analysis of the special case, in which ω equals one.

Additionally, equation (8) makes it possible to predict next year's earnings in addition to next year's abnormal earnings. Since the current year book value and abnormal earnings are known variables at date *t*, equation (8) can be reformulated as follows.

$$E_t\left[\tilde{x}_{t+1}\right] = rb_t + \omega x^a_t + v_t \qquad \textbf{(10)}$$

Equation (10) implies that the expected value of earnings of the next year can be calculated based on current year's book value of equity, abnormal earnings and other information.

2.4. Market Value in Ohlson Model

Based on CSR, DDM and LID, a new equation which leads to the value of a firm's equity can be formulated as follows.

$$P_t = b_t + \alpha_1 x^a_t + \alpha_2 v_t \qquad \textbf{(11)}$$

where

$$\alpha_1 = \omega / (1 + r - \omega) \geq 0$$
$$\alpha_2 = (1 + r) / (1 + r - \omega)(1 + r - \gamma) > 0$$

We prefer to give the formula directly rather than to go into the details of calculations. The derivation of this market value equation is demonstrated in the Ohlson's original work in appendix 1 (Ohlson, 1995).

Equation (11) shows that the value of a firm equals to its book value plus abnormal earnings weighted with a factor and other information which stands for effects uncovered by accounting.

ω and γ are the persistence parameters in the (x_t^a, v_t) process; larger values make P_t, the market value, more sensitive to realizations of abnormal earnings and other information (Ohlson, 1995). We analyze different cases by placing extreme values for two parameters of the model.

- $\omega = 0, \gamma = 0$: The abnormal earnings have minimum or zero persistence and they are random with an expected value of zero. The market value matches with the book value. The abnormal earnings of the current year are not a good predictor for the abnormal earnings of next year (Spremann, 2004, my translation).
- $\omega = 1,\ \gamma = 0$: The abnormal earnings have a strong persistence. They, alone, determine goodwill and suffice in prediction of future abnormal earnings (Spremann, 2004, my translation).

Up to date several researches have been conducted and different possible cases have been empirically analyzed. One of these empirical studies conducted on 50.000 the US-companies by Dechow, et al. (1999) estimates two parameters, ω and γ, as 0.62 and 0.32 respectively.

3. Empirical Evidence from Austria

3.1. Dataset and Methodology

In this part an empirical analysis is conducted on the end-of-accounting year data gathered from ATX companies on Vienna Stock Exchange within a time period from 2000 to 2008 while the data of first seven years is used to estimate the model parameters. So, the model is calibrated cross-sectionally on the basis of historical data and tested on the sample data of 2008. The used data is collected mainly from the website of Vienna Stock Exchange[1], Reifeisen Bank's website[2] and annual reports of particular firm.

Austrian Traded Index consists of 20 best performing companies, so called blue chips, on Vienna Stock Exchange from different branches. The composition of the index is revised twice a year.

Since the companies in banking and insurance branches are implementing different accounting methods they are excluded from further analysis. Moreover, companies offered to public not before the estimation period, so 2000, are omitted in this study since they do not provide the numbers needed. Another firm that must be eliminated from the research is Austrian Airlines as it shows substantial losses in its

[1] http://en.wienerborse.at/indices/
[2] http://www.rcb.at/aktienoesterreich.0.html

annual reports for the period of research. In *Table 1* the names of 12 companies, which are included in empirical analysis, are given.

	Company
1	Andritz AG
2	BETandWIN.com Interactive Entertainment AG
3	EVN AG
4	Flughafen Wien AG
5	Mayr-Melnhof Karton AG
6	OMV AG
7	RHI AG
8	Schoeller-Bleckmann Oilfield Equipment AG
9	Telekom Austria AG
10	Verbund AG
11	Voestalpine AG
12	Wienerberger AG

Table 1: **List of companies used in empirical analysis**

As mentioned earlier Ohlson Model (1995) requires several fundamental assumptions in order to be implemented on a real life sample. While implementing the model, many assumptions are made in literature which have been criticized by Loa and Lys (2000). We will avoid such assumptions, which can result in misleading inferences, while trying to keep the analysis as simple as possible. Our first assumption is about interest rate, which must be proper to discount the abnormal earnings. Based on the fact that equity of a firm is to be discounted in this model, we should employ a "cost of equity" rate. In Hail and Leuz (2005) the cost of equity for 40 different countries are estimated and for Austria they submit a cost of equity of 11.21%.

Furthermore, we replace the parameter γ in equation (9) with a constant value, say one. This assumption implies a linear information dynamic of abnormal earnings while it considers the "other information term" as a constant value affecting next period's abnormal earnings. So, LID becomes $\tilde{x}^a_{t+1} = \omega x^a_t + v_t$, where we assume $v_{t+1} = v_t$.

3.2. Results

Using least-square analysis, we estimate the persistence parameters in LID. The best fitting linear relation to the data available takes the following form where ω equals to 0.60245 and the other information term, v_t, to 53.5689.

$$\tilde{x}^a_{t+1} = 0.60245 x^a_t + 53.5689 \qquad \textbf{(12)}$$

	ω	v_t
Estimated value	0.60245	53.5689
St. deviation	0.074	13.4458
t statistic of ω	8.167758 ($T_{crit.}$=1.9944 for α=0.05)	
St. dev. of $\tilde{x}^a_{t+1}$	112.7985	
R^2	0.488	
F statistic	66.712 ($F_{crit.}$=3.9777 for α=0.05)	
Degree of freedom	70	

Table 2**:** Test statistic of calibrated model

The estimated model shows an explanatory power of 48.8% due to estimated coefficient of correlation. The t-test statistic approves that the coefficient ω is significantly different from zero and F-test verify that the estimated relation between next and current year's abnormal earnings is adequate. The test statistics related with calibrated model can be found above, in *Table 2*.

Given the equation above we can now specify the model parameters, α_1 and α_2, in market value equation in (11) and calculate intrinsic value of equity of each company for 2008 with the following relation.

$$P_{2008} = b_{2008} + 1.182 * x^{a}_{2008} + 19.466 * v_{2008} \quad \textbf{(13)}$$

Firm	P_{2008}	V_{2008}	$D=P_{2008}-V_{2008}$	P_{2008}/V_{2008}
Andritz AG	1688.45	944.32	744.13	1.788
BETandWIN.com AG	1153.76	434.91	718.85	2.653
EVN AG	3870.14	1798.50	2071.64	2.152
Flughafen Wien AG	1829.46	666.75	1162.71	2.744
Mayr-Melnhof Karton AG	1926.36	1116.06	810.30	1.726
OMV AG	10660.51	5593.54	5066.97	1.906
RHI AG	1327.85	426.38	901.47	3.114
Schoeller-Bleckmann AG	1312.74	347.77	964.96	3.775
Telekom Austria AG	2800.63	4738.00	-1937.37	0.591
Verbund AG	4403.05	10028.48	-5625.43	0.439
Voestalpine AG	5546.42	2472.58	3073.85	2.243
Wienerberger AG	3245.62	986.51	2259.11	3.290
Average	3313.749	2462.816	850.933	
Standard Deviation	2698.598	2937.475	2629.389	
t statistic			1.1211	
t value for 5% significance level			2.2010	

Table 3: Comparison of companies' Values with Ohlson Model and Market Capitalization

In Table 3, market values of ATX-companies calculated with Ohlson Model are listed while, as benchmarking data, the market capitalization, number of outstanding shares multiplied with end of year stock price of corresponding company, is used. One can infer that based on the provided data Ohlson Model generally results with a higher value than the value determined in the stock exchange. Hence, the market values of companies are overestimated by the model. However, t-statistic value implies that difference between two datasets is not to be considered statistically significant, despite the fact P/V ratio are mainly greater than 1 and even 3 in some cases.

The overestimated market values rest on the fact that stock prices on ATX have experienced vast declines in value because of the catastrophic financial crisis that shook off the whole financial markets at the end of 2008. It has diminished the prices up to 70 percent on most of the stocks on ATX in comparison with the prices in the previous year. Since the Ohlson Model relies on accounting data, the crisis will have a "delayed impact" on book value, while it "directly" affects stock markets.

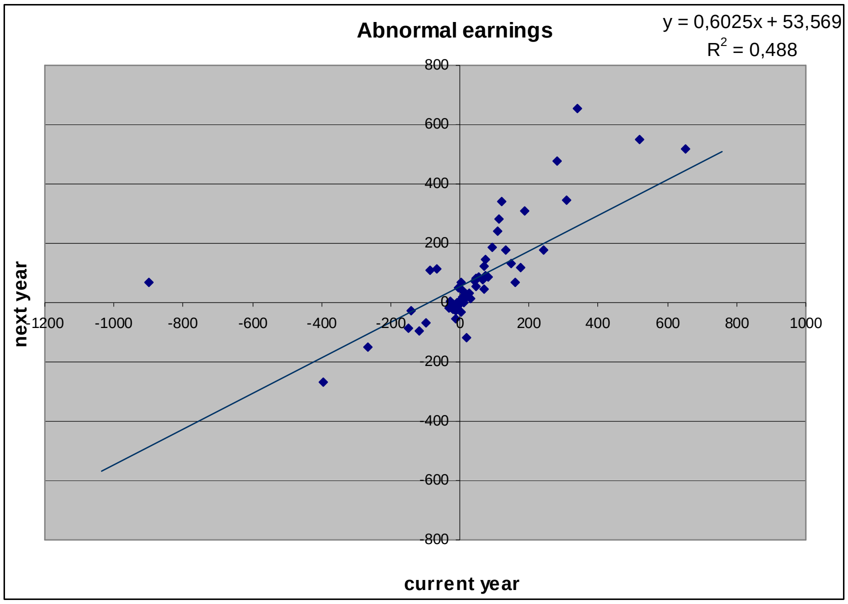


Figure 1: Abnormal earnings: comparison current and next year

Another factor which helps to explain the inconsistency in the estimations is unsatisfactory performance of some companies. As is depicted in Figure 1, annual abnormal earnings of many companies are negative and this reduces the estimation power of the model essentially. Thus, we conclude that cost of equity offsets the earnings, or even reverses them to economic losses.

In addition, the size of the dataset can be increased, in order to increase adequacy of the model. Instead of end of year data, collecting day-to-day accounting numbers would enhance the model's explanatory power.

Finally, a controversial improvement possibility is to group firms into branches or to generate firm specific models with firm's accounting numbers and firm specific cost of capital rates. But this, however, will complicate the model calibration.

3.3. Summary

Beside many critics, Olhson Model (1995) is accepted and applied in finance widely. Bridging firm's accounting numbers to its market value is the main advantage of the model while estimation of parameters which cannot be gathered from accounting data, such as "other information", represents a major obstacle on the empirical studies.

The model generated in this work has a moderate explanatory power on the market values of ATX companies. This limited adequacy relies basically on the financial crisis in the model's validation period, 2008. Other factors which may enhance the consistency of the model are increasing the sample size, regrouping the firms according to their field of activity and estimation branch or even firm specific parameters.

[1] Cupertino, César Medeiros and Lustosa, Paulo Roberto B., The Ohlson Model of Evaluation of Companies: Tutorial for Use. Brazilian Business Review, Vol. 1, No. 1, 2004. Available at SSRN: http://ssrn.com/abstract=70476.

[2] Dechow, Patricia M., Hutton, Amy P., Sloan, Richard G., An Empirical Assessment of the Residual Income Valuation Model. Journal of Accounting and Economics, Vol. 26, Issues 1-3, pp 1-34, (1999).

[3] Feltham, Gerald A., Ohlson, James A., Valuation and Clean Surplus Accounting for Operating and Financial Activities. Contemporary Accounting Research, Vol. 11, No. 2, pp 689-731, (Spring 1995).

[4] Gregory, Alan, Saleh, Walid and Tucker, Jon P., A UK Test of an Inflation-Adjusted Ohlson Model. Journal of Business Finance & Accounting, Vol. 32, No. 3-4, pp. 487-534, April 2005. Available at SSRN: http://ssrn.com/abstract=708657.

[5] Gordon, Myron J. "Dividends, Earnings and Stock Prices". Review of Economics and Statistics 41: 99–105 (1959).

[6] Hail, Luzi and Leuz, Christian, International Differences in the Cost of Equity Capital: Do Legal Institutions and Securities Regulation Matter? (December 2005). ECGI - Law Working Paper No. 15/2003; Rodney L. White Center for Financial Research Working Paper No. 17-04; AFA 2005 Philadelphia Meetings. Available at SSRN: http://ssrn.com/abstract=641981 or DOI: 10.2139/ssrn.641981.

[7] Lee, Charles M.C., Myers, James N. and Swaminathan, B., What is the Intrinsic Value of the Dow? (January 17, 1997). Available at SSRN: http://ssrn.com/abstract=376 or DOI: 10.2139/ssrn.37.

[8] Lo, Kin and Lys, Thomas Z., The Ohlson Model: Contribution to Valuation Theory, Limitations, and Empirical Applications (February 2000). Sauder School of Business Working Paper. Available at SSRN: http://ssrn.com/abstract=210948 or DOI: 10.2139/ssrn.210948.

[9] Ohlson, James A., The Theory of Value and Earnings, and An Introduction to the Ball-Brown Analysis. Contemporary Accounting Research, pp 1-19, (Fall 1991).

[10] Ohlson, James A., Earnings, Book Values, and Dividends in Equity Valuation. Contemporary Accounting Research, Vol. 11, No. 2, pp 661-687, (Spring 1995).

[11] Ohlson, James A., Positive (Zero) NPV Projects and the Behavior of Residual Earnings (June 2000). Available at SSRN: http://ssrn.com/abstract=232741 or DOI: 10.2139/ssrn.232741.

[12] Ota, Koji, A Test of the Ohlson (1995) Model: Empirical Evidence from Japan (November 1, 2001). Available at SSRN: http://ssrn.com/abstract=287513 or DOI: 10.2139/ssrn.287513.

[13] Spivey, Michael F., McMillan Jeffrey J., Using the Edwards-Bell-Ohlson Model to Value Small and Entrepreneurial Type Businesses (2003), http://www.aoef.org/papers/2003/mcmillan.pdf, (25.06.2009, 01:54).

[14] Spremann, Klaus., Valuation - Grundlagen moderner Unternehmensbewertung., München., Oldenbourg, 2004.

DELTA HEDGING AND RETURN

Miklavž Mastinšek
University of Maribor ,Faculty of Economics and Business
Razlagova 14, 2000 Maribor, Slovenia
mastinsek@uni-mb.si

Abstract: The problem of discrete-time hedging by a continuous-timeBlack-Scholes delta is studied. The paper considers a variation of hedging which depends on the time length of the rehedging interval. The question of transactions costs and the hedging return is considered.

Keywords: derivatives, delta hedging, transaction costs

1 INTRODUCTION

In a recent financial crisis the need for more regulation of trading financial derivatives appeared. Many European countries, U.S., and others have taken some immediate measures (like for instance temporarilly banning the short selling), in order to regulate the trading of financial derivatives. One of the reasons for a worldwide sensitivity of regulations is due to the huge value of these transactions.

In a recent report Bank for International Settlements (BIS) of Basel, Switzerland reveals that the global notional (nominal) amount outstanding of over the counter (OTC) derivatives in first half of 2008 has reached the astonishing value of 683,7 trillion U.S. $; see [1]. Fortunately these amounts provide only a measure of market size and not the true risk.

In order to reduce the risk for such highly leveraged contracts different hedging strategies are usually applied. The delta hedging is most widely used dynamic hedging technique used in practice. The delta hedging of options is a strategy with the intention to reduce (hedge) the risk associated with price movements in the underlying asset by offsetting long and short positions. In practice this means that from time to time the position in the underlying has to be readjusted and thus transaction costs has to be considered. Transaction costs due to delta hedging are obviously dependent on the frequency of hedging. More frequent hedging means more precise hedging (smaller hedging error), however the total transaction costs can easily increase significantly.

The option valuation problem with transaction costs has been considered extensively in the literature. In many papers on option valuation with transaction costs the discrete-time trading is considered by the continuous-time framework of the Black-Scholes-Merton partial differential equation (BSM-pde) ; see e.g. [4], [6], [10]. Since in continuous-time models the hedging is instantenuous, hedging errors appear when applied to discrete trading.

It is known that transaction costs can be included into the Black-Scholes-Merton equation by considering the appropriately adjusted volatility; see e.g. [4], [7], [10].

When the hedging is in discrete time, then over the time interval $(t, t+\Delta t)$ the number of shares N is kept constant while at the time point $t+\Delta t$ the number of shares is readjusted to the new value N’. Over that period of time the value S of the underlying changes to $S+\Delta S$.

The proportional transaction costs depend on the difference $|N'-N|$ which is usually approximated by the gamma term, in general the largest term of the associated Taylor series expansion. In the case when other partial derivatives of delta are not small compared to the gamma, higher order approximations can be considered. For a suitable choice of N which incorporates the time sensitivity of the delta, the expected proportional transaction costs can be reduced. We can show that in that case the order of the mean and the variance of the hedging error or return can be preserved.

2 TRANSACTION COSTS

Let us briefly give the details. For simplicity assume that the time Δt between successive rehedgings is constant. Suppose that at time t the portfolio consists of a long position in the option and a short position in N units of shares with the price S. Then over the time interval (t,t+Δt) the number of shares kept constant is usually given by: N= V_S(t,S); see e.g. [2], [3], [9]. At the time point t+Δt the portfolio is rehedged and the number of shares is adjusted to N' = V_S(t+Δt,S+ΔS). If N is given by $N = V_S(t,S)$, then the proportional transaction costs at rehedging t+Δt are equal to:

$$TC = \frac{k}{2}\left|N'-N\right|(S+\Delta S) = \frac{k}{2}\left|V_S(t+\Delta t, S+\Delta S) - V_S(t,S)\right|(S+\Delta S) \qquad (1.1)$$

where k represent the round trip transaction costs measured as a fraction of the volume of transactions; for the details see e.g. [6].

The absolute value of the difference ΔN=|N'-N| is usually approximated by $|V_{SS}\Delta S|$, in general the largest term of the Taylor series expansion.

If S=S(t) follows the geometric Brownian motion, then over the small noninfinitesimal interval of length Δt its change can be approximated by:

$$\Delta S = S(t+\Delta t) - S(t) \approx \sigma S Z\sqrt{\Delta t} + \mu S \Delta t\,, \qquad (1.2)$$

where Z is normally distributed variable with mean zero and variance one; in short Z~N(0,1), for the details see e.g. [5]. In that case the first order approximation of ΔN is given by the gamma term:

$$\Delta N = \left|N'-N\right| = \left|V_{SS}(t,S)\sigma S Z\sqrt{\Delta t}\right| \qquad (1.3)$$

see e.g. [Le].

Let us assume now that the discrete time adjusted hedge depends on the length of the rehedging interval and is of the form:

$$N = V_S(t+\alpha\Delta t, S) \quad where \quad 0 \le \alpha \le 1 \qquad (1.4)$$

For α=0, this is the ordinary Black-Scholes delta. For α=1 this is the Black-Scholes delta at the next rehedging time.

In that case the proportional transaction costs are equal to:

$$\begin{aligned}\Delta N &= \left|N'-N\right| = \\ &= \left|V_{SS}(t,S)\Delta S + (1-\alpha)V_{St}(t,S)\Delta t + \tfrac{1}{2}V_{SSS}(t,S)\Delta S^2 + O(\Delta t^{3/2})\right|\end{aligned} \qquad (1.5)$$

It can be shown that for different values of α lower expected transaction costs can be obtained; see e.g. [8]. Consequently this will affect the hedging error. We will show that the order of the mean and the variance of the hedging error in remains the same for $0 \le \alpha \le 1$.

3 HEDGING RETURN

We will now consider more closely the change of the value of a portfolio Π over the time interval (t,t+Δt). Suppose that the portfolio Π at time t consists of a long position in the option and a short position in N units of shares with the price S :

$$\Pi = V - NS \qquad (2.1)$$

We assume that the amount equivalent to the portfolio value can be invested in a riskless asset. Let us define the hedging error or the hedging return ΔH as the difference between the return to the portfolio value ΔΠ and the return to the riskless asset.

By assumption the price of the underlying follows the geometric Brownian motion so that (1.2) holds. Then the following result can be obtained:

Proposition 1 *Let σ be the annualized volatility and r the annual interest rate of a riskless asset . Let V(t,S) be the solution of the Black-Scholes-Merton equation:*

$$V_t(t,S) + \frac{1}{2}\sigma^2 S^2 V_{SS}(t,S) + rSV_S(t,S)) - rV(t,S) = 0 \ , \tag{2.2}$$

If the approximate number of shares N held short over the rebalancing interval of length Δt is equal to:

$$N(t) = V_S(t + \alpha\Delta t, S) \quad , \tag{2.3}$$

for some $0 \le \alpha \le 1$, then the mean and the variance of the hedging error is of order $O(\Delta t^2)$.

Proof Let us consider the return to the portfolio Π over the period (t, t+Δt), $t \in [0, T_0 - \Delta t]$, where T_0 is time at option expiry. By assumption over the period of length Δt the value of the portfolio changes by:

$$\Delta\Pi = \Delta V - N\Delta S \tag{2.4}$$

as the number of shares N is held fixed during the time step Δt.

First we consider the change ΔV of the option value V(t,S) over the time interval of length Δt. By the Taylor series expansion the difference can be given in the following way:

$$\begin{aligned} \Delta V &= V(t+\Delta t, S+\Delta S) - V(t,S) = \\ &= V_S(t,S)(\Delta S) + \frac{1}{2}V_{SS}(t,S)(\Delta S)^2 + \frac{1}{6}V_{SSS}(t,S)(\Delta S)^3 + \\ &\quad + V_t(t,S)(\Delta t) + O(\Delta t^2) \end{aligned} \tag{2.5}$$

Note that

$$V_S(t + \alpha\Delta t, S) = V_S(t,S) + \alpha V_{St}(t,S)\Delta t + O(\Delta t^2) \tag{2.6}$$

By (2.4) it follows:

$$\begin{aligned} \Delta\Pi &= \Delta V - N(t)\Delta S = V_t(t,S)(\Delta t) + \left[V_S(t,S) - N(t)\right](\Delta S) + \\ &\quad + \frac{1}{2}V_{SS}(t,S)(\Delta S)^2 + \frac{1}{6}V_{SSS}(t,S)(\Delta S)^3 + O(\Delta t^2) \end{aligned} \tag{2.7}$$

When the number N of shares is equal to:

$$N = V_S(t + \alpha\Delta t, S) \approx V_S(t,S) + \alpha V_{St}(t,S)\Delta t \qquad , \tag{2.8}$$

we get:

$$\begin{aligned} \Delta\Pi &= V_t(t,S)(\Delta t) + \frac{1}{2}V_{SS}(t,S)(\Delta S)^2 + \\ &+ \frac{1}{6}V_{SSS}(t,S)(\Delta S)^3 - (1-\alpha)V_{St}(t,S)\Delta t\Delta S + O(\Delta t^2) = \\ &= V_t(t,S)(\Delta t) + \frac{1}{2}V_{SS}(t,S)(\sigma^2 S^2 Z^2 \Delta t + 2\sigma\mu S^2 Z \Delta t^{\frac{3}{2}}) + \\ &+ \frac{1}{6}V_{SSS}(t,S)\sigma^3 S^3 Z^3 \Delta t^{\frac{3}{2}} - (1-\alpha)V_{St}(t,S)\sigma S Z \Delta t^{\frac{3}{2}} + O(\Delta t^2) \end{aligned} \tag{2.9}$$

By assumption Z~N(0,1) so that $E(Z)=E(Z^3)=0$ and $E(Z^2)=1$. Hence the expected value of $\Delta\Pi$ is equal to:

$$E(\Delta\Pi) = V_t(t,S)\Delta t + \frac{1}{2}V_{SS}(t,S)(\sigma^2 S^2 \Delta t + O(\Delta t^2) \tag{2.10}$$

By assumption the amount Π can be invested in a riskless asset with an interest rate r. Thus over the rehedging interval of length Δt the return to the riskless investment is equal to:

$$\begin{aligned}&\Pi r\Delta t = (V(t,S) - NS)r\Delta t = \\ &= \left[V(t,S) - V_S(t+\alpha\Delta t, S)(S)\right]r\Delta t + O(\Delta t^2) = \\ &= \left[V(t,S) - V_S(t,S)(S)\right]r\Delta t + O(\Delta t^2) =\end{aligned} \quad (2.11)$$

The hedging error is equal to: ΔH=ΔΠ-ΠrΔt By (2.10) and (2.11) the expected value of ΔH is equal to:

$$\begin{aligned}&E(\Delta H) = E(\Delta\Pi - \Pi r\Delta t) = V_t(t,S)\Delta t + \frac{1}{2}V_{SS}(t,S)(\sigma^2 S^2 \Delta t) - \\ &\quad - \left[V(t,S) - SV_S(t,S)\right]\rho\Delta t + O(\Delta t^2)\end{aligned} \quad (2.12)$$

Therefore, when V(t,S) satisfies at t+Δt the BSM equation, the hedging error can be written as:

$$\begin{aligned}&\Delta H = \frac{1}{2}V_{SS}(t,S)(\sigma^2 S^2 (Z^2 - 1)\Delta t + 2\sigma\mu S^2 Z \Delta t^{\frac{3}{2}}) + \\ &\quad + \frac{1}{6}V_{SSS}(t,S)\sigma^3 S^3 Z^3 \Delta t^{\frac{3}{2}} + O(\Delta t^2)\end{aligned} \quad (2.13)$$

Hence the mean and the variance of ΔH are zero to the order of $O(\Delta t^2)$:

$E(\Delta H) = O(\Delta t^2)$ and $V(\Delta H) = O(\Delta t^2)$. □ (2.14)

References

[1] Bank for International Settlements, Monetary and Economic Department, »OTC derivatives market activity in the first half of 2008. http://www.bis.org/publ/otc_hy0811.pdf.

[2] Black F. and Scholes M., »The pricing of options and corporate liabilities«, J. Pol. Econ. 81, (1973) ,637-659.

[3] Boyle P. and Emanuel D., »Discretely adjusted option hedges«, J. Finan. Econ. 8 (1980), 259-282.

[4] Boyle P. and Vorst T., »Option replication in discrete time with transaction costs«, J. Finance 47 (1992), 271-293.

[5] Hull J.C., Option, Futures & Other Derivatives, Prentice-Hall, New Jersey, (1997).

[6] Leland H.E., »Option pricing and replication with transaction costs«, J. Finance 40 (1985), 1283-1301.

[7] Mastinšek M. "Discrete-time delta hedging and the Black-Scholes model with transaction costs", Math. Meth. Oper. Res. 64 (2006), 227-236.

[8] Mastinšek M. "Expected transaction costs and the time sensitivity of the delta". *SOR '07 proceedings*. Zadnik Stirn L.(ur.), Drobne S.(ur.) Ljubljana: Slovenian Society Informatika (SDI), Section for Operational Research (SOR), 2007, str. 307-312.

[9] Merton R.C., »Theory of rational option pricing«, Bell J. Econ. Manag. Sci. 4 (1973), 141-183.

[10] Toft K.B., »On the mean-variance tradeoff in option replication with transactions costs«, J. Finan. Quant. Analysis, Vol. 31, 2 (1996), 233-263.

APPLYING MARKOV CHAINS IN BONUS SYSTEM

[1]**Ivana Simeunović,** [2]**Aleksandar Zdravković** and [3]**Jelena Minović**
[1,3]Union University of Belgrade, Belgrade Banking Academy
[2]Institute of Economic Sciences, Belgrade
Zmaj Jovina 12, 11 000, Belgrade, Serbia
{isimeunovic, aleksandar.zdravkovic, jminovic}@ien.bg.ac.rs

Abstract: The process of determining the premium inevitably bears a certain risk. It has stochastic character. One way in the procedure of insurance premium calculation is experience based calculation of premium. Among all the methods in practical research, No Claims Discount System – NCD System, holds the important place. A basic idea is contained in analysis of accumulated claims process that is one of the most common applications of Markov chains.

Keywords: Bonus System (No Claims Discount System), Markov chains, claim.

1 INTRODUCTION

One of the methods used in calculating the insurance premium is experience-based premium rating.

This method relies on the hypothesis that in order to calculate the insurance premium it is necessary to analyze losses of the insured in the past.

Reasons for introducing such method of premium calculation derive from the fact that, even when risks can be classified according to their factors of identification, certain heterogeneity within a risk family has been retained. Thus, there are many situations in which the amount of incurred damage per a policy is almost equal to amount expected for the whole family of risk, while on the other hand, there are many examples of insurance policy whose experience will be significantly different from the expected value. We will try to resolve this problem by more precise classification of risk families using their additional factors. Unfortunately, this approach would soon turn to be quite impractical, if applicable at all. A more acceptable resolution is experience-rating of premium, where the premium does not only depend on experience of the whole family, but on individual risk which is being rated.

Most frequently used methods of experience-rating of the premium in practice is the bonus system (NCD system) and credibility theory.

2 BONUS SYSTEM

When calculating the premium, the insurer's decision is often founded on consideration of the number of claims submitted by the insurance holder in the past.

This method, known as ***bonus system***, grants discount on the main amount of the premium which resulted from the achieved result of the insurance holder.

When bonus system is applied, the insurance holder is granted a discount on a usual premium, which directly depends on the number of years without claims. In the procedure of determining whether the individual insurance holder has the right to be granted a discount, it is common that the damage which is caused by the responsibility of some other person has to be ignored, due to which the bonus is often called the discount for non-existence of guilt.

Application of the described system is common in those types of insurance in which realization of risk greatly depends on behavior and characteristics of the insurance holder, as it is the case in motor vehicles insurance. Almost all insurance companies have used this system. Also, in obligatory insurance the bonus should be justified in advance for the following year, according to the period for which the insurance contract has been signed. This method can be applied in collective life assurances and health insurance.

Elements of bonus system are the following: discount category and transition rules of passing from one category to another. When categories are concerned, they refer to the number of years without indemnity claim. Transfer from one category to another is defined by the insurer according to previous experience of the insurance holder.

For example, there is a bonus system with three categories (category 0 – without discount, category 1 – 20% discount and category 2 – 30% discount). Category 0 includes these insurance holders who pay full amount of premium (in practice it will differ from case to case depending on personal characteristics of the insurance holders, e.g. age, gender, etc). Category 1 entails the insurance holders who pay 80% from the full amount of premium, while the last category entails policy portfolio of these insurance holders who pay 70% of the full amount of the premium.

Transition rules from one category to another are as follows: if the insurance holder has no indemnity claims during the year he is transferred to the category of bigger discount, while if there is one or more claims, the insurance holder is transferred to the lower category.

Bonus system in motor vehicles insurance, where the amount of premium is based on previous claims, is one of the basic applications of the Markov chains.

3 APPLICATION OF MARKOV CHAINS IN BONUS SYSTEM

3.1 Main features of Markov chains

Any stochastic process represents a model of some random event which is time dependant. As a random variable describes some random event, so the stochastic process is a family of random variables X_t, for each time moment t.

The total number of random events is marked with S. According to the example of the analyzed insurance company, which offers three different levels of discount, there is the set S = (0, 1, 2) where each element represents the level of appropriate discount.

The example which is being analyzed represents one stochastic process which has a discrete number of random events.

It is said that a stochastic process has Markov's characteristics if the number of random events in the future could be predicted exclusively through current number of random events, without the usage of the data from the past, which can be written in the following way [3]:

$$P\left[X_t \in C \middle| X_{S_1} = x_1, X_{S_2} = x_2, ...X_{S_n} = x_n, X_S = x\right] = P\left[X_t \in C \middle| X_S = x\right] \qquad (1)$$

for all $S_1 < S_2 < .... < S_n < S < t$, all random events $x_1, x_2.........x_n, x \in S$ and all sub-sets $C \subseteq S$.

Markov's process with discrete number of random events and discrete time is called Markov's chain.

3.2 Application in insurance

Discount status of the insurance holder makes one Markov's chain. If there are three levels of discount, as it is described in the given example, the set of random events would be: S = (0, 1, 2).

If we assume that probability that the insurance holder would not have any claim during one year equals $\frac{4}{5}$ then Markov's chain which describes this process can be represented by a transition graph or by a transition matrix.

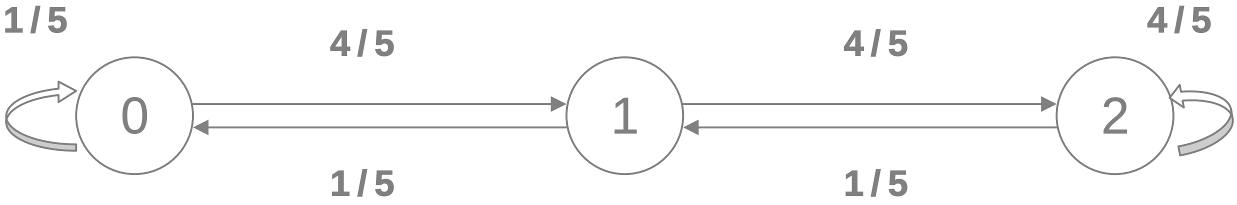


Figure 1: Transition Graph No. 1

It is evident that for any value of probability $p > 0$, the arrow points out that direct transfer form one random event into another is possible. Value $\frac{4}{5}$ represents probability that the insurance holder will not have claim, thus providing opportunity for his transfer to the level of bigger discount. However, the insurance holder will descend form a certain level of discount to lower level, according to the adverse probability:

$$q = 1 - p = 1 - \frac{4}{5} = \frac{1}{5} \tag{2}$$

Transition matrix represents square matrix in the form of $n \times n$, where n represents the number of random events in the set S which is the set of the total number of random events.

In this way we can form the transition matrix of the insurance holder from category *i* into category *j* year in year out.

$$P = \begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdot & \cdot & p_{0n} \\ p_{10} & p_{11} & p_{12} & \cdot & \cdot & p_{1n} \\ p_{20} & p_{21} & p_{22} & \cdot & \cdot & p_{2n} \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \cdot & \\ p_{n0} & p_{n1} & p_{n2} & & & p_{nn} \end{bmatrix}$$

where p_{ij} represents the above mentioned probability of transition of the insurance holder from *i* into *j* category.

In the analysis of different levels of discount with three random events, we would have the following:

$$P = \begin{bmatrix} \frac{1}{5} & \frac{4}{5} & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} \\ 0 & \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

Transition matrix provides easy calculation of probability of maximum discount. So the insurance holder who is on a zero level of discount achieves the maximum amount of discount with greater probability:

$$p_{02} = p_{01} \cdot p_{12} = \frac{4}{5} \cdot \frac{4}{5} = \frac{16}{25} \tag{3}$$

Let us now analyze the insurance company for motor vehicle insuring which grants four levels of discount to its insurance holders:

Table 1: Bonus system with four categories

Category	**Discount**
0	0
1	30
2	45
3	65

Transition rules for moving from one category into another are as follows:

If the insurance holder submits claim in the current year, his discount status descends one level down provided that he had no claims in the previous year, that is two levels down if he had claims in the previous year. Besides, non-existence of claims results in ascending to higher level of discount, or keeping up maximum discount if the insurance holder has already reached that position in ranking.

It can be noted that such a chain of random variables X_i - categories of discount does not make Markov's chain in the set of random events S = (0,1,2,3), having in mind the following:

$$\begin{aligned} P(X_{n+1} = 0 | X_n = 2, X_{n-1} = 1) = 0 \\ P(X_{n+1} = 0 | X_n = 2, X_{n-1} = 3) > 0 \end{aligned} \tag{4}$$

However, it is necessary to apply analysis of another category of discount $X_n = 2$.

This category can be reached in two ways:

1) by non-existence of the claim from category $X_n = 1$ or
2) if the insurance holder who is at the third level of discount submits claim during the current year.

Yet, the possessing of this information will make the observed chain of random variables a Markov's chain. The following signs have been introduced:

$2\uparrow$: the insurance holder is granted 45% discount stating that there was no claim during the previous year.

$2\downarrow$: the insurance holder is granted 45% discount, but he submitted claim in the previous year.

Finally, in order to form Markov's chain, we have to assume that the probability of the insured case also equals 1/5 this time.

Graph of transition for the observed process can be presented in the following way:

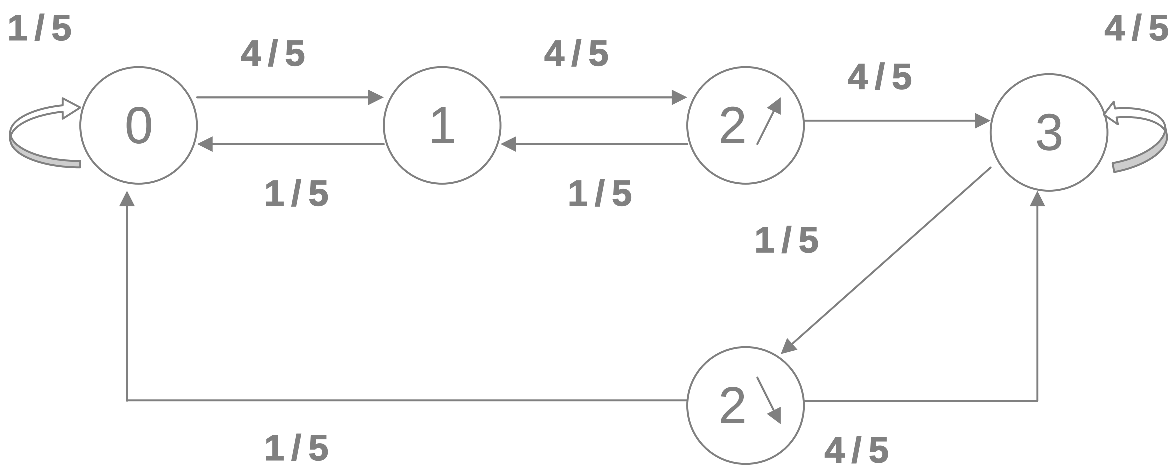


Figure 2: Transition graph No.2

Transition matrix for the stated process would be the following:

$$P = \begin{bmatrix} \frac{1}{5} & \frac{4}{5} & 0 & 0 & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 & \frac{4}{5} \\ \frac{1}{5} & 0 & 0 & 0 & \frac{4}{5} \\ 0 & 0 & 0 & \frac{1}{5} & \frac{4}{5} \end{bmatrix}$$

The stated matrix is 5×5 for 5 different random events can be achieved, or

$S = (0,1,2\uparrow,2\downarrow,3)$.

Using this matrix or the above mentioned graph enables an easy calculation of all desired probabilities of transition of the insurance holder from one category to another. So, the probability that the insurance holder who is at the zero level of discount will reach the maximum discount, is equal to

$$p_{03} = p_{01} \cdot p_{12\uparrow} \cdot p_{2\uparrow 3} = \frac{4}{5} \cdot \frac{4}{5} \cdot \frac{4}{5} = \frac{64}{125} \tag{5}$$

It is important to mention that Markov's chains and transition matrix can be used for estimation of the expected number of the insurance holders at all levels of discount in any year. Let us mark the expected number of the insurance holders in the first year of the observed time chain with p_0, assuming that all insurance holders start with paying the total amount of premium without discount. That means that p_0= 1. Expected number of the insurance holders in the first year at all levels of discount can be represented by vector $p_1 = (1,0,0,...)$. In the second year, vector which represents the expected number of the insurance holders at all levels would be $p_2 = p_1 \cdot P$, where P represents transition matrix.

Finally, a conclusion can be drawn that the expected number of insurance holders in year n at all levels of discount is calculated by multiplying the appropriate vector from the previous year with the transition matrix which describes the Markov's chain of random process.

Application of bonus system suffers serious shortages. Namely, in deciding whether to submit claim or not the insurance holder will often resolve problem on his own, because of the consequences implied in calculation of the premium for the next year. In other words, the insurance holder will not submit claim if it is lower than the amount of the future increased premium. This situation is unfavorable for the insurer who, in that case, will not be able to rank certain risk.

On the other hand, with the application of bonus, the premium is adjusted to individual risk to a great extent, which is a tactical and psychological measure, for it is proved that the insurance holders who are granted a bonus in advance have far less claims than other insurance holders [2]. So to speak, the probability of damage often depends on the level of discount of the individual insurance holder.

Let us now analyze various situations which may occur if bonus system is applied in calculation of the premium and which resulted from the various decisions taken by the insurance holders on submitting claims.

The main reason for not submitting the claim refers to the level of the increased premium in future. This fact can be easily understood if we compare differences in the amounts of premium in case when the insurance holder either submits the claim or decides not to do so.

If we look back at the example which was analyzed above we can find out the following:

The insurance company offers three levels of discount: 0%, 20% and 30% and the full amount of premium is 500 EUR.

If the insurance holder does not submits claim in the first year (or in the following years) his future premiums would amount as follows:

400€ 350€ 350€

On the other hand, if there is a damage claim in the first year, future premiums would amount as follows:

500€ 400€ 350€

The difference in premiums which the insurance holder will pay in the second example is:

$$(500 + 400 + 350) - (400 + 350 + 350) = 150 \text{ €} \tag{6}$$

In this way, the differences in premiums in all categories of discount can be calculated. For example, the insurance holder who is in category with 20% discount, that is at the first level of discount, and who at that moment had an accident and suffered damage, would consider the following two situations:

I If he does not submit a claim, the amount of his future premium would be as follows:
- 400 € (because he is in the first category of discount)
- 350 € (having assumed that there was no damage claim in the previous year)
- 350 € (assuming that he would not suffer damage)

II If he submits claim, the insurance holder will pay the following three premiums:
- 400 € (because he is at the first level of discount at the beginning)
- 500 € (because he submitted a claim, and goes back to the zero level of discount)
- 400 € (having assumed that in the previous period he did not suffer damage, so that he can return to the first level of discount)

The difference in premiums for the three-year period which is the result of different decisions taken by the insurance holder amounts to:

$$(400+500+400)-(400+350+350)=200 \tag{7}$$

Note: In the analysis of the difference in the amount of premiums – (6) and (7), that are the consequences of different decisions on the application the insured damage caused, time dependence of the value of money is not included.

Calculated differences are the result of the observation of the future period until the year when the insurance holder would reach maximum discount. The number of these years would significantly influence the decision whether or not to submit the claim.

However, there is a significant difference between probability that the insurance holder would suffer damage and probability that he would submit the claim.

In the case described above, it can be concluded that the insurance holder would submit the claim if its amount is higher than 200 EUR. The conclusion can be generalized in the following way:

X stands for a random variable defined as: X – claim amount. Y refers to difference in premiums caused by different decisions of the insurance holders on whether to submit claim. Assuming that the distribution of a random variable X is known which determines the claim amount, probabilities of submitting claim can be easily calculated at certain levels of discount. We would have the following:

$$P(\text{ submitting claim } \mid \text{ claim (accident) }) = P(X > Y) \tag{8}$$

4 CONCLUSION

One of the reasons which justifies and encourages the use of bonus method is very simple, so to say, automatic calculation of premium. In other words, the insurance holder who submits fewer claims, would pay less than the insurance holder who submits claims more often. Although this seems very simple, it would soon turn out that the amounts of premiums which the insurance holders pay are not proportional to the claim amounts. The reason for this lies in a small number of discount categories and in small differences among granted discounts.

Application of bonus represents an inevitable part of certain premium systems, as it is the motor third party liability. The main aim which should be achieved by application of this method is decrease in claims. Numerous research show that all insurance holders who were granted a bonus have fewer number of claims if compared with other insurance holders.

Advantage of using bonus system reflects in possibility to remove smaller claims. Its application reduces claims costs, or even more important, it reduces costs of their processing. Reducing the number of claims decreases the costs which the insurance holder has to pay.

The use of bonus system justifies competitiveness in calculating insurance premium and contributes to its more correct calculation.

References

[1] Bowers, N., Gerber, H., Hickman, J., Jones, D., Nesbit, C., 1986. Actuarial mathematics", The society of actuaries, Itasca.

[2] Hossak, I.B., Polard, J.H., Zehenwirt, B., 1999. Introductory statistics with applications in general insurance 2end edition, Cambridge University Press cop.

[3] Krdžić, I., 2007. Matematičko-statistički osnovi utvrđivanja premije imovinskog osiguranja, magistarska teza, Beograd.

[4] Marović B., Avdalović V., 2003. Osiguranje i upravljanje rizikom, Birografika, Subotica.

[5] Rolski T., Schmidli V., Schmidt V., Teugels J., 1998. Stohastic process for insurance and finance, Wiley.

[6] Simeunović, I., 2008. Iskustveno određivanje premije-Bonus sistem, Industrija, Ekonomski institut, Beograd.

[7] Vaughan, E., Vaughan, T., 2000. Osnove osiguranja, upravljanje rizicima, Mate, Zagreb.

REAL OPTION ANALYSIS - DECISION MAKING IN VOLATILE ENVIRONMENT

Ivana Tomas, dipl.oec.,
Faculty of Economics, University of Rijeka,
Ivana Filipovića 4, 51000 Rijeka, Croatia
E-mail: ivana.tomas@efri.hr

Josipa Višić, dipl. oec.,
Faculty of Economics, University of Split,
Matice hrvatske 31, 21000 Split, Croatia
E-mail: josipa.visic@efst.hr

Abstract: Since the capital investments absorb significant amount of cash and determine future business activity, valuation of investment project is considered to be one of the most important decisions companies must make. For the last two decades developed countries have been using contemporary methods of valuation of investment projects that include the possibility of adjusting future decisions depending on the conditions in the business environment.

Capital investment efficiency assessment cannot be a reliable or an objective basis for making investment decisions unless it also takes into account real option analysis. Although real option approach questions reliability of traditional methods of valuation of investment projects, it should primarily serve as a supplement rather than a substitute for traditional methods of valuation of investment projects aiming at more successful investment decisions making.

Keywords: investment, risk, uncertainty, discounted cash flow, net present value, real option.

1 INTRODUCTION

During the process of evaluating acceptability of an investment project, traditional discounted cash flow method presume that company will hold its assets passively i.e. they ignore the adjustments company could make after the project has been accepted and implemented. Adjustments which allow flexibility of changing old decisions when project related conditions change, are known as *real options*. The term *real option* has been introduced by Steward Myers in year 1977. According to Myers, evaluation of investment possibilities using net present value method ignores the value of the option which arises from the uncertainty which is inevitable in every project. A decade later, real option method has spread to investment decision making, partially due to contribution of numerous authors, such as: Dixit (1989), Dixit and Pindyck (1994), Pindyck (1988, 1991), Brennan and Schwartz (1985), Kemma (1988, 1993), Sick (1989), Trigeorgis (1988, 1991, 1993, 1995, 1996), Ingersoll and Ross (1992), Myers and Maid (1990), Maboussin (1999), etc.[19].

Unlike traditional methods of investment projects valuation which treat investment project as a static process determined by basic variables and projected cash flow, contemporary methods regard an investment as a dynamic process that gives the management the flexibility in decisions making involving future investment projects. In fact, binding the chosen scenario to the life cycle of the investment project is not consistent with "real life" conditions, so it is almost certain that management will react to certain occurrences that could appear after the acceptance of investment project. If the unexpected circumstances that occur are favorable, management will try to use the given opportunities in order to increase net present value of a project. Also, if unexpected circumstances are unfavorable it is hard to believe that there will be no reaction of managers [14].

According to Dixit and Pindyck [9] net present value of a project is generally easy to calculate but it is based on false assumption that investment is irreversible and impossible to postpone. In fact, even if a project has a positive net present value, it does not necessarily mean it should be implemented right away. Sometimes delaying the implementation of a project can additionally improve its value. Techniques of real options evaluation can assess components of managerial adaptability that are hard to perceive and which are neglected or underestimated by traditional approach [18]. Traditional approach does not include existence of additional adjustments to circumstances that could not be perceived during analysis of financial efficiency of a project. Therefore, it is impossible to ignore the assumption that traditional (conventional) project evaluation methods actually underestimate the value of a project [14].

2 TYPES OF OPTIONS ON REAL ASSETS

When making investment decisions, real option approach includes risk and uncertainty management with the help of the various options that can be used or abandoned, depending on future changes in markets and technology. Greater the uncertainty surrounding the investment project, greater is the possibility that the option will be exercised, thus making its value higher. Some of the available real options are: 1) option to expand, 2) option to abandon, 3) option to change inputs and outputs and 4) timing option.

1) Option to expand - Option to expand allows expansion of production if market conditions become favorable. It appears in the infrastructural and strategic industries, high technology industry, research and development, IT and pharmaceutical industry. The value of this sort of project comes not from within, but from the opportunities that growth provides. For example, the first generation of products of high technology even with a negative net present value, can serve as a foundation for lower costs and proven quality in products of next generations. Option to expand has characteristics of the call option. It offers a limited loss in the amount of negative net present value of initial suboptimal (pilot) investment. Limitation of the maximum loss reduces the risk of the investment project which would occur if overall investment in the amount of optimal capacity was immediately taken. In this way, the option to expand appears as a kind of call option, which provides limited loss in unfavorable circumstances, and significant profit prospects in favorable [14].

2) Option to abandon - If the project causes financial loss company can activate the option to abandon the project. While the abandonment may seem like an act of cowardice, this option often saves the company from large financial losses. Therefore the value of the project can be increased if it has the option to abandon [5].Option to abandon can be evaluated as an American put option on the value of the project with the cost of realization; this is practically equal to the rest of the value, or the value of the best alternative use of the assets. When the present value of the project sinks below the liquidation value, assets can be sold, which is actually a realization of the put option. This option is very important for the large, capital intensive projects such as nuclear plants, air and rail traffic.

3) Option to change inputs and outputs - Many manufacturing sectors have built-in flexibility of production, depending on changes in demand. The possibility of using different inputs (option to change inputs) in order to manufacture a certain output is significant in sectors such as agriculture, power generation and chemical industry [15]. In industries where the demand is variable and dominated by a small series of products, such as automotive and electronic industries, the ability to adjust is offered with option to produce different output with the same input. Option to change the inputs is particularly valuable in terms of frequent changes in oil price (companies no longer want to depend on one source of raw materials so they resort to the use of cheap gas boilers or dual gas/oil boilers that are less sensitive to the oscillations of the oil price). Additional cost of investments in flexible system is justified if the option to switch energy sources has a greater value of fixed variants of usage of energy [5] and [1].

4) Timing option - Timing options as well as other options are directed to reduction of the risk of investing in real investment projects facing significant uncertainty. While waiting, investment project is frozen in the existing stage of its life, with the intent to be activated when/if more favorable circumstances appear. There is no commitment for management to activate the investment project if favorable circumstances do not appear [14]. When decision on investment can not be delayed, or when it is a "now or never" investment, net present value (traditional tool) of the project equals the value of the project calculated using real option methods. If there is no possibility to delay investment, the standard deviation as a measure of volatility of a project does not influence the calculation of the value of an option and in this case, the value of the project is the same as the call option [15]. However, if there is a possibility to delay an investment, net present value of the project will differ from the value of the project obtained by real option method due to following reasons: the time value of money (money today is worth more than same amount in the future) and the value of assets invested (if asset value increases, the decision on investment is valid, and if asset value decreases, the project is rejected) [15].Timing option can be used when considering the large investment projects that show a large probability of realizing a negative net present value. Therefore, it is necessary to compare the value of the lost cash flow caused by delay in investment with the possibility of obtaining valuable information [14].

3 EVALUATION OF REAL OPTIONS

Traditional methods of discounted cash flows are relatively easy to understand since the theoretical framework of this method is clear, and the concept of time value of money is the only prerequisite for their understanding. In comparison with the traditional tools of projects evaluation, real option analysis is considerably more complex and requires a higher degree of mathematical understanding.

In the real option approach, investment opportunities are observed as financial options (derived securities). Namely, the option (call or put) represents the right (but not the obligation) of purchase or sale of any security, on a pre-determined price. Following the logic of financial options, the investor has the opportunity (the right, but not the obligation) to invest in a project. If you invest, you realized that option. If you do not "enter" in the project, the option remains the unused portions of the right since the minimum price of the option is 0 (zero),

$$C = \text{Max}\,(S-X,0),$$

where:

C = the value of call option
S = market price of one share,
X = price of option's realisation

Investment opportunity, being right, but not the obligation of investment in a project, cannot diminish the value of the project (minimum value is equal to zero). Therefore, the value of real options (investment opportunities) is always positive, or zero, and can only increase the value of investment projects. With all the above mentioned, several additional features, which are offered to the investor, make the investment project more flexible and more valuable. [17].

If there was a call option sufficiently similar to investment project, its value would indicate the value of an investment project. Hence, when setting up models of real option valuation, it is necessary to translate variables of European call option directly into “real” investment analogs as shown in Table 2:

Table 1: Comparison of financial and real options variables

CALL OPTION	***VARIABLE***	***INVESTMENT OPPORTUNITY***
Stock price	$\mathbf{S_0}$	Present value of project's Free Cash Flow
Exercise price	**X**	Expenditure required to acquire project assets
Time to expiration	**T**	Length of time the decision may be deferred
Risk-free interest rate	**r**	Time value of money
Standard deviation	σ	Riskiness of project assets

Source: Luehrman, Timothy A., 1998., p. 53. [11].

The analogy between the realization of financial call options, and taking investment, refers to the analogy of owning call options and the ownership of investment opportunity (a project). When the investor decides to invest, he is actually exercising the option. In this case, the investment cost is a price of an option's exercise, while the assets in which he invests, i.e. the future value of investments shown through discounted cash flow, is a share price equal to the reference option [15].

Real option valuation method of investment projects is actually the extension of the theory of financial options on real property. Then as for financial options, the value of real options depends on the five basic variables and significant sixth, which includes the lost cash flows due to the competition [6] and [12].

Depending on available input data, there are several ways which make possible to reach the value of option that includes investment project. For the calculation of the value of options it is necessary to have the option valuation model - a formula with which it is possible to obtain the value of options. The two best known and most widely used Option calculators are Black & Scholes option valuation model and the Binomial option valuation model.

3.1 Black&Scholes option valuation model

The most common model based on partial differential equations and used for the evaluation of financial and real options is the Black & Scholes option evaluation model [11]. Fisher Black, Myron Scholes and Robert Merton set in 1973. a well known Black & Scholes formula for evaluation of derivatives. Formula can be used for valuing options on stocks, currencies, real property, and it is used by options traders, investment bankers and financial managers.The basic model is based on the following assumptions [4] and [3].No payment of dividends in the period to options maturity;

- The model does not take into account transaction costs or taxes;
- Non-risk interest rate is constant through time;
- Options can be carried out only at its maturity (European option);
- Return on shares has the lognormal distribution, which means that it is a natural logarithm of return on a stock ln (1 + r) in the form of the normal curve.

This model was initially developed for assessing the value of call options on a stock. The theory of financial options, within which is developed Black & Scholes model for valuing financial options needs to be adapted to be applied to real property. Initial model can be applied to evaluating investments in real property since the investment project, ie, investment opportunity, is similar to call option because it allows the right, but not the obligation to invest. Value of call option (C) shown with Black&Scholes formula:

$$C = S_0 * N(d_1) - \frac{X}{e^{rt}} N(d_2)$$

$$d_1 = \frac{\left[\ln\left(\frac{S_0}{X}\right) + \left(r + \frac{1}{2}\sigma^2\right) * T\right]}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

Given variables in the equations are: C = value of call options, S0 = present value of future cash flows of assets in which we invest, X = price realization of option or the cost of the investment, r = non-risk interest rate for continuous compounding, T = time of maturity of the option σ = standard deviation, d_1, d_2 = deviation from the expected value of the normal distribution, $N(d_1)$ and $N(d_2)$ = the probability that a standarized, normally distrubuted random variable will be less than or equal to *d*, ln = natural logarithm e = 2.71828, base of natural logarithm [16].

The easiest way to calculate the value of options is by using Black & Scholes equation in which is easily to identify S0, X, T and r variables, while the $N(d_1)$ and $N(d_2)$ can be calculated using Microsoft Excel. Standard deviation (σ), which expresses the risk that occurs when investing in stocks, can be found in the financial markets or in the financial statements. However, when evaluating options that appear when investing in real property, determining the value of standard deviation is more difficult. According to Orsag, the possibility of perception of risk through the distribution of probabilities is possible to conduct in scenario analysis based on a small number of cases, computer simulations or by breaking the life cycle of the project on the specific phases of decision-making tree, thus determining specific scenarios of operation of each stage [14]. Moreover, Luehrman lists several approaches of evaluation of standard deviations [11]. One of them is the preference value of the standard deviations, where the value of standard deviation will be higher in projects that have greater market risk and using more discount rates. In the last 15 years, the standard deviation of market portfolio of U.S. shares, which are included in the formation of the most known joint-stock index was 20%. With investment projects, it can be expected even higher standard deviation; on the American market, depending on the risk, it is between 30% and 60% a year. Another approaches are related to determining applied volatility by using historical data and simulation of the standard deviations using Monte Carlo simulation techniques.

3.2 Binomial option valuation model

The second most popular model is the Binomial option valuation model. Binomial model looks like a decision tree in which the possible values of the basic property change depending on time of option's maturity. This model tracks the movement of asset prices as a binomial process in which assets can move in two possible directions, i.e. may fall or increase. The changes in the property value are marked with *u* and *d* factors, where u> 1 and d <1 [7].Basic assumptions of the Binomial option valuation model are as follows [2]:

- The market is efficient: all the relevant information is available simultaneously to all investors and each investor is acting rationally;
- There are no tax or transaction costs;
- Non-risk interest rate is constant through time;
- Share price (the value of real property) follows the multiplicative binomial process in discrete time.

The Binomial model shows that as the uncertainty clears in the future, management can make appropriate decisions at that time by comparing the expected payoff with the investment cost.

Initial point S_0 in the Binomial model shows the current value of the underlying asset. Probability of changing asset value in the future indicates the p. Conversely, the probability of falling asset value is expressed with 1-p. In the first step (node) of the binomial model asset value can move in two directions, up to (S0u) or down to (S0d). The next (second) step results in three possible assets values such as (S0u2, S0ud, S0d2), the third time step in four (S_0u^3, S_0u^2d, S_0ud^2, S_0d^2) etc. The last step in the Binomial model indicates the range of possible assets values at the end of the of the options life [10].

Up and down factors, u and d, depends on the volantility of the underlying asset and they can be expressed as follows:

$$\mathbf{u = \exp(-\sigma\sqrt{\delta T})}$$

also equation can be rewritten as:

$$\mathbf{d = \frac{1}{u}}$$

Through every time period there is a probability p that asset value will grow for percentage d, respectively the probability (1-p) that the assets will fall for percentage d:

$$\mathbf{p = \frac{\exp(r_f * \delta * T) - d}{u - d}}$$

where r is risk-free rate corresponding to the option life and $\delta * T$ is the time associated with each time step of the binomial tree.

The inputs required for setting Binomial model and calculating the option value are: the present value of the underlying asset (S_0), present value of implementation cost of the option (X), time to expiration in years (T), volatility of the natural logarithm of the underlying free cash flow returns in percent (σ), risk-free rate or the rate of return on a riskless asset (r) and the time-steps or time scale between steps ($\delta * T$) [13].

4 PROJECT VALUATION USING REAL OPTIONS (case study)

An example of investment project evaluation using the Black & Scholes and Binomial model is shown below.

Problem:

Pharmaceutical company is considering development of new product that would complement existing products of company. Previous experience with similar products have shown that companies can wait for a maximum of five years with the launching of new product without suffering significant losses of revenue. Discounted cash flow methods (DCF) that uses an appropriate risk-adjusted discount rate have shown that the present value of expected future net cash flows from the exploitation of new product amounts to 200 million, while the investment cost to develop and market product is expected to be 240 million. Uncertainty of future cash flows (annual volatility) is estimated to be 30%, and the annual risk-free interest rate over the option's life is 5%. What is the value of the option to wait (delay)?

1) *Black&Scholes model*[1]

The option to wait or delay appears as a kind of a call option, which offers the possibility of acquiring a positive net present value in the future, due to circumstances that affect the formation of the net present value of the project [14]. The following input parameters are known: S_0 = 200 million it, X = 240 million it, σ = 30%, r = 5%, T = 5 years.

[1] Results obtained using Excel and Risk[@] software application.

$$d_1 = \frac{\left[\ln\left(\frac{S_0}{X}\right) + \left(r + \frac{1}{2}\sigma^2\right) * T\right]}{\sigma\sqrt{T}} = \frac{\left[\ln\left(\frac{200}{240}\right) + \left(0.05 + \frac{1}{2} * 0.30^2\right) * 5\right]}{0.30\sqrt{5}} = 0.43629926$$

$$d_2 = d_1 - \sigma\sqrt{T} = 0.43629926 - 0.30\sqrt{5} = -0.2345211\ 3$$

$$C = S_0 * N(d_1) - \frac{E}{e^{rt}} * N(d_2)$$

$$= 200 * N(0.436299\ 26) - \frac{240}{2.71828^{0.05*5}} * N(-0.23452\ 113) = \underline{\underline{57.755 \text{ million}}}$$

2) *Binomial Model*

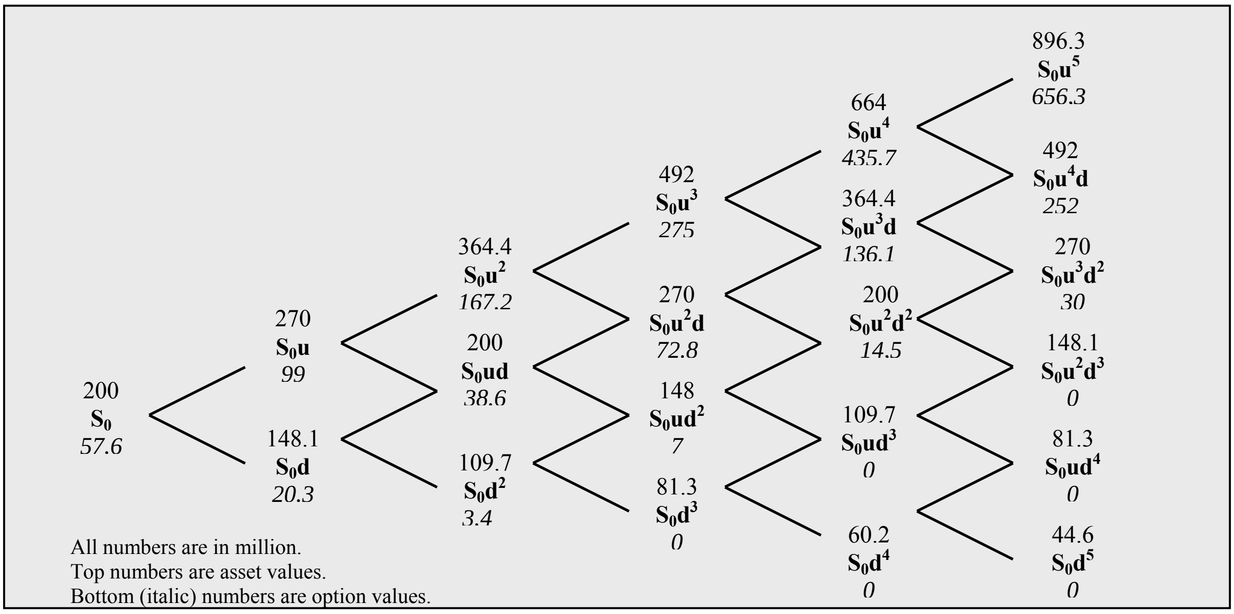
After identifying inputs required for setting up model that are the same as in the Black & Scholes model with the exception that $\delta = 1$ years, it is necessary to calculate, the u, d and p parameters:

$$u = \exp(\sigma\sqrt{\delta T}) = \exp(0{,}30 * \sqrt{1}) = 1.349859$$

$$d = \frac{1}{u} = \frac{1}{1.349859} = 0.740818$$

$$p = \frac{\exp(r * \delta * T) - d}{u - d} = \frac{\exp(0.05 * 1) - 0.740818)}{1.349859 - 0.740818} = 0.509741$$

Then, it is necessary to create a Binomial tree and calculate the asset values on each node of the Binomial tree, using one-year time interval. Time of options maturity can be divided into several phases. Upon completion of each phase, management has option whether to invest in product development at that point or delay its implementation and wait until next time period. The upper numbers on the Binominal tree present expected future asset values over the option life and bottom numbers indicate option values as it is shown in Figure 1.



Source: Modified according to Kodukula & Papadesu

Figure 1: Binomial Tree for Option to Wait

When preparing Binomial tree it is necessary to present the value of expected cash flows arising from investing in development of new product, S_0, multiply with the up factor u and down factor d to obtain S0u and S0d. Moving to the right, with the same procedure it is necessary to calculate the expected value of cash flows for every node of the Binomial tree until the last step [10]. For example; S_0u = 200 million*1.349859 = 270 million; S_0d = 200 million*0.740818 = 148.1 million.

At the end of the second year, introduction of new product is expected to generate cash flow between 364.4 million and 109.7 million, and at the end of the fifth year possible values of expected cash flows vary between 896.3 million and 44.6 million.

Once have learnt the value of expected future cash flows at each node of the Binomial tree which are shown in the form of the above values, it is necessary for each node to calculate the value or price of the options (below italic values in the scheme). The option values are calculated from the extreme right values in the scheme according to the initial values to the left ("backward induction"). On each node there is the possibility of investing in the development of new product or deferral of investment to further. At node S_0u^5 expected asset value is 896.3 million. If option is exercised in the fifth year, and investment cost of developing new product is 240 million, then net asset value of the introduction new product is: 896.3 million – 240 million = 656.3 million. But if we delay realization of option and wait until next time period, the revenues will be zero because option expires (becomes worthless) at the end of the fifth year due to the impact of competition and other influential factors on the market. Hence, at node S_0u^5 the option value is 656.3 million and rational decision will be not to wait but rather invest in the development of new product.

Expected asset value at node $S_0u^2d^3$ is 148.1 million but the option value at this node is zero because the investment of 240 million is resulting in a net loss of 91.8 million. In these circumstances, rational decision would not join the investment in the development of new product.

Furthermore, at the intermediate node S_0u^4 we can calculate the expected asset value for keeping the option open as discounted weighted average of potencial future option value [10].

$$\begin{aligned}&\left[\mathbf{p}\left(\mathbf{S_0u^5}\right)+(\mathbf{1-p})*(\mathbf{S_0u^4d})\right]*\mathbf{exp(-r*\delta t)}\\&=\left[\mathbf{0.509741*(656.3\ million)+(1-0.509741)*(252\ million)}\right]\\&\mathbf{*\ exp(-0.05)*(1)=435.73\ million}\end{aligned}$$

If the option is exercised and we invest 240 million in developing new product, the net asset value would be 424 million (664 million - 240 million). However, holding the options open until the next period (fifth year) gives possibility of realizing higher asset value (435.73 million). Therefore, it is better to continue to wait, rather than to exercise the option.

Similarly, at the node S_0ud^3, the expected asset value for keeping the option open is zero. Payoff at this node is 109.7 million and if the option is exercised by investing 240 million, it would result in a net loss of 130.3 million.

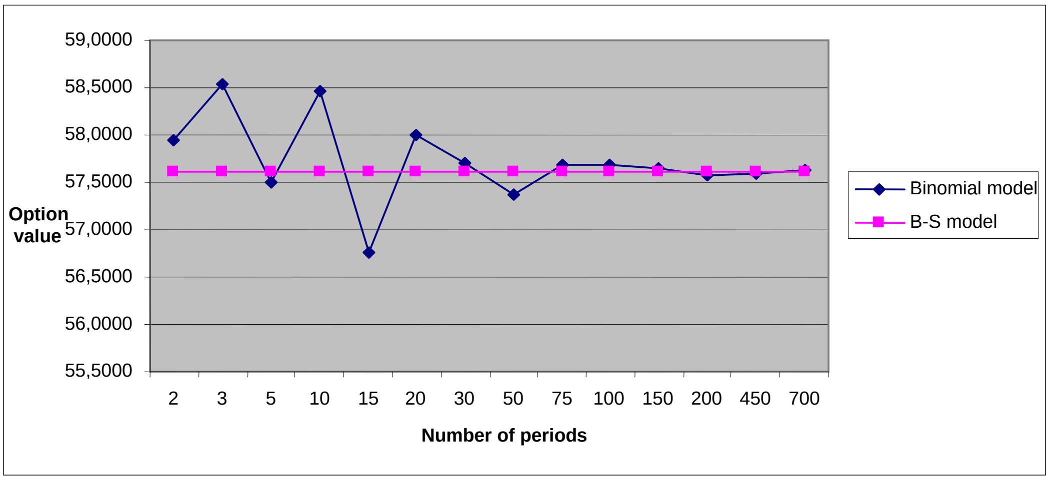
$$\left[\mathbf{0.509741*(0\ million)+(1-0.509741)*(0\ million)}\right]*\mathbf{exp(-0.05)*(1)=0\ million}$$

The same procedure has given the option values until the time 0, where we can perceive the value of the introduction of new product of 57.6 million, which is approximately value provided by Black & Scholes option valuation model. Also, as it can be seen in the table below, as the number of periods to maturity option grows towards infinity, binomial formula, with certain assumptions converge to Black & Scholes formula [8] and [1].

Table 2: Values obtained with the Binomial and Black-Scholes model

n	Binomial model	B-S model
	57.4924	57.61051914
2	57.9536	57.61051914
3	58.5437	57.61051914
5	57.4924	57.61051914
10	58.4719	57.61051914
15	56.7524	57.61051914
20	57.9970	57.61051914
30	57.7095	57.61051914
50	57.3743	57.61051914
75	57.6879	57.61051914
100	57.6807	57.61051914
150	57.6521	57.61051914
200	57.5683	57.61051914
450	57.5986	57.61051914
700	57.6220	57.61051914

Source: author's calculations according to Aljinović, Marasović, Šego, 2008, p. 217.



Source: author's calculations according to Aljinović, Marasović, Šego, 2008. p. 217.

Chart 1: Convergence Binomial option valuation model with Black-Scholes model

The Binomial model value is evident from the comprehensibility of model by showing that as the uncertainty clears over time, decision-makers are able to make appropriate decisions on implementation, or rejecting the project simply comparing the expected cash flows at each node of the Binomial tree with investment costs which are performed in order to implement the project.

5 DISCOUNTED CASH FLOW METHODS VS. REAL OPTION ANALYSIS

One of the most popular methods of discounted cash flows used in evaluating investment projects is the method of net present value (NPV) at which decisions about the acceptance or rejection the project largely depend on the adequacy of estimation of project's cash flow but also on the selection of appropriate discount rates which are used in discounting future net cash flows.

If we compare the decisions based on discounted cash flow methods and real option approach we can see that the introduction of new product from the previous evaluated example using discounted cash flow methods (DCF), would result in a negative net present value of 40 million (expected payoff of 200 million less the investment cost of 240 million). This would lead to the rejection on investment in a new product. However, traditional methods of valuation, which still occupy the main spot in the evaluation of investment projects, ignore the value of options that may come in the life cycle of the project so the traditional analysis should complement with optional methods of evaluation of projects.

Real option approach values covered options to wait or delay, which serves as a supplement to the method of net present value and increases the value of an investment project. The project has real option value of approximately 57.6 million created by the option characheristics of the project related to high uncertainty.

Aforementioned, total value of the project represents the sum of net present value of the project and option to wait. Therefore, decision based on real option analysis increases the value of project (i.e. introduction of new product) whose total value is positive and approximately 17.6 million (net present value of -40 million + option value of 57.6 million). Real options approach provides additional value to project and offers managers a choice of decisions also helping to be more rational in their decision making.

6 CONCLUSION

Among the contemporary methods of assessment and evaluation of investment projects, important role has the real option analysis, whose evaluation is based on the analogy of financial options to real options. All above mentioned real options have equal motive - limiting bad business results. With the growth of uncertainty related to future, option value increases and that affects initial decision on acceptance or rejection of a project. Decision on rejection made on basis of traditional methods can be altered if the option value is high enough. Likewise, decision on acceptance of a project can be changed if compensatory value of an option is higher than the lost cash flow; but, if an investment is very prosperous or completely uninteresting to investors real option analysis will not change the outcome.

Supporters of contemporary methods emphasize imperfections of traditional methods assessment and evaluation of investment projects, such as choice of adequate discount rate, static approach, subjectivity in determining expected cash flows and regarding an investment to be "now or never" decision without evaluating potential possibility of postponing investments. Yet, significance of traditional methods has never been denied. However, many business decisions are part of so called “gray area” which demands rational reasoning, making real option method an indispensable tool in decision making process. It is necessary to accentuate that *real option method should only be used as a supplement, rather than a substitute for discounted cash flow method*.

References

[1] Amram, M., Kutilaka, N., 1999: Real Options, Managing Strategic Investment in an Uncertain World, Harvard Business School Press, Boston, Massachusetts, 181, 216 p.

[2] Aljinović, Z., Marasović, B., Šego, B., 2008: Financijsko modeliranje, Zgombić&Partneri, Split – Zagreb, 189 p.

[3] Bierman, H., Smidt, S., 1993: Capital Budgeting Decision, Eighth Edition. Upper Saddle River (New Jersey): Prentice Hall, pp. 471-473.

[4] Black, F., Scholes, M., 1973: The pricing of Options and Corporate Liabilities, Journal of Political Economy, Vol. 81, May-June, 640 p.

[5] Brealey, R.A., Myers, S.C, Marcus, A.J., 2007: Osnove korporativnih financija, 5. izdanje, MATE, Zagreb, pp. 255-257.

[6] Copeland, T.E., Antikarov, V., 2001: Real Options: A Practitioner s Guide, First Edition, New York, Texere, 7 p.

[7] Cox, J., Ross, S. A., Rubinstein, M., 1979: Option Pricing: A Simplified Approach ,Journal of Financial Economics, Rochester, Vol. 7., No. 3., pp. 229 – 263.

[8] Damodaran, A, 2000.: The Promise of Real Options, Journal of Applied Corporate Finance, Vol. 13., No. 2., 96 p.

[9] Dixit, A.K., Pindyck, R.S., 1995: The options approach to capital investment, Harvard Business Review, May–June, pp. 105-116.

[10] Kodukula, P., Papadesu, C., 2006: Project Valuation using Real options, J. Ross Publishing, Florida, 70, 78, 79 p.

[11] Luehrman, T.A., 1998: Investment Opportunities as Real Options: Getting Started on the Numbers, Harvard Business Review, July - August, No. 4., 10, 53, 139 p.

[12] Leslie, K. J., Michaels, M. P., 1997: The Real Power of Real Options, McKinsey Quarterly, No. 3., 9 p.

[13] Mun, J., 2006: Real Option Anaysis; Tools and Tehniques for Valuation Strategic Investments and Decisions, Second Edition, John Wiley&Sons, New Jersey, 129 p.

[14] Orsag, S., 2006: Izvedenice, HUFA, Zagreb, pp. 532-552.

[15] Pušar, D., 2004: Real option metoda u ocjeni investicijskih projekta, Magistarski znanstveni rad, Ekonomski fakultet Sveučilišta u Rijeci, Rijeka, pp. 51-55, 167-169.

[16] Ross, S.A., Westerfield, R.W., Jaffe, J., 2005: Corporate Finance, Seventh Edition, McGraw-Hill Irwin, New York, 634 p.

[17] Rovčanin, A., 2005: Opcioni pristup vrednovanju kapitalnih ulaganja, Ekonomski pregled, Ekonomski institut Zagreb, Zagreb, Vol. 56, No. 7-8., 547 p..

[18] Schwartz, E. S., Trigeorgis, L., 2004: Real Options and Investment under Uncertainty, The MIT Press, pp. 1-5.

[19] H.T.J. Smit, L. Trigeorgis, 2006: Real options and games: Competition, alliances and other applications of valuation and strategy, Review of Financial Economics, No. 15, 96 p.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section VI:

Location and Transport

THE PARALLEL REGRESSION-FUZZY SYSTEM FOR LAND APPRAISAL

Marija Bogataj*, Danijela Tuljak Suban and Samo Drobne*****
* University of Ljubljana, Faculty of Economics, Ljubljana and MEDIFAS, Šempeter pri Novi Gorici, Slovenia, e-mail: marija.bogataj@ef.uni-lj.si
** University of Ljubljana, Faculty of Maritime Studies and Transport, Portorož, Slovenia, e-mail: danijela.tuljak@fpp.uni-lj.si
*** University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia, e-mail: samo.drobne@fgg.uni-lj.si

Abstract: This article describes an effort to design parallel regression-fuzzy system to estimate real estate market value, especially for non built-up urban plots. The data set consists of data on sales of urban land in the land market of Ljubljana and it's surrounding in 2006, and it contains parameters describing typical non built-up area property features and the current sale (transaction) price. The study explores the parallel use of fuzzy reasoning to assess better the value of real estate, which has been previously assessed by exponential regression model. The results are compared with those obtained using a traditional multiple regression model only. The paper also describes possible future research in this area. This method is interesting for real estate appraisers, real estate companies, and bureaus because they provide better overview of location prices.

Keywords: land appraisal, land market, fuzzy logics, regression analysis.

1 INTRODUCTION

In land and other real estate market value assessment, many attributes are often used to determine a property's fair value. Some of them are very often also well correlated, especially those where accessibility and transportation costs play important role. In the past such real estate value assessments have often been supported by multiple regression-based methods. Linear or some other regression models have been used. In the last years an increasing interest in use of non-conventional methods for real property value assessment has been detected. The most commonly studied non-conventional methods are neural network-based land appraisal methods being previously presented by [3] and later by many others like [1]. The ability of such methods to interpret the numerous and complex interactions of common attributes is one of the main motivations for using neural networks in land and other real estate market value assessment. Fuzzy logic has been proposed as an alternative method by [2] first, but empirical studies have been presented in details much later by [4] and by some other authors.

In this paper we describe the design and implementation of a parallel regression-fuzzy approach to estimate market value (price) for non built-up urban land. We demonstrate that fuzzy pricing model can improve regression analysis of real estate market data by introducing non-smoothness in continuous regression analysis. Fuzzy pricing model could generate lower euro pricing errors, but in combination with regression analysis it assures a greater precision and extrapolates better from volatile pricing environments. It is capable of modelling complex nonlinearities.

2 MATERIALS AND METHODOLOGY

The data set consists of data on sales of urban land in Ljubljana and its' surrounding in 2006. Data contain parameters describing typical non built-up urban land. In our analysis, we used a small sample of 209 not yet built potentially residential properties (empty plots).

In our study, we explored the use of fuzzy reasoning to assess better real estate value previously assessed by exponential regression model. Firstly, we used regression methods to analyse the prices of non built-up urban plots. We got the regression model of non built-up urban land price per m^2 in Ljubljana in 2006. The model gave us an unexpected result: the coefficient of distance to sub-centre was positive, what was against the findings in the location theory. We decided to investigate the results by fuzzy reasoning to estimate how buyers "feel" distance to sub-centre. Actually, we investigated each of analysed factors by fuzzy reasoning.

In fuzzy approach, all input linguistic variables are defined with three terms (nizka, srednja, visoka), which mean short, medium, and long distance. The output linguistic variables are defined by five terms (zelo nizka, nizka, srednja, visoka, zelo visoka), which mean very low, low, middle, high and very high prices. All fuzzy sets are defined by a membership function of trapezium shape, with mostly four limit points. Continuous changing of limit points location influenced improvement of the results.

3 THE REGRESSION RESULTS FOR TOTAL SAMPLE DATA

Results of modelling market value (price) for non built-up urban land in Ljubljana and its' surrounding area using regression analysis show that prices are very volatile.

In our analysis, where we analysed data for 209 non built-up plots in Ljubljana and its' surroundings, the model was:

$$y = a \cdot \prod_{i=1}^{3} X_i^{\alpha_i} \tag{1}$$

or

$$\log y = \log a + \sum_{i=1}^{3} \alpha_i \log X_i \tag{2}$$

where y is the price per m^2 of non built-up urban land in Ljubljana, X_i is studied coefficient (X_1 is relative distance to central business district (CBD), X_2 is relative distance to the nearest sub-centre, X_3 is relative distance to the nearest line of existing sewage system), and a and α are parameters estimated in the regression analysis. We got the regression statistics, given in Table 1 and Table 2.

From Table 2, the model for estimating the price of non built-up urban land per m^2 in Ljubljana in 2006 is (in Slovenian tolars; $240\,\text{SIT} \cong 1\,€$):

$$y = 10^{4.279} X_1^{-0.489} X_2^{0.065} X_3^{-0.119} = 19.011 \cdot X_1^{-0.489} X_2^{0.065} X_3^{-0.119} \tag{3}$$

It means, that not built-up plot in the average distance to CBD, in average distance to suburb and in average distance to sewage system has price approximately $10^{4.279}$ SIT ($\cong$ 80 €) per m^2 and it is falling by distance to CBD for 4.6 % ($1 - 1.1^{-0.489} = 0.046$) when distance is increasing for 10 % in average. However, it also decreases with distance to sewage system, but increases with the distance to sub-centre, which is rarely case in reality.

The latest result on distance to sub-centre, which is against the other theoretical findings, should be investigated further. From Table 2, we can see that P-value for X_2 is very high, but we do not wish just to skip this factor from the analysis. Therefore, we decided to investigate this phenomenon by fuzzy reasoning to estimate how buyers really “feel” distance to the sub-centre.

Table 1: Regression statistics for all available data

Multiple R	0.595
R Square	0.354
Adjusted R Square	0.345
Standard Error	0.283
Observations	209

Table 2: Regression coefficients for all available data

	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	4.279	0.027	155.30	1.3E-21
L_CENTRE $\equiv X_1$	-0.489	0.092	-5.30	2.6E-07
L_SUBCENTRE $\equiv X_2$	0.065	0.061	1.06	0.29
L_SEWAGE $\equiv X_3$	-0.119	0.022	-5.40	1.6E-07

While regression analysis gives R Square 0.354, coefficient of determination in fuzzy approach, when optimising slops of membership functions, is equal to 0.377.

We have found out that the limit point position of the first term “nizka” has very strong impact on division of population into two subsets with different behaviour. Namely, we introduced the simple membership functions with continuously adaptive slopes, as follows.

Introducing the substitution for the distance to CBD by $dX_{1,i}$:

$$dX_{1,i} = \log\left(\frac{X_{1,i}}{\overline{X}_1}\right),$$

where the distribution of distances to CBD $dX_{1,i}$ is presented in Figure 1.

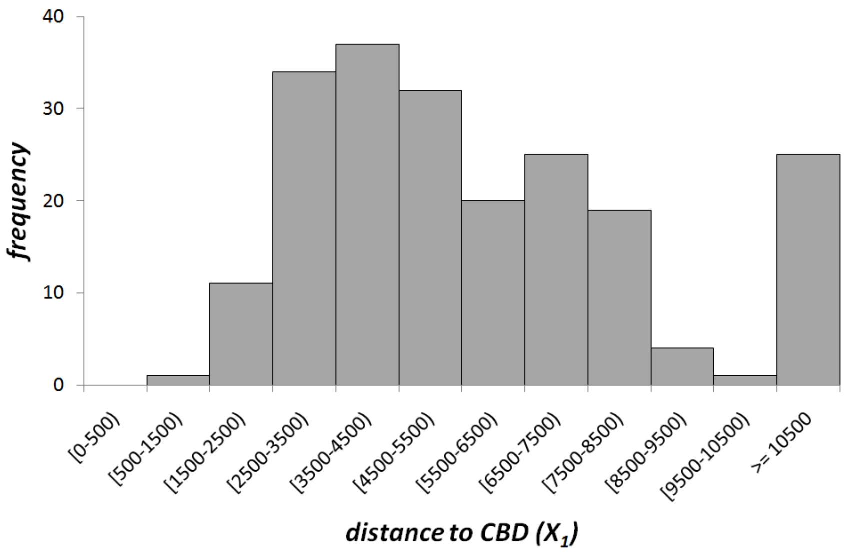


Figure 1: The distribution of distances to CBD (L_CENTRE $\equiv X_{1,i}$,in meters)

The membership function at optimal slope of "nizka" – that means that the buyers "feel" the plots in the relative distance less than 0.5 ($0.5 \equiv 4500\,\mathrm{m}$) being mostly in short distance – was as presented by Figure 2.

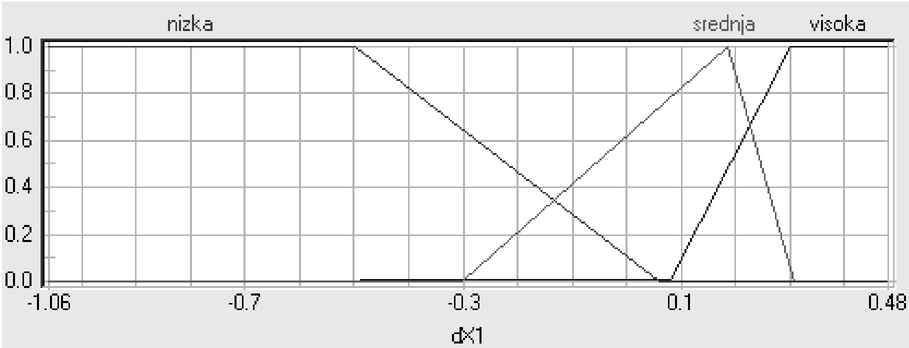


Figure 2: The membership function at optimal slope for distance to CBD (L_CENTRE $\equiv X_{1,i}$)

The slopes have been simultaneously optimized for criterion function being determination coefficient, which was supposed to be maximal.

For distances to sub-centres $dX_{2,i}$, we got the following substitution:

$$dX_{2,i} = \log\left(\frac{X_{2,i}}{\overline{X}_2}\right),$$

where distances to sub-centre had the distribution presented in Figure 3.

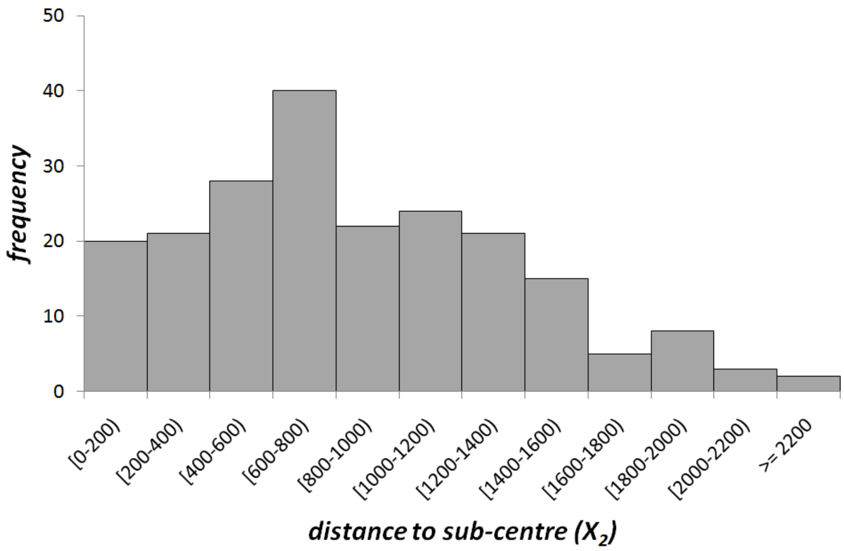


Figure 3: The distribution of distances to sub-centre (L_ SUBCENTRE $\equiv X_{2,i}$,in meters)

Many other factors, not included in our analysis, are well correlated with distance to sub-centre and this factor replaces them in the model. Optimal slopes at membership functions are presented below. Figure 4 shows that as long as the distance to sub-centre $dX_{2,i}$ is less than -0.17, which is close to 600 m, the buyers feel it mostly as "nizka", it means short.

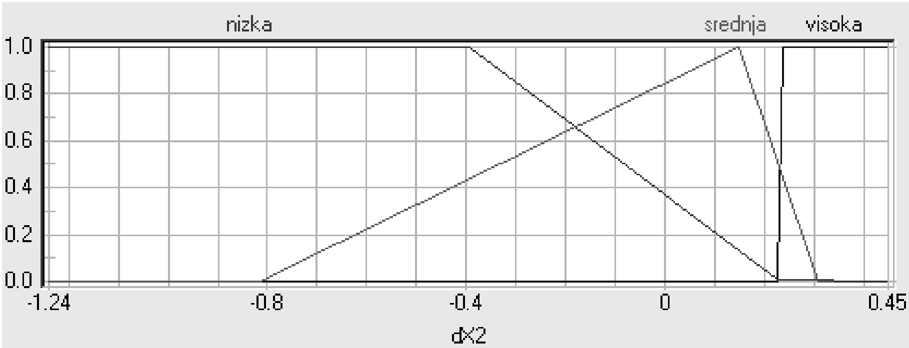


Figure 4: The membership function at optimal slope for distance to sub-centre (L_ SUBCENTRE $\equiv X_{2,i}$)

Third factor, which significantly influences land price, is distance to available connection to the sewage system. Figure 5 shows the distribution of distances to sewage system**.**

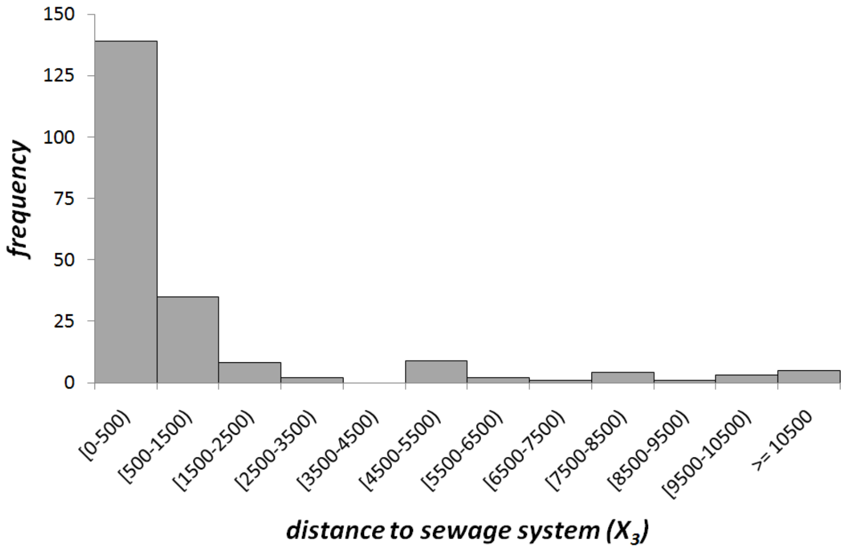


Figure 5: The distribution of distances to sewage system (L_ SEWAGE $\equiv X_{3,i}$,in meters)

Membership function has the following form for normalised variable:

$$dX_{3,i} = \log\left(\frac{X_{3,i}}{\overline{X}_3}\right).$$

Figure 6 shows the membership function at optimal slope for distance to sewage system. Here, 70 % of all distances to sewage system are lower than 500 m. In membership function it equals to -0.38, and 20 % of all plots was in the radius between 500 and 1500 m from the sewage system, where the value dX_3 is 0.09.

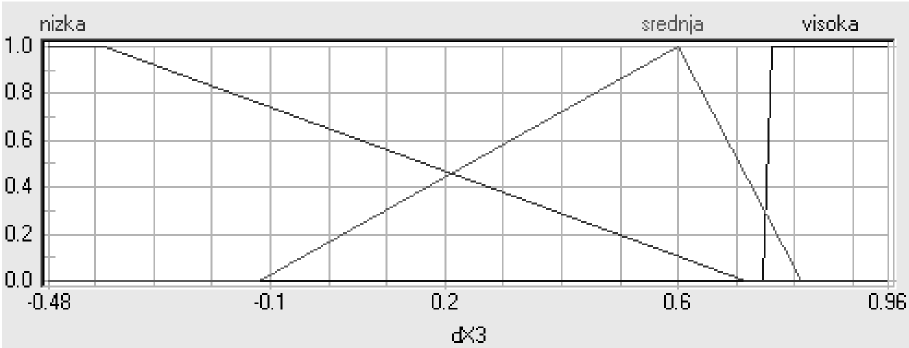


Figure 6: The membership function at optimal slope for distance to sewage system (L_SEWAGE$\equiv X_{3,i}$)

Figure 7 shows the membership function for fuzzy prices, where $dY_i = \log(Y_i)$. Figures 8, 9 and 10 show the distribution of transaction prices of non built-up plots in Ljubljana and surroundings in 2006, where the prices are reported in tolars ($240\,\text{SIT} \cong 1\,€$).

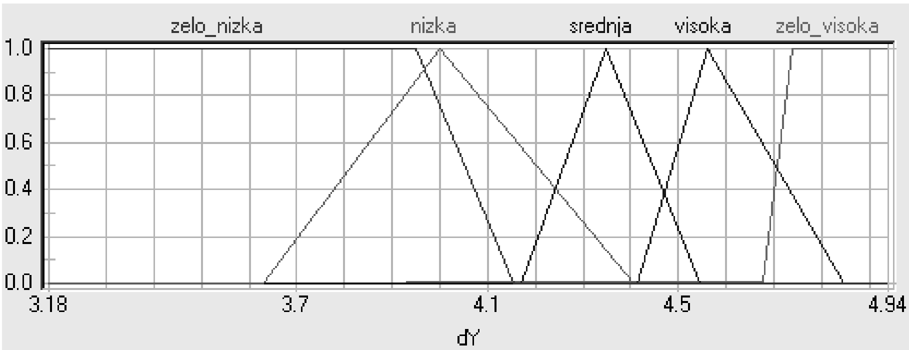


Figure 7: The membership function at optimal slope for transaction prices of non built-up plots ($dY_i = \log(Y_i)$)

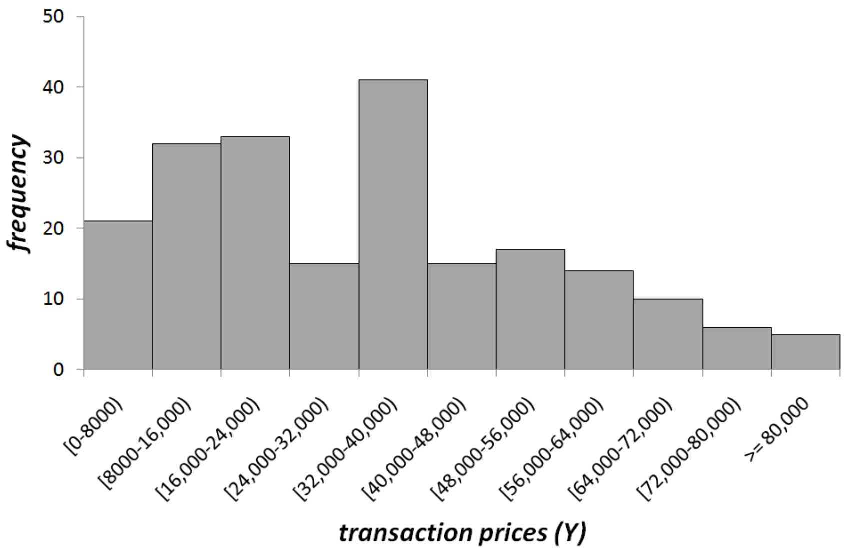


Figure 8: The distribution of transaction prices of non built-up plots (in SIT)

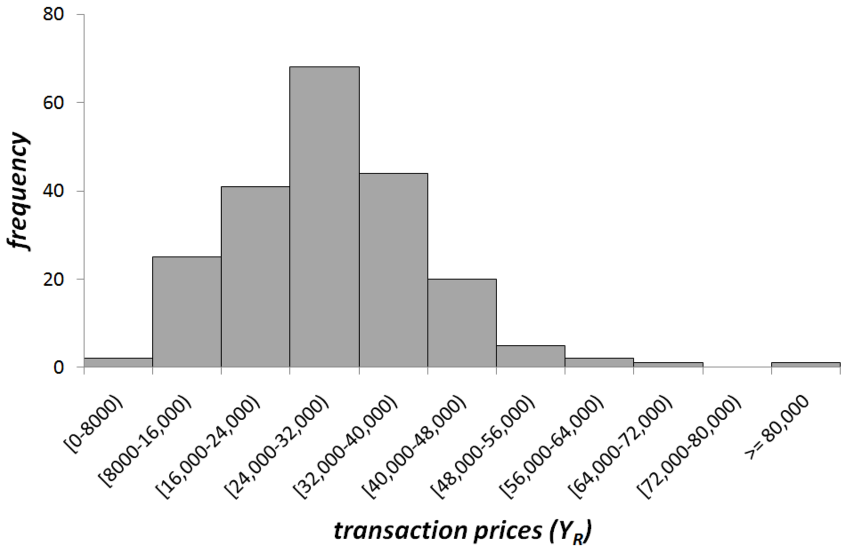


Figure 9: The distribution of transaction prices of non built-up plots evaluated by regression analysis (in SIT)

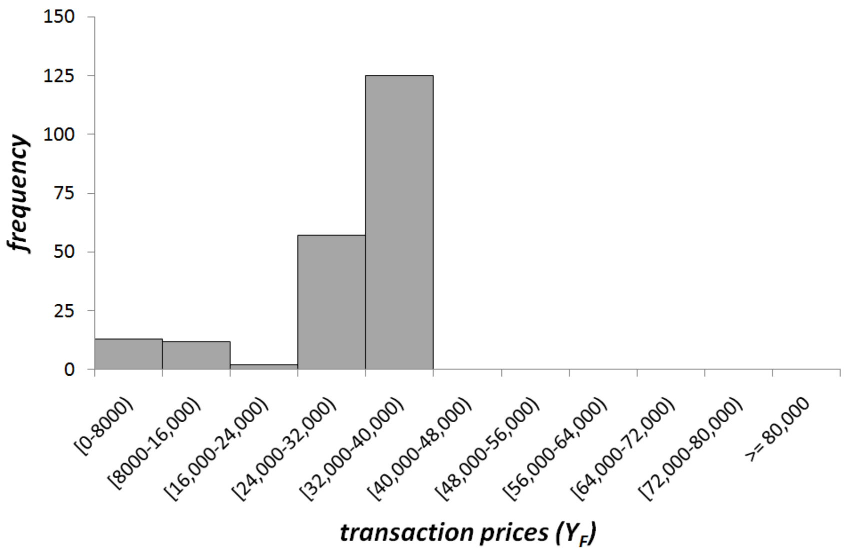


Figure 10: The distribution of transaction prices of non built-up plots evaluated by fuzzy logics (in SIT)

4 EXPERIMENTAL RESULTS OF PARALLEL APPRAISAL

Fuzzy logics divides non built-up land plots into two subsets, those with the price between 8000 and 24,000 tolars per square meter, and those between 32,000 and 48,000 tolars per square meter. The results of comparison between fuzzy (F) and land market data (transaction value) are presented in contingency table (Table 3), where it is obvious, that fuzzy logics approach underestimates higher prices. It means that we are not able to detect high speculative transactions influencing fuzzy appraisal, while regression analysis equally differs for higher than for lower estimations (as it is expected according to the methodology used in regression analysis).

Table 3: Contingency between fuzzy estimates (F) and real transaction prices

	$M_F - 2\sigma$	$M_F - \sigma$	M_F	$M_F + \sigma$	$M_F + 2\sigma$
$M - 2\sigma$	12	15	15		
$M - \sigma$	1	15	47		
M		14	51		
$M + \sigma$		7	25		
$M + 2\sigma$			7		

From the Table 3 we can see that fuzzy reasoning does not recognise extra attractive locations which are influencing land market value. The situation is different by using regression analysis (see Table 4).

Table 4: Contingency between regression estimates (R) and real transaction prices

	$M_R - 2\sigma$	$M_R - \sigma$	M_R	$M_R + \sigma$	$M_R + 2\sigma$
$M - 2\sigma$	2	29	11		
$M - \sigma$		30	29	4	
M		31	34		
$M + \sigma$		6	23	2	1
$M + 2\sigma$	1	6	7		

While regression analysis enables smooth pricing, fuzzy logics show that buyers do not think about more or less accessible plots "continuously". They "fill" value of land changing by distance without having well founded measure for distance. We can suppose that the regression functions for those plots, which are "mostly supposed" to have better accessibility, are different of those which are "mostly supposed" to have worse accessibility Therefore the sample of plots has been divided into two parts of different sizes depending on the value of distance to sub-centre. From Figure 4 it follows that first subset consists of plots for which L_SUBCENTRE<600 m. We got the regression functions for two subsets, as presented in Tables 5, 6, 7 and 8.

For the first subset of data (data for non-built up plots with short distance to sub-centre) we got higher multiple R and no significant coefficient L_SUBCENTRE. That can now be easily explained: for all those plots buyers are concentrated on the fact to be close to the sub-centre. But the second subset of data (data for non-built up plots with high distance to the sub-centre) shows quite high influence of L_SUBCENTRE too, being negative, as expected in real life application – as follows from the location theory and could be interpreted wrong in first regression analysis, given in Table 1 and Table 2. Furthermore, the influence of accessibility to sewage system is shown to be important as well (Table 8).

Using parallel regression-fuzzy approach we can investigate smoothness of all factors usually analysed in the regression analysis for land appraisal.

Table 5: Regression statistics for subset of non-built up plots with short distance to sub-centre (L_SUBCENTRE<600 m)

Multiple R	0.603
R Square	0.364
Adjusted R Square	0.335
Standard Error	0.338
Observations	70

Table 6: Regression coefficients for subset of non-built up plots with short distance to sub-centre (L_SUBCENTRE<600 m)

	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	4.299	0.110	39.09	2.3E-47
L_CENTRE $\equiv X_1$	-0.628	0.179	-3.507	0.001
L_SUBCENTRE $\equiv X_2$	0.089	0.176	0.506	0.615
L_SEWAGE $\equiv X_3$	-0.055	0.043	-1.290	0.201

Table 7: Regression statistics for subset of non-built up plots with big distance to sub-centre (L_SUBCENTRE $\geq$ 600 m)

Multiple R	0.557
R Square	0.310
Adjusted R Square	0.295
Standard Error	0.254
Observations	140

Table 8: Regression coefficients for subset of non-built up plots with big distance to sub-centre (L_SUBCENTRE $\geq$ 600 m)

	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	4.290	0.040	107.22	2.8E-133
L_CENTRE $\equiv X_1$	-0.399	0.123	-3.24	0.002
L_SUBCENTRE $\equiv X_2$	-0.252	0.143	-1.77	0.079
L_SEWAGE $\equiv X_3$	-0.166	0.028	-5.99	1.7E-08

5 CONCLUSIONS

The regression method is the most widely used to obtain econometric models. It is well known that classical regression model can be used to assign market value based on a plot's characteristics even if the transaction value (price) of a specific plot is not observed. Regression analysis sometimes gives unexpected results like in our case, where coefficients of distance to sub-centres have been calculated as being positive, what is against the location theory. But the prices are not changing smoothly in reality. For this reason, we decided to study, how people "feel" the distance through the fuzzy reasoning. We have investigated each of three factors by fuzzy reasoning: relative distance to central business district, relative distance to the nearest sub-centre and relative distance to the nearest line of existing sewage system. When turning the attention to the distance to the sub-centres, we can find that people “feel” the distance being less than 600 m as mostly short, not influencing price reduction. That is why we divided total sample of plots into two subsets: those, for which the distance to the sub-centre is less than 600 m – they are, according to the membership function, supposed mostly to be in short distance – and others. Regression analysis on these two samples leads us to the expected and also theoretically appropriately explained results.

We can conclude that fuzzy reasoning helps us to find the reasons for unexpected results of regression analysis in the procedure of land appraisal. Fuzzy approach with continuous changing of limit points enables us to divide properly the data in subsets which give more reliable results, while the hyper-surface of transaction prices becomes unsmooth.

References

[1] Bee-Hua, G. Evaluating the Performance of Combining Neural Networks and Genetic Algorithms to Forecast Construction Demand: The Case of the Singapore Residential Sector. *Construction Management and Economics*, 2002, 18(2):209–217.

[2] Byrne, P. Fuzzy Analysis: A Vague Way of Dealing with Uncertainty in Real Estate Analysis. *Journal of Property Valuation & Investment*, 1995, 13(3):22–41.

[3] Do, Q. and G. Grudnitski. A Neural Network Approach to Residential Property Appraisal. *Real Estate Appraiser*, 1992, 58:38–45.

[4] Guan, J., J. Zurada, and A. Levitan. An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment. *Journal of Real Estate Research*, 2008, 30(4):395 –421.

FIRST AID STATIONS LOCATION PROBLEM AND AN IMPACT OF THE SET OF CANDIDATES ON THE RECOURSE OF A PATIENT

Marta Janáčková and **Alžbeta Szendreyová**
Faculty of Management Science and Informatics, University of Žilina,
Univerzitná 2, 010 26 Žilina, Slovak Republic,
Marta.Janackova@fstroj.uniza.sk, Alzbeta.Szendreyova@fri.uniza.sk

Abstract: This paper deals with a medical emergency system design by using methods of mathematical programming. The problem is based on optimal locations of the first aid stations in the Slovak Republic area. The important criterion in provision of service is a kilometrage.

Keywords: uncapacitated location problem**,** emergency system design.

1 INTRODUCTION

From mathematical or programming point of view the problem how to design the first aid station locations belongs to the location problems. Some aspects must be taken into account when quick medical service is provided. Financial decisions and a number of locations must be regarded but on the other side, it is necessary to ensure the accessibility of the service for the potential patients. Herewith, we usually follow that in case of need, the ambulance vehicle gets to the patient in a specified time. This demand is often hard to fulfill. The response speed is influenced by a current situation on the roads, by an overlapping of various acute needs for interventions etc. Sometimes, the service (transfer to the hospital) has to ensure a subsidiary ambulance vehicle from a distant position. Thus, the service is provided on the certain level of quality.

We can evaluate the quality of services using0 different criteria. It is possible to design a system for ambulance-stand in order to minimize the average travel time of the ambulance vehicle from the ambulance-stand to the potential patient, which belongs to its service territory. However, this criterion disadvantages potential patients from the small and remote areas of the service territory. The minimization of the maximum travel time of the closest ambulance vehicle to a patient is another criterion how to evaluate the service quality. Unfortunately, this criterion requires more ambulance-stands. The set of places, where the ambulance-stands can be located, influences the locations of ambulance-stands as well. The use of each mentioned criterion leads to a specific model.

In this paper we solve an uncapacitated location problem with maximal service distance and with prescribed maximal number of locations. The criterion of service quality will be the distance, which the ambulance vehicle has to pass over to achieve the patient. For the system design we will use more types of sets of candidates for the locations of the ambulance-stands. We will study the impact of the set cardinality on the solution quality.

2 MODEL

The emergency system design comes out from the assumption that the ambulance vehicles can be placed in well-accessible road network nodes. Let I be the set of possible locations. Customers (patients) are located in the dwelling places, eventually in the network nodes. The dwelling places form a set J. Each dwelling place is characterized by a number of customers' b_j, $j \in J$. The problem is to locate a given number p of ambulance-stands (centers) at some nodes from the set I in order the ambulance vehicle to come from the place $i \in I$ to the place $j \in J$ in the prescribed period T_{max}.

The segment between node i and j is evaluated at the coefficients c_{ij} for each location $i \in I$ and each place $j \in J$. One possibility is to evaluate the segments with real distances d_{ij} between nodes i and j of the road network. In case of time evaluated segments it is enough to re-count the distances d_{ij} to the time t_{ij} for $i \in I$ and $j \in J$. When creating a matrix with coefficients c_{ij} the quality of individual types of roads can also be taken into account. If the time t_{ij} exceeds the allowed maximum T_{max}, it is suitable to replace the evaluation with penalty, which is big enough.

We also assumed that the emergency service must visit the place j b_j-times. In the case that all customers from the place j are planed to be served from the same location i, the coefficient c_{ij} can be modified by product of the time t_{ij} and the number of customers b_j of the place j. So, $c_{ij} = b_j t_{ij}$ for $i \in I$ and $j \in J$.

The decision on locating or not locating the ambulance vehicle at a place i will be modeled by a variable y_i, which takes the value of 1 if a ambulance-stand is placed at the position i and it takes the value of 0 otherwise. A zero-one variable z_{ij} models the decision on an assignment of patients from the dwelling place j to the ambulance location at the place i. The variable z_{ij} takes the value of 1, if the patients from the dwelling place j are satisfied from the ambulance-stand i and it takes the value of 0 otherwise. The model of the problem follows:

$$\textit{Minimize} \quad \sum_{i \in I} \sum_{j \in J} c_{ij} z_{ij} \tag{1}$$

$$\textit{Subject to} \quad \sum_{i \in I} z_{ij} = 1 \qquad \text{for } j \in J \tag{2}$$

$$z_{ij} \le y_i \qquad \text{for } i \in I,\ j \in J \tag{3}$$

$$\sum_{i \in I} y_i \le p \tag{4}$$

$$y_i \in \{0,1\} \qquad \text{for } i \in I \tag{5}$$

$$z_{ij} \in \{0,1\} \qquad \text{for } i \in I,\ j \in J, \tag{6}$$

where the used coefficients have the following meaning:

c_{ij} … valuation of the segment between node i and j ,
p … required number of locations,
I … the set of possible ambulance locations,
J … the set of dwelling places.

The objective function (1) corresponds to the sum of products of times and numbers of patients (customers). Constraints (2) guarantee that a patient from dwelling place $j \in J$ is assigned to the exactly one of possible ambulance locations. The constraints (3) ensure that the variable y_i takes the value of 1, when at least one customer is assigned to the location i. The constraint (4) puts the limit p on the number of location.

3 NUMERICAL EXPERIMENTS

For numerical experiments, we used a software product which was implemented in the University of Zilina. This software exploits the procedure *BBDual*. This procedure solves the uncapacitated location problem. It is based on the branch and bound method and it works with the real-sized networks very well [1].

Our algorithm enables to select the speed of ambulance vehicle following the type of a road communication, but we used the speed of 60 km/h for all types of road. An assignment was regarded as successful, if the customer from the dwelling place j is reached by the ambulance from the place i in 15-minutes.

The whole problem was solved on the real network of the Slovak Republic, which consists of 2916 dwelling places with potential patients. The positions and the numbers of inhabitants of the dwelling places are known. Following the actual state we made a decision to locate maximal 223 first aid center locations. The edges - segments of the roads were sorted into four types according to the quality [2], [3].

We constructed the sets of the candidates for ambulance-stand by choice of the villages of the Slovak Republic following the next rules: we created 3 types of sets of candidates each of them in 5 different sizes, namely with 400, 800, 1200, 1600 and 2000 candidates. These sets differed from each other not only in their cardinalities but also in the approach of their selections of the possible ambulance-stand candidates. The villages were sorted in descending sequence according to the numbers of their inhabitants. The candidates were selected from the first K villages with the biggest number of inhabitants. K took the values of 500, 1000, 1500, 2000 and 2500. In the approach *A,* there were 80% ambulance placement randomly chosen from the set of K villages with the biggest number of inhabitants. In the approach *B,* the possible candidates for the ambulance placement were selected as the first 60% of villages from the set of K villages with the biggest number of inhabitants and one half from the remaining 40% of K villages with the biggest number of inhabitants was selected by randomly generation. In the approach *C* the candidates were selected as the first 60% of villages from the set of K villages with the biggest number of inhabitants and the rest of candidates were obtained by random choice generated from the remaining villages of the Slovak Republic to complete the set of candidates.

We observed values of the objective function (after re-count in kilometers) and the computational time, that was measured in seconds in all instances. Only for information, we also reported the numbers of optimal locations. The results in particular cases of the given series of sets didn't differ too much. That is why we report the average values. An impact of the customers set cardinality on value of the objective function is in the table 1.

Table 1: An impact of the cardinality of the candidates sets on the objective function

Number of candidates	Average value of objective function	Average computational time [s]	Number of infeasible tasks (from 15 problems)
400	18,842,385	59	1
800	17,946,241	328	2
1200	17,616,022	517	2
1600	17,683,984	3,823	1
2000	18,008,371	3,060	1

We noticed bigger differences of the results according to the way of candidate. The results are presented in the table 2.

Table 2: An impact of candidate selection of on the objective function

	Number of candidates				
	400	**800**	**1200**	**1600**	**2000**
Selection of type A	21,241,524	19,371,574	19,255,466	18,594,381	18,662,139
Selection of type B	17,449,845	17,440,280	16,750,785	16,773,587	17,354,602
Selection of type C	17,835,787	17,026,870	16,841,817	16,687,924	17,062,443

In some cases the optimization process was not finished successfully. It means that the process was not successful in reaching exactly p ambulance locations. In some cases, there were proved neither the dependence on the cardinality of candidate sets nor the dependence on the way of selection of the candidate sets.

4 CONCLUSIONS

The results show that the problem can be successfully solved in the most cases even if large instance is used. When the set of the candidates for ambulance-stand was bigger, the values of the objective function were a little bit better, as we expected. It is surprising that the selection of candidates by approach *A* induced higher values of the objective function in all cases. In the sets of types *B* and *C,* there were also included villages with a small number of inhabitants, but 75% elements of the candidates sets are chosen from the biggest dwelling places. That fact certainly influences on result.

We can claim that ascending cardinality of the candidate sets has a positive impact on accessibility of the service. The way of the selection of the candidates set has some bigger impact on the values of the objective function.

We supplemented the algorithm for selection of the candidate sets by hand-made selection of candidates. This approach enables to add or to remove a candidate for ambulance-stand based on its geographical location in a real map. The selected locations are graphically distinguished. It is also possible to segment the map template and to fill up the selection of candidates according to the sector requirements. This approach to candidate set selection together with algorithm *BBdual* produce a useful tool for ambulance location problem design.

References

[1] Janáček, J., Buzna, Ľ.: An acceleration of Erlenkotter-Körkel's algorithms for the uncapacitated facility location problem. In: Annals of Operations Research. - ISSN 0254-5330. - Vol. 164, No. 1 p. 97-109. (2008).

[2] Janáček, J.: Time Accessibility and Public Service Systems. In: Proceedings of the conference "Quantitative Methods in Economics", Bratislava, pp 57-63, (2006).

[3] Jánošíková, L.: Emergency Medical Service Planning. Communications –Scientific Letters of the University of Žilina, Vol. 9, No 2, pp 64-68, (2007).

Acknowledgment: This work was supported by project "Centre of excellence for systems and services of intelligent transport" ITMS 26220120028

Project Part-Financed by the European Union

European Regional Development Fund

Operačný program VÝSKUM a VÝVOJ

Modeling the fiscal impact of Electronic Fee Collection System

Tomaž Kramberger[1],

Uroš Kramar[2],

Andreja Čurin[3].

University of Maribor, Faculty of Logsitics,

Mariborska cesta 7, 3000 Celje.

Abstract

The European transport infrastructure is extremely important for the economic development, the workforce mobility and competitiveness of the European economy. It is becoming increasingly difficult to depend solely on public funding of transport infrastructures, especially due to the growing budget deficit that many EU member state are facing and due to the need for more private investments in infrastructure projects.

In Europe the decision to implement toll roads is a transport-policy goal. Choosing the appropriate technology for establishment of road tolling is not an easy decision and it does not depend only on the character of the technology itself but also on the efficiency of the system. One needs to take into account that the aim is to maximize potential toll incomes and benefit the national economy.

For this purpose the mathematical model was built and some simulations were performed on several instaces to show comparison betwen systems based on different types of technology. We have simulated the model on several different instances. The results show us that DSRC technology used on moderate lenght of toll roads is most efficient.

[1]tomaz.kramberger@uni-mb.si
[2]uros.kramar@fl.uni-mb.si
[3]andreja.curin@fl.uni-mb.si

1 Introduction of the topics

Toll roads are at least 2700 years old [13]. Tolls already had to be paid in the seventh century BC by travelers using the Susa-Babylon highway in Mesopotamia. In his works also Aristotle refers to tolls in some Arab countries and India. Later in the Middle Ages, the Germanic tribes charged tolls to travelers across some mountain passes in the Alps, in the 14th century and 15th century tolls were regularly used in the Holy Roman Empire.

Many modern European roads were originally constructed as toll roads in order to recoup the costs of construction. In 14th century, in England, some of the most heavily used roads were repaired with money raised from tolls. Turnpike trusts were established in England beginning in 1706, and were ultimately responsible for the maintenance and improvement of most main roads in England and Wales, until 1870s. Most trusts improved existing roads but some new ones, usually only short stretches of road, were also built.

The way toll roads are funded and operated differs from country to country. Some of these toll roads are privately owned and operated, while others are owned and operated by the government or owned by the governemnt and privately operated. Toll roads are an important part of road management. Toll revenues are an important source of financing the construction and maintenance of new roads as well as the pricing policy in this field is an important part of the transport policy of particular countries. Apart from regulating the supply of transport services an adequate combination of policies must also deal with pricing policies. In the market economy this presents an effective way of how to better manage the demand for transport services and achieve more effective use of infrastructure. By implementing more efficient payment systems, economic instruments (taxes, fees or emission trading systems) can encourage users of transport services either to use cleaner vehicles or transport modes (walk on foot or by bicycle) and use less congested infrastructures, or to travel at different times. To this end, the pricing policy in this area presents an effective way for achieving sustainable mobility.

The European transport infrastructure is extremely important for the economic development, the workforce mobility and competitiveness of the European economy [5, 11, 10, 12]. Whether these objectives will be met or not depends on the degree to which the users are burdened with the costs for infrastructure usage and on the funding possibilities concerning new investments. It is becoming increasingly difficult to depend solely on public funding of transport infrastructures, especially due to the growing budget deficit that many of the EU member states are facing and due to the need for more private investments in infrastructure projects.

In Europe the decision to implement toll roads is a transport-policy goal [7, 3, 8, 9]. An optimal toll road system is the one that on one hand is acceptable and fair to the users of motorways and on the other hand efficient and economical to the society as a whole. The basic premise when planning toll roads is the so called fair toll system principle, i.e. equal burdening of all motorway users based on the driven distance. One needs to take into account that the aim is to maximize potential toll incomes and benefit the national economy.

2 Definition of the problem

Optimal toll roads system introduced in all EU countries has to be in accordance with European directive 2003/58/EC and with the findings of the projects MEDIA, CESARE II in CESARE III [1]. Additionally, the final solution will have to take into account recommendations by expert groups, which prepare acts for implementation of the directive 2004/52/EC [17]. If we sum up all requirements into a few lines, we can state the following. The final solution has to [2]:

- enable toll payments without stopping or slowing down the user,
- enable interoperability and be applicable in all EU member states,
- to a certain degree ensure control over misuse and enable collection of debt.

If we consider the long history of tolling it is understandable that through the ages many different ways of toll collection have been developed. The most basic division of toll collection is the division into toll collection with the *use of an obstacle* and the *free flow toll collection*. As the name implies, in the former an obstacle must be placed in front of the vehicle, which stops it from continuing its run and the run is enabled after the toll is paid. Free flow toll collection does not affect the vehicle's travel. In this case the obstacle is virtual. With the aid of technology the obstacle identifies the vehicle and charges the set toll while the vehicle continues unhindered with its journey. A third notable toll collection type is the so called vignette system. Through purchase of a vignette the driver acquires the right to use the roads for the duration of the vignette.

Based on demands presented above we can ascertain that obstacle based toll collection is not acceptable for the present time. Therefore the only adequate system is the free flow toll collection which can be further divided into three subsystems:

- satellite system [15](Global Navigation Satellite System - GNSS),
- microwave system (Dedicated short-range communications - DSRC),
- purchase of permits through the internet or through public terminals.

The main purpose of the satellite system which is used for toll collection is to precisely and reliably determine the location of the vehicle that is subject to toll collection. In case of the GNSS/CN toll collection system the users use built-in OBU [4] in their vehicles with built-in GPS receiver. As the name already implies the system uses two technologies – global navigation and mobile technology, which is used for the transfer of data. While the dedicated short-range communications (DSRC) for positioning uses wireless mid-range communication which enables communication between in the vehicle built-in OBU and electronic equipment RSE [5] at road edges (when the device restores communication with the receiver at road edge, the vehicle is indubitably positioned) and it was deliberately developed for the use in road transport [14]. From the technological point of view it is a subset of RFID technology which is well known for the use in logistics.

Considering all above mentioned it is clear that choosing the appropriate technology for establishment of road tolling is not an easy decision and it does not depend only on the character of the technology itself but also on the efficiency of the system. Efficiency of the road tolling can be measured regarding the extent it suits the above mentioned demands and how much money contributes into the national budget [6] . The factors that influence the efficiency of road tolling are following:

- rate per kilometer of toll road,
- number of vehicles using toll roads,
- total length of toll roads,
- fixed costs for development of the toll road system,
- variable costs for operation of the system,
- etc.

The factors that influence the efficiency of the system are both, the ones that depend on the selected technology and the ones that do not depend on the selected technology. The rate of toll road and number of vehicles are not influenced by the technology, while other factors are influenced directly or indirectly. Fixed (CAPEX) and variable costs depend on the selected technology and are rising in accordance with increase of length of toll roads.

[4] On Board Unit

[5] Road Side Equipment

[6] Or into the purses of concessionaires when the road is not administered by the state.

In continuation of the article we will present the model for calculation of toll road profit in selected time frame. We will compare the use of different systems, which are based on different technologies while the inputs are the same. Further we will show how the use of different technologies reacts on different input parameters (different lengths of toll roads, different number of vehicles, etc.). We will show the mathematical model build for this purpose and gained results.

3 Building the mathematical model

3.1 The idea of the model

The amount of tolls collected for a specific road section depends on the number of vehicles on that section, toll rates[7] and number of kilometers covered on that section. Should the amount of tolls collected on a specific road section surpass the operational costs for this same section, then the difference between these two amounts is considered a profit which is to be optimized. The operational costs of a system vary from system to system; they depend on selected tolling technology, data transfer costs, infrastructural costs, etc. Mathematically this would be equal to:

$$Inc = COST + R$$

Whereas Inc is income from toll collection and $COST$ are all setting up costs of the system as well as all operational costs, Inc can also be expressed as $n \cdot T \cdot s$, whereas n is the number of vehicles on this section, T is the rate and s the number of kilometers covered on this section. Similarly $COST = CAPEX + OPEX$. This means the following relationship is to be pursued:

$$n \cdot T \cdot s \gg COST$$

The greater is the difference between the values Inc and $COST$, the greater the financial flow R into the toll road fund is.

3.2 Model structure

As evident from the description of the basic idea of the model the main components of the model are:

- fixed costs needed for development and establishment of the system ($CAPEX$),
- costs for operation of the system ($OPEX$),
- income from road tolling (Inc).

As evident the model of the profit can be mathematically expressed as follows:

$$Profit = \sum_{i=1}^{k} (Inc - O_{yb}) \cdot Tg^{(i-1)} \cdot Inf^{(i-1)} - C_{yb} \tag{1}$$

[7]Euro/km, it differs regarding the cathegory of vehiclel n road

Table 1: Fixed and variable costs of ETC system.

	$CAPEX$
TS	costs of building the tolling stations,
ES	costs of building the control stations
CS	costs of setting-up the central system
PS	costs of building the points of sale
OBU	costs of OBU units,
WAN	costs of WAN network,
MM	costs of setting-up the cartographic system
	$OPEX$
$OOTS$	operational costs of the tolling stations,
$OOES$	operational costs of the control stations
$OOCS$	operational costs of the central system
OPS	operational costs of the points of sale
$OOBU$	operational costs of OBU units,
$OWAN$	operational costs of WAN network,
OMM	operational costs of the cartographic system

where Inf is inflation and Tg yearly traffic growth $O_{yb} = OPEX$ (yearly) and $C_{yb} = CAPEX$ (yearly)[8] .

In order to calculate the value of profit according to presented model we have to define also its components:

- $CAPEX$

  Fixed yearly costs depends on the items presented in the Table 1.

Formally CAPEX can be expressed as:

$$C_{yb} = (TS + ES + CS + PS + OBU + WAN + MM) \tag{2}$$

- $OPEX$

  Similarly as $CAPEX$ also $OPEX$ can be expressed formally:

$$O_{yb} = OTS + OES + OCS + OPS + OOBU + OWAN + OMM \tag{3}$$

.

  From the equation and the Table 1 it is evident that variable costs in the model depend on the operational costs of items defined in CAPEX.

- Inc

  Income from road tolling depends (glej tabelo2) on the number of vehicles that use the toll road, number of kilometers that these vehicles cover and the rate defined for the covered kilometer of

[8] Yearly $CAPEX$ and yearly $OPEX$ are calculated in a way that total amount is divided by number of years, which depends on selected time frame of modeling.

Table 2: Income from road tolling depends on several parameters.

	Inc
n	number of vehicles using toll roads/day (with repetition),
T	rate/covered kilometer,
s	number of kilometers covered/vehicle,
Num	No. of vehicles/day in traffic (without repetition),
Tvr	expected percentage of violators,
$PTvr$	expected percentage of violators who will cover their penalty fee,
Pen	penalty fee in Euros*.

*Described item collected from the penalty fee can be added to the income only when the manager is the penalty fee collector. This kind of regulation is known in Austria where the violator is fined by the manager Asfinag. If the violator does not cover the penalty fee the case is handed over to the police who consider it as the minor offence and the penalty fee is added to the income of state budget.

the toll road. In the detailed calculation it is needed to consider different categories of vehicles as well as different categories of roads and appurtenant rates. In general Inc is calculated through:

$$Inc = n \cdot T \cdot s + Num \cdot Tvr \cdot PTvr \cdot Pen \tag{4}$$

.

4 Simulation

The computer simulation is based on the mathematical model presented in previous chapter. In the mathematical model we will simulate profit regarding diffrent input instances. We named the simulation with the same input instance the scenario. Regarding the input we have 5 different scenarios[18].

4.1 Scenarios

Among all possible scenarios we selected the following. As a basic scenario we selected road tolling of motorways for cargo vehicles and private vehicles. Slovenia which is our test country has 607 km of motorways. For the second scenario we selected road tolling of motorways, first category roads which are parallel with motorways and transit roads, in total length of 1098 km. Third scenario scoops all the roads from the second scenario and the road tolling of second category roads in length of 466 km. Fourth scenario describes so called vignette systems for private vehicles in which the owner of the vehicles pays yearly fee for using the toll roads. In this case the fee does not depend on covered kilometers but on the time of use. The road tolling for cargo vehicles stays the same as in the first three scenarios.

In the fifth scenario the input stays the same only the extent of toll roads narrows to the motorways as in the first scenario. In the table 3 you can find input of all five scenarios in detail.

5 Results analysis

After the analysis of calculation of profit upon all five scenarios we gained the results as seen in the Figure 1. From the presented column diagram it is evident that already in the first year the orange

Table 3: Comparison of input in different scenarios.

model Item	scenario1	scenario2	scenario3	scenario4	scenario5
Road tolling of cargo vehicles	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Road tolling of private vehicles	$\surd$	$\surd$	$\surd$	$-$	$-$
Vignette system for private vehicles	$-$	$-$	$-$	$\surd$	$\surd$
Total length of toll roads; (s)	$607km$	$1089km$	$1555km$	$1089km$	$607km$
Number of OBU for private vehicles	$3mio$	$3mio$	$3mio$	$-$	$-$
Number of OBU for cargo vehicles	300.000	300.000	300.000	300.000	300.000
Number of sold half-year vignettes	$-$	$-$	$-$	$2mio$	$2mio$
Number of sold yearly vignettes	$-$	$-$	$-$	$1mio$	$0,75mio$
Yearly traffic growth, (Tg); [%]	$7,3$	$7,3$	$7,3$	$7,3$	$7,3$
Number of tolling stations, $\frac{s}{8,5}=x$.	72	394	890	394	72
Share of control stations; (%).	15	15	15	15	15
number of vehicles using toll roads/day (with repetition); (n), $PV/3,5-7t/>7$t.	105345, 2450, 16230	210260, 5156, 18560	220720, 5656, 19560	210260, 5156, 18560	105345, 2450, 16230
rate / covered kilometer, private vehicles; (T); $[EUR/km]$.	$0,077$	$0,077$	$0,077$	$-$	$-$
rate / covered kilometer, cargo vehicles$3.5-7.5t$; (T); $[EUR/km]$.	$0,141$	$0,141$	$0,141$	$0,141$	$0,141$
rate / covered kilometer, cargo vehicles$>7,5$t; (T); $[EUR/km]$.	$0,205$	$0,205$	$0,205$	$0,205$	$0,205$
time of modeling	10 years	10 years	10 years	10 years	10 years

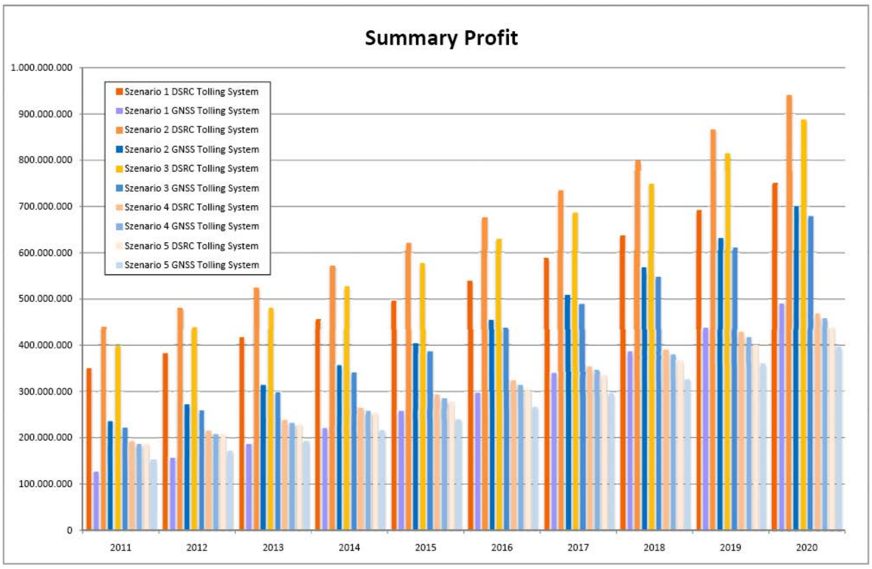


Figure 1: Comparison of five different scenarios using two different technologies.

column which presents the second scenario that uses the DSRC technology rises the highest. Less than ten percent lower is the column of the third scenario which uses the DSRC technology as well. If we look years ahead we can see that profit is absolutely increasing (because of the traffic growth and also because of the expected inflation) however the difference in the absolute value of profit regarding the particular scenario stays the same. If we compare the profit regarding the use of different technologies it is evident that GNSS technology does not defeat the DSRC technology in any of scenarios. Of course we can not assert that regarding the profit the GNSS technology can not defeat the DSRC technology . Our simulations show that in case of road tolling of higher length of lower category roads with shorter sections (i.e. 1 km) the GNSS technology approaches the DSRC technology. Thus in the third scenario both of technologies expressed as percentage are the most equalized. If in this case we would increase the length of low category roads up to 6000 km the GNSS technology would beat the DSRC technology. This would happen because the items TS and ES from the equation 2 would increase too much.

Considering the scenarios on which we performed the simulation we can confirm that the most efficient is the scenario 2 using the DSRC technology. If we look at the Figure 2 it is evident that using the DRSC technology brings 300 millions larger profits already in the first year. In next ten years in which the investments render the amount increases up to 330 millions EUR. The difference in profit appears because of the bigger cost of building and the use of GNSS system. The relation among costs is evident from the Figure which shows that the CAPEX and OPEX costs when using the GNSS technology are significantly bigger.

6 Conclusions

As evident from the facts and presumed conditions, presented in the article, for the purpose of road tolling, the DSRC technology is much more efficient than the GNSS technology. However the technology itself is not disputable for quite some time as all the technologies used for the purpose of road tolling at present are on a very high level of development. More important factors when deciding for the road tolling system are cost efficiency of the system, reliability of the system, capacity of the system, capability of the system to collect as much toll as possible etc. [6, 16].

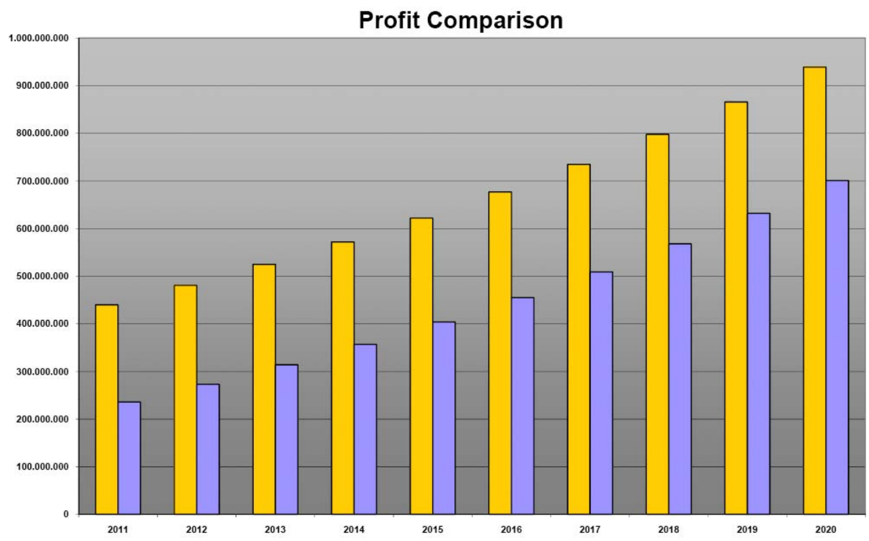


Figure 2: Results of the calculation for the scenario 2

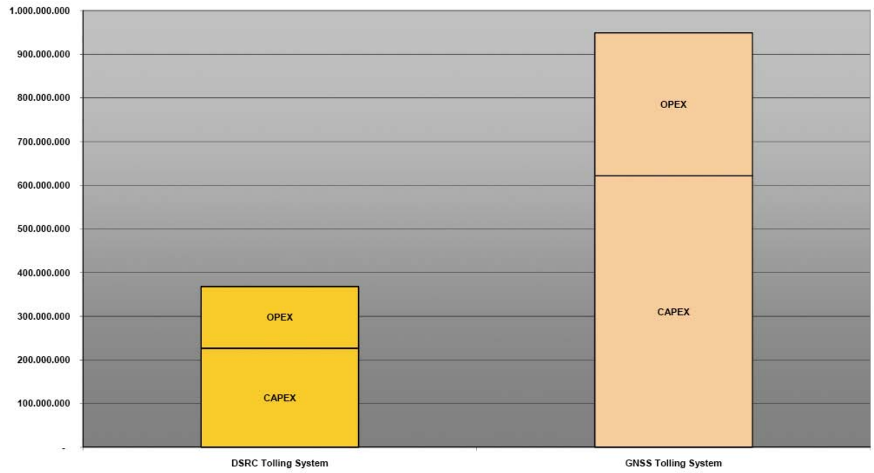


Figure 3: Comparison of CAPEX in OPEX for the scenario 2.

The system founded on the GNSS technology depends on a high degree on the GPS navigation system on which the operators and concessionaires or particular countries where the road tolling is performed, do not influence. Also the transfer of data depends on a high degree on the mobile telephony providers. While the system operation on the basis of the DSRC technology entirely depends on the road tolling manager which aslo owns the infrastructure and is responsible for collecting the toll.

Nevertheless that cost analysis showed bigger efficiency of the system based on DSRC technology, this sometimes is not sufficient to make a decision. The decision usually includes also the elements such as implementation of different transport policies, from detour of traffic to desired roads and many others. Therefore the decision for selecting the road tolling system is and will be political.

References

[1] CESARE III PROJECT, Interoperability of electronic fee collection systems in Europe, http://www.asecap.com/pdf_files/D1.2-RevisedCesareModel-9October2006-Final.pdf.

[2] Commission of the European Communities (1998a). Fair Payment for Infrastructure Use: A phased approach to a common transport infrastructure charging framework in the EU. Brussels.

[3] Commission of the European Communities (1998b). Fair Payment for Infrastructure Use. White Paper.

[4] Commission of the European Communities (1998c). The Common Transport Policy; Sustainable Mobility: Perspectives for the Future. Brussels.

[5] Commission of the European Communities (2006). Keep Europe moving: sustainable mobility for our continent. Brussels.

[6] Commission of the European Communities (2008a). Communication from the Commission to the European Parliament and the Council, Greening Transport. COM(2008) 433 final. Brussels.

[7] Commission of the European Communities (2008b). Communication from the Commission to the European Parliament, Council, the Economic and Social Committee and the Committee of the Regions. Strategy for the internalisation of external costs. COM(2008) 435 final. Brussels.

[8] Commission of the European Communities. (2008c). Proposal for a Directive of the European Parliament and of the Council amending Directive 1999/62/EC on the charging of heavy goods vehicles for the use of certain infrastructures. COM(2008) 436 final. Brussels.

[9] Commission of the European Communities. Directorate-General for Transport- DG 7 (1995). Towards Fair and Efficient Pricing in Transport. Green Paper.

[10] Ecoplan. (2001). Alpnet, Work Package 1, Trans-Alpine Crossing - An Owerview.

[11] European Parliament, Committee on transport and tourism (2008a). Reporter: Gabriele Albertini. Report on sustainable European transport policy, taking into account European energy and environment policies (2007/2147(INI)). Brussels.

[12] European Parliament, Committee on transport and tourism (2008b). Reporter: Saďd El Khadraoui: . Working document on the proposal for a Directive of the European Parliament and of the Council amending Directive 1999/62/EC on the charging of heavy goods vehicles for the use of certain infrastructures. Brussels.

[13] Gilliet, Henri (1990). Toll roads-the French experience. Transrouts International, Saint-Quentin-en-Yvelines.

[14] Harvey J. Miller and Shih-Lung Shaw (2001). Geographic Information Systems for Transportation. Oxford University Press. ISBN 0195123948.

[15] High-tech truck toll system finally launched in Germany, *http://www.computerworld.com/mobiletopics/mobile/story/0,10801,98679,00.html.*

[16] Official Journal of the European Communities (2004). Directive 2004/52/EC of the European Parliament and of the Council of 29 April 2004 on the interoperability of electronic road toll systems in the Communityi. Brussels.

[17] Official Journal of the European Communities (2006). Directive 2006/38/EC of the European Parliament and of the Council of 17 May 2006 amending Directive 1999/62/EC on the charging of heavy goods vehicles for the use of certain infrastructures. Brussels.

[18] Resolucija o prometni politiki Republike Slovenije (RePPRS) (Intermodalnost: čas za sinergijo) v: Uradni list RS ¨t. 58/2006 z dne 06.06.2006 [online]: Http://zakonodaja.gov.si (pridobljeno januar, 2009)

VERY LARGE STREET ROUTING PROBLEM WITH MIXED TRANSPORTATION MODE

Peter Matis
Department of Transportation Networks, Faculty of Management and Informatics, University of Žilina, Slovak Republic, 010 26, e-mail: peter.matis@fri.uniza.sk

Michal Koháni
Department of Transportation Networks, Faculty of Management and Informatics, University of Žilina, Slovak Republic, 010 26, e-mail: michal.kohani@fri.uniza.sk

Abstract: Servicing a large number of customers in a city zone is often a considerable part of many logistics chains. The capacity of one delivery vehicle is limited, but, at the same time, it usually serves a large number of customers. This problem is often called a Street Routing Problem (SRP). New heuristics for solving very large SRP is evaluated based on real data. This paper presents several approximations of length for an SRP with mixed transportation mode and compares them with published approximations used for VRP or Traveling Salesman Problems (TSP). The heuristics was tested in five real world instances ranging from 12000 to 26000 customers.

Keywords: SRP; VRP; heuristics; approximations

1 INTRODUCTION

Postal delivery in agglomerations with large numbers of customers is one of the most difficult operational problems faced by local delivery companies. This paper describes a several methods for the postal delivery in towns with more than 100000 inhabitants, with the object of designing service districts and vehicle collection routes subject to a number of operational constraints.

We want to find and evaluate solutions for cases with more than 10000 customers. In such large agglomerations, service can be performed using mixed transportation mode. The service personnel is driven by car to the first point of the service area, serves the area on foot, and then returns to the depot by car. To find a good solution for these cases, we separated the problem into two steps. In the first step, customers are aggregated to natural clusters and length of SRP for these clusters is estimated. In the second step, a special case of CVRP is solved to find a good car routes.

Our experimental sample was from seven large cities and surrounding villages in Slovakia. The customers were houses in these cities and full street infrastructure is available. The cities were Bratislava, Kosice, Banska Bystrica, Zilina, Trnava, Presov, Nitra. The number of customers for these four agglomerations varies from 11000 to 29000.

Data was collected manually from source maps ZM 1:10000, purchased from the Geodetic and Cartographic Institute of Slovak Republic. Some of the data was collected using GPS receivers. This data is not publicly available at the moment.

The remainder of this paper is arranged as follows. First, a definition of the problem is given. Then, new aggregation methods are presented and a method for approximation of route length for SRP is presented and compared to existing methods. The possible solution for the special case of CVRP for the transportation of postmen to districts is presented.

2 DEFINITION OF A PROBLEM

In today's postal delivery exists a trend to centralize operations of distribution to a small number of centers. This creates a large number of customers that are served from one central depot. Centralizing all delivery operations to one depot for a large number of customers has some advantages and disadvantages. One of the key advantages to centralization is the ability to use more advanced and expensive technologies for preparation before the deliveries are picked up. The preparation of deliveries takes usually more than 10% of the total working time of postmen. By shortening this time combined with good transportation methods, one can allow more time for a delivery and/or reduce number of postmen.

Figure 1 presents possible service areas around the city of Banska Bystrica and its 11414 customers. The area includes Banska Bystrica and 17 surrounding villages.

Figure 1: Customers and depot in the large Banska Bystrica region

The goal is to find good solutions for the SRP in such a large, centralized region. For each customer, there is an average daily service time. Postmen can be driven by car to the starting point of their service route and, at the end of the working day they can be driven back to depot. Deliveries are made on foot. The car's driver is also a postman and serves one district. Transportation expenses are relatively low compared to other expenses. Delivery time is limited for each postman. Time gaps between the physical end of the deliveries and the arrival of the car for returning the postman back to the depot should be tight.

3 THE AGGREGATION OF CUSTOMERS TO NATURAL CLUSTERS AND APPROXIMATION OF ROUTE LENGTH

One of the key problems in solving a very large SRP with mixed transportation mode is to aggregate customers into natural clusters. Each cluster is then treated as one district for the SRP and is served by one postman. The size of the cluster is determined by estimated length of the SRP route, the travel speed of the postman, the delivery time for customers in the cluster, and the total time available for serving the cluster.

Although plenty of work seems to be done in the field of aggregation in general, there is not a specific developed method for SRP aggregation. By aggregating customers, the problem size decreases. On the other hand, one must assume that errors are introduced as customers may be aggregated that do not belong to the same cluster in the optimal solution. In reality, optimality is not the goal, but getting a good solution in a short amount of time is crucial. If aggregation can help in getting such a good, but not necessarily optimal, solution quickly, it will be of interest.

The aggregation of customers on a daily basis is a daunting task simply because there are too many customers and their demands change from day to day. Therefore, assignments made in the previous day might not be applicable for the following day. In order to tackle this problem, we create permanent districts. The logic is that a postman then can know his district well and lower possible errors in delivery while also shortening delivery time.

For aggregation of customers to natural clusters, we used part of Fuzzy Cluster Heuristics (FCH) [6]. This heuristics can be implemented in the following steps:

1. Estimate the minimum number of clusters p needed for serving all the customers using the following formula

$$p = \downarrow \left(\frac{\sum_{c=1}^{N} CST_c + \frac{\sum_{s=1}^{M} D_s}{V}}{SWT} \right) \tag{1}$$

   where symbol $\downarrow$ represents nearest smaller integer number, CST_c is an average service time for the customer c, N is a number of customers, M is a number of street segments where customers are located, D_s is a length of street segment s, V is an average speed of the postman walking on foot, SWT is the available service time of one postman.
2. Locate p medians, so they are uniformly distributed in the serviced area.
3. Create p clusters of customers around these medians using clustering by "fuzzy c-means" (FCM). The membership of customers in each cluster is set as a triangle fuzzy number. The triangle fuzzy number is based on route distance between the cluster median and customer.
4. For each cluster, approximate the route length and estimate the average service time.
5. If there are many clusters (more than 20%) which have their service time over the SWT, increase the number of clusters p and go to step 2, otherwise end the algorithm with the resulting clusters and these are used as SRP districts.

In the literature there are some good samples as to how other authors estimate the length of TSP or VRP [1]. In this algorithm, we used an approximation of length for the SRP route of one postman. In our case, this is actually the length of special TSP, where the postman does not necessarily need to return back to the point where his route started. The topology of network for SRP is specific, and we used the following estimation of the length for one postman's path:

$$TSP(D) = 1.445 * SD \tag{2}$$

where *SD* is a total distance of all the street segments that are served by one postman. This formula was created from a simple regression using sample data with 78 service clusters compared to the best results from heuristics described by Matis [6].

4 MATHEMATICAL MODEL OF SRP WITH MIXED TRANSPORTATION MODE

Mathematical model of SRP with mixed transportation mode after aggregation step can be similar to the mathematical model of Capacitated Vehicle Routing Problem (CVRP) as described by Janacek [3].

$$\min \sum_{r\in R}\sum_{i\in J'}\sum_{j\in J', j\neq i} c_{ij}x_{ijr} \qquad (3)$$

s.t.

$$\sum_{r\in R}\sum_{i\in J', i\neq j} x_{ijr} = 1, \qquad j \in J \qquad (4)$$

$$\sum_{i\in J', i\neq j} x_{ijr} = \sum_{i\in J', i\neq j} x_{jir}, \qquad j \in J', r \in R \qquad (5)$$

$$\sum_{j\in J}\sum_{i\in J', i\neq j} x_{ijr} \leq K_r, \qquad r \in R \qquad (6)$$

$$\sum_{j\in S}\sum_{i\in S, i\neq j} x_{ijr} \leq |S| - 1, \qquad r \in R, S \subseteq J, |S| > 2 \qquad (7)$$

$$x_{ijr} \in \{0,1\} \qquad r \in R, i \in J', j \in J', i \neq j \qquad (8)$$

Here *J* is a set of all clusters, $J' = J \cup \{D\}$ is a set of all clusters and a depot, *R* is a set of all available vehicles and each vehicle can be used only once. c_{ij} is the travel cost from cluster *i* to cluster *j*, K_r is the capacity (number of passengers) of vehicle *r*. x_{ijr} is a decision variable that takes value 1 if the link from *i* to *j* is used by vehicle *r* and cluster *j* is served by this vehicle, 0 otherwise.

Equation (3) is the objective function, which minimizes the total travel costs, (4) ensures that all customers are left exactly once, and (5) ensures that all customers are entered the same number of times as they are left. Equation (6) ensures that total number of postmen in one vehicle does not exceed vehicle capacity, Equation (7) ensures that sub-tours not involving depot are avoided and (8) gives the domain for the decision variables.

5 SOLUTION OF DISTRIBUTION OF POSTMEN TO DISTRICTS

As presented in the model there is need for series of equations (7) that ensure avoiding sub-tours not involving depot. These equations do not allow solving real size problem to optimum because there is too many of them. An additional constraint that we did not include in the model is that the vehicle has to return to the depot using the same route.

We used a TABU search (TABU) metaheuristic [6] that was changed to this specific case. The heuristic was named the TABU for Postman Collection heuristic (TABUPC).

A major difference between the original TABU and TABUPC is that a number of serviced clusters in the route are limited by maximum number of passengers of car. For one trip's length calculation, we only use lengths of segments from depot until the last customer. The car has to return back to the depot using the same route.

The algorithm is based on ideas from Gendreau [2]. The initial solution is generated by simple best position heuristics. Every customer is inserted into the first tour that has enough capacity left over. The customer is inserted into the tour in the best possible way. If no tour with enough free capacity can be found, a new tour is created.

Moves are performed by applying an operator that removes a customer from one route and inserts him into another tour. The insertion of a customer into a tour is done so as to minimize the increase in the length of the tour, but without changing the order of the customers already in the tour. The move is evaluated according to evaluation function:

$$e(s) = D(s) + \mu Q(s) + \lambda \overline{Q}(s) \qquad (9)$$

where $e(s)$ is a value of the evaluation function for a solution *s* after the move, $D(s)$ is the total distance of all tours after the move, μ, λ are positive parameters, μ is dynamically adjusted during the search, $Q(s)$ is the total violation of capacity constraints for the tours, $\overline{Q}(s)$ is the total capacity that is not used for cars used for the tours.

To diversify the search, any solution that has higher value of $e(s)$ compared to solution before the move is given a penalty $p(s) = \omega\delta_{ik}$ that is added to $e(s)$. Here, δ_{ik} is the number of times when the attribute (*i*, *k*) has been part of a *good solution*, that is, a solution that is feasible and has a total length less than η times the length of the best solution found so far. The parameter ω is used to control the intensity of the diversification. These penalties are used to lead the search into less explored parts of the solution space whenever a local optimum is found.

The TABU search starts from the initial solution and moves, at each iteration, to the solution that is nontabu and minimizes $e(s)$. In each step, the attribute (*i*, *k*) that was removed from *A*(*s*) is now declared tabu for several iterations. During these iterations, it is not allowed to move customer *i* back to tour *k*. By the use of a simple aspiration criterion, a tabu move can still be chosen if this leads to a solution that is the best found thus far in the search and is feasible. After each move, the values of the parameters μ, λ are adjusted. If the current solution is feasible, the value of μ is decreased to make it less costly to visit an infeasible solution. In the opposite case, μ is increased to lead the search back into the feasible region of the solution space. If the current solution is feasible and has a total length less than η times the length of the best feasible solution found so far during the search, and the number of iterations performed has reached 200, the solution is considered *good*. Whenever a good solution is found, the values for the δ_{ik} attributes of the found solution are incremented. The search continues until a defined number of moves have been performed.

Figure 2 shows natural clusters of customers with their centers and routes of cars serving these centers. In each car there is the maximum of 5 passengers and 40 clusters are served by 8 cars.

Figure 2: Centers for the natural clusters in large Zilina region served by cars and routes for 8 cars

5 CONCLUSIONS

There are expanding activities in the cities that can be presented as street-based tasks. We introduced some methodology for solving a specific SRP with a very large numbers of customers. We used samples from seven cities for verification of our methodology and algorithms.

We presented a new heuristic for creation of natural clusters and estimation of the route's length. We have not compared results of the heuristics with any other results because there is no publication dealing with this problem. From the small sample, we can conclude that an average calculation time for this part of the solution is 232sec using a PC with Dual Core Processor Intel Pentium. For most of the practical applications, this is acceptable and results are promising.

We created a variation of a TABU search heuristic for the distribution of postmen to their service districts using cars. Results show that the heuristic can be used also for the cases with a heterogeneous vehicle park.

References

[1] Figliozzi M.A., Planning Approximations to the Average Length of Vehicle Routing Problems with Varying Customer demands and Routing Constraints. Transportation Research Record: Journal of the Transportation Board, No 2089, p. 1 – 8.

[2] Gendreau M., Crainic T.G., Cooperative Parallel Tabu Search for Capacitated Network Design., J. Heuristics 8 (6), 601-627, (2002)

[3] Janáček J., Optimization in transport network, EDIS-ŽU, Žilinská univerzita v Žiline, 2003, 248 p.

[4] Matis P., Management of street routing problems using decisions support system, Communications 3, Zilinska univerzita v Ziline, 2006, p. 5 – 8

[5] Matis P., The relationship between quantitative and qualitative measurements in solving of street routing problems, 15th Internatinal Scientific Conference on Mathematical Methods in Economics and Industry, Herlany 2007, p. 144 – 152.

[6] Matis P., Decision support system for solving the street routing problem, TRANSPORT 2008 23(3), ISSN 1648-4142, p. 230 – 236.

[7] Oppen J. – Løkketangen A., Arc routing in a node routing environment, Computers & Operations Research Volume 33, Issue 4, 2006, s. 1033–1055

[8] Poot A. – Kant G. – Wagelmans A., A Saving based method for real-life vehicle routing problems, Journal of the Operational Research Society, Vol. 53, 2002, p. 57-68

Acknowledgment: This work was supported by project "Centre of excellence of information sciences and knowledge systems" ITMS 26220120007

Project Part-Financed by the European Union

European Regional Development Fund

Operačný program VÝSKUM a VÝVOJ

THE LOGISTICS OF HEAVY AND SMALL PACKAGES

Lorenzo Ros*, M^{a} Victoria de la Fuente*, Marija Bogataj**
*Universidad Politécnica de Cartagena, Campus Muralla del Mar
30202 Cartagena, (Spain)
**University of Ljubljana e-mail: marija.bogataj@ef.uni-lj.si

Abstract: The objective of this paper is to present the development of an algorithm for cargo of small but heavy packages assignment in logistics industry. The paper describes how it could increase efficiency in logistics from initial cycle of activities through several days to the end of job in a ceramic industry and similar productions of heavy and small items when time and costs play the role in criterion function indirect through efficiency and profit. Real life application of algorithm developed here has been running on the time horizon of more than week. Though the results of first steps (initial solution) of algorithm are not so good as the results of already known algorithms for transportation assignments, the algorithm is improving the value of criterion function rapidly during further iterations dealing with the sequences of daily assignments, what is major improvement in applications for such kind of algorithms, known up to now. The algorithm was well accepted novelty at beneficiary in ceramics industry.

Keywords: logistics, supply chain management, cargo assignments, transportation costs, algorithm.

1 INTRODUCTION

The characteristic of ceramic products is that the volume of product is relatively small but the pieces are heavy, the setup times are high, and large batches are manufactured. The companies have to invest time and resources in dealing with the information directly related to the grouping of merchandise and customer orders. The process of dealing with this information includes the analysis and thorough knowledge of a series of variables such as transport types, the maximum capacity that can be transported, the analysis of distribution routes, the location of client-companies, the suppliers to be visited on each route, etc. The lack of appropriate tools to support optimal logistics of such items has led to the development of different techniques to facilitate both the calculation and implementation of these tasks. An analysis of the tools which support decision making in logistics of such cargo appearing in the relevant literature consists of three areas of study and three kinds of tools have been developed up to now:

(a) Tools which deal with problems of assignment,
(b) Tools which study problems of capacity, and
(c) Tools which cover the problems of scheduling.

The aim of this paper is to present the development of an algorithm for improving efficiency of road transport in ceramic industry, in order to group and deliver the merchandise contained within client orders in the shortest possible time and with the lowest possible costs. As such, it is necessary to study how the orders or product batches might be grouped and/or divided so as to make them suitable for subsequent loading and distribution. These issues are vitally important due to the variability of the loads associated with transport routes as well as because of variable capacities of available means of transport. This problem is traditionally known as “the knapsack problem”, which combined with the generalized assignment problem (GAP), presents the mathematical base for the development of the proposed algorithm. Many authors worked on this problems like described in Martello and Toth [15] and in many other papers.

Over the last decades different techniques have been developed for the generalized assignment problem (GAP). These have been defined as the assigning of n tasks to m agents with the objective of minimizing operation times and costs. In the continued search for optimal solutions to GAP the use of heuristics has been of special importance. However, the attempt to use these methods in real systems has led to the transformation of classic GAP into a problem of multilevel assignment (MGAP), introducing concepts of associated levels of efficiency to the development of the tasks of the agents. This new problem, first studied in late seventy's was reconsidered by Osorio and Laguna [17] and was later revised by Jaszkiewicz [12] in order to improve the efficiency of MGAP agents through the application of metaheuristics with multiple objectives. New perspectives of the problem have been introduced by: Li and Curry [13] using taboo searches; McLay and Jacobson [16] using dynamic programming approach; and Zhang and Ong [27] by focusing the study of BIGAP through heuristics and applying vehicle routing method (studying profit and cost). One of direct applications of MGAP has been posed recently by Chu [4] who also studied the real problem of loading a lorry. In this study, Chu compares the results obtained from the mathematical model with those of a heuristic algorithm. The realization of a series of experiments yielded almost optimal results by applying heuristics. These results were not only an improvement of results obtained with the mathematical model but were also calculated more quickly. A similar study for "knapsack sharing problem" has been done by Yamada et al. in the number of papers, recently in [26]. The authors developed exact algorithms (branch-and-bound, binary search, combinational optimization) for the knapsack problem, subject to the constraints of fixed capacity and number of items. The results were compared with an existing heuristic algorithm. Better results have been obtained and also better than in Chu [4], especially when dealing with problems on a large scale.

Complementary methods are based on stochastic dynamic programming and Markov Chains approaches as proposed by Li et al. [14], where the algorithm finds a generalised solution to the scheduling of work and prompt delivery to clients. The polynomial algorithm developed general solution which was vague and poorly outlined. Only if the variables (number of clients, routes, locations, etc.) were restricted, the algorithm simplifies and reaches the optimal solution. Having stochastic setting, recently Albareda-Sambola et al. [1] analysed stochastic GAP with recourses (work which is scheduled and processed with determined associated probability). The solutions are reached through Branch-and-Cut algorithms, of which the BCS (simultaneous branch and cut) is the most efficient in identifying an optimal solution. If known solutions of knapsack problem are used for assignment of orders to vehicles, a problem of scheduling arises, as described in [23], [2], [4], [19], [6], and [18]. Yalaoui and Chu [25], and later Tahar et al. [22], used linear programming and heuristic approach, to get very practical tools dealing with real-life problems. Also work of Chang and Lee [3], later revised by He et al. [11], followed this criterions. These authors focus the study of the two-stage scheduling problem on the coordination of the production and delivery schedule, subject to a limited number of vehicles, and with the measurement of customer service level. They suggest that solutions might be found through polynomial algorithms. Within this group of problems, it is also worth pointing out the solutions developed by Dunstall and Wirth [7], [8], and Wang and Chen [24], in regard to the problem of machinery scheduling (in parallel, in series, equal and/or different) in order to minimize working times. The studies carried out suggest the development of a B&B algorithm, to which a series of heuristics (SWPT, SWMPT, LWST, LBT, LSU ...) might be applied, thus enabling the programming of a series of work (grouped into families) that would be processed by identical machinery in parallel. These studies managed to develop simple, effective and efficient heuristics for scheduling of work carried out by identical machines. Bur in logistics the identical objects are not the case in general. Studying the problems of capacity, scheduling, and assignment, we have been drawing attention to the following aspects: **(a)** The development of exact mathematical models has been traditionally used in the study of these problems. However, it is very difficult to find optimal solutions, due to the complexity of both the real problem as well as the mathematical model that represents it. **(b)** The use of other mathematical techniques, such as B&B and heuristics, offers other approaches in the analysis of these problems. These techniques facilitate the search for solutions, reducing the calculation time and simplifying the real problems. As such their application is being widely studied by the scientific community. **(c)** Many contributions quoted so far, have been focused on finding solution through the use of algorithms or calculations carried out in stages, often two stages like in [20], [11], [1], etc. These methods avoid the need of complete recalculation of algorithm. Developing the transport assignment algorithm presented in this paper, the authors started from the results of Ferreira and Pradin [9], who rely on a method of two phases, firstly selecting and assigning families of pieces (phase 1) and secondly establishing a new division and reallocation in order to minimize possible inefficiency (phase 2). Here the idea of Osorio and Laguna [17], who introduced in their algorithm different levels of efficiency of multiple resources or agents after the assignment, has also been considered. Below, the improved procedure and solution for assignment of loads and orders to transportation means will be discussed. This assignment is determined according to different capacities of transport options, collecting orders, minimizing division of orders, and improving efficiencies in loading.

2 DESCRIPTION OF THE ALGORITHM

The algorithm presented in this paper, is based on the technique of Branch and Bound previously developed by Laursen and later described and upgraded by Gendron and Cranic, [10], and Clausen [5]. This technique when applied in conjunction with decision rules to a combined problem of transport, capacity and assignment will enable the development of an algorithm that facilitates the collection of specified merchandise in client orders in the shortest possible time and with the lowest possible costs. For the case presented, the analysis of the information collected by the logistic operator was necessary. This information which already existed within the company is not always available to the field of logistics, which required our additional effort. The algorithm will verify the possibility of coordinating objectives which were previously seen as incompatible, such as an optimisation of logistic resources and the reduction of transport costs in addition to the costs involved in warehousing for the logistic operator. In this sense, the objective of the algorithm is to maximise the efficiency of the use of different means of transport, as well as the loading obtaining a maximization of profits for all those agents involved in the value chain. In profit optimization all other requirements are better consider as partial costs minimization approach.

2.1 Notation

We introduce the following notation:

- $\mathbf{R_k}$: k-th of existing routes, chosen by the logistic operator, according to the location of suppliers, geography, infrastructure, etc. For instance, when the logistic operator defines 6 routes (R1, R2, ..., R6) which allow to pick up all the suppliers' orders received daily, then indicator k varies from 1 to 6 (k = 1,2,3,4,5,6).
- $\mathbf{T_i}$ **:** i-th type of vehicle for road transportation, classified according to it's maximum load capacity, and their availability for use by the logistic operator. For example, if the logistic operator is going to use 5 different type of vehicles, (T1, T2, ..., T5), then i varies from 1 to 5 .
- $\mathbf{S_j}$**:** j-th supplier, mainly one of ceramic manufacturers, who use the services of a logistic operator and who is included in the logistic operator's area of interest.

- $\mathbf{S_{jk}}$: j-th supplier assigned to road k by the logistic operator, where the suppliers' assignment to each route is carried out according to the proximity between supplier and route.
- $\mathbf{O_{mj}}$**:** m-th order issued by j-th supplier (ceramic manufacturer). For instance, if supplier 2 issues 3 orders to be picked up by the logistic operator, then j=2 and m=1, 2, 3 (where indicator l points out the number of orders, and indicator j shows which supplier issues that orders). Therefore, Om2=[O_{12}, O_{22}, O_{32}] will be all the orders issued by supplier 2.
- $\mathbf{L_n}$: A special type of orders, called "little orders", with an associated load of less than 3000 kg (L1) or remainders of non-assigned orders that will be detailed during developing the algorithm, in step 3.6 of point 2.3 (L2,L3,L4). All of these little orders will be grouped together in order to be assigned to a means of transport.
- $\mathbf{C_i}$**:** The cost assigned to i-th means of transport (T_i), generally by size and load capacity. As an example, if the logistic operator only uses transport means T2 and T4, during a day, the costs he must consider are C_2 y C_4.

The **algorithm constraints** which are established with the assignment of the transport and its orders are directly related to the variables previously mentioned.

- The number of available road transport vehicles of type i: $\overrightarrow{N^o(T_i)}=[T_i]$ (c-1)
- Maximum load of type i transport (Ti): $Capacity\ (T_i)=[c_i]$ (c-2)
- The maximum number of combinations of routes that can be carried out by the different T_i:

$$\overrightarrow{ComR(T_i)}=[r_i] \qquad (c\text{-}3)$$

- The maximum no. of trips to base that i-th vehicle can make:

$$\overrightarrow{Tripstobase(T_i)}=[v_i] \qquad (c\text{-}4)$$

- The maximum number of suppliers that can be visited by vehicle i on route k, is denoted by w_{ik}. There exists the matrix of maximal number of suppliers per road. [w_{ik}]=Matrix $[Transport \times routes]$. (c-5)

2.2 The rules for the assignment of available transportation means to chosen transportation routes

In order to realise the allocation of vehicles to the routes a series of Rules of Assignment have been determined, thereby establishing a prioritized order together with its application to the route.

- *RULE 1:* Assignment of the means of road transport in decreasing order of load associated to each of routes.
- *RULE 2:* Assignment of a vehicle to a route if its capacity is greater or equal to the load picked up on the route.

$$Load(R_k) \leq Capacity\ (T_i)$$

- *RULE 3:* Division of orders if the load to be collected on a route is greater than the capacity of the assigned vehicle.

$$Load(R_k) > Capacity\ \ (T_i)$$
$$Load(R_k) = Load_1(R_k) + Load_2(R_k)$$
$$Load_1(R_k) = Capacity\ \ (T_i)$$

- *RULE 4:* Division of the orders if the load to be collected for one supplier only is greater than the capacity of the allocated vehicle.

$$Load(S_{jk}) > Capacity\ \ (T_i)$$

$$Load(S_{jk}) = Load_1(S_{jk}) + Load_2(S_{jk})$$
$$Load_1(S_{jk}) = Capacity\ (T_i)$$
$$Load_2(S_{jk}) = \sum_{non-assigned} Load(O_{mj})$$

- *RULE 5:* The division of orders if the load to be collected on a route is less than the capacity of the assigned vehicle, but fails to meet the constraint of the maximum suppliers to visit per route $(n^o \max S_{jk})$.

$$Load(R_k) < Capacity(T_i)$$
$$Load(R_k) = Load_1(R_k) + Load_2(R_k)$$
$$Load_1(R_k) \quad < Load(n^o \max S_{jk})$$
$$Load_2(R_k) = \sum_{non-assigned} Load(S_{jk})$$

- *RULE 6:* The assignment of vehicles for the collection of possible little orders (routes with a load less than 3000 kg.) and the remainder of orders through the combination of loads not yet assigned to the different routes.

The rules defined for assignment should be applied with differing priority depending on the load to be collected:

- The rules of assignment 1 and 2 are the most important and are applied simultaneously.
- Once the rules have been applied, rules 3, 4 and 5 are applied. These rules are mutually exclusive and depend upon the results of the application of rules 1 and 2.
- Finally rule 6 is applied.

2.3 The stages of algorithm

STAGE 1: THE GROUPING OF SUPPLIERS BY ROUTE.

The different transport routes are defined by the logistic operator; such routes are decided in relation to the location of the supplying companies (manufacturers). The supply companies are assigned to a particular route according to their proximity to each route.

The allocation of each supply company (S_j) to a route (R_k) can be noted as S_{jk}:

STAGE 2: CALCULATION OF DAILY LOADS.

At this stage of the algorithm the daily loads for each supplier are calculated, denoted by *Load*(S_{jk}), as well as the daily load for each route, *Load*(R_k), as defined in stage 1.

Step 2.1. Calculation of the load according to all orders issued by each supplier:

$$Load(S_{jk}) = \sum_{l=1}^{d} Load(O_{mj}) \quad \text{(units: kg)}$$

Step 2.2. Calculation of the load according to all suppliers assigned to each route:

$$Load(R_k) = \sum_{j=1}^{b} Load(S_{jk}) = \sum_{j=1}^{b} \sum_{m=1}^{d} Load(O_{mj}) \qquad \text{(units: kg)}$$

STAGE 3: ASSIGNING THE MEANS OF TRANSPORT.

At this stage the means of transport needed to collect all the loads of the received orders are assigned to the different defined routes. It is at this point where the rules of allocation defined in Section 2.2 are applied. For the application of the rules of load allocation the following steps need to be followed:

Step 3.1. Selection of the route with the greatest load and assignment of the vehicle T_i with a capacity similar to this load (application of rules 1 and 2 simultaneously, as described in section 2.2):

T_i is assigned if $Load(R_k) \leq Capacity\ (T_i)$, $\forall\ i \rightarrow i=\{1....a\}, \forall\ k \rightarrow k=\{1....c\}$

Step 3.2. In case when the load of a route is greater than the capacity of the non assigned transport:

$$Load(R_k) > Capacity\ (T_i)$$
$$\forall\ i \rightarrow i=\{1....a\}$$
$$\forall\ k \rightarrow k=\{1....c\}$$

the division of the route's load is undertaken as indicted:

$$Load(R_k) = Load_1(R_k) + Load_2(R_k)$$

According to the following procedure:

$$Load_1(R_k) = Capacity\ (T_i)$$
$$Load_2(R_k) = Load(R_k) - Load_1(R_k)$$
$$\forall\ i \rightarrow i=\{1....a\},\ \forall\ k \rightarrow k=\{1....c\}$$

Step 3.3. If the load to be transported for any individual supplier is greater than the capacity of the assigned vehicle:

$$Load(S_{jk}) > Capacity(T_i)$$
$$\forall\ i \rightarrow i=\{1....a\},\ \forall j \rightarrow j=\{1....b\},\ \forall\ k \rightarrow k=\{1....c\}$$

The division of the order is carried out in the following way (as rule 4 points out):

$$Load(S_{jk}) = Load_1(S_{jk}) + Load_2(S_{jk})$$
$$\forall j \rightarrow j=\{1....b\},\ \forall\ k \rightarrow k=\{1....c\}$$

and where: $Load_1(S_{jk}) = Capacity(T_i)$

$$\forall\ i \rightarrow i=\{1....a\},\ \forall j \rightarrow j=\{1....b\},\ \forall\ k \rightarrow k=\{1....c\}$$
$$Load_2(S_{jk}) = \sum_{non-assigned} Load(O_{mj})$$
$$\forall j \rightarrow j=\{1....a\}, \forall\ k \rightarrow k=\{1....c\},\ \forall\ m \rightarrow m=\{1....d\}$$

Step 3.4. In the event of a load assigned to a transport route *Load(R$_k$)* being less than the capacity of the allocated transport T_i, but failing to meet the constraint of the maximum number of suppliers S_{jk} to be visited (in exceeding that number), the load is divided in the following manner:

$$Load(R_k) = Load_1(R_k) + Load_2(R_k)$$

where:

$$Load_1(R_k) = Load(n^{o} \max S_{jk})$$
$$Load_2(R_k) = \sum_{non-assigned} Load(S_{jk})$$
$$\forall j \rightarrow j=\{1....b\},\ \forall\ k \rightarrow k=\{1....c\}$$

Step 3.5. Having completed the allocation of transport to the route with the greatest load, the allocation of vehicles continues with the route with the next largest load. For this, steps 3.1, 3.2, 3.3, and 3.4 are carried out. This process should be continued until all the routes with loads greater than 3000kg have been processed.

Step 3.6. The allocation of little orders and remainders of non assigned orders:

L_n= [L_1, L_2, L_3, L_4]

L_1 (or little orders) are defined as having an associated load of less than 3,000kg, and as such are the last to be assigned within the calculation of the algorithm:

$L_1 = Load\ (S_{jk}) \leq 3000kg$

Grouped together with small orders are the remaining orders (that is, those which have been left without the assignment of a means of transport). These remaining orders (L_2, L_3, L_4) arise from the application of the following rules:

i. L_2 is defined from application of rule 3 in step 3.2.:

$$L_2 = Load_2(R_k) = Load(R_k) - Load_1(R_k) \leq 3000 \text{ kg}$$

(where $Load_2(R_k)$ is obtained when $Load(R_k) > Capacity(T_i)$)

ii. L_3 is defined from application of rule 4 in step 3.3.:

$$L_3 = Load_2(S_{jk}) = \sum_{non-assigned} Load(O_{mj}) \leq 3000 \text{ kg}$$

(where $Load_2(S_{jk})$ is obtained when $Load(S_{jk}) > Capacity(T_i)$)

iii. L_4 is defined from application of rule 5 in step 3.4:

$$L_4 = Load_2(R_k) = \sum_{non-assigned} Load(S_{jk}) \leq 3000 \text{ kg}$$

(where $Load_2(R_k)$ is obtained when $Load(R_k) < Capacity(T_i)$)

Having determined the different types of little orders to be collected in this stage (L_1, L_2, L_3, L_4), those that remain (without assignment) T_i will be assigned to the possible groups of little orders as described below:

$$Capacity\ (T_i) \geq \sum_{n=1}^{4} L_n$$

$$\forall\ i \rightarrow i=\{1....a\},\ \ \forall\ n \rightarrow n=\{1...4\}$$

And while considering the constraints applying to the combination of routes (c-3), the number of trips to base (c-4), and the suppliers to be visited (c-5).

STAGE 4. CALCULATION OF COST

With all the orders received by the logistic operator now assigned to the selected vehicles T_i (Stage 3), attention is now turned, in this section, to the calculation of the total cost (Ct) associated to the use of the selected vehicles T_i :

$$Ct = \sum_{i=1}^{a} \delta_i * C_i \qquad \text{where } \delta_i = \begin{cases} 0, & \text{if } T_i \text{ is not used} \\ 1, & \text{if } T_i \text{ is selected} \end{cases}$$

STAGE 5. CALCULATION OF THE EFFICIENCIES IN THE ASSIGNMENT OF THE MEANS OF ROAD TRANSPORT

To check an optimal assignment of vehicles in this stage, a calculation of the efficiency associated to each vehicle is carried out. The definition of efficiency is dependent upon the requirements of the logistics operator; here it is defined as the use of a means of transport in relation to the maximum number of suppliers that this means of transport can visit in a route.

Step 5.1. Calculation of the efficiency(α) for each selected T_i in Stage 3:

$$T_i \rightarrow \alpha_i \ , \ \ \alpha_i \in [0,1] \ , \ \ \forall_i \rightarrow i=\{1....a\}$$

where:

$$\alpha_i = \frac{Suppliers.Visited.by.T_i.in.R_k}{Maximum.N^o.of.Suppliers.to.Visit.in.R_k\_(w_{ik})}$$

$$\forall i \rightarrow i=\{1....a\}$$

$$\forall k \rightarrow k=\{1....c\}$$

Step 5.2. If the efficiencies calculated for each T_i are equal or greater than 0.6, the assignment is final, as this is a level considered acceptable by the logistic operator. As such, this ends the calculation with the algorithm.

Step 5.3. If efficiency calculation gives the result less than 0.6, the assignment of means of transport T_i should be repeated, (returning once more to the calculation of Stages 3 and 4).

3 EXPERIMENTAL RESULTS

To check and develop the algorithm various tests have been carried out using historical data of the logistic operator for different days in the past. This includes information on the various means of transport and their maximum capacities, as well as, the routes determined by the logistic operator, and the suppliers located on the individual routes.

The details relating to the logistic operator are:

- Existing routes: R_k = [A, B, C, D, E, F]
- Type of transport means: T_i = [T_1, T_2, T_3, T_4, T_5, T_6, T_7]
- Transport load capacity, number of vehicles available, the combination of the maximum routes that each vehicle can make, the number of empty trips to base to unload that each vehicle can make, and the daily cost of use or hire of each of the vehicles are given in the table 1 of input data:

Table 1: Input data

VEHICLE	T1	T2	T3	T4	T5	T6	T7
CAPACITY (kg)	25.000	16.000	8000	5500	4500	4000	1100
Number of transport means	2	2	2	2	2	2	1
Maximum number of routes	3	3	4	4	4	4	6
Number of trips to base	2	2	3	3	3	3	4
Costs Euros/day	300	220	170	160	160	160	125

Table 2: Maximum number of suppliers that can be visited by each transport according to route:

	A	B	C	D	E	F
T1	12	10	8	6	6	9
T2	12	10	8	6	6	9
T3	15	14	12	6	6	14
T4	15	14	12	6	6	14
T5	17	18	13	6	6	15
T6	17	18	13	6	6	15
T7	18	18	14	6	6	15

The activities of a logistic operator specifically include the following:

- Daily collection of all orders received during the previous day.
- Considering the level of efficiency (acceptable program if efficiency is evaluated by greater than 0.6.
- Special treatment of the vehicle T7, which belongs to the company (a van with a reduced capacity). As this vehicle is considered a company asset and need not be rented it is regarded having fixed cost, although it is in continuous use its level of efficiency tends to be low due to its use in collecting scattered orders.

The results of two of the tests carried out are shown below:

Example 1: Data relating to 3/1/2006

The day before 3/1/06, 32 orders from different suppliers were received, as shown in total amount in table 3, where the corresponding suppliers' orders have been recorded according to the routes established by the logistic operator.

Table 3: Orders received on 3/1/2006, grouped for each of the routes.

Route	A	B	C	D	E	F
Load (Kg)	9245	7252	5485	217	-	63.412
Nº Suppliers to visit	3	8	2	1	-	18

The first iteration of the algorithm reveals that for a total cost of 1495 € there is a level of efficiency less than 0.6 in the majority of the assigned transport. After the calculation of a new iteration, the new assignment enables an increase in the level of efficiency of all the logistic operator's vehicles, and the corresponding cost falls to 1115 € .

Example 2: Data relating to 19/1/2006

The day before 19/1/06, 61 orders from different suppliers were received, in amount as can be seen in table 4, where the corresponding suppliers' orders have been added according to the routes determined by the logistic operator.

Table 4: Orders received on 19/1/2006, grouped for each of the routes.

Route	A	B	C	D	E	F
Load (Kg)	40.906	24.402	4.577	12.517	1.885	23.898
Nº Suppliers to visit	14	22	4	2	1	18

The first iteration of the algorithm resulted in the assignment of various transports (T1, T3, T6) with efficiencies of less than 0.4 (unacceptable), and entailing a cost of 1985 € for the defined assignment. The calculation of a second iteration improves the use of the transport, increasing efficiencies and eliminating the assignment of vehicle T3. Although the modification of the first assignment reduces the cost (1645 €), it decreases the efficiency of T2, and as such calls for continued iteration in order to meet the constraint of obtaining efficiencies of greater/higher than 0.6. This constraint was met after a third iteration and resulted in a cost of 1485 € for the new assignment.

4 CONCLUSIONS

The development of the exact mathematical models has been the traditional way to study similar kind of assignment problems. However, as we have seen, it is difficult to obtain the optimal solutions, due to the complexity of both the real problem as well as the mathematical problem that represents it.

The use of alternative mathematical techniques, such as B&B and heuristics, contribute to the analysis of these problems. These techniques facilitate the search for solutions, reducing the time taken in calculations, and simplifying the formalization of real problems, and as such their application is being widely studied by the scientific community. The analysis and study of the problems associated with the work of the logistic operator has influenced the creation of a tool and a calculation procedure (algorithm) which optimises the use of available transport resources both those owned by the company as well as those obtained from elsewhere.

The goal that we pursue with the development of the algorithm for the assignment of transport at a logistic operator is to reduce the complexity of the variables used in the assignment of the means of transport. A goal has been achieved. The number of variables has been reduced from a group of ten, that the logistic operator originally worked with, to five variables given by the algorithm. This has been achieved without reducing the information considered in decisions making. The algorithm is focused on optimising the use of the means of land transport allowing improvements in the collection and grouping of orders for later distribution. The assignment is made in relation to the capacity, categorising each possible transport for the collection of orders in order to minimize the involvement of each vehicle and as such reducing the inefficiencies of the load, thereby achieving the main aim, that of reducing the costs of transport. The use of the algorithm, by a logistic operator, offers also the following indirect advantages, which has not been associated to the formalization of the previous problem:

a. Once the information corresponding to the orders and the specific characteristics of the range of products is analysed and systemised an increase in the use of resources follows.
b. The service offered to clients is improved by reducing lead times.
c. The logistic resources are rationalised thereby minimising the underutilisation of resources and reducing the associated costs.
d. The algorithm deals with the study of the type of client as well as the orders and the requirements of the logistic services offered.

The application of the algorithm in a real situation of a Logistic Operator was carried out on orders received (loads to be distributed) over a number of days. The application of the algorithm with real data has demonstrated that the algorithm does not always, at first, produce an optimal result. This is due to the underutilisation of some of the selected means of transport and as such surpasses the values considered acceptable. However, after a second or third iteration of the application of the algorithm to the means of transport under consideration, the algorithm demonstrated, simultaneously, the following aspects:

a. Maximise the use of transport.
b. Increase the efficiencies of the means of transport of the logistic operator.
c. Minimize the costs derived from the use of the means of transport.

References

[1] Albareda-Sambola M., van der Vlerk M., Fernández E. (2006). Exact solutions to a class of stochastic generalized assignment problems. EJOR 173, 465-487.

[2] Caccetta L., Hill S.P. (2001). Branch and Cut Methods for Network Optimization. Mathematical and Computer Modelling 33, 517-532.

[3] Chang Y.C., Lee C.Y. (2004). Machine scheduling with job delivery coordination. EJOR, 158, 470-487.

[4] Chu C. (2005). A heuristic algorithm for the truckload and less-than-truckload problem. EJOR, 165, 657-667.

[5] Clausen J. (1999). “B&B Algorithms – Principles and Examples”. www.imada.sdu.dk

[6] Climer S., Zhang W. (2006). Cut-and-solve: An iterative search strategy for combinatorial optimization problems. Artificial Intelligence, 170, 714-738.

[7] Dunstall S., Wirth A. (2005). Heuristic methods for the identical parallel machine flowtime problem with set-up times. Computers & Operations Research, vol.32, pp.2479-2491.

[8] Dunstall S., Wirth A. (2005[a]). A comparison of branch-and-bound algorithms for a family scheduling problem with identical parallel machines. EJOR. Vol.167, pp.283-296.

[9] Ferreira J.F., Pradin B. (1993). A methodology for cellular manufacturing design. International Journal of Production Research, 31 (1), 235-250.

[10] Gendron B., Cranic T.G. (1994). Parallel Branch-and-Bound Algorithms: Survey and Synthesis. Operations research, 42 (6), 1042-1066.

[11] He Y., Zhong W., Gu H. (2006). Improved algorithms for two single machine scheduling problems. Theoretical Computer Science 363, 257-265.

[12] Jaszkiewicz A. (2004). On the computational efficiency of multiple objective metaheuristics. The knapsack problem case study. EJOR, 158, 418-433.

[13] Li V.C., Curry G.L. (2005). Solving multidimensional knapsack problems with generalized upper bound constraints using critical event tabu search. Computers & Operations Research, 32, 825-848.

[14] Li C., Vairaktarakis G., Lee C. (2005). Machine scheduling with deliveries to multiple customer locations. EJOR, 164, 39-51.

[15] Martello, S., Toth, P. (1981). An algorithm for the generalized assignment problem. Proceedings of the 9th IFORS Conference, Hamburg, Germany.

[16] McLay L., Jacobson S. (2007). Algorithms for the bounded set-up knapsack problem. Discrete Optimization 4, 206-212.

[17] Osorio M., Laguna M. (2003). Logia cuts for multilevel generalized assignment problems. EJOR 151, 238-246.

[18] Pereira M., Valério J.M. (2007). A branch-and-price algorithm for scheduling parallel machines with sequence dependent setup times. EJOR 176, 1508-1527.

[19] Ralphs T.K. (2003). Parallel branch and cut for capacitated vehicle routing. Parallel Computing, 29, 607-629.

[20] Realff M.J., Kvam P.H., Taylor W.E. (1999). Combined analytical and empirical learning framework for branch and bound algorithms: the knapsack problem. Artificial intelligence in engineering, 13, 287-300.

[21] Rouillon S., Desaulniers G., Soumis F. (2006). An extended branch-and-bound method for locomotive assignment. Transportation Research – Part B, 40, 404-423.

[22] Tahar D., Yalaoui F., Chu C., Amadeo L. (2006). A linear programming approach for identical parallel machine scheduling with job splitting and sequence-dependent setup times. Int.J. Prod.Econ 99, 63-73.

[23] Toth P. (2000). Optimization engineering techniques for the exact solution of NP-hard combinatorial optimization problems. EJOR, 125, 222-238.

[24] Wang X., Cheng TCE. (2005). Two-machine flowshop scheduling with job class setups to minimize total flowtime. Computers & Operations Research, 32, 2751-2770.

[25] Yalaoui F., Chu C. (2003). An efficient heuristic approach for parallel machine scheduling with job splitting and sequence-dependent setup times. IIE Transactions 35 (2), 183-190.

[26] Yamada T., Fujimoto M., (2006). An exact algorithm for the knapsack sharing problem with common items. European Journal of Operational Research, 171, 693-707.

[27] Zhang C., Ong H. (2007). An efficient solution to biobjective generalized assignment problem. Advances n Engineering Software 38, 50-58.

OR AND SIMULATION IN COMBINATION FOR OPTIMIZATION

Nico M. van Dijk†, Erik van der Sluis†
René Haijema‡, Assil Al-Ibrahim† and Jan van der Wal†
† University of Amsterdam, Faculty of Economics and Business
Roetersstraat 11, 1018WB, Amsterdam, the Netherlands
{n.m.vandijk, h.j.vandersluis, a.alibrahim, j.vanderwal}@uva.nl
‡ Wageningen University and Research Centre, Logistics, Decision and Information sciences,
Hollandseweg 1, 6706 KN Wageningen, the Netherlands
r.haijema@wur.nl

Abstract: This paper aims to promote and illustrate the fruitful combination of classical Operations Research and Computer Simulation. First, an instructive example of parallel queues will be studied. This example already shows the necessary combination of queueing and simulation. Next, two more complex 'real life' applications show that this OR-Simulation combination can be most useful for 'practical optimization'.

Keywords: call centers, blood platelet production, train scheduling, simulation.

1 INTRODUCTION

Simulation, or more precisely as will be meant throughout this paper: discrete event simulation, is well known as a most powerful tool for process and performance evaluation in a vast majority of fields. Standard applications are found in the production sector, the service industry (call centers, administrative logistics), transportation (public transportation systems, road traffic, airports, harbours, maritime, express delivery systems and so on) and most recently health care.

In most of these applications simulation is necessarily required, as analytic techniques, most notably OR (Operations Research)-techniques such as queueing analysis and mathematical programming, are insufficient due to:

- the complexity of the system,
- the underlying simplifying assumptions required,
- the various types of (non-exponential) stochastics involved.

Clearly, if an optimization problem can be parameterized, such as typically for capacity determination, different simulation search approaches can be suggested to expedite and automate the search for optimal values. An elegant exposé of such methods can be found in [3]. Nevertheless, it remains to be realized that simulation is nó optimization tool by itself.

This is where OR might contribute in either of two directions:

i) To suggest scenarios based upon OR-results and insights.
ii) To provide an OR-optimization techniques.

A combination of OR and simulation might then become beneficial:

- Simulation for its evaluation,
- OR for its optimization.

The advantages of this combination for optimization are schematically represented in Table 1

Table 1: Combined Advantages.

OR – Advantages	Simulation - Disadvantages
Optimization By techniques Also by insights	Simple models Strict assumptions
OR - Disadvantages	**Simulation - Advantages**
Evaluation By scenarios By numbers only	Real-life stochastics Real-life complexities

By this paper both directions of this combination will be illustrated by specific practical applications: direction i) in section 2 and direction ii) in sections 3 and 4. The applications all rely upon practical and recent research and schematically concern the following topics and combinations with simulation:

Section	Topic	Combination
2	Pooling in call centers	SIM + Q
3	Blood banks	SIM + DP
4	Railways	SIM + Q + DP
Legend: SIM: Simulation; Q: Queueing; DP: Dynamic Programming		

The paper collects and is based on separate papers for each of these applications. Its primary focus is to convey and promote the message of the practical potential of a combined OR-Simulation approach. This message as well as the instructive example (section 2.1) can also be found in [6]. Sections 2.2 and 2.3 and both the more complex real life applications in sections 3 and 4 are new.

2 TO POOL OR NOT

Should we pool servers or not? This seems a simple question of practical interest, such as for counters in postal offices, check-in desks at airports, physicians within hospitals, up to agent groups within or between call centers. The general perception seems to exist that pooling capacities is always advantageous.

2.1 An instructive example (Queueing)

This perception seems supported by the standard delay formula for a single (exponential) server with arrival rate λ and service rate μ: $D = 1 / (\mu - \lambda)$. Pooling two servers thus seems to reduce the mean delay by roughly a factor 2 according to $D = 1 / (2\mu - 2\lambda)$.

However, when different services are involved in contrast, a second basic result from queueing theory is to be realized: Pollaczek-Khintchine formula. This formula, which is exact for the single server case, expresses the effect of service variability, by:

$\boldsymbol{W_G} = ½\,(1+c^2)\,\boldsymbol{W_E}$ with $c^2 = \sigma^2 / \tau^2$ and

$\boldsymbol{W_G}$ the mean waiting time under a general (and E for exponential) service distribution with mean τ and standard deviation σ.

By mixing different services (call types) extra service variability is brought in which may lead to an increase of the mean waiting time.

This is illustrated in Figure 1 for the situation of two job (call) types 1 and 2 with mean service (call) durations $\tau_1 = 1$ and $\tau_2 = 10$ minutes but arrival rates $\lambda_1 = 10\ \lambda_2$. The results show that the unpooled case is still superior, at least for the average waiting time $\boldsymbol{W_A}$. Based on queueing insights, more precisely Pollaczek-Khintchine's formula, a two-way or one-way overflow scenario can now be suggested for further improvement as also illustrated in Figure 1.

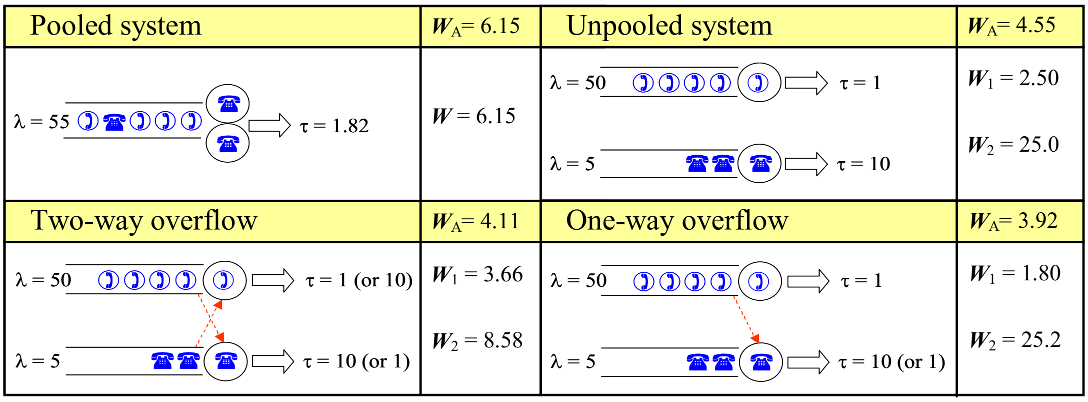


Figure 1: Pooling scenarios.

A Combined approach

To achieve these improvements simulation is necessarily required as overflow systems cannot be solved analytically. A combination of queueing for its insights to suggest scenarios and of simulation for evaluating these scenarios thus turns out to be fruitful.

Call centers (large number of servers)

Similar results can also obtained for larger number of servers, say with 10, 50 or 100 servers, such as arising in realistic call centers. This is illustrated in Table 1. The one-way overflow scenario turns out to be superior to both the pooled and the unpooled scenario for realistically large numbers of call centers agents. (Here the mix ratio of short and long services is similar as in the example above. For further details, see [5].)

Table 2: Results for two server groups each with s servers.

	Pooled	Unpooled			One-way overflow			
s	W_P	W_1	W_2	W_A	W_1	W_2	W_A	% overflow
1	11.53	4.49	45.24	8.18	3.40	45.47	7.20	5.1%
5	1.76	0.78	7.57	1.41	0.53	7.75	1.20	4.9%
10	0.71	0.34	3.52	0.63	0.21	3.66	0.52	4.5%
20	0.26	0.15	1.44	0.26	0.08	1.54	0.21	4.0%
50	0.05	0.04	0.38	0.07	0.02	0.42	0.05	2.9%

It can also be shown that a simple priority rule, particularly a preemption scenario, for short (type 1) jobs, generally seems to perform quite well and to be 'optimal' among simple scenarios. Unfortunately, preemption (interruption) of service will generally be unacceptable or impractical.

For practical interest, therefore, we aim to investigate whether an improvement over a simple overflow or non-preemptive rule can be obtained by more sophisticated non-preemptive rules using threshold values on the queue lengths.

2.2 Threshold policies

To this end, let m_i the number of type i jobs waiting, i = 1, 2. A threshold rule is then described by:

Thr$(\theta_1^1, \theta_2^1; \theta_1^2, \theta_2^2; \Omega_1, \Omega_2)$:

A server of type 1 serves jobs from queue 2 if either
(i) $m_2 \geq \theta_2^1 \wedge m_1 < \theta_1^2$ or (ii) $m_1 = 0 \wedge m_2 \geq \Omega_2$;
otherwise, it serves jobs from queue 1.

A server of type 2 serves jobs from queue 1 if either
(i) $m_1 \geq \theta_1^1 \wedge m_2 < \theta_2^2$ or (ii) $m_2 = 0 \wedge m_1 \geq \Omega_1$;
otherwise, it serves jobs from queue 2.

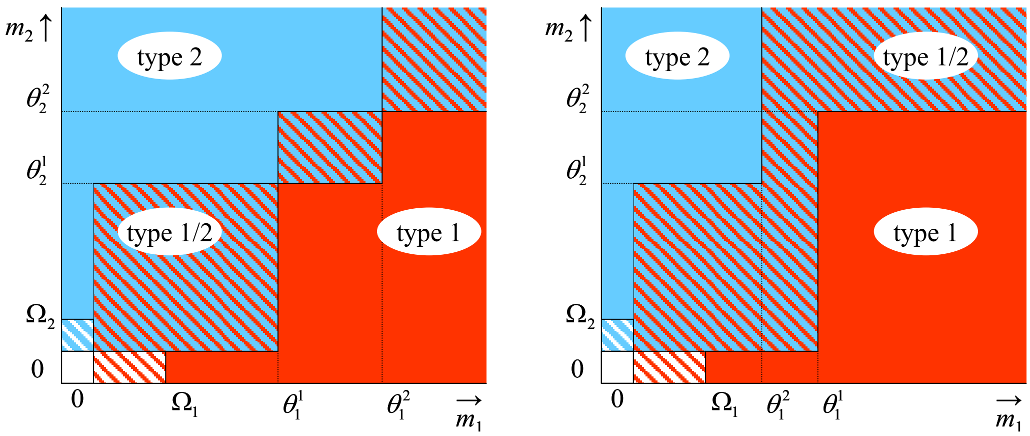


Figure 2: Queue length dependent priorities, with $\theta_1^1 < \theta_1^2$ (left), $\theta_1^1 > \theta_1^2$ (right).

By the θ_i-values, calls are thus given priority in case their queue length becomes too large. In case a queue becomes empty, overflow to an idling server is limited for both types. The threshold rule is illustrated in Figure 2. The red (blue) area consists of states where type 1 (2) jobs are given priority. In states covered by the dashed areas, an idling server takes a job from its own queue. By OR and simulation we then seek

$$W_A^{**} = \min_{\theta_1^1, \theta_2^1; \theta_1^2, \theta_2^2; \Omega_1, \Omega_2} W_A\{\mathbf{Thr}(\theta_1^1, \theta_2^1; \theta_1^2, \theta_2^2; \Omega_1, \Omega_2)\}$$

As shown in *Table* this leads to further improvement.

Table 3. Optimal threshold values.

	NP1	Thr$(\theta_1^1,\theta_2^1;\theta_1^2,\theta_2^2;\Omega_1,\Omega_2)$		T(θ_1)		Pooled
s	W_A	W_A	$(\theta_1^1,\theta_2^1;\theta_1^2,\theta_2^2;\Omega_1,\Omega_2)^*$	W_A	θ_1^*	W_P
1	4.58	4.36	3, ∞, 3, ∞, 1, 5	4.37	3	11.53
2	2.44	2.24	3, ∞, 3, ∞, 1, 2	2.27	4	5.27
3	1.63	1.49	4, ∞, 4, ∞, 1, 2	1.51	4	3.33
4	1.19	1.10	4, ∞, 4, ∞, 1, 3	1.12	4	2.34
5	0.93	0.86	4, ∞, 4, ∞, 1, 3	0.88	5	1.76
10	0.38	0.37	4, ∞, 4, ∞, 1, 2	0.38	1	0.71
15	0.21	0.21	1, ∞, 1, ∞, 1, 1	0.21	1	0.40
20	0.14	0.14	1, ∞, 1, ∞, 1, 1	0.14	1	0.26
30	0.07	0.07	1, ∞, 1, ∞, 1, 1	0.07	1	0.13

Conclusion

$\mathbf{T}(\theta_1)^* \equiv \mathbf{Thr}(\theta_1,\infty;\theta_1,\infty;1,1)^* \approx \mathbf{Thr}(\theta_1^1,\theta_2^1;\theta_1^2,\theta_2^2;\Omega_1,\Omega_2)^*$ for small number of servers s. For all s, a simple single threshold rule $\mathbf{T}(\theta_1)^*$ will be nearly optimal. This rule, which is easy to implement in practice, generally performs quite well.

2.3 Strict improvement

Alternatively, rather counter intuitively we might even be able to obtain policies which are superior to the pooled case for both type 2 and type 1 jobs (in other words for which the average waiting time of any job strictly improves the pooled case), by a double optimization procedure as by steps 1 and 3 below. To this end, among the threshold policies $S(\theta_1, \theta_2) \equiv \mathbf{Thr}(\theta_1, \theta_2, \theta_1, \theta_2, 1, 1)$, an S(Opt)-rule is determined that takes into account the waiting times of either job types by:

Step 1: $\min_{\theta_1,\theta_2} \max\{\boldsymbol{W}_1[S(\theta_1,\theta_2)], \boldsymbol{W}_2[S(\theta_1,\theta_2)]\}$
This leads to an optimal threshold combination $(\theta_1, \theta_2)^*$
and overall average waiting time under $(\theta_1, \theta_2)^*$: $\boldsymbol{W}_A[S(\theta_1, \theta_2)^*]$.

Step 2: $\boldsymbol{W}_A[S(\theta_1, \theta_2)^{**}] = \min_{\theta_1,\theta_2} \boldsymbol{W}_A[S(\theta_1,\theta_2)]$
s.t. $\max\{\boldsymbol{W}_1[S(\theta_1, \theta_2)], \boldsymbol{W}_2[S(\theta_1, \theta_2)]\} < \boldsymbol{W}_{\text{Pooled}}$.

Step 3: If $(\theta_1, \theta_2)^{**}$ exists then $\boldsymbol{W}_A[S(\text{Opt})] = \boldsymbol{W}_A[S(\theta_1, \theta_2)^{**}]$,
otherwise $\boldsymbol{W}_A[S(\text{Opt})] = \boldsymbol{W}_A[S(\theta_1, \theta_2)^*]$.

Figure 3 shows the relative and strict improvements (mean waiting time reduction) that can so be obtained for both type 1 and type 2 jobs over the pooling scenario for $k = 10$ and $s = 1$ to 20. The improvements are only in the order of a few % but consistently outside 95% confidence intervals with a range of 0.5%. The 100% line represents the pooled case.

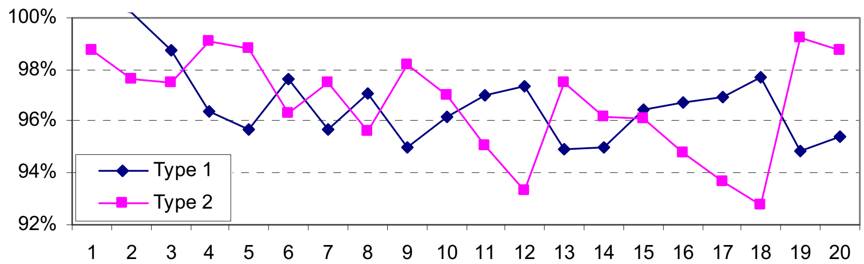


Figure 3. Relative improvements over the pooled scenario.

3 BLOODMANAGEMENT

3.1 Problem motivation

Blood management is a problem of general human interest with a number of concerns and complications. Our problem of interest will concentrate: on the production and inventory management of blood platelets. Here there are a number of conflicting aspects.

On the one hand, the demand is highly 'uncertain' and despite planned surgeries (if such information is used) roughly 50% (at week basis) still remains to be unpredictable. Clearly, as lives may be at risk, shortages are to be minimized.

On the other hand, as the supply is voluntary while such shortages may take place, blood is to be considered as highly precious. Any spill, by outdating, of blood (products) is thus highly ‘undesirable’ if not to be avoided at all. As an extra complicating factor, blood platelets (thrombocytes) have a limited life-time or rather ‘shelf life’ of at most 6 days, while red blood cells and plasma in all sorts of blood types can be kept for months up to over a year. In addition, regular production of a platelet pool takes about one day. Hence production volumes should be set carefully.

3.2 Combined approach

In [2] a combined ‘new’ approach for the blood platelet inventory problem has therefore been followed, which combines OR and simulation by the following steps:

- Step 1: First, a stochastic dynamic programming (SDP) formulation is provided.
- Step 2: The dimension of the (SDP) formulation is then reduced (downsized) by aggregating the state space and demands so that the downsized (SDP) problem can be solved numerically (using successive approximation). That is, the optimal value and an optimal strategy is determined for the downsized SDP.
- Step 3: Then, as essential tying step, this optimal policy is (re)evaluated and run by simulation in order to investigate the structure of the optimal strategy. Therefore we ‘register’ the frequency of (state, action)-pairs for the down-sized problem.
- Step 4: By a heuristic search procedure a simple practical near to optimal order-up-to strategy is then derived for a rule that resembles the structure at the ‘simulation table’.
- Step 5: The quality (near-to-optimality) of this practical simple order-up-to strategy is then also evaluated by simulation.

3.3 OR and Simulation step

As the technical (mathematical) details of steps 1 and 2 are somewhat ‘standard’ but also ‘complicated’ and worked out in detail in [2], let us just restrict to a compact presentation of the essential OR and Simulation step 1 and step 3.

3.3.1 Step 1 OR-approach (SDP)

To give an SDP formulation, the state of the system is described by $(d, \boldsymbol{x})$ with

d: the day of the week ($d = 1,2,\ldots,7$) and
$\boldsymbol{x} = (x_1, x_2,\ldots, x_m)$ the inventory state with x_r = the number of pools with a residual life time of r days (maximal m = 6 days) (A pool is one patient-transfusion unit containing the platelets of 5 different donations).

Let

$\boldsymbol{V}_n(d, \boldsymbol{x})$: represent the minimal expected costs over n days when starting in state $(d, \boldsymbol{x})$.

The optimal inventory strategy and production actions are then determined by iteratively computing (solving) the SDP-equations for $n = 1,2,\ldots$.

$$\boldsymbol{V}_n(d,\boldsymbol{x}) = \min_k \left[c(\boldsymbol{x},k) + \sum p_d(b)\, \boldsymbol{V}_{n-1}(d+1, t(\boldsymbol{x},k,b))\right]$$

with

k	the production action,
$c(\boldsymbol{x}, k)$	the one day costs in state $\boldsymbol{x}$ under production k,
$p_d(b)$	the probability for a (composite) demand b,

$t(\boldsymbol{x}, k, b)$ the new inventory state depending on k, b, $\boldsymbol{x}$, and some issuing policy, and $V_0(d, \boldsymbol{x}) \equiv 0$.

However, for a realistically sized problem for one of the Dutch regional blood banks the computational complexity of this SDP for a one-week iteration already becomes of an order 10^{14}, which makes the computation prohibitively large.

Therefore, we have downsized the demands and inventory levels by aggregating the pools in quantities of 4. This strongly reduces the computational complexity, so that an optimal strategy can be computed for this downsized problem by the optimizing actions of the SDP. However, in practice one needs a simple rule and this optimal strategy has no simple structure. For example, it prescribes the following production volumes on Tuesday at the following states, which all have the same total inventory level of 14 pools, but of varying ages.

Table 4: Optimal Productions by SDP.

Production	Inventory (old,…, young)
7	(0, 0, 5, 0, 0, 9)
8	(0, 0, 6, 0, 0, 8)
9	(0, 0, 8, 0, 0, 6)
10	(0, 6, 2, 0, 0, 6)
10	(5, 0, 3, 0, 0, 6)

3.3.2 Step 3 Simulation

In order to derive a simple order-up-to strategy which only depends on the total predicted inventory, the actual platelet production-inventory process is therefore simulated for 100,000 replications so as to register how often which total predicted final inventory level (I) and corresponding action occurs under the optimal strategy (as determined by SDP) for the downsized problem. As an illustration, for a particular day of the week and the dataset of the regional blood bank, this led to the 'simulation table' in Table 5.

Table 5: Simulation Frequency table of (State, Action)-pairs for Tuesdays from Simulation of Optimal SDP Solution for 100,000 weeks.

I	2	3	4	5	6	7	8	9	10	11	12	13	14	cum.
Order-up-to														
23													4	4
22												28		28
21											96			96
20										267				267
19								2	748	3				753
18							18	1928	31	1				1978
17						6331	4490	353	26	1				11201
16					8260	2078	783	7						11128
15	3131	14123	20926	23646	10087	2593	39							74545
:														
0														
cum.	3131	14123	20926	23646	18347	11002	5330	2290	805	272	96	28	4	100000

For example, it shows by row 15 and column 7 that during the 100,000 replications 2593 times a state was visited with a total final inventory (I) of 7 followed by a production decision of 8 (order-up-to 15). Order up-to-level 15 occurs in 74.5% of the states visited, however often a higher production is optimal. The order-up-to level can be seen as a target-inventory level for Wednesday mornings.

We conclude that a simple order-up-to rule might perform well. By investigating the states at which the optimal production volume is higher we have derived an even better rule that closely resembles the optimal production strategy.

3.4 Results

Applying this approach to data from a Dutch regional blood bank, we could draw the following conclusions:

1. A simple order-up-to rule could reduce the spill from roughly 15 to 20%, as a figure that also seems rather standard worldwide, to less than 1% (while also shortages were reduced and nearly vanished).
2. The combined SDP-Simulation approach led to accuracy within 1% of the exact optimal value for the downsized problem.

For detailed results we refer to [2].

4 RAIL-TRACK SCHEDULING

4.1 Motivation

An example of yet another class of decision problems is found in rail-track scheduling. Despite their length rail tracks still are a scarce resource that has to be shared in an intelligent way. As an example, consider the junction as depicted in Figure 4. Two or more trains may enter the junction more or less simultaneously. It has then to be decided which train to admit first.

To a certain extent the basis of this decision problem is deterministic but in practice it is also highly stochastic due to stochastic arrival times, delays and speed differences. Accordingly, the problem has the flavour of both a scheduling and a queueing problem.

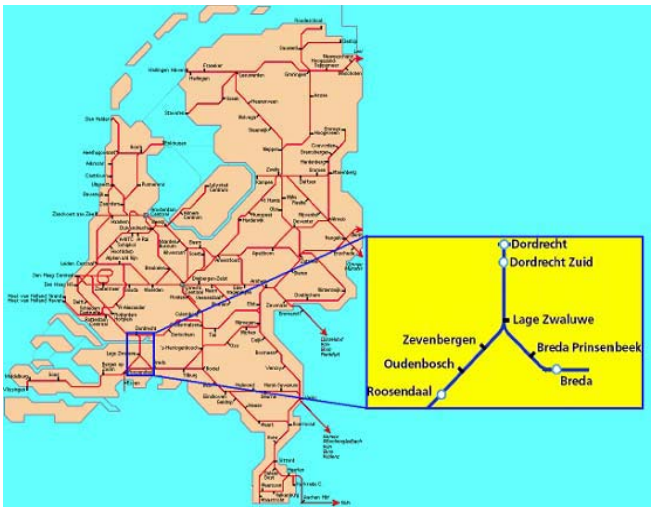


Figure 4: The Railway Infrastructure of The Netherlands with selected junction for the test case.

4.2 Approaches

OR-approach

This track conflict problem can partially be regarded as a ‘standard’ OR-scheduling problem. More precisely as a job shop problem with blocking. By identifying trains as jobs and tracks as machines, an ‘optimal train order’ for a track can be found by a branch-and-bound technique. It is a job-shop problem with blocking as an occupied track section blocks a successive train to enter that section. Trains at the preceding section can thus be delayed. The job-shop formulation, however, uses fixed handling times without delays and variability's.

Simulation–approach
As delay aspects and the variability of travel times are crucial for the track conflict problem in the first place, a stochastic approach is necessarily required. Simulation would thus be in place despite the fact that it does not optimize at all. Indeed, in [4] simulation is used for analyzing a junction. In this study each train is assigned a dynamic priority. The dynamic priority can be a function of the train type, its experienced delay, the delay caused by acceleration and possible other conflicts. However, optimization is not involved.

Combined approach
In [1], therefore, a more extended combination of OR and simulation is suggested. To include both queueing (time) and scheduling (optimization) aspects a Semi-Markov Decision Process (SMDP) is formulated.

SMDP-formulation
The Semi Markov Dynamic Programming (SMDP) formulation for the stochastic junction-track scheduling problem essentially takes into account the stochastic nature and different durations of transitions. It has the form:

$$\boldsymbol{V}_{n+1}(A,v,d) = \min_k \begin{Bmatrix} c(A,v,d) + \sum_{(A,v,d)'} P^k\left[(A,v,d);(A,v,d)'\right]\boldsymbol{V}_n(A,v,d)'[\tau/\tau^k(A,v,d)] \\ +[1-\tau/\tau^k(A,v,d)]\boldsymbol{V}_n(A,v,d) \end{Bmatrix}$$

where a state (A, v, d) represents a state of the form:

$(A, v, d) = (A_1, v_1; A_2, v_2; d_1, d_2, \ldots, d_N)$ with
A_ℓ denoting the trains in queue ℓ ($\ell = 1, 2$) with
v_ℓ indicating whether the trains are moving ($v_\ell = 0$) or not ($v_\ell = 1$)
d_j the train type which is moving in the j^{th} position past the junction.

The cost $c(.,.,.)$ incurs the time that the trains are spending in the sub-network up to the next transition. Here

$P^k\left[(A,v,d);(A,v,d)'\right]$ represents the transition probability from a state (A,v,d) into $(A,v,d)'$

$\tau^k\left[(A,v,d);(A,v,d)'\right]$ is its expected duration, and

$\tau = \min_{(A,v,d),k} \tau^k(A,v,d)$ the expected time in (A,v,d) the expected time in (A,v,d), with

$$\tau^k(A,v,d) = \sum_{(A,v,d)'} P^k\left[(A,v,d);(A,v,d)'\right]\tau^k\left[(A,v,d);(A,v,d)'\right].$$

Simulation-SMDP approach
A combined approach can now be suggested, which combines simulation with the SMDP optimization algorithm in a number of steps, as briefly outlined below.

Step 1: (Simulation) Trains are generated for the junction sub-network according to a global train schedule but with a number of stochastic elements to include initial randomness (Poisson arrivals) and speed differences. The trains are simulated until a conflict is detected. The simulation run is interrupted and the conflict is registered.

Step 2: (SMDP optimization) The SMDP module is activated. This module first collects all train information (the positions, types, speeds and routes assigned). Next, all feasible train orders are detected and enumerated. A cost function is defined which leads to a cost value for a given order decision. Next, by the SMDP algorithm the 'best' train order is determined which leads to a minimal local delay.

Step 3: The delays are (re)evaluated by simulation under the SMD-optimal decisions (train orders). Step 1 is repeated under the new train orders. The simulation is resumed up to a next conflict. This procedure is iterated up to some stop criterion.

In short, simulation is used to capture queueing, to generate conflicts and to evaluate decisions made while SMDP is used to capture stochastic durations and to determine optimal scheduling decisions (train orders).

4.3 Application results

An example

In cooperation with ProRail (the Dutch Railway operator) the approach has been applied to a small but complicating and generic junction within The Netherlands, as shown in Figure 7. The junction has 12 arriving trains per hour, 6 fast passenger intercity trains (IC) and 6 slow freight trains (FR); half of the trains on each one of the arriving tracks. After the junction there are 5 positions (which reflect a distance of more than 13 km). The FR trains need 170 seconds to accelerate from speed 0 to speed 80km/hr, while the IC trains only need 30 seconds to reach the speed of 120 km/hr.

Table 6: Results by simulation and the SMDP-simulation rule for the FCFS, IC-FR (priority to passenger trains), and FR-IC (priority to freight trains).

	12 trains per hour			
Discipline	Delay IC (secs)	Delay FR (secs)	Avr Delay (secs)	Number conflicts (per hour)
FCFS	182	86	134	2.6
IC-FR	164	109	137	2.6
FR-IC	175	48	111	2.3
SMDP	162	51	106	2.2

To verify that the combined SMDP-simulation approach outperforms simple practical rules like the FCFS (First Come First Served) rule or a strict priority rule for passenger or for freight trains, the approach is compared with these rules by simulation. Table 3 shows the results. The values are average delays per train type over 12 days at 15 hours a day. The results show that the SMDP-simulation approach almost captures the quality for passenger trains as by strictly prioritizing passenger trains and for freight trains as by strictly prioritizing freight trains.

Note that there is no other way to evaluate the different rules other then by simulation. Also note that even for experienced and intelligent train schedulers it is impossible to generate and compare all strategies. The SMDP-algorithm computes presumably optimal decisions. Clearly, in practice these decisions can be overruled in the light of other information and expertise of the schedulers, which cannot be included in the SDMP-simulation model. Accordingly, the combined SMDP-simulation approach should primarily be seen as a valuable tool to support practical train scheduling.

References

[1] Al-Ibrahim, A., Van der Wal, J. (2005). A Semi Markovian Decision approach for train conflict resolution. Submitted.

[2] Haijema, R., Van der Wal, J., Van Dijk, N.M. (2007). Blood platelet production: Optimization by dynamic programming and simulation. Computers & Operations Research 34, 760-779.

[3] Krug, W. (2002). Modelling, Simulation and Optimisation for Manufacturing, Organisational and Logistical Processes. Erlangen, Deutschland, Gruner Druck GmbH.

[4] Sahin, I. (1999). Railway traffic control and train scheduling based on inter-train conflict management. Transportation Research Part B 33, 511-534.

[5] Van Dijk, N.M., Van der Sluis, E. (2008). To Pool or not to Pool in Call Centers. Production and Operations Management 17, 1-10.

[6] Van Dijk, N.M., Van der Sluis, E. (2008). Practical optimization by OR and simulation. Simulation Modelling Practice and Theory 16, 1113-1122.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section VII:

Environment and Human Resources

MULTICRITERIA JOB EVALUATION WITHIN THE CENTRE OF LIFELONG EDUCATION

Zoran Babić, Zdravka Aljinović and **Ivana Tadić**
University of Split, Faculty of Economics
Matice hrvatske 31, 21 000 Split, Croatia
{babic, zdravka.aljinovic, itadic}@efst.hr

Abstract: The paper develops a model for the job evaluation, in particular – evaluation of lecturers and their work within the context of lifelong education. After selecting and defining seventeen job factors (criteria), the problem of defining the weight of each particular job factor had to be solved. Assessment of the importance of the proposed criteria is obtained by combining the ranking method with the rating method, which included appraisals of a number of experts. Estimated criteria weights are then used in the TOPSIS method, which is in this case slightly modified. The result of this modeling is unique value (value of the so called relative closeness index) between 0 and 1, which assesses each particular lecturer i.e. his work, and serves as a coefficient for the lecturer's payment. The model is suitable for every job evaluation where job factors (criteria) are defined.

Keywords: job evaluation, lifelong education, ranking method, rating method, TOPSIS.

1 INTRODUCTION

The paper follows from the job evaluation problem within the Centre of lifelong education at the Faculty of Economics, University of Split. The Centre engaged many lecturers and specialist from the fields of economics, finance, IT, statistics and others. All seminars, themes, lectures and lecturers are evaluated by the committee of experts. The problem is: How to make an objective evaluation without clearly defined job factors - criteria and their importance? Also, when criteria and their importance have been defined, how to get an objective picture of each particular lecturer and his/her work and how to determine their basic salary. This paper gives answers to these questions. The model developed in the paper can be used in every job evaluation where job factors and their importance are defined.

The paper is organized as follows: section 2 presents the elements of the job evaluation theory and defines seventeen job factors in this particular job evaluation process. In the third section, the ranking and rating methods for assessment of previously introduced multiple criteria are presented. Estimated criteria weights are then used in the fourth section, where the TOPSIS method is applied in evaluation of lecturers' work. The model results with the value of relative closeness index which presents the unique grade of lecturers' work. The main conclusions are outlined in section 5.

2 JOB EVALUATION

Job evaluation represents a method of identifying the relative value of each particular job in relation to all jobs within an organisation. The aim of job evaluation is to determine the pay structure as well as differences that exist within the base pay for the different job positions and according to different job requirements. Job evaluation ranks all the jobs within the organisation in the hierarchy that reflects the relative worth of each [2]. Job evaluation has direct implication on determining employees' basic salary which stimulates, facilitates and awards individual employee's performance at work. On the other hand, job evaluation has also an indirect implication on reinforcing organisational culture and implementation of organisational strategic goals.

2.1 The methods of job evaluation

There are different methods used to evaluate jobs, all subjective to a certain degree. The most common job-based systems are: 1) job ranking or ordering method, 2) job grading or job classification method and 3) the point system method. The first method arranges jobs in a simple rank order, from the highest to the lowest, while not taking into consideration the specific criteria. Job ranking or ordering method represents the quickest, simplest, easiest to understand and the oldest job evaluation method which depends on the evaluator's knowledge and the degree of objectivity [2], [4]. The second mentioned method uses a number of job related factors, such as education, job experience or responsibility to determine grades or classes of the job. In addition, it requires benchmarking of job description for each grade or class. There are also some problems in the usage of this method such as the need to describe each grade or class, which is an extremely complex task in large organisations. Furthermore, there is a problem of assigning jobs to particular grades or classes, which also depends on the evaluator's degree of objectivity [2], [4].

The point method, which represents the analytical method of job evaluation, is the most commonly used method. Job is quantified by a set of different job factors by allocating points to each factor. Further, each factor with its precise definition is divided into different levels also with their belonging definitions. Points are added up by levels of each particular job factor and its total determines the job's relative worth to the organisation.

The most usual steps for the point method are as follows [1]:

1. Selecting and defining the set of job factors;
2. Selecting and defining particular factors within defined groups;
3. Defining different levels of job factors;
4. Defining the weight (ponder) of particular job factors;
5. Defining the weight (ponder) of particular levels of each job factor;
6. Defining the maximum score that can be related to each job, as well as maximum number of different levels of particular job factors.

Defining different levels of particular job factors has to be as objective as possible. While full objectivity will be achieved in some situations, it is still to be expected that evaluator will often be subjective, due to deficiency of objective evaluating factors. For instance, education, as an objective factor, can be defined as *competence gained through regular education* and can represent an objective factor for job evaluation. Example of education evaluation by different levels can be seen in the following table. Hereafter, levels have to be separately defined in order to allocate points more precisely.

Table 1: Example of point system – objective evaluation of job factor

EDUCATION	LEVEL	POINTS
PhD degree	5	250
Masters degree	4	200
Bachelors degree	3	150
Secondary school	2	100
Elementary school	1	50

Source: [4], pp. 447.

As already stated, sometimes evaluation of particular job factors will be based solely on the evaluator's knowledge and subjectivity. For instance – responsibility, as a subjective factor, can be defined in the following way: *"Responsibility includes different factors required for successful job accomplishment, such as: quality of knowledge transfer, appropriate decision* *making, accomplishment of regular tasks and duties in time and coordination of work with course participants."*

Table 2: Example of point system – subjective evaluation of job factor

RESPONSIBILITY	LEVEL	POINTS
More than required	3	90
Required	2	60
Less that required	1	30

Source: Authors

2.2 Criteria for job evaluation in lifelong education

For the purpose of job evaluation in lifelong education, seventeen job factors are selected and defined. These are further classified into ten most frequently used job factors (Table 3) and other seven possible job factors (Table 4), which are to some extent, specific to this particular job evaluation process.

Table 3: Criteria from 1 to 10: The most frequently used job factors

JOB FACTORS (THE MOST FREQUENTLY USED)	
JOB FACTORS	**DEFINITION OF JOB FACTORS**
1. Competence gained through regular education	Minimum of theoretical knowledge required for regular job accomplishment, gained through formal education and measured by levels of education (education degrees).
2. Competence gained through additional education	Additional knowledge required for regular job accomplishment, gained after formal education (courses or seminars) at certified institutions, measured by degrees, certificates and duration.
3. Working experience	Knowledge gained through work in workplace, performing the same or similar jobs in profession, measured by duration of minimum experience required for successful job accomplishment.
4. Mental effort	Mental effort required for performing the necessary tasks, such as: thinking, decision making, problem solving and performer's preliminary preparation required for unobstructed business process.
5. Physical effort	Effort resulting from long hours spent speaking and standing.
6. Complexity of work	Set of performer's mental activities, processes and decisions required for successful job accomplishment.
7. Responsibility	Responsibility includes different factors required for successful job accomplishment, such as: quality of knowledge transfer, appropriate decision making, accomplishment of regular tasks and duties in time and coordination of work with course participants.
8. Group size	Number of participants (within a group) required for unobstructed business process, implying unobstructed lectures, discussions and individual interactions.
9. Usage of technical equipment	Usage of technical equipment in order to enhance the process or teaching and learning by making lectures more contemporary and diversified.
10. Course length	Minimal course length required for unobstructed process of lecturing and acquiring the presented contents and skills for its implementation in everyday life, measured through required number of lessons.

Source: Authors

Table 4: Criteria from 11 to 17: Other possible job factors

OTHER POSSIBLE JOB FACTORS	
11. Course implementation in practice	Teaching method applied in order to adjust processed material to participants. This provides participants with the necessary skills for successful implementation of the acquired knowledge in everyday business practice.
12. Importance of the course for the social community	Organising the course in a way which will make sure that it contributes significantly to the social community both with its presented contents as well as through participants' individual implementation of acquired knowledge and skills in everyday business, usually measured through performance analysis.
13. Importance of the course for the Faculty	The course is organised and the topics selected by the potential lecturers (as a form of their individual confirmation as scientists) with the aim of recruiting as participants important members of the society, who will in their turn be able to contribute to the recognition of the Faculty within the given local community.
14. Lecturing	The way in which lecturer organises his/her lessons which incorporates: content (material) preparation, lecturer preparation, understanding and adjustment of the teaching material to the course participants in order to provide unobstructed interaction among course participants, as well as the implementation of acquired knowledge and skills in the business practice.
15. Used materials	The contents of the used materials should correspond to the course topics which in their turn are in accordance with the current business trends as well as a range of business situations the participants will be faced with in the future.
16. Creativity	The method in which lecturer adjusts lectures to emerging business situations and responds to participants' requirements and questions with adequate business examples using his/her knowledge, skills and abilities.
17. Relations	The way in which lecturer develops a relationship with course participants which is reflected in the possibility of tutorials for discussing any possible problems or dilemmas resulting from the course.

Source: Authors

3 ASSESSING THE MULTIPLE CRITERIA

There are many methods for evaluation of the criteria weights and their importance. Majority of the researchers agree that these should be simple and easily understandable for decision makers. Accordingly, in this paper two simple methods for the final importance of criteria are applied.

Assessment of the importance of the proposed seventeen criteria (Tables 3 and 4) is obtained with the ranking method and rating method, where the judgements of experts are incorporated. A very important aspect of this type of assessment is the fact that introducing more experts usually results in different ranking based on each individual assessment. This is why these two methods are necessary to synthesise different assessments. The final importance of the criteria, i.e. the normalized criteria weights are obtained by combination of the methods, using relation (5).

3.1 Ranking method

Suppose that there are n criteria A_j (j = 1, 2, ..., n) to be evaluated by r experts E_k (k = 1, 2, …, r). Every expert has to rank all criteria according to its importance in such a way that the most important criteria gets n-1, the next important criteria gets n-2,… while the least important criteria will be ranked as 0.

These ranks are manipulated as follows:

$$R_j = \sum_{k=1}^{r} R_{jk} \qquad (1)$$

where R_{jk} is the rank (number from 0 to n-1) assigned to criterion j by judge k; R_j is the sum the ranks across judges for each criterion j.

The criteria weights are now determined by:

$$w_j = \frac{R_j}{\sum_{j=1}^{n} R_j} \qquad (2)$$

where w_j is the weight of criterion j across all judges.

This ranking method is simple and least time-consuming for the judges. Since each judge produces only a set of integers, it is not possible to develop a set of weights for each judge; only the weights for the composite of all the judges' ranks can be obtained.

In this case, 11 experts ranked 17 criteria with the ranks (values) between 16 and 0, according to the criterion importance. It is then easy to obtain the values of R_j and the normalized criteria weights (Table 5).

Table 5: Experts ranks, the sum of the ranks across judges and the normalized criteria weights for 17 criteria

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	R_j	λ_j
C1	4	14	7	13	1	16	15	15	16	11	14	112	**0,082232**
C2	1	13	6	2	16	15	2	16	14	15	3	100	**0,073421**
C3	11	9	8	9	14	3	9	14	15	9	9	101	**0,074156**
C4	15	7	10	15	10	4	16	4	10	8	8	99	**0,072687**
C5	9	3	5	0	6	0	0	3	3	7	0	36	**0,026432**
C6	3	8	9	14	5	6	14	0	1	6	13	66	**0,048458**
C7	13	5	11	16	4	1	13	8	0	10	12	81	**0,059471**
C8	10	2	4	6	3	5	4	7	2	5	7	48	**0,035242**
C9	8	0	3	1	0	14	3	6	9	3	6	47	**0,034508**
C10	14	1	0	5	11	7	1	1	7	4	1	51	**0,037445**
C11	7	16	12	11	13	10	10	13	11	16	4	119	**0,087372**
C12	5	15	1	12	12	12	6	12	13	0	5	88	**0,064611**
C13	6	12	2	10	15	9	8	2	12	1	16	77	**0,056535**
C14	12	11	16	7	9	13	12	11	8	13	15	112	**0,082232**
C15	16	6	13	4	8	11	7	10	4	2	2	81	**0,059471**
C16	2	10	15	3	7	8	11	9	5	12	11	82	**0,060206**
C17	0	4	14	8	2	4	5	5	6	14	10	62	**0,045521**
												1362	**1**

3.2 Rating method

The criteria are presented to each of the judges who are requested to give ratings for each criterion. The rating values are usually real and range from 0 to 10 or 100. In this case the rating values are from 0 to 10, where the higher the value, the more important the criterion. More than one criterion can have the same rating. The weight for each criterion is now derived in the following manner:

$$w_{jk} = \frac{\rho_{jk}}{\sum_{j=1}^{n} \rho_{jk}} \qquad (3)$$

With the relation (3), firstly the normalized criteria weights are obtained for each judge and then relation (4) gives the average weight for each criterion. Namely, in (4) nominator represents the sum of the weights for criterion j according to all experts (k = 1,..., r), while denominator is simply the sum of all experts' weights:

$$w_j = \frac{\sum_{k=1}^{r} w_{jk}}{\sum_{j=1}^{n}\sum_{k=1}^{r} w_{jk}}, \qquad (4)$$

where ρ_{jk} is the rating by judge k to criterion j, w_{jk} is the normalized weight for criterion j from judge k and w_j is the final (total) weight for criterion j.

In this case experts gave the ratings shown in Table 6 where in the last column the criteria weights obtained by rating method, i.e. by relations (3) and (4), are presented.

Table 6: Experts ratings and the criteria weights for 17 criteria

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	w_j
C1	8	9	7	5	8	8	8	10	10	8	8	**0,071402**
C2	4	9	6	7	8	7	6	8	5	9	10	**0,062647**
C3	3	7	8	6	7	7	6	5	5	7	9	**0,055366**
C4	8	7	5	4	5	10	6	3	6	4	8	**0,051889**
C5	7	4	2	3	1	6	5	1	4	5	7	**0,035048**
C6	8	5	8	8	6	10	7	6	5	4	10	**0,061249**
C7	7	10	8	8	5	9	9	9	9	7	10	**0,072353**
C8	6	6	4	6	2	7	4	4	6	6	6	**0,044989**
C9	5	8	5	3	5	6	3	3	4	7	5	**0,042476**
C10	4	4	3	4	2	6	3	7	4	5	7	**0,038504**
C11	5	10	10	8	10	8	6	9	8	10	7	**0,073017**
C12	5	6	10	9	10	9	8	2	8	10	8	**0,068028**
C13	9	6	7	9	9	9	8	6	9	6	8	**0,069447**
C14	9	10	10	6	5	8	8	10	7	8	10	**0,072305**
C15	4	9	6	5	4	7	2	9	6	9	8	**0,054105**
C16	7	8	7	7	7	9	9	5	8	8	10	**0,067519**
C17	6	8	8	4	5	9	6	8	5	7	10	**0,059655**

The final criteria weights are then obtained by relation (5) which is advisable in the case when criteria weights are estimated with more than one method [3].

$$w_j{}^0 = \frac{\lambda_j w_j}{\sum_{j=1}^{n} \lambda_j w_j}, \forall j \tag{5}$$

The final criteria weights calculated according to relation (5) are given in Table 7, where the criteria weights are multiplied by 100 to make it easier to interpret them as percentage values.

It can be seen that in this case such combining of different weights increases the high weights and decreases low weights. If someone finds it inappropriate, the simple mean of weights can be used.

Table 7: Final criteria weights

w_i^0	w_i^0 *100
0,095187	9,518653
0,074568	7,456773
0,066561	6,656061
0,061145	6,114505
0,015018	1,501794
0,048116	4,811619
0,069758	6,975786
0,025704	2,570397
0,023763	2,376269
0,023373	2,337344
0,103423	10,34234
0,071255	7,125548
0,06365	6,364955
0,096391	9,6391
0,052164	5,216449
0,0659	6,590013
0,044024	4,402399
1	100

4 THE TOPSIS METHOD IN JOB EVALUATION

Previously estimated criteria weights are now used in the TOPSIS method which is in this case slightly modified.

Yoon and Hwang (1980, 1981) developed the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method based upon the concept that the best alternative is the one with the shortest distance from the ideal solution and the farthest from the negative-ideal solution [3]. So, we are looking for an alternative, the most similar to the ideal alternative and the least similar to the negative-ideal alternative, where similarity is defined in the terms of distance.

In our case the ideal alternative is a person (lecturer) with the best ratings across all criteria (10), while the negative-ideal alternative is a lecturer with the worst ratings across all criteria (1). Of course, in real life these ideal and negative-ideal mostly don't exist, but what we want to find out is how much an investigated alternative is 'distant' from these two. For the alternative, the relative closeness (RC) index has to be calculated as follows:

$$RC_i = \frac{S_{i-}}{S_{i+} + S_{i-}}, \tag{6}$$

where S_{i-} is the weighted distance from the negative-ideal alternative and S_{i+} is the weighted distance from the ideal alternative. Distances are calculated as Euclidean distances where deviations are weighted with earlier calculated criteria weights:

$$S_{i-} = d(A_i, A^-) = \sqrt{\sum_{j=1}^{n} \left[w_j \cdot (t_{ij} - t_j^-) \right]^2} \,, \tag{7}$$

$$S_{i+} = d(A_i, A^+) = \sqrt{\sum_{j=1}^{n} \left[w_j \cdot (t_{ij} - t_j^+) \right]^2} \,, \tag{8}$$

where A_i presents an investigated alternative, A^+ and A^- are the ideal and negative-ideal alternatives, w_j is the criterion weight.

Apparently, $RC_i = 1$ if $A_i = A^+$, and $RC_i = 0$ if $A_i = A^-$. An alternative is closer to the ideal solution as RC_i approaches to 1.

When the criteria are expressed in different units or different scales, the TOPSIS method requires the transformation of attributes in order to transform the various attribute dimensions into nondimensional attributes, which allows comparison across the attributes. One way is to take the outcome of each criterion divided by the norm of the total outcome vector of the criterion at hand. The columns of the decision matrix are presented by:

$$\overline{X_j} = \frac{X_j}{\|X_j\|} = \frac{X_j}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, \; j = 1,2,..., \; n. \tag{9}$$

When only one alternative (lecturer) is investigated and its relative closeness index is calculated, the decision matrix consists of three columns (investigated, ideal and negative-ideal alternatives) and of a number of rows depending on the number of observed criteria (Table 8).

Now, by using previously exposed formulae, it's easy to calculate for the alternative A that

S_{i-} = 128.6362, S_{i+} = 150.2967 (where the weights are multiplied by 100) and

RC_A = 0.461172.

For the alternative B, for which it can be seen from the decision matrix that its ratings are better (higher), it is calculated S_{i-} = 167.8802, S_{i+} = 79.9899, and relative closeness index is RC_B = 0.67729.

It is obvious that alternative B is better because of higher relative closeness index. It is also obvious according to its distance from ideal and negative-ideal alternatives: smaller S_{i+}, higher S_{i-}, which results in higher *RC* index.

Table 8: Decision matrix for two alternatives A and B, 17 criteria

wj*100	antiideal	ideal	A	B
9,518653	1	10	2	6
7,456773	1	10	4	8
6,656061	1	10	5	9
6,114505	1	10	6	6
1,501794	1	10	7	7
4,811619	1	10	4	7
6,975786	1	10	9	9
2,570397	1	10	10	10
2,376269	1	10	6	6
2,337344	1	10	1	6
10,34234	1	10	1	6
7,125548	1	10	5	8
6,364955	1	10	6	6
9,6391	1	10	7	7
5,216449	1	10	8	8
6,590013	1	10	9	9
4,402399	1	10	8	8

5 CONCLUSION

In the paper, the model for job evaluation based on the TOPSIS method is developed. In this case we evaluated the lecturers, specialists and their activities within the process of lifelong learning, but the model can be applied to any job evaluation provided that job factors are selected and defined. It is also applicable in any situation where unique mark for some alternative is needed. Besides that, the model is very simple and understandable for decision makers; it gives results with the assistance of Excel, a very simple and accessible package.

References

[1] Bahtijarević-Šiber, F., 1999. Management ljudskih potencijala. Golden marketing, Zagreb

[2] DeCenzo, D.A., Robbins, S.P., 2005. Fundamentals of Human Resource Management, John Wiley and Sons, Inc., New York

[3] Hwang, C.L., Yoon, K., 1981. Multi Atribute Decision Making. Methods and Applications, Springer-Verlag, Berlin-Heidelberg- New York

[4] Stone, R.A., 2005. Human Resource Management, John Wiley & Sons Australia, Ltd., Sidney

[5] Tabucanon, M.T., 1988. Multiple Criteria Decision Making in Industry, Elsevier, Amsterdam

[6] Triantaphyllou, E., 2000. Multi-Criteria Decision Making Methods: A Comparative Study, Kluwer Academic Publishers, Dordrecht/ Boston/ London.

DELIMITATION OF FUNCTIONAL REGIONS USING LABOUR MARKET APPROACH

Samo Drobne*, Miha Konjar and Anka Lisec***
* University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia
e-mails: {samo.drobne;anka.lisec}@fgg.uni-lj.si
** e-mail: mkonjar@yahoo.com

Abstract: Despite its long history, the concept of functional regions does not enjoy a common definition; neither in its use as an analytical term nor in its upsurge as a political one. The article explores labour market approach to define the functional regions in Slovenia. For this purpose, the conceptual issues underlying the operation of spatial labour markets are analysed. Furthermore, method for delimitation of functional regions by using only data for inter-municipal daily commuting of working population is suggested. The method was tested to analyse how so defined functional regions fit to proposed provinces (administrative regions) in Slovenia.

Keywords: commuting, labour market, functional region, administrative region, Slovenia.

1 INTRODUCTION

A functional region is a region characterised by a high frequency of intra-regional economic interaction, such as intra-regional trade in goods and services, labour commuting and household shopping [4]. The basic characteristic of a functional region is the integrated labour market, in which intra-regional commuting as well as intra-regional job search and search for labour is much more intensive than the inter-regional counterparts. So, the border of a labour market region is a good approximation of the borders of a functional region [3].

For the delimitation of functional labour market regions commuting conditions are used in most OECD countries [6]. Commuting conditions like distance, closeness, commuting thresholds, travel times determine the magnitude of the commuting flows between areas. On the basis of commuting flows, a functional region can then be defined as a region in which a large proportion of workers both live and work.

In practise, two different concepts to delimit travel-to-work-areas are used: (a) delimitation around a centre, and (b) delimitation using algorithms or cluster analysis based on a combination of distance, closeness, commuting thresholds, travel times, etc.

It should be noted that certain centre-based definitions, normally, do not represent a division into regions or an exhaustive breakdown of the national territory but correspond to areas of extended urban influence, that is to say that those portions of the national territory which lie outside this area of influence are all considered being rural areas. In delimitation based on centres, particular care needs to be taken in the definition of these centres. While some countries identify centres according to the population or level of employment, others take account of commuting conditions. In the latter case, the centre must be “self-sufficient”, which means that the number of workers living and working there is higher than the number of workers commuting to another centre, or it must attract a number of workers that is substantially higher than the number of workers leaving the centre to work outside.

In the countries, where functional regions are estimated by aggregating municipalities, those regions are therefore fully compatible with basic territorial units, the level which serves as the reference for censuses and also for other types of survey and data collection. As a result, most of the relevant statistics and indicators used for territorial analysis are available at the level of functional regions. But, if the administrative regions do not follow the functionally linked labour market areas, the labour market policy (and other policies affected by it), which are generally defined for administrative regions, may be less effective [2].

In Slovenia, 42 local labour systems ("micro-regions") and 17 regional labour systems ("mezzo-regions") were delimited for 2006 [7]. Those systems were defined using data from Statistical Register of Employment and some extra social and economic data to define central municipalities. The main aim of that work was to define and delineate functional urban areas.

The aim of this paper is to present method for delimitation of functional regions by using **only** data on inter-municipal daily commuting. The method, suggested here, follows centre-based labour market approach, which uses one-way commuter flows. Moreover, the method has been tested to analyse how functional regions fit to proposed provinces (administrative regions) in Slovenia.

2 LABOUR MARKET APPROACH

According to [4] there are three different levels of interaction that may be used in estimating the extension of a functional region using labour market approach.

Assume two (regional) centres indexed i and j connected by a line. The first endpoint is labelled by i and the second by j. Moreover, x denotes an intermediate point between i and j. At a location (x) the commuting frequency to the centre i is $f_i(x)$. The functional region consists of all geographic locations that satisfy one of the following three conditions.

The first condition would be to include all locations with any commuting to the centre i. The extension of the functional region i, $\mathbf{FR}_i$, is defined by $\mathbf{FR}_i = \{x : f_i(x) > 0\}$.

A second alternative condition would be to use a cut-off frequency $\dot{f}$ (larger than zero) for inclusion

$$\mathbf{FR}_i = \{x : f_i(x) \geq \dot{f} > 0\}. \quad (1)$$

The cut-off frequency gets rid of the very few long distance commuters. This rule does not allow extremely low interaction. With the second condition one obtains a smaller functional region than when using the first condition.

The third criterion for defining a functional region is to consider neighbouring central places and to calculate the breakpoint between the different central places. The border is found where the attraction is equal to both of the closest cores. This is formally described by

$$\mathbf{FR}_i = \{x : f_i(x) \geq f_j(x)\}. \quad (2)$$

Moreover, combinations of the first, second and third criterion is possible; hence:

$$\mathbf{FR}_i = \{x : f_i(x) \geq f_j(x) \text{ and } f_i(x) \geq \dot{f}\}. \quad (3)$$

In theory, the borders of functional regions are exactly defined and the theoretical borders will not necessarily follow borders of administrative regions. In practice, the functional region consists of smaller areas that have been aggregated. This means that the smallest geographical area, for which there exist commuting data, influences how close the estimated functional region will be to the theoretical functional region.

3 MATERIALS AND METHODOLOGY

The suggested method for identification of functional regions follows centre-based labour market approach. It uses one-way commuter flows of inter-municipal working population.

In our application we used municipalities as the smallest geographical areas to aggregate them into the functional regions. Data on inter-municipal commuting to work were acquired from Census 2002 [11]. There are newer data available for labour commuting in the Statistical Register of Employment. However, there are two main reasons why we decided for the old ones: firstly, only the Census data distinguish between daily and weekly commuting (mobility), and secondly, municipalities of destination are not well defined in the Statistical Register of Employment (companies do not always report proper information for the locations, where their employees work).

In 2002, there were a total of 287,272 inter-municipality commuters between 192 municipalities in Slovenia. Most of them (more than 74%) used cars either as a driver or passenger and only 9% of working population commuted to work by bus and less than 1.5% of them by train [1]. In the application of model (3), we analysed 36,672 ($192^2 - 192$) inter-municipality connections. In addition, we tested how so defined functional regions fit to the proposed six or eight administrative regions (provinces) in Slovenia [5, 8, 9, 10]. In Fig. 1 and Tab. 1 and 2, there are proposals for six and eight administrative regions and centres in Slovenia.

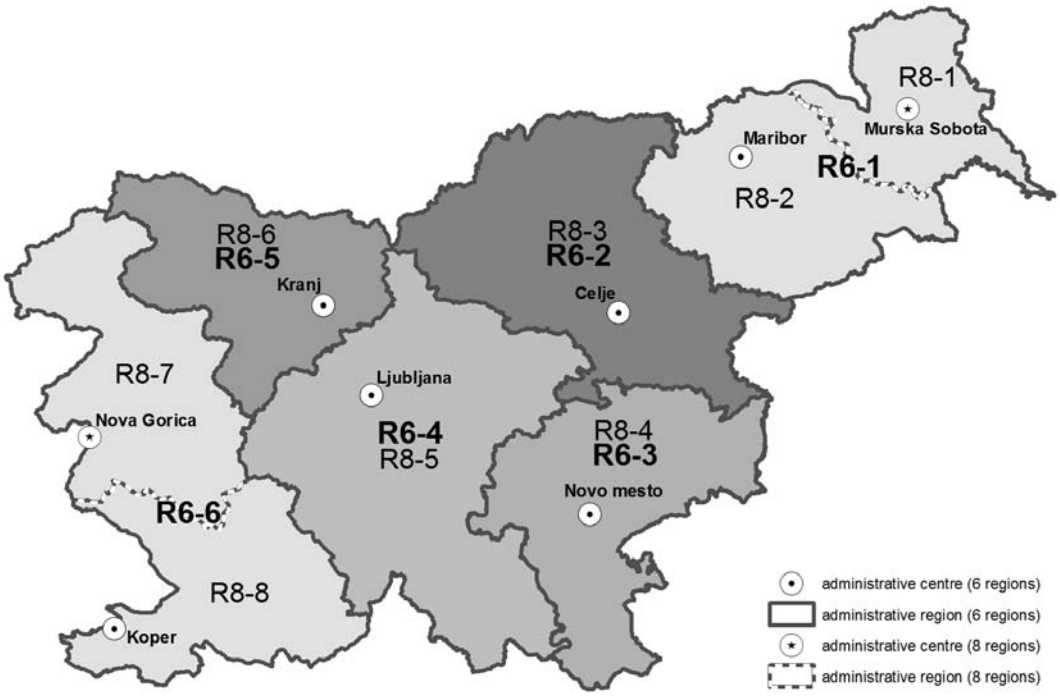


Figure 1: Proposal for six and eight administrative regions (centres) in Slovenia.

Table 1: Six administrative regions and administrative centres in Slovenia.

Id_R6	***Administrative region***	***Administrative centre***
R6-1	Severovzhodna Slovenija	Maribor
R6-2	Savinjskošaleška	Celje
R6-3	Jugovzhodna Slovenija	Novo mesto
R6-4	Osrednjeslovenska	Ljubljana
R6-5	Gorenjska	Kranj
R6-6	Primorska	Koper

Table 2: Eight administrative regions and administrative centres in Slovenia

Id_R8	***Administrative region***	***Administrative centre***
R8-1	Pomurska	Murska Sobota
R8-2	Podravska	Maribor
R8-3	Savinjska	Celje
R8-4	Dolenjska	Novo mesto
R8-5	Osrednjeslovenska	Ljubljana
R8-6	Gorenjska	Kranj
R8-7	Goriška	Nova Gorica
R8-8	Primorska	Koper

First, we analysed commuting conditions of municipalities to the proposed administrative centres according to the model (1). Here, we analysed different cut-off frequencies ($\dot{f}$) defined by the percentage of workers – daily commuters who work in the central municipality. As already noted above, the cut-off frequency get rid of the very few long distance commuters, which is appropriate in the "local" labour market approach. But, in our research, we analysed "larger" functional regions consisting, in average, of 32 $(192/6)$ and 24 $(192/8)$ municipalities. For this reason we examined different breakpoint values. In order to delimit functional regions around the administrative centres, we used model (2), where we combined municipalities into chains to the potential functional centres.

As a first stage of the suggested method, municipalities that are strongly self-sufficient should be identified. A municipality was considered strong self-sufficient if it fulfilled two conditions. The first condition was based on the sum of commuting frequencies of working population from the municipality i to (other) municipalities j, $\sum_j f^{(i)}$ and can be given as

$$f^{(i)}(\%) = \frac{\sum_j f^{(i)}}{\sum_i f_i + \sum_i f_j} \leq \dot{f}^{(i)}(\%), \tag{4}$$

where $f^{(i)}(\%)$ is the percentage of workers who lived in the municipality i and daily commute to other municipalities j, $\sum_i f_i$ is total number of workers who live and work (stay) in the municipality i, $\sum_i f_j$ is total number of commuters who leave to other municipalities j from the municipality i, and $\dot{f}^{(i)}(\%)$ is a cut-off relative frequency.

The second condition for self-sufficient municipality was based on number of working places in the municipality i, w_i, which were calculated using only data on inter-municipal daily commuting. The municipality became self-sufficient if the number of working places in the municipality was at least $\dot{w}$:

$$w_i = \sum_i f_i + \sum_j f_i \geq \dot{w} \ , \tag{5}$$

where $\sum_j f_i$ is total number of commuters who come from other municipalities j to municipality i, and $\dot{w}$ depended on the number of proposed administrative centres (six centres in six proposed regions, $R_A = 6$, or eight centres in eight proposed regions, $R_A = 8$):

$$\dot{w} = \begin{cases} \dot{w}_6 & \text{if } \mathrm{R_A} = 6 \\ \dot{w}_8 & \text{if } \mathrm{R_A} = 8 \end{cases} . \tag{6}$$

When self-sufficient (central) municipalities were defined, we created chains of municipalities from self-sufficient municipalities (centres) till condition (2) was satisfied. The chains are formed for three types of municipalities: (a) the municipalities, that were directly connected with their maximum commuting flow of working population to the proposed centre, were automatically placed to that centre; (b) municipalities that are not directly connected with their maximum commuting flow to the proposed centre, but they are connected with their maximum commuting flow to non self-sufficient municipality, which is than connected to the one of the analysed centre (chains have been determined iteratively); and (c) the pairs of municipalities, which presented each other the destination of the maximum flows, have been connected to the region, in which the direction of the second maximum flow was oriented.

4 RESULTS

The results of the analyses of basic commuting conditions, using model (1) and criterion that 5% of daily commuters work in the central municipality (one of the six or eight administrative centres), are presented in Fig. 2. Note, that there are municipalities from which at least 5% of working population commute to two centres. In particularly the municipalities in the surrounding of Kranj often present municipalities of origin from where a lot of working population commute to Kranj as well as to Ljubljana.

From Fig. 2 the Slovenian municipalities which have a very weak connection to any proposed administrative centre (not coloured) are also evident. Further analysis showed that they were connected (in sense of daily commuting) to other municipalities in the region. But, there is a big cluster of such municipalities in the north-west to the administrative centre of Celje. Those municipalities are not connected, according to mentioned conditions, to any administrative centre in Slovenia. The results below explain that phenomena.

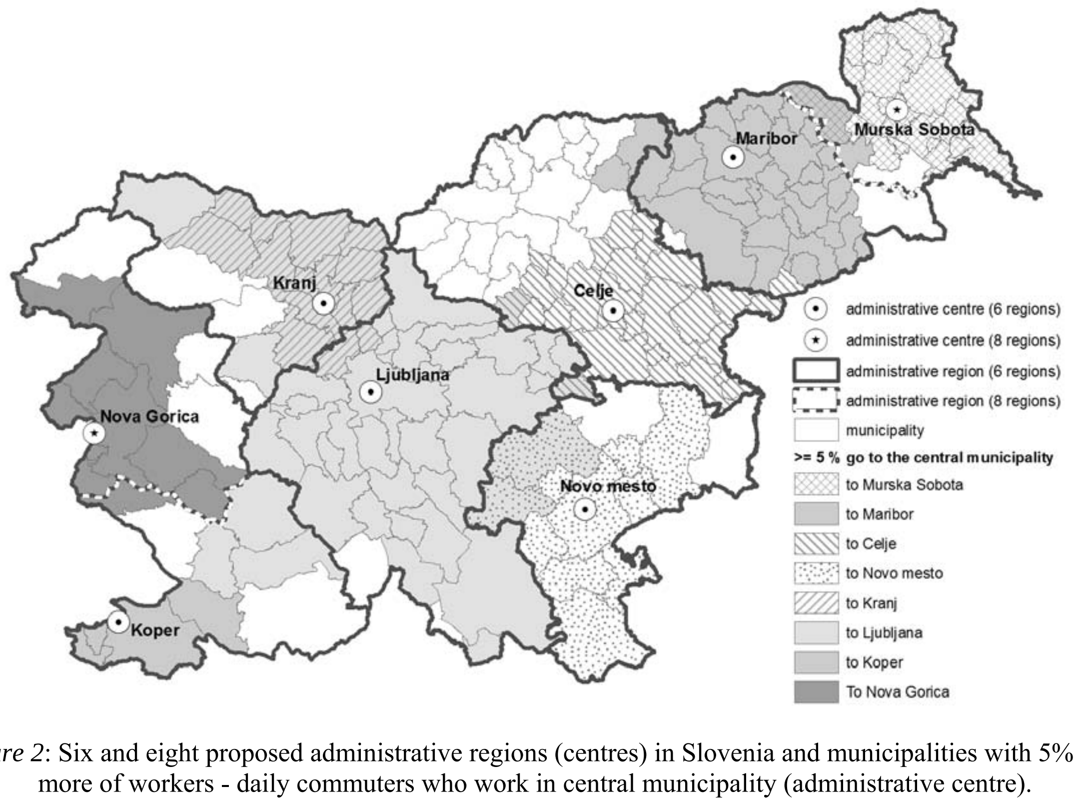


Figure 2: Six and eight proposed administrative regions (centres) in Slovenia and municipalities with 5% and more of workers - daily commuters who work in central municipality (administrative centre).

In our application, the municipality was considered as strong self-sufficient municipality if less than 35% of the working population commuted from the municipality and if there were more than 20,000 of working places in the municipality (for proposal of six regions) or more than 15,000 (for proposal of eight regions). We have examined different breakpoint values and found these values to be the most appropriate to test six or eight administrative centres in Slovenia. Tab. 3 and 4 show functional centres, the percentage of daily commuters who work in the centre, and number of working places for to two cases of regionalisation of Slovenia.

When we applied models (4, 5 and 6), which define self-sufficiency of the municipality (centre), for Census 2002 data of inter-municipal working commuters, we discovered disparity of proposed eight administrative centres and potentially defined nine functional centres. In Tab. 3 and 4, potential self-sufficient centres of six respectively eight functional regions are listed. Comparing lists of administrative and potential functional centres (in pairs: Tab. 1 and 3, and Tab. 2 and 4), we can notice that the proposal of six administrative centres is well prepared, but if we ruminates on introduction of eight provinces (administrative regions) in Slovenia, we should reflect the model once more. According to our results, a new, ninth functional centre Velenje occurred, if Nova Gorica and Murska Sobota should be included in the list of functional centres. The last statement can be confirmed by the fact of homogeneity of the last group of functional centres when looking at the criterion of number of working places in the self-sufficient centre (Murska Sobota and Nova Gorica). According the criterion of number of working places, Murska Sobota and Nova Gorica is followed by Krško respectively Slovenj Gradec with instantly less working places (10,511 working places in Krško and 9468 in Slovenj Gradec). On the other hand, a new centre Velenje forms a separate group with already existing centres of Novo mesto and Koper.

Table 3: Six central municipalities in Slovenia by number of working places

Code of municipality	***Municipality***	***Percentage of workers who work in the municipality of residence***	***Number of working places***
61	Ljubljana	90.3	175,043
70	Maribor	83.7	58,760
11	Celje	76.1	28,214
52	Kranj	65.8	26,497
85	Novo mesto	87.2	22,990
50	Koper	78.9	20,913

Table 4: Nine central municipalities in Slovenia by number of working places

Code of municipality	***Municipality***	***Percentage of workers who work in the municipality of residence***	***Number of working places***
61	Ljubljana	90.3	175,043
70	Maribor	83.7	58,760
11	Celje	76.1	28,214
52	Kranj	65.8	26,497
85	Novo mesto	87.2	22,990
50	Koper	78.9	20,913
133	Velenje	82.8	19,631
84	Nova Gorica	71.6	16,985
80	Murska Sobota	80.8	16,139

Having defined the self-sufficient municipalities respectively the functional centres, the municipalities as the smallest geographical areas were aggregated into the functional regions, based on data about the inter-municipal commuting to work. For this purpose, three steps interaction was applied when creating chains of municipalities (see methodology). In Fig. 3 and 4, the results of delimitation on six respectively nine functional regions in Slovenia are presented.

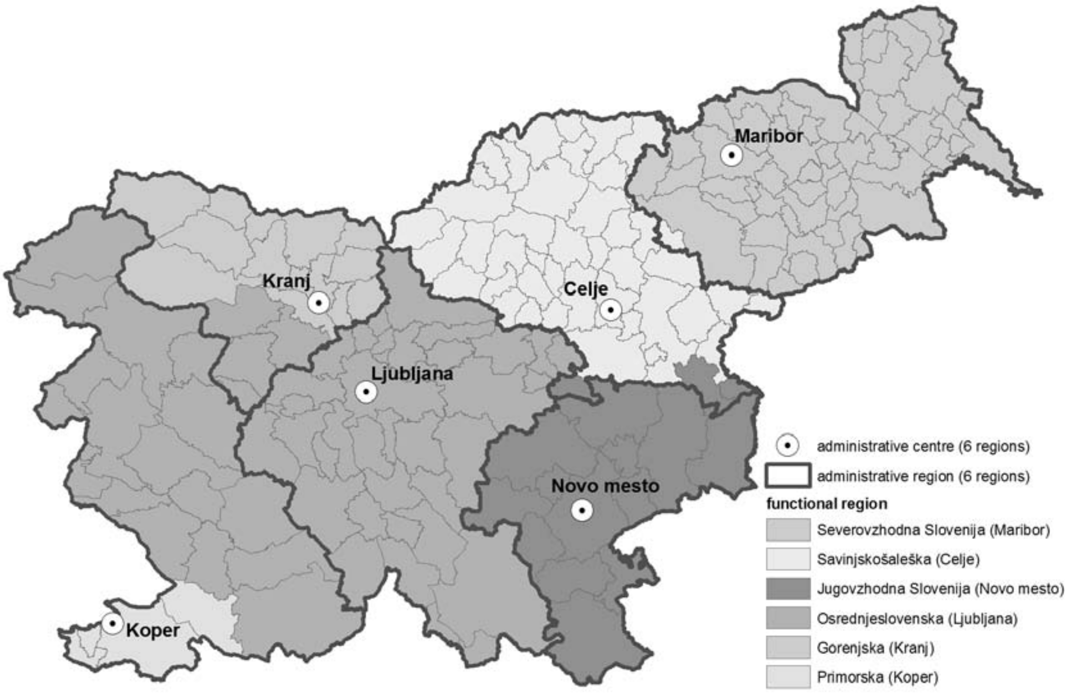


Figure 3: Six administrative and six functional regions in Slovenia

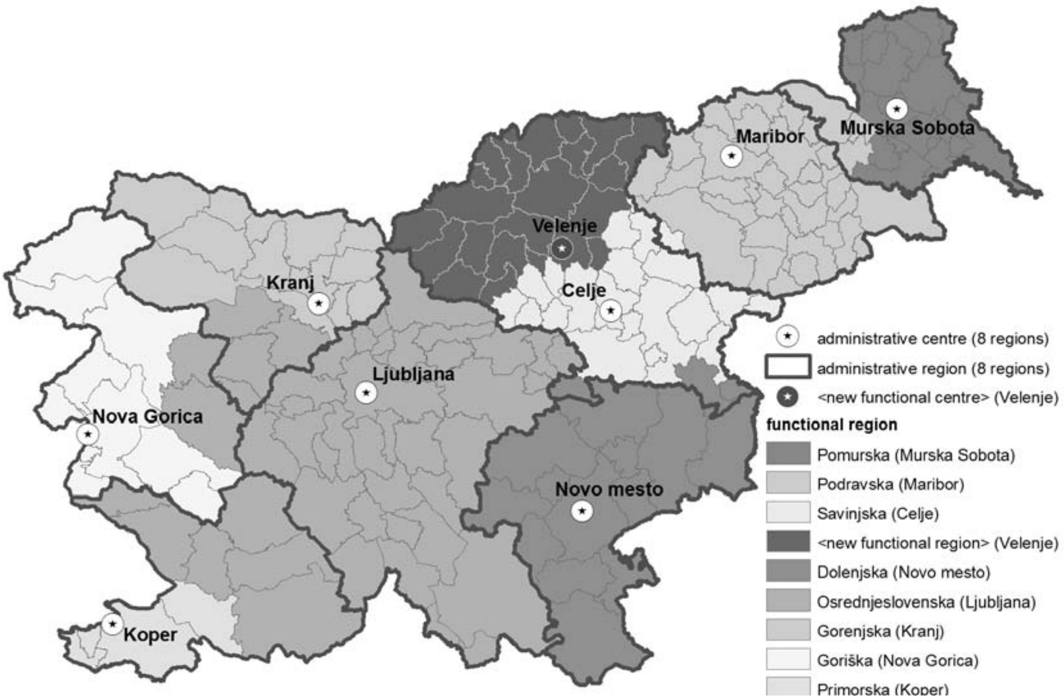


Figure 4: Eight administrative and nine functional regions in Slovenia

5 CONCLUSIONS

In the paper, we suggested method for delimitation of functional regions by using only data on working inter-municipal daily commuting. Since the regionalisation is an important topic of political debate in Slovenia, we decided to identify functional regions at the level, which could be comparable to the proposed administrative regions (provinces). For this reason, the number of functional regions was limited to six and eight.

Based on labour market data (daily commuting) we firstly identified the functional centres (centres of the self-sufficient municipalities). However, the six administrative centres of the regions are comparable to the six functional centres, while for the case of eight administrative regions additional functional centre appeared which could not be overlooked.

For defined functional centres, the functional regions had to be identified (six and nine respectively). As we have already noted, the borders of functional regions can be exactly defined in theory. But in practice, where the functional regions consist of smaller areas that had to be aggregated (municipalities, in our application), the additional method based on building chains of the municipalities was applied. Following the idea of chains of municipalities, based on daily commuting data, the entire territory of Slovenia was covered by functional regions. These regions differ a little bit from the proposals of the administrative regions, in particular for the case of eight administrative and nine functional regions.

According to the results of our research, the proposal for six administrative regions in Slovenia is “stronger” than the model of eight ones. Following the results of suggested and used method, the eight administrative regions are not well defined – in that case nine functional regions would be much more appropriate when considering commuting conditions of inter-municipal working population. This finding should have a significant influence on proposition of new administrative delimitation of provinces in Slovenia.

References

[1] Bole D. (2004): Daily Mobility of Workers in Slovenia. *Acta geographica Slovenica*, 44(1):25-45.

[2] Cörvers F., M. Hensen and D. Bongaerts (2009): Delimitation and Coherence of Functional and Administrative Regions. *Regional Studies*, 43(1):19-31.

[3] Karlsson C. (2007): *Clusters, Functional Regions and Cluster Policies*. CESIS Electronic Working Paper Series. KTH, Stockholm, Sweden. (http://www.infra.kth.se/cesis/documents/WP84.pdf).

[4] Karlsson C. and M. Olsson (2006): The identification of functional regions: theory, methods, and applications. *Ann Reg Sci*, 40:1–18.

[5] Lavtar R. (Ed.) (2004): Dokumenti in študije o pokrajinah v Sloveniji, Zbornik, Ministrstvo za notranje zadeve, Ljubljana. (Documents and Studies about Provinces in Slovenia, Proceedings, Ministry of the Interior).

[6] OECD (2002): *Redefining Territories – The functional regions*. Organisation for Economic Co-operation and Development. Paris, France.

[7] Pichler Milanović N., D. Cigale, M. Krevs, P. Gostinčar, A. Černe, A. Zavodnik Lamovšek, T. Žaucer, M. Sitar, V. Drozg and J. Pečar (2008): *Strategy for a Regional Polycentric Urban System in Central-Eastern Europe Economic Integrating Zone*. RePUS project, Final report. University of Ljubljana, Faculty of Arts, Ljubljana.

[8] Pogačnik A., A. Zavodnik Lamovšek, S. Drobne, B. Trobec and K. Soss (2009a): Analiza konceptov regionalizacije Slovenije s predlogom območij pokrajin , Končno poročilo, Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo; Ljubljana. (Analysis of Concepts of Regionalisation of Slovenia with the Proposal of Provinces, Final Report, University of Ljubljana, Faculty of Civil and Geodetic Engineering).

[9] Pogačnik A., A. Zavodnik Lamovšek, S. Drobne, B. Trobec and K. Soss (2009b): Analiza modelov pokrajin (3, 6, 8) po izbranih kazalnikih : dodatek h končnemu poročilu. Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo, Ljubljana. (Analysis of Models of Provinces (3, 6, 8) by Selected Parameters, Appendix to Final Report, University of Ljubljana, Faculty of Civil and Geodetic Engineering).

[10] Pogačnik A., A. Zavodnik Lamovšek, S. Drobne, T. Žaucer, B. Trobec, N. Pichler Milanović and M. Štefula (2009c): Analiza razvojnih virov in scenarijev za modeliranje funkcionalnih regij, Drugo poročilo, Univerza v Ljubljani, Fakulteta za arhitekturo, gradbeništvo in geodezijo, Ljubljana. (Analysis of Development Scenarios for Modelling of Functional Regions, University of Ljubljana, Faculty of Civil and Geodetic Engineering).

[11] SORS (2009): *Persons in Employment - Daily Commuters by Municipality of Residence and Municipality of Place of Work, Municipalities, Slovenia, Census 2002*. Statistical Office of Republic of Slovenia. Ljubljana. (www.stat.si/pxweb/Database/Census2002/Municipalities/Population/Activity/Activity.asp)

WOOD FUEL SUPPLY MODEL – COSTS, ENERGY USAGE AND CO_2 EMISSIONS

Gregor Guna[1], **prof.dr. Vincenc Butala**[1] and **prof.dr. Marija Bogataj**[2]
University of Ljubljana, Faculty of Mechanical Engineering[1], Faculty of Economics[2]
Aškerčeva 6[1], Kardeljeva ploščad 17[2], SI-1000 Ljubljana, Slovenia
gunagregor@yahoo.com, vincenc.butala@fs.uni-lj.si, marija.bogataj@ef.uni-lj.si

Abstract: In this article we study the problem of wood fuel supply in order to satisfy demand at heating plant with minimal costs and to evaluate environmental impacts, namely energy usage and the accompanying CO_2 emissions. This problem is very complex for a supply manager due to the characteristics of the production, processing and transport of wood fuel. The purpose of this paper is to describe the cost optimization supply model based on linear programming. The model is made in MS Excel and applied to potential supply of Trbovlje heating plant.

Keywords: wood fuel supply, optimization model, cost optimization, energy usage, CO_2 emissions, wood chips, wood pellets

1 INTRODUCTION

Due to the complex supply chain and related costs, the use of wood as fuel is limited. This costs can be largely reduced by choosing the most cost effective transport, source, place and time of processing and storage. Because of impact on environment we should not overlook energy consumption and CO_2 emissions of whole wood fuel supply chain.

The aim is to build a support tool for decision makers, a model for wood fuel supply. The solution of model will be the most cost effective supply and as such the basis for the calculation of energy usage and CO_2 emissions by different phases od supply chain, by months and total, for annual supply. We focus on the two most frequently used forms of wood fuel, wood chips and wood pellets. For better overview, the results will be shown on numerous graphs.

Different types of modeling wood fuel supply were already made in the past. Basically, the models are divided to those that deals with only one user and those which supply more customers within known area. Among the first, Eriksson and Bjorheden already cost optimized the supply in 1989 [11]. Simple and transparent approach for solving the supply of the district heating plant with the forest residues were developed by Gronault and Rauch. The solution of their model presents the minimum supply costs and answeres the question if there is enough stock of wood biomass in observed region for all consumers [4]. Scandinavian authors are the most active in optimizing wood fuel supply. Gunnarsson, Rönnqvist and Lundgren studied the forest residues supply chain. This model considers the problem of selecting the optimal harvesting site, choosing appropriate type of transport, storage and it includes seasonality [5]. Carlsson and Rönnqvist minimized the total cost of wood fuel supply by multiphase linear problem. In this case, the target consumer was paper mill, therefore the wood heat value was not considered [9]. Kanzian et al used GIS (Geographic Information System) and linear programming in biomass supply optimization. The model includes four different scenarios of supplying 9 or 16 customers in a known region. The optimization is based on minimizing the supply costs over larger time period [7]. Help for planning wood fuel supply offers the EDSS (Environmental Decision Support System), based on GIS. It optimizes supply costs by linear programming and includes different energy conversions of biomass. This decision model observes several types of wood biomass, but does not take into account the seasonality of biomass and does not consider greenhouse gas emissions [3].

2 WOOD FUEL RESOURCES AND CHARACTERISTICS

2.1 Wood fuel resources

For energy purposes we can use forest wood fuel, unpolluted waste wood and residues of wood processing [9, 10]. By planning the wood fuel supply, we have to consider only those resources and quantities that are actually available and not all the capacities because of economic or physical reasons, restrictions [12].

2.2 Wood fuel characteristics

For processing, transport and combustion technologies the characteristics of wood fuel are important. The quality of wood fuel is determined by various standards (SIST, ÖNORM, CEN / TC).

Heat value of wood is its main characteristics, measured in kWh/kg, MJ/kg, kWh/m^3 or MJ/m^3. The average value of dry wood ranges from 18.1 to 19.3 MJ/kg. Because of water content in air-dried wood, the heat value falls to 15 MJ/kg, and of fresh wood to 8 MJ/kg [1]. The density and water content in the wood are the main factors that reduces the heat value of wood. We determine wood moisture on the basis of dry weight according to SIST EN 13183-1:2003 [13]. Density of solid dry wood ranges from 320 to 750 kg/m^3 [3] and it is of great importance for calculating the amount of stored energy per volume.

Considering the different forms of wood biomass in whole supply chain, it is necessary to know the problem of measurement. All forms of wood biomass in the model are measured by loose volume (pm^3 or nm^3). The advantage of measurement in volume and not by weight is by small dependence of water content in wood (the volume difference of wet or dry wood is only 5%) [6]. To convert volumes between the various forms, we use factors. For example, from 1 m^3 of solid wood we produce about 1 nm^3 wood pellets or about 3 nm^3 wood chips.

3 WOOD FUEL SUPPLY CHAIN

Modern logistics is a process, linking the whole supply chain and all material flows with a goal to fulfill the needs of the end user on time with minimum costs [2]. Modeling the wood fuel supply and applying it can show the potential for cost reductions and possibility for minimizing the environmental impact. By planning the wood fuel supply of large heating plant we need to involve supply manager, consumer, haulage companies, terminal operators and transport companies.

3.1 The phases in wood fuel supply chain

For modeling the wood fuel supply we need to analyse all phases of supply chain. These are logging, harvesting, collecting, processing, storage and transport. The technological equipment, labour, location of logging or collecting wood biomass, the capacities and final form of wood fuel determines the sequence of operations, phases [8]. Each phase has its own characteristics and variables that determines the total costs, energy usage and CO_2 emissions of supply chain.

3.1.1 Logging, harvesting and collecting

Logging and harvesting technology of small scale producer is based on manual work and in most cases is carried out by tractor, larger companies uses harvesters [8]. Collecting of wood biomass takes place on the forest road, where the forest fuel wood is loaded on the truck or processed into wood chips. This phase includes the collection of waste wood and residues of wood processing.

3.1.2 Processing of wood biomass

Processing of wood biomass into wood chips or pellets is preferably done before transporting it over long distances due to better utilization of transport. Processing can be done in the forest, before transport to the terminal or at the terminal. This phase includes drying the wood biomass to the required wood moisture.

Wood chips are pieces of wood, sized from 3 to 5 cm, up to 10 cm. During the storage, wood chips do not dry, therefore the required moisture of wood chips has to be achieved before processing. Chips are produced in the forest by mobile chippers or at the terminal. Processing wood biomass at the terminals is more productive. Bulk density of wood chips ranges from 190 to 270 kg/nm^3, the heat value is ca. 3 to 4 GJ/nm^3 [10]. Characteristics of wood chips must meet the technology of combustion.

Wood pellets are standardized wood fuel made by compressing fine particles of wood into cylindrical shape, (diameter 4 to 20 mm, length 100 mm). Production of pellets consists of the following phases: grinding of wood raw material into small particles; drying wood raw material to wood moisture of 8 to 10%; compression of small particles into the pellets; cooling; separation and the decomposition of residues, which are returned in the process; storing pellets in bags or in silos [14]. By pressing wood into pellets, we increase the density and the heat value while reducing the cost of transport and storage. Bulk density of pellets ranges from 600 to 700 kg/nm^3, and it is two to three times greater than the density of wood chips [10]. Lower water content improves combustion. The heat value of wood pellets ranges from 8 to 11 GJ/ nm^3 [15].

3.1.3 Storage

Terminals are used to balance seasonal variation of supply and demand and also to offer more chipping possibilities. At terminals, non-chipped, chipped and pelleted wood can be stored. Terminal equipment consists of cranes, chippers, crushers, pellet mills and sorting devices. Storage of wood biomass raises the whole supply cost, in literature is estimated 26 % raise of the storage costs [7].

3.1.4 Transport

The type of transport used depends on the form of wood biomass. The transport cost depends on the distance, on the price of fuel used and of loading/unloading time. Water transport and piping is rarely used. On a distance less than 150 km, trucks offer the most economical and flexible transportation, especially when transporting wood biomass from large numbers of sources is required. In our model we included the transport by truck and train, which is mainly used in Slovenia.

4 MODEL

The model is based on different supply routes with three initial, raw forms of wood biomass: forest wood fuel, waste wood and residues of wood processing. It considers two end forms of wood biomass, wood chips and pellets. Beside minimizing the supply cost with the model, we want to evaluate the energy usage and actual greenhouse gas (CO_2 emissions). The optimization is based on the method of linear transport programming. The solution provides us the volumes of delivered wood biomass from different supply routes. Optimization is based on enrolled data. We have already set the fuel consumption, energy usage and CO_2

emissions per volume unit of wood fuel for some typical machines used in whole wood fuel supply chain. This data can easily be adjusted.

4.1 Model structure

The basis for the formulation of the model is chart, shown in Figure 1, presenting the amount of wood fuel supply, consumption and safety stock. Monthly consumption of wood fuel (wood chips or wood pellets) y_j in nm^3 and monthly safety stock s_j in nm^3 is determined by heating plant. Index i denotes certain supply route (in model we set 6 individual supply routes) and j denotes the time (months, from 1 - January to 12 - December). Safety stock is usualy determined as a level that should eliminate the risk for scarcity of fuel and is therefore the safety stock level. It is roughly estimated as volume of wood fuel, combusted in 5 to 10 days. Initial quantity of safety stock in January is equal to Decembers volume. The monthly volume of wood fuel delivered to the heating plant by specific supply route is marked with $x_{i,j}$. For all supply routes, we must know the associated costs. With the type of fuel used (petrol, diesel, electricity), time usage of different machines per volume of wood fuel and by multiplying with x_{ij} value, we get the energy usage and CO_2 emissions on monthly level for individual supply route.

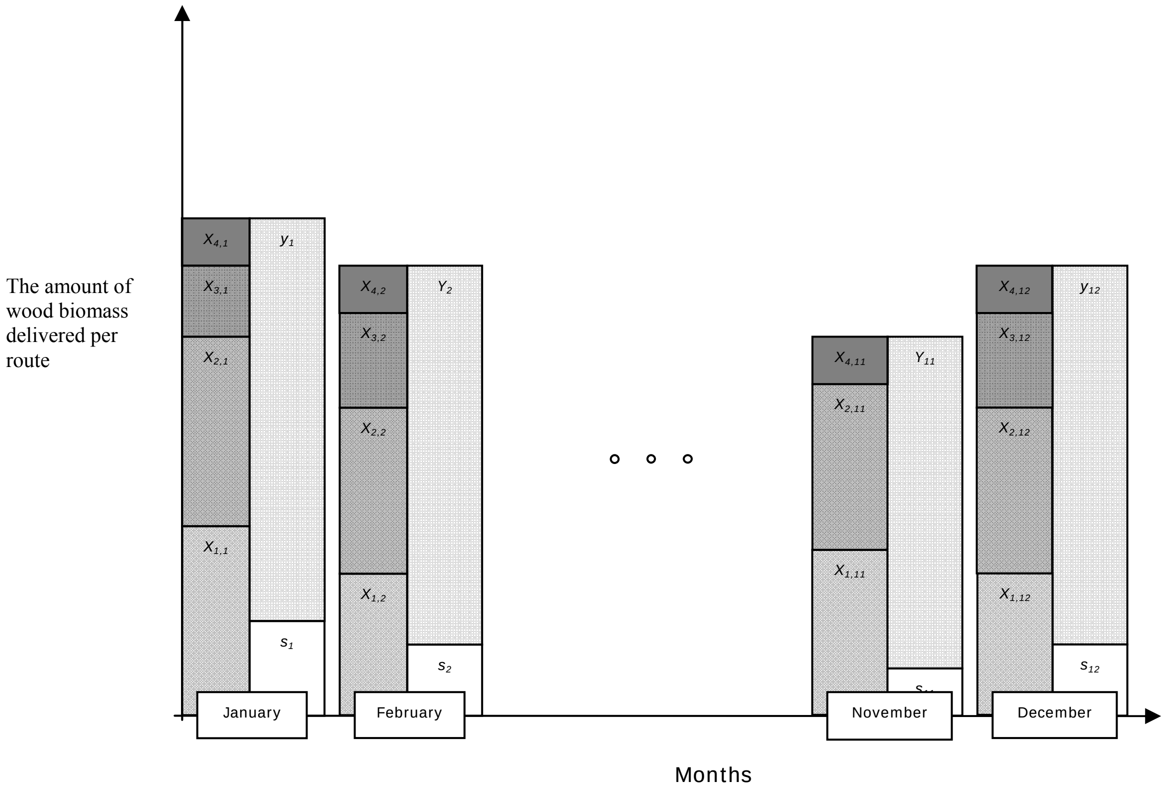


Figure1: Graph of wood fuel supply to the heating plant.

Monthly volume (demand) of delivered wood biomass (1) in nm^3 (wood chips or pellets) to the heating plant are formulated as :

$$\sum_i x_{ij} = y_j + s_j \quad (1)$$

Objective function of supply cost optimization (2) is:

$$\min_i \left(\sum_j c_i x_{ij} \right) \quad (2)$$

The restriction (3) of model is noted below, where d_j expresses monthly demand of wood biomass in nm^3, delivered to heating plant:

$$\sum_i x_{ij} = d_j \quad (3)$$

The cost of supply c on individual supply route i is c_i . It is a sum of costs of different phases of whole supply chain (4), equal for all months, measured in €/nm^3:

$$c_i = f_{l,i} + f_{wc,i} + f_{wp,i} + t_{f,i} + t_{ww,i} + t_{r,i} + t_{cp,i} + t_{ch,i} + t_{ph,i} + s_{u,i} + s_{wc,i} + s_{wp,i} \quad (4)$$

Logging, harvesting and gathering wood biomass cost is noted as f_l, processing cost of wood chips as f_{wc} and processing cost of wood pellets as f_{wp}.

The transport cost of forest wood biomass to terminal is noted as t_f , transport cost of wood waste to terminal as t_{ww}, transport cost of wood processing residues to terminal as t_r, wood chips transport to wood pellets processing site as t_{cp}, wood chips transport cost from processing site to heating plant as t_{ch} and wood pellets transport cost to heating plant as t_{ph}.

The storage cost of unprocessed wood biomass is noted as s_u, storage cost of wood chips as s_{wc} and storage cost of wood pellets as s_{wp}. All these costs considers equipment costs (investment, depreciation, insurance), the costs of maintenance and repairs, the costs of fuel, the cost of labor and management.

We already know the energy usage of wood chips or wood pellets production E_i in MJ/nm^3 and CO_2 emissions of wood chips or wood pellets production e_i in kg $CO2/nm^3$ for individual supply route. By multiplying those values with x_{ij}, we get the energy usage and CO_2 emissions on monthly level for wood chips or wood pellets supply, as it is shown in (5) and (6).

$$E_{i,j} = \sum_i E_i x_{i,j} \qquad \forall j \in J \qquad [MJ] \quad (5)$$

$$e_{i,j} = \sum_i e_i x_{i,j} \qquad \forall j \in J \qquad [kg\ CO_2] \quad (6)$$

4.2 Supply model in MS Excel

The model is set in MS Excel and divided into four sections. It can be used for wood chips or wood pellets supply of heating plant individualy or for comparing these supplies.

In the first section we have to evaluate the expected amount of heat demand for each month, from January to December and to determine the volume of security stock of wood fuel. For calculating the wood fuel demand volume we must input the heating value, bulk density and wood moisture of wood fuel.

In the second part of the model we already defined most frequently used machines in wood fuel supply chain, their capacities, energy usage and CO_2 emissions.

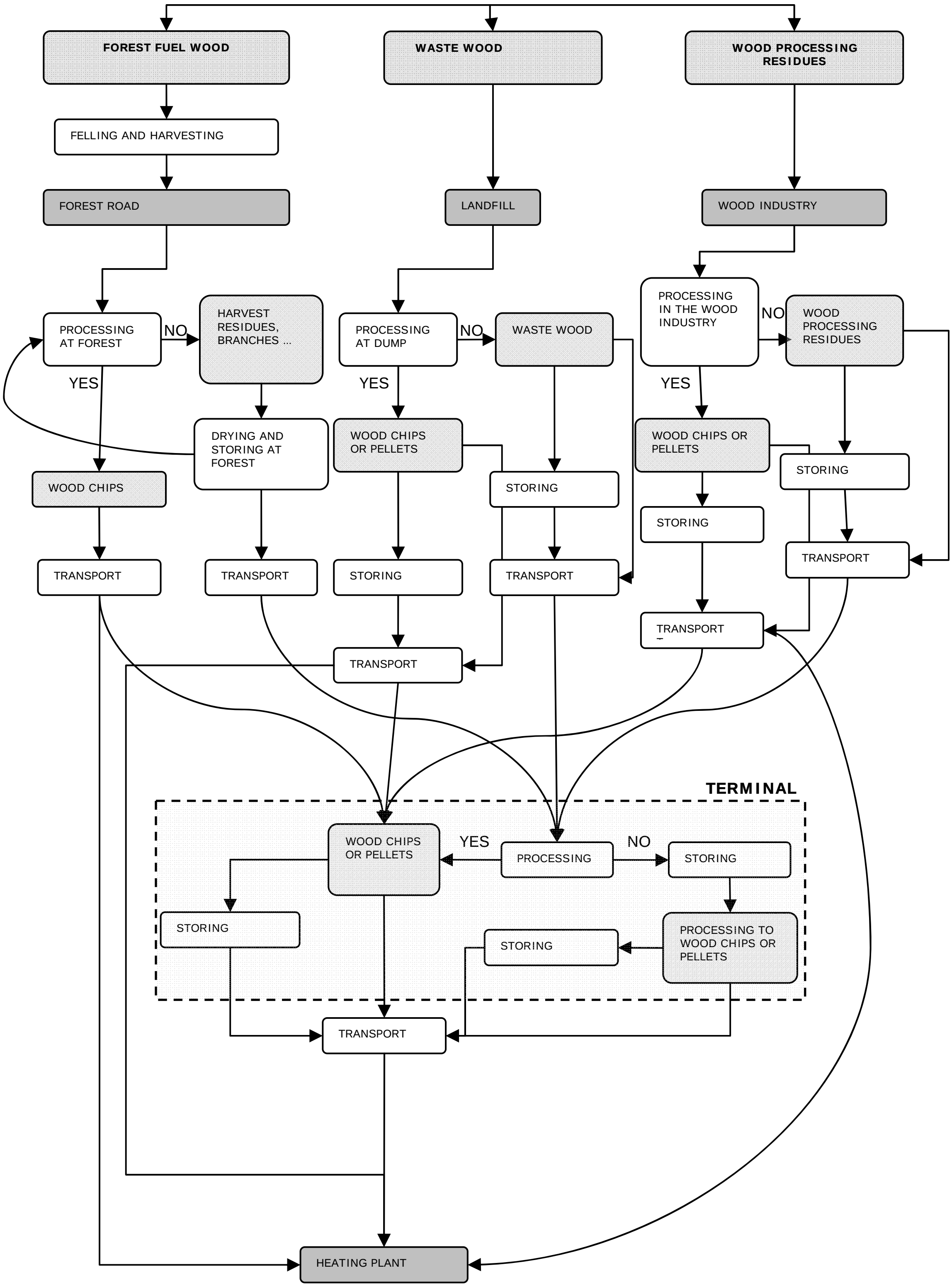


Figure 2 : Flow chart of all possible supply routes from three different sources of wood biomass to a heating plant.

The third section deals with individual supply route. We have to evaluate time use per volume of wood fuel of different machines, involved in different phases of supply chain. The modeling is based on the volume of the final form of wood fuel. By conversion factors, shown in Table 1, we can use the model with three different input raw wood biomass forms.

Table 1: Indicative volume conversion factors (evaluation of author).

Raw wood biomass	Wood chips (nm^3)	Wood pellets (nm^3)
1 nm^3 forest fuel	2,15	0,5
1 nm^3 waste wood	2,3	0,6
1 nm^3 residues of wood processing	2,5	0,7

With the use of these factors we can follow the individual supply route and insert data about capacity of specific source, costs (harvesting, storage, processing), type of transportation and distances. For better overview of different supply routes we can use the Figure 2.

At the last, fourth stage we use tables in MS Excel for monthly cost optimization of wood fuel supply. Based on the previous three stages and with the help of optimization tool Solver, we get the solution. Basic Excel Solver uses a Simplex method and is intended to solve not-too-complex optimization problems. Parameters in Solver are set to minimize the supply cost. The delivery of each route is set to be greater than 0 and should not exceed the capacity of individual supply route, the total delivery is the same as user's monthly demand.

Solution contains cost optimized monthly supply in €, energy usage in GJ and CO_2 emissions in t. Model warns us if delivery is unsatisfactory and displays the results in tables by month and divided by stages. It draws the following charts:

- energy usage for the whole wood chips supply chain and separately for the whole wood pellets supply chain,
- CO_2 emissions by whole wood chips supply chain and separately by whole wood pellets supply chain,
- comparison of costs, energy usage and emissions of CO_2 and
- whole wood chips and pellets supply chain costs by phases in relative parts.

4.3 Model aplication

With the intention to introduce the capabilities of the model, we aplicated it to the wood fuel supply of Trbovlje heating plant. The heating company demands are shown in Table 2 below.

Table 2: Energy production, wood chips and pellets demand by heating plant [17].

	Jan	Feb	Mar	Apr	May	Jun	Jul	Avg	Sep	Oct	Nov	Dec
Energy (MWh)	8400	8000	6500	4000	2000	1500	1000	1000	3000	4500	6500	8000
W. chips (nm^3)	7529	7171	5826	3585	1793	1345	896	896	2689	4034	5826	7171
W. pellets (nm^3)	4200	4000	3250	2000	1000	750	500	500	1500	2250	3250	4000

As it is seen from Table 2, the demanded volume of wod chips is quite larger compared to demanded volume of wood pellets. This is because of higher heating value of wood pellets per volume unit. The solution will compare supply of Trbovlje heating plant with these two forms of wood fuel.

We chose four different supply routes; localy situated forestry company Gaber, localy situated wood-processing company Lipa, a distant forestry company BiH and the regional landfill Odvoz. We have defined their monthly capacities, costs and machines used.

4.4 Results

The results of cost optimization are as follows. In case of wood chips supply we fully use the capacities of Gaber, Lipa and Odvoz, capacities of BIH are used partly. In case of wood pellets supply we fully use the capacities of Gaber, Lipa and BIH. Supplying the wood pellets from Odvoz is used partly. The wood chips supply cost on annual level is 129400 €, the wood pellets supply cost is 98300 €.

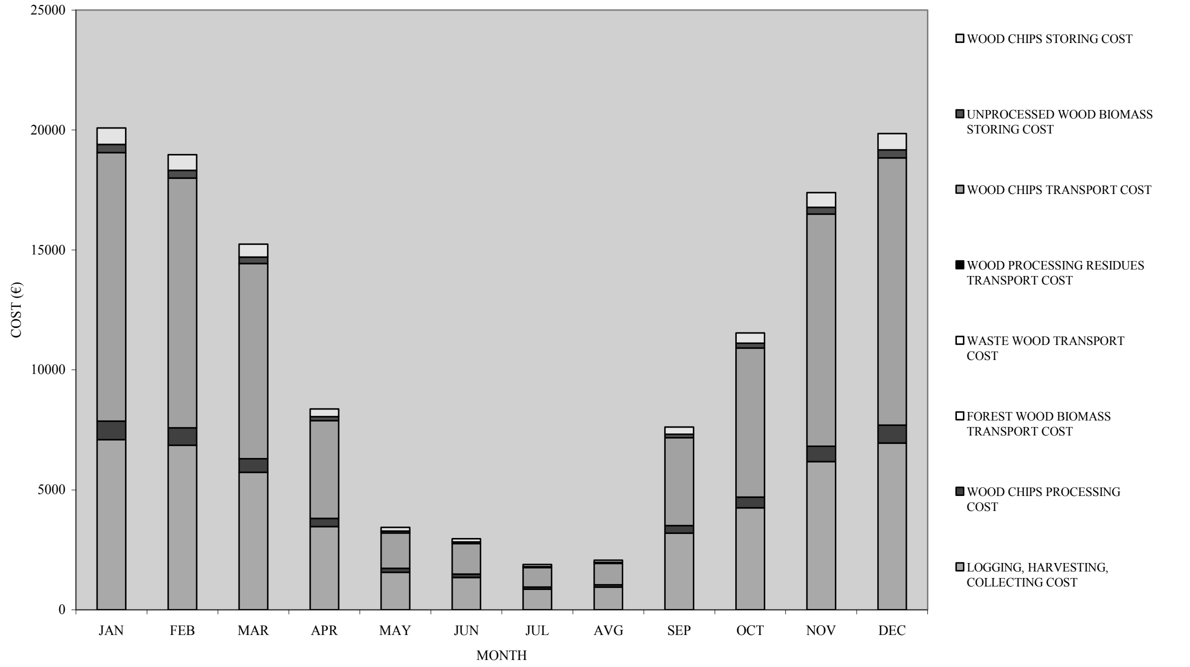


Figure 3: Graph of wood chips supply chain costs of Trbovlje heating plant, divided into different phases. Other graphs, displaying wood chip and wood pellets supply chain energy usage and CO_2 emissions are shown in the same way.

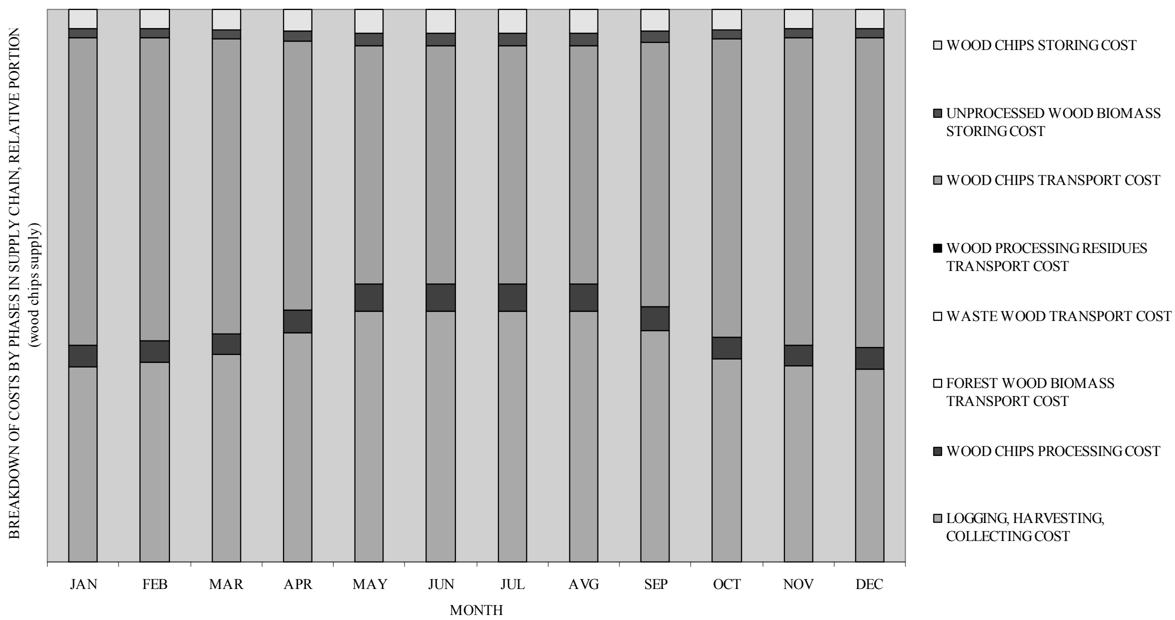


Figure 4: Graph, showing breakdown of Trbovlje heating plant whole wood chips supply chain costs by phases in relative parts.

In the wood chips supply shown in Figure 3 and 4 above, we can see that the transportation, logging, harvesting and collecting presents the largest share of whole supply chain cost, the same as with wood pellets. There is a significant difference in supply chain energy usage between wood chips and pellets. Energy usage in the whole wood chips supply chain was 322.0 GJ per year and in the case of whole wood pellets supply chain 588.9 GJ per year. The increased energy usage in the production of pellets results in higher whole supply chain CO_2 emissions, 67.5 t CO_2 per year, while the wood chips supply chain does not emitte large quantities of CO_2, only 30.9 t per year. Wood pellets production is characterized by high energy consumption due to the need of drying, reducing size and granulation of the material. We had no deficit of delivered wood chips or wood pellets. Differences in cost, energy usage and CO_2 emissions between wood chips and wood pellets supply are shown in Figure 5.

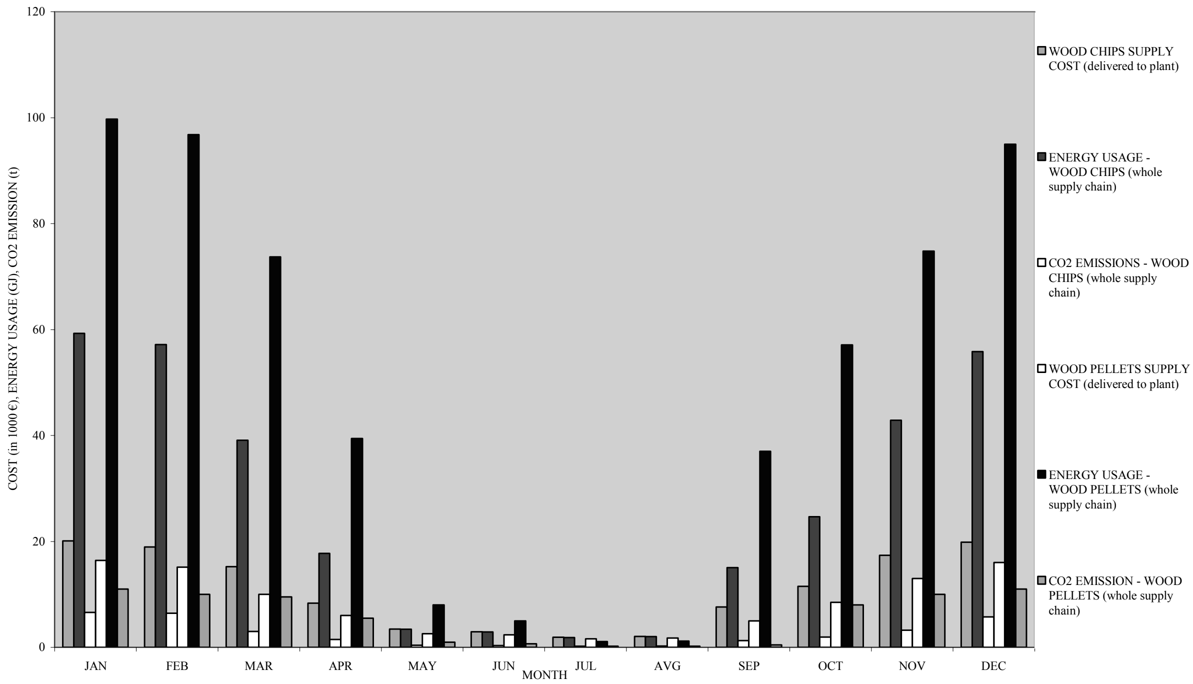


Figure 5: Wood chips and wood pellets supply of Trbovlje heating plant, comparison.

In Figure 5, it is shown that calculated total supply cost of demanded wood chips to Trbovlje heating plant is a bit higher comparing to supply cost of demanded wood pellets, mainly due to the larger transport volumes of wood chips. The increased energy usage in the production of pellets results in higher CO_2 emissions of whole supply chain as seen in Figure 5.

5 CONCLUSION

In order to optimize the supply cost by choosing the most cost efficient supplier, we use the method of linear programming. The user obtaines solution by entering the necessary data into the model. The solution of supply is the basis for energy usage and CO_2 emission calculations. The media often missleads the public stating that wood combustion and heating does not contribute additional greenhouse gas emissions. By this model we can evaluate emissions and as it is shown in our application, emissions of the whole supply chain are quite high. The model can also be used to compare the total supply costs, energy usage and emissions of wood chips and pellets. With graphic presentation of solution, we can easily evaluate which phase presents a larger part of costs, uses larger amounts of energy and emitte high quantities of CO_2. With the help of this tool, we can also select the best location of heating plant. The advantage of the model is in its design in Excel tables, with the possibility of simple improvements, upgrades. The proposal for further work on the model is to upgrade it as a standalone program with links to other databases and the possibility of determining the distance (or optimization) with the help of a geographic informational system.

References

[1] Carlsson, D., Rönnqvist, M., 2005. Supply chain management in forestry – case studies at Södra Cell AB, European journal of operational research, 163, pp. 589–616.

[2] Chopra, S., 2001. Supply chain management: Strategy, planing and operation. Upper Sadlle River, New Jersey, Prentice Hall, 2001, 449 p.

[3] Frombo, F. et al, 2009. Planning woody biomass logistics for energy production: A strategic decision model. Biomass and bioenergy 33, pp. 372–383.

[4] Gronalt, M., Rauch, P., 2007. Designing a regional forest fuel supply network, Biomass and Bioenergy 31, pp. 393–402.

[5] Gunnarsson, H., Ronnqvist, M., Lundgren, J. T., 2004. Supply chain modelling of forest supply, European journal of operational research 158, pp. 103–123.

[6] Hakkila, P., 1989. Utilization of Residual Forest Biomass. Springer-Verlag Berlin Heidelberg, 568 p.

[7] Kanzian, C. et al, 2009. Regional energy wood logistics – optimizing local fuel supply, Silva Fennica 43, pp. 113–128.

[8] Krajnc, N., Krajnc, R., 2003, Tehnologije pridobivanja lesnega kuriva, Gozdni obnovljivi viri 4, pp. 9–11.

[9] Operativni program rabe lesne biomase kot vira energije (OP ENLES 2007 – 2013), 2007. Ministrstvo za okolje in prostor.

[10] Pogačnik, N., Krajnc, R., 2000. Potenciali lesne biomase uporabne v energetske namene. Gozdarski vestnik 7-8, pp. 330–332.

[11] Summary of summer school OPET CHP/DHC ENV-3, Jyvaskyla, 2003. Pridobljeno 17. 4. 2009 s spleta: *www.opet-chp.net/download/wp3/summerschoolreport.pdf.*

[12] Rosillo-Calle, F. et al, 2007. The biomass assessment handbook – Bioenergy for a sustainable environment, Earthscan London, 269 p.

[13] Straže, A., 2009. Metode določanja lesne vlažnosti, Biotehniška fakulteta, oddelek za lesarstvo. Pridobljeno 11. 5. 2009 s spleta: www.*les.bf.uni-lj.si*

[14] Stritih, U., Butala, V., 2002. Proizvodnja pelet v Sloveniji, EGES 1, pp. 2–4.

[15] Swigon, J., Longauer, J., 2005. Energy consumption in wood pellets production, Folia forestalia Polonica 36, pp. 77–83.

[16] Van Belle, J. F., Temmerman, M., Schenkel, Y., 2003. Three level procurement of forest residues for power plant Biomass and bioenergy 24, pp. 401–409.

[17] mr. Krajnik, 19. 2. 2009. Operator at Trbovlje heating plant.

APPLYING A'WOT METHOD TO PRIVATE FOREST MANAGEMENT: CASE STUDY ON CHAMBER OF AGRICULTURE AND FORESTRY OF SLOVENIA

Špela Pezdevšek Malovrh , Janez Krč, Lidija Zadnik Stirn
University of Ljubljana, Biotechnical Faculty,
1000 Ljubljana, Slovenia
spela.malovrh@bf.uni-lj.si, janez.krc@bf.uni-lj.si, lidija.zadnik@bf.uni-lj.si

Abstract: The article is related to institutional problems of sustainable forest management, new technological conditions, optimal economic and ecological effects, and satisfaction of forest owners and society as a whole regarding forestry. In particular, the article deals with an influence of the Chamber of Agriculture and Forestry of Slovenia (C.A.F.S.) on private forest management. The strengths, weakness, opportunities and threats (SWOT) of the institutional/owners' organization in relation to the problems of private forest management are determined. By the use of AHP analysis we study which SWOT factors are more influential for the strategic, tactical and operational private forest management guidelines, generated by nongovernmental institutions, in our case the C.A.F.S.

Keywords: private forest management, institutional influence, Chamber of Agriculture and Forestry of Slovenia, SWOT analysis, multiple criteria decision making, analytic hierarchy process

1 INTRODUCTION

Since Slovenia gained its independence in 1991, the institutional organization of forestry has undergone a lot of changes. Forestry in Slovenia experienced organizational changes with separation of public from business activities (after 1993). As a consequence of these changes, forest management is organized in a top-down way, in which the public interest is more important than the interest of forest owners. Nowadays, the disadvantages in the realization of organizational changes are seen in the non-optimal organization of institutions, expensive institutional activities, inadequate organization of private forest owners, technological setbacks, the fact that younger forest development phases are frequently neglected, which leads to poor growth of stable and quality forests, the production capacity of private forest sites not being exploited enough, only 60% of wood which could be cut according to forest management plans is actually cut, high costs of timber removal and harvest because of fragmentation of forest holdings, long-term stagnation of the prices of biomass, etc., [6]. According to Pezdevšek Malovrh and Krč [5], successful forest management, based on environmental friendly, multifunctional and sustainable principles, and which ensures the public interest as well as the interest of forest owners, can be implemented only through cooperation between owners and institutions.

In Slovenia, institutions and organizations on state level as well as on local level are active in influencing and improving the private forest management. Some of them control the management through law and legal instruments, the others stimulate it through financing and co-financing, and the third improve the educational level through education and counselling. The extent of various tasks that these institutions are performing is changing, too, due to organizational changes in forestry and general changes in the socio-economic environment, [3].

In this paper we analyze a nongovernmental organization, i.e., the Chamber of Agriculture and Forestry of Slovenia (C.A.F.S.) which was established on the basis of the Chamber of Agriculture and Forestry Act (OG RS, Nr. 41-2025/99), and its influence on private forest management. The aim of this analysis is to improve private forest management and to reduce the existing drawbacks in the current private forest management practices. The data for this analysis were obtained through questionnaires. The questions were set up with the aim to determine the institutional (C.A.F.S.) influence on private forest management in the light of sustainable management, new technological conditions, optimal economic and ecological effect and, not least, to satisfy the requirements of forest owners and society as a whole. The employees who are working in the C.A.F.S. sector of forestry consulting were interviewed. In the first part of the interview they were asked to express their opinion about the strengths, weakness, opportunities and threats of their activities on private forest management. In the second part of the questionnaire they were asked to give pairwise comparisons of SWOT factors, [15]. The AHP analysis, ([7], [8]), carried out on the surveys' data generates the importance (rank) of the SWOT factors and thus, gives the information which of them are important for the most suitable management of private forests, according to the set goals.

2 METHODOLOGY USED

The method employed in this paper is A'WOT method [9]. This method, which could be assigned as a hybrid method consisting of SWOT analysis and AHP method, is in this paper used to analyze the multiple criteria problem which is presented in a hierarchical structure.

2.1 Hierarchical multiple criteria decision making

Decision making deals with strategies (decisions, alternative paths, alternatives, different projects, etc.) and objectives (goals) in the perspective of a changing environment. Further, it requires careful consideration and evaluation of the external and internal factors. Very important is also the assessment of opportunities, threats, strengths and weaknesses of the strategic paths under consideration.

Multiple criteria decision making is based on the fact that the choice of a solution is affected by numerous criteria, the importance of which varies and which are of hierarchical structure, i.e., they can be presented at different levels. A decision making problem is thus broken down into smaller subordinate problems (parameters, criteria, attributes), and these are then assessed separately for each parameter. The final assessment is obtained by means of a specific combining procedure, [12]. As evident from Špendl et al. [12], the model is formed by means of parameters (attributes), X_i. These are variables which represent subordinate problems of a decision making process (attributes which define the quality of alternatives). Then the utility function F is initiated. Function F represents the rule according to which the values of individual parameters are combined to form variable Y. This variable represents a final assessment of alternatives.

In actual cases we know what we prefer, but we are unable to assign a certain value to a solution (the utility function F is not known). What is needed, then, is a procedure which converts preference relation into utility function. One of the procedures which make this possible is the AHP method.

2.2 Analytic hierarchy process (AHP)

In problems dealing with multiple and conflictive objectives (goals, factors) of the alternatives, and above all with objectives of different importance Saaty's analytic hierarchy process, assigned as AHP method is employed to determine the best alternative. AHP can incorporate mixed data that may include both qualitative and quantitative judgments, and is capable of analyzing multiple factors (parameters, attributes, criteria). AHP is based on a gradual mutual comparison of two objectives (pairwise comparison) at the same level. A scale from 1 to 9 is used for making the comparison, while the reciprocals of these values tell that if objective k has one of reasonable assumptions of the above nonzero numbers assigned to it when compared with objective j, then j has the reciprocal value when compared with k, ([7], [8]). Comparisons between individual objectives are gathered in a pairwise comparison matrix $A = [a_{kj}]$, (k = 1, 2, ... K, j = 1, 2, ... K) if there are K objectives. Each objective k is associated with a weight w_k. The weights ratio of the objectives k and j is written as intensity of importance:

$$a_{kj} = \frac{w_k}{w_j} \quad (1)$$

Since, in practice, we never encounter perfectly consistent estimations [7], we prove the consistency as described in Winston [14], using the consistency index. Further, by (1) defined vector of weights $w = (w_1, w_2, \dots, w_K)$ is calculated with multiple squaring of matrix A to the satisfactory exponent, i.e., A, A^2, $(A^2)^2$, etc. and then the lines are summed up and the values normalized, [13]. The vector of weights $w = (w_1, w_2, \dots, w_K)$ is therefore scaled between 0 and 1, $\sum w_k = 1$, and calculated by the following equation:

$$w_k = \frac{\sum_{j=1}^{K} a_{kj}}{\sum_{k=1}^{K}\left(\sum_{j=1}^{K} a_{kj}\right)} . \quad (2)$$

Vector w can also be obtained by searching for eigenvalues λ of matrix A: $Aw = \lambda_{max} w$, where λ_{max} is the maximum eigenvalue of matrix A and w is the corresponding eigenvector. In practical cases, this tends to be a complex calculation procedure, [1]. The eigenvalues which correspond to the eigenvector could also be obtained as, [14]:

$$\lambda_{\max} = \frac{1}{K} \sum_{i=1}^{K} \frac{(Aw)_i}{w_i} \quad (3)$$

The measure of inconsistency is then defined by the difference ($\lambda_{\max}$ - K). It is expressed by the consistency index CI:

$$CI = (\lambda_{\max} - K) / (K - 1). \quad (4)$$

Inconsistency CR is calculated as:

$$CR = CI / RI \quad (5)$$

where a random index RI is given in tabular form, [14]:

K	1	2	3	4	5	6	7	8	9	10
RI	0	0	0,58	0,90	1,12	1,24	1,32	1,41	1,45	1,51

If CR < 0.1, the matrix A is sufficiently consistent. In the opposite case, the matrix should be corrected; otherwise the results will not be correct. For the purpose of this paper we used AHP as implemented in the program Expert Choice.

2.3 SWOT analysis

SWOT analysis means analysis and assessment of comparative strengths and weaknesses of a strategy in relation to competitive strategies, and environmental opportunities and threats which the strategy under consideration may face. SWOT analysis is, as such, a systematic study and identification of those aspects of the strategy that best suit, in our case, sustainability, maximal expected profit, refers to ecological objectives, and respects the forest owner’s acceptance of the examined alternatives. SWOT should be based on logic and relational thinking so that the selected strategy improves the strategy’s strength and opportunities and at the same time reduces the weaknesses and threats, [15], [16].

Strength is a distinct superiority (competitive advantage) of technical knowledge, financial resources, skill of the people, image of products and services, access to best network, of discipline and morale. Weakness is the incapability, limitation and deficiency in resources such as technical, financial, manpower, skills, image and distribution patterns of the alternative under examination. It refers to constraints and obstacles of the alternative. Corporate weaknesses and strengths are a matter of how the alternative can achieve best results compared to other, similar competitive alternatives. Weaknesses and strengths of the alternative present internal forces and factors required to be studied and assessed with the goal to evaluate and rank the alternatives under consideration.

Opportunities and threats are the external factors of the examined strategies. These factors are changing with the change of governmental, industrial, monetary and market policies, including the changes of legal and social environments. Environmental opportunity is an area in which the particular strategy would enjoy a competitive advantage. Proper analysis of the environment, identification of new market, new and improved customer groups and new relationship could present opportunity for the strategy. Threat is an unfavourable environment for the strategy. Increased bargaining power of users and suppliers, quick change of government policy, rules and regulations may pose a serious threat to the strategy.

SWOT analysis is nowadays very important for decision making. Such analysis can be undertaken effectively through a brainstorming session with the participation of experts and users of the environment, land, firm, etc. involved in the strategy. SWOT analysis has many advantages. Within SWOT, internal and external factors are analyzed and summarized in order to attain a systematic decision situation. There are also several shortcomings of using SWOT. SWOT results in listing and quantitative examination of internal and external factors, and groups the factors in strength, weakness, opportunity and threat groups, but it is not able to identify or analytically determine the most significant factor or group in relation to the examined strategy. In order to get qualitative information, to yield analytically determined priorities for the factors and groups included in SWOT analysis and to make them commensurable we have integrated SWOT analysis with AHP when measuring the institutional influence on private forest management. The problem is organized in a hierarchical structure around the concept of objectives (in our case SWOT groups: strengths, weaknesses, opportunities, threats), and attributes (in our case SWOT factors), within a two-level hierarchy (Figure 1). The first level is viewed as objective/group level. These groups are not directly measurable by themselves, but are presented by factors which are found at the second level. The factors define the effect of the SWOT group.

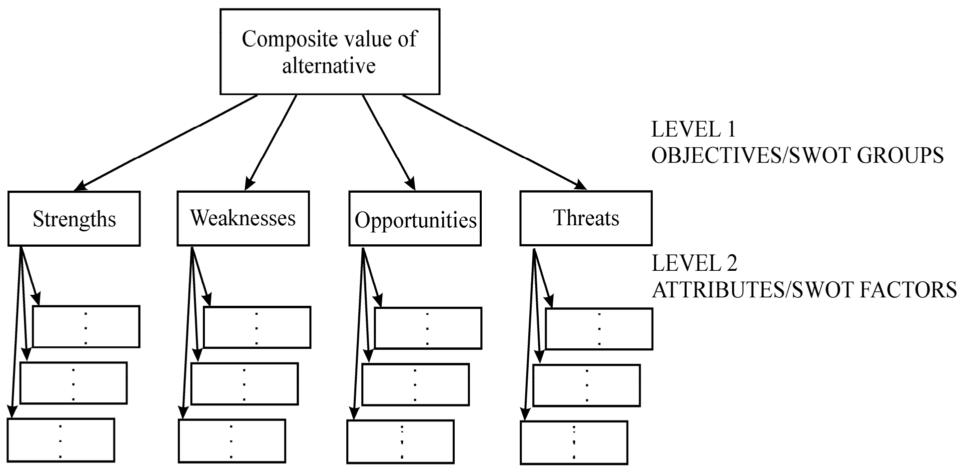


Figure 1: Factor's and group's hierarchy for composite value of institutions in SWOT, [15]

2.4 A'WOT method

A'WOT method is performed according to the following steps, [9]:

Step 1: SWOT analysis is carried out. The relevant factors of the external and internal environment are identified and included in SWOT analysis.

Step 2: Pairwise comparisons between SWOT factors are carried out within each SWOT group. Using these comparisons, the relative local priorities (importance) of the factors are computed using the eigenvalue method of AHP, i.e., formula (2). The consistency is proved using formulas (3), (4) and (5).

Step 3: Pairwise comparisons are made between the four SWOT groups. The factor with the highest local priority is chosen from each group to represent the group. These four factors are then compared and their relative priorities are calculated using again the eigenvalue method of AHP (formula (2)). These priorities are the scaling factors of the four SWOT groups and are used to calculate the overall (global) priorities of the independent factors within groups. This is done by multiplying the factors of local priorities (defined in Step 2) by the value of the corresponding scaling factor of the SWOT group.

Step 4: The results are employed for the generation of management strategies and for the evaluation of the process under consideration. New goals may be set, strategies defined and implementations planned to take into consideration the most important factors.

3 PRACTICAL CASE

C.A.F.S. is a nongovernmental organization. Its task is to protect and to represent interests of agriculture, forestry and fishery, to consult the individuals and the legal entities that perform agricultural, forestry and fishing activities and to accelerate economic and environmental friendly agriculture, forestry and fishing (OG RS, Nr. 41-2025/99).

The Chamber of Agriculture and Forestry Act (OG RS, Nr. 41-2025/99), requires compulsory membership for all forest and agricultural land owners whose cadastral income exceeds a prescribed limit (approx. 85€/ha in 2009). In the frame of the C.A.F.S., a consulting sector for forestry is functioning. Its part is also a forestry advisory service. The latter is a professional service; its activities are accessible to all members. The forestry field within the C.A.F.S. is not sufficiently covered as 13 regional units have only 3 advisors (data from 2009). Thus, the required tasks in the field of forestry are completed only in departments in Maribor, Ljubljana and Nova Gorica. For more efficient activities of forestry advisory service, according to the program that the C.A.F.S. asks for, at least 70 employees would be needed, [11]. The second problem regarding forest management is caused by owners with small forest property who have low cadastral income and therefore they are not members of the C.A.F.S., [2]. All these and similar forest management problems are of multiple objectives and thus, we used A'WOT model to identify and to solve them.

Step 1 of A'WOT analysis

SWOT analysis involved five employees from the forestry advisory sector of the C.A.F.S. The interviewers are assigned as A1, A2, A3, A4 and A5. We interviewed them in autumn 2007. Before conducting the interviews, we studied their activities with the help of the annual report [10], and work program [11]. The results of these surveys are given in Table 1.

An important strength of the C.A.F.S. according to employees is wood certification. Wood certification enables the export of wood and its products in the EU, because importers require an internationally recognized certificate, which tells that a product is made from wood that is sustainably sourced. The strength of the C.A.F.S. is also the fact that it is a nongovernmental organization which works for the owners' benefits with an interdisciplinary approach, with direct counselling mostly about their rights, advantages of associating and a joint sale of timber, education possibilities in the field of forestry, about possibilities how to get subsidies for silvicultural work in forests and about the circumstances on the wood market. C.A.F.S. also offers lectures about the management of private property, subsidies in forestry and National forestry program 2007-2013. Employees have professional knowledge and can offer professional help to forest owners regardless of the public interest. These strengths separate the C.A.F.S. from the public forestry service.

Weakness of the C.A.F.S. is the lack of employees. The employees are consequently in an inferior position compared to agricultural consultants and are not able to cover the entire country, which reflects in limited power in influencing private forest management and in not taking the initiative in associating private forest owners to public forestry service. Consequently their work is not specialized enough, so they tend to be ignored by the government when passing proposals and remarks on the legislation.

Increasing the number of employees is an opportunity that will influence the acceptance of new the laws, and will offer professional help by organizing local associations. Through associations of forest owners, the C.A.F.S. will easily perform education and counselling and will provide circumstances for joint timber sale. Further, a net for more efficient communication of the C.A.F.S. with private forest owners could be established through associations. Bilateral cooperation between wood industry and forest owners could help forest owners to achieve positive effects in forest management and timber sellers to reduce costs. Internal threats of the C.A.F.S. are connected with personnel deficiency. They only cover Ljubljana, Maribor, Nova Gorica, so it is hard for them to work on the local basis. Since the government is unprepared to increase the number of employees, the C.A.F.S. impact on private forest management is much smaller than that of the public forestry service. The main threat of the outside environment that affects their activity is connected with the decrease of the rural population, weak cooperation of private forest owners, and their inadequate knowledge about the timber market.

Table 1: SWOT factors generated by employees of C.A.F.S.

STRENGHTS S1:Wood certification S2: Counselling and education of members and public S3: Interdisciplinary approach, forest as a part of a farm S4: Cooperation in committees for evaluation of damages S5: Nongovernmental organization S6: Professional knowledge S7: Work for forest owners' interests **S**	**WEAKNESSES** S1: Problems with employment S2: Disregarding their comments on new laws S3: Allowing public forestry service to take initiative to associate private forest owners S4: Inferior position of forestry consultants towards agricultural consultants S5: Limited power and presence to influence management S6: Insufficient funds **W**
O **OPPORTUNITIES** S1: More activities for associating forest owners S2: Possibility to influence accepting new laws S3: Improving employment capacities and cooperation with associations S4: Professional support of the interest of local associations S5: Introducing good practice from abroad S6: Establishing a network with forest owners S7: Promoting the use of wood and adding added value to products made from wood	**T** **THREATS** S1: Decreasing numbers of farmers as forest owners S2: Forest owners are not connected S3: Inferior position towards other forest institutions S4: Insufficient presence in public S5: Nonexistent political will for enlargement

Step 2 of A'WOT analysis

Pairwise comparisons between SWOT factors are carried out within each SWOT group. Employees were asked to make their judgment about pairwise comparisons between SWOT factors, i.e., pairwise comparisons of seven factors of strengths, pairwise comparisons of six factors of weaknesses, pairwise comparisons of seven factors of opportunities and pairwise comparisons of five factors of threats. From individual pairwise comparisons we calculated the geometric mean for each SWOT factor. The results are shown in Table 2 and Table 3.

Table 2: Pairwise comparisons for SWOT groups:

STRENGHTS																					
	S1: S2	S1: S3	S1: S4	S1: S5	S1: S6	S1: S7	S2: S3	S2: S4	S2: S5	S2: S6	S2: S7	S3: S4	S3: S5	S3: S6	S3: S7	S4: S5	S4: S6	S4: S7	S5: S6	S5: S7	S6: S7
A1	1/6	1/6	4	1/3	1/3	1/3	1	3	4	4	3	3	1/2	1	3	2	1/2	1/4	1/3	1/3	1
A2	4	4	4	1/4	1/4	1/4	4	6	1/6	1/6	1/7	7	1/4	1/4	1/4	1/7	1/7	1/7	1/4	1/4	1/4
A3	1/5	5	1/7	7	1	1/7	7	1/5	7	3	1/2	1/2	1	1	1/5	5	5	1/3	1/3	1/5	1/5
A4	6	1/3	1/2	1/5	1/2	1/2	1/5	1/3	1/5	1/4	1/7	1	1	3	1/3	1/2	3	1/3	3	1	1/3
A5	1	1/9	1/9	1	1/6	1	1/7	1/6	9	1/8	1/2	1	9	1	1	9	1	1	1/9	1/9	1

WEAKNESSES															
	S1:S2	S1:S3	S1:S4	S1:S5	S1:S6	S2:S3	S2:S4	S2:S5	S2:S6	S3:S4	S3:S5	S3:S6	S4:S5	S4:S6	S5:S6
A1	5	7	2	2	1	2	3	2	1/3	2	1/3	1/3	1/3	1/5	1/3
A2	7	7	7	7	1/9	5	6	5	1/5	5	5	1/9	1/3	1/9	1/9
A3	1/3	9	7	7	1	6	7	3	1/3	1/3	1/3	1/5	1/2	1/5	1/5
A4	1/5	5	5	5	5	7	7	1	7	1/3	1/3	1	4	3	1
A5	9	9	9	9	6	9	1	1	9	1	1/9	1	1	1/9	1/9

OPPORTUNITIES																					
	S1: S2	S1: S3	S1: S4	S1: S5	S1: S6	S1: S7	S2: S3	S2: S4	S2: S5	S2: S6	S2: S7	S3: S4	S3: S5	S3: S6	S3: S7	S4: S5	S4: S6	S4: S7	S5: S6	S5: S7	S6: S7
A1	1/2	1/9	1	3	2	4	1/5	3	5	5	5	3	6	3	2	3	1	2	1	1	1
A2	7	1/7	1/7	7	1/7	1/3	1/6	1/6	1/5	1/6	1/6	6	6	6	6	6	6	1/5	1/6	1/7	7
A3	1/5	1/5	2	2	1/3	1/3	1/3	4	2	2	1/2	5	6	7	4	3	1/5	1/5	1/3	1/5	3
A4	1/4	1/3	1/3	3	3	3	1	3	3	2	2	2	2	3	1	2	3	3	1	1/2	1
A5	1	1/5	1/5	6	1	1	1/5	1/5	4	1/5	1/5	2	5	1	1	5	1	1	1/5	1/5	1

THREATS										
	S1:S2	S1:S3	S1:S4	S1:S5	S2:S3	S2:S4	S2:S5	S3:S4	S3:S5	S4:S5
A1	5	1/3	1/5	1/5	1/3	1/5	1/9	1/3	1/5	1/3
A2	1/7	1/3	1/5	1/9	1/5	1/5	1/5	1/5	1/5	6
A3	3	2	1/2	1/3	1	2	1/3	1/3	1/3	1
A4	1/6	1/5	1/5	1/5	1/3	1/5	1/3	1	1	1/2
A5	1	9	1/5	5	9	1	5	1/9	1/3	9

Table 3: Comparison matrices for SWOT groups

Strengths	S1	S2	S3	S4	S5	S6	S7
S1	1	0,956	0,658	0,662	0,651	0,370	0,359
S2	1,046	1	0,956	0,725	1,531	0,574	0,447
S3	1,519	1,046	1	1,332	1,024	0,944	0,549
S4	1,511	1,380	0,750	1	1,451	1,014	0,331
S5	1,537	0,653	0,977	0,689	1	0,392	0,284
S6	2,702	1,741	1,059	0,986	2,551	1	0,441
S7	2,787	2,237	1,821	3,022	3,519	2,268	1

Weaknesses	S1	S2	S3	S4	S5	S6
S1	1	1,838	7,236	5,356	5,356	1,337
S2	0,543	1	5,194	3,882	1,974	1,069
S3	0,138	0,192	1	0,964	0,459	0,394
S4	0,186	0,257	1,037	1	0,659	0,275
S5	0,186	0,506	2,174	1,515	1	0,231
S6	0,747	0,934	2,536	3,629	4,324	1

Opportunities	S1	S2	S3	S4	S5	S6	S7
S1	1	0,705	0,177	0,452	3,764	0,824	1,043
S2	1,417	1	0,268	1,037	1,888	0,922	0,698
S3	5,632	3,727	1	3,898	4,644	3,277	2,168
S4	2,208	0,964	0,256	1	3,519	1,292	0,751
S5	0,265	0,529	0,215	0,284	1	0,406	0,309
S6	1,2129	1,0845	0,3051	0,7740	2,4595	1	1,8384
S7	0,9582	1,4310	0,4611	1,3303	3,2271	0,5439	1

Threats	S1	S2	S3	S4	S5
S1	1	0,814	0,758	0,232	0,367
S2	1,229	1	0,660	0,407	0,397
S3	1,320	1,516	1	0,259	0,280
S4	4,317	2,460	3,866	1	1,465
S5	2,728	2,516	3,565	0,683	1

From data in Table 3 we calculated weights – utility vector (w) for individual SWOT factors. Pairwise comparison matrices of individual SWOT groups were squared to the satisfactory exponent and then the lines were summed up and the values normalized [13]. Thus, by the use of formula (2) we calculated weights for individual factors of SWOT groups (Tables 4, 5, 6 and 7), and by the use of formulas (3), (4) and (5) the inconsistency ratios.

Table 4: Weights for strengths factors

Wood certification	0,0814
Counselling and education of members and public	0,1089
Interdisciplinary approach, forest as a part of a farm	0,1329
Cooperation in committees for evaluation of damage	0,1256
Nongovernmental organization	0,0913
Professional knowledge	0,1683
Work for forest owners' interests	0,2917

Table 5: Weights for weakness factors

Problems with employment	0,3597
Disregarding their comments for new laws	0,2156
Allowing public forestry service to take initiative to associate private forest owners	0,0536
Inferior position of forestry consultants towards agricultural consultants	0,0577
Limited power and presence to influence management	0,0842
Insufficient funds	0,2292

Table 6: Weights for opportunity factors

More activities for associating forest owners	0,0961
Possibility to influence accepting new laws	0,1048
Improving employment capacities and cooperation with associations	0,3705
Professional support of the interest of local associations	0,1308
Introducing good practice from abroad	0,0456
Establishing a network with forest owners	0,1252
Promoting the use of wood and adding added value to products made from wood	0,1270

*Table 7***:** Weights for threats factors

Decreasing numbers of farmers as forest owners	0,0923
Forest owners are not connected	0,1133
Inferior position towards other forest institutions	0,1163
Insufficient presence in public	0,3818
Nonexistent political will for enlargement	0,2964

Step 3 of A'WOT analysis

The factor with the highest local priority is chosen from each group to represent the group. So strength is represented by factor "Work for forest owners interests", weakness by factor "Problems with employment ", opportunities by factor "Improving employment capacities and cooperation within associations" and threats by factor "Insufficient presence in public". These four factors were then compared by the employees within the same procedure as described in Step 2. We calculated weights for individual SWOT groups (Table 8).

Table 8: Weights for SWOT groups

	S	W	O	T		
S	1	1,600	0,215	3,129		0,2098
W	0,625	1	0,341	1,878	→	0,1552
O	4,644	2,930	1	2,980		0,5337
T	0,320	0,533	0,336	1		0,1014

CI=0,0813; CR=0,0903

Step 3 of A'WOT analysis
It is administered in the conclusions.

4 CONCLUSIONS

The analysis carried out according to A'WOT shows that it is necessary for the C.A.F.S. to take into consideration all four parts of SWOT analysis. The most important aspect is that the C.A.F.S. utilizes opportunity factors. This is also the reason why we decided to use an offensive approach to strategic planning. This approach represents a combination of strategic guidance: extension of personnel scope, enlargement of power of forestry consulters, as it is only with a good network that will cover the entire Slovenia that the C.A.F.S. could become a strong and equal partner to other institutions and forest owners. Only so could they implement additional measures in the field of counselling to private forest owners, education of forest owners, as well as offer technical help by organizing different co-operations among forest owners. C.A.F.S. must emphasize its strengths, being a nongovernmental organization with good knowledge, to which the Chamber of Agriculture and Forestry Act (OG RS, Nr. 41-2025/99) assigned the task to represent the interests of forestry and to counsel to individuals and legal persons that are involved with forestry. Active counselling and education of members is the key to a more effective management of private forests. Improved cooperation between the C.A.F.S., public forestry service and the Association of private forest owners on the state level and the societies on the local level should lead to a partnership between them. As a result of that the C.A.F.S. will have a more important role in associating private forest owners and the initiative of cooperation will not be limited to the public forestry service. With the establishment of such relationships the C.A.F.S. will have a greater role in influencing forest policy [4].

References

[1] Grošelj, P., Zadnik Stirn, L., 2009. Primerjava metod lastnih vektorjev, LLSM in DEA za računanje vektorja uteži v modelih AHP. *Uporab. inform. (Ljubl.)*, 2009, let. 17/2, pp. 79-87

[2] Jeromel, J., 2004. Zadružništvo kot možnost povezovanja lastnikov gozdov: Diplomsko delo, UL, BF, Ljubljana, 55 p.

[3] Medved, M., Pezdevšek Malovrh, Š., 2006. Associating of small-scale forest owners in Slovenia. In: Wall, Sarah (eds.). Small-scale forestry and rural development: the intersection of ecosystems, economics and society : proceedings of IUFRO 3.08 conference. Dublin; Galway: CONFORD, National Council for Forest Research and Development: Galway-Mayo Institute of technology, pp. 282-288.

[4] Medved, M., 2006. Vloga Zavoda za gozdove Slovenije pri povezovanju lastnikov na lokalnem nivoju, *Gozd. vestn.*, 2006, letn. 64, št. 10, pp. 462-475

[5] Pezdevšek Malovrh, Š., Krč, J., 2006. Evaluation of the influence of institutional subjects on private forest management in Slovenia. In: *Formec 2006: proceedings*. [S. l.: s. n., 2006], pp. 258-265.

[6] Resolution on national forest program, 2008, Association of forestry societies of Slovenia, Ministry of agriculture, forestry and food of the Republic of Slovenia, Ljubljana.

[7] Saaty, T.L., 1994. Fundamentals of decision making and priority theory. RWS Publication, Pittsburgh.

[8] Saaty, T.L., 2005. The analytic hierarchy and network processes for the measurement of intangible criteria and for decision-making. In: Figueira et al.(eds.), Multicriteria decision analysis, Springer Int. Series in Operations Research and Management Science, pp. 345-406.

[9] Pesonen, M., Ahola, J., Kurttila, M., Kajanus, M., Kangas, J., 2001. Applying A'WOT to forest industry investment strategies: case study of a finnish company in North Amerika, in Anaylitic hierarchy process in natural resources and environmental decision making, Schmoldt et al. (eds.), Kluwer Academic Publishers, Dordrecht, pp. 187-198.

[10] Annual report of Chamber of agriculture and forestry of Slovenia 2006, Ljubljana, 61 p.

[11] Program of Chamber of agriculture and forestry of Slovenia 2007, Ljubljana, 284 p.

[12] Špendl, R., Rajkovič, V., Bohanec, M., 1996: Primerjava kvalitativnih in kvantitativnih odločitvenih metod: DEX in AHP pri ocenjevanju projektov. Zbornik posvetovanja, Organizacija in management, XV. posvetovanje organizatorjev dela 1996, Univerza v Mariboru, Fakulteta za organizacijske vede, Kranj, pp.190-199.

[13] Taha, H.A., 1997. Operations Research. Prentice Hall, New Delhi.

[14] Winston, W.L., 1994. Operations research: Applications and algorithms. Duxbury Press, Bellmont, CA.

[15] Zadnik Stirn, L., 2006. Evaluation of environmental investment projects using a hybrid method (SWOT/AHP). In: V. Boljunčič (eds.), Proceedings KOI 2006, 11th International conference on operational research, Pula, Croatia, September 27-29, Croatian Operational Research Society: Faculty of Economics and Tourism, Pula, pp. 245-255.

[16] Zadnik Stirn, L., Pezdevšek Malovrh, Š., Mihelič, M., Krč, J., 2008. A multicriteria approach to private forests management regarding institutional factors. *Nauk. vìsn. NLTU Ukr.*, 2008, vypusk 18.8, pp. 175-188.

APPLICATION OF MULTI-ATTRIBUTE DECISION METHODS IN ROAD INFRASTRUCTURE MANAGEMENT

Aleksander Srdić and **Jana Šelih**
University of Ljubljana, Faculty of Civil and Geodetic Engineering
Jamova 2, SI-1000 Ljubljana, Slovenia
{asrdic, jselih}@fgg.uni-lj.si

Abstract: During its service period, the road infrastructure is subjected to deterioration due to various environmental and mechanical loads. As a consequence, the performance of the network is decreasing. Road managers have to ensure that the system provides the required service, therefore they have to take appropriate decisions regarding maintenance, repair, rehabilitation or demolition of various assets that are a part of the network system. The decisions should be based on data regarding the existing condition, risk of the use, life cycle costs, age and other factors. In practice, one can observe that due to the complexity of the problem, the optimal choice of planned interventions is often left to the road managers' subjective judgment.
The present research aims at development of a multiple criteria decision support system to a) determine the priority ranking of maintenance, repair and rehabilitation projects, and to b) to support the choice between rehabilitation and demolition and replacement option for a single asset. The proposed decision support system selects among all assets forming a part of the infrastructure system a set of assets to be repaired or rehabilitated, and takes into the account the budget constraint option. Whole life cycle costs of a single asset subjected to various scenarios are also considered as the basis for the decision between rehabilitation and demolition and repalcement. Results are presented for a case study of a group of 27 overpasses on a selected a motorway section. They show that the selection of criteria employed in the decision process is crucial in obtaining the targeted goals. The selected criteria should therefore reflect all aspect of sustainable development, namely, social, environmental and economic conditions related to the selected group of assets. The calculated life cycle costs of a single asset indicate that when the asset is a reasonably good shape, rehabilitation is strongly preferred to its demolition and replacement with a new one .

Keywords: decision support system, maintenance, rehabilitation, demolition**,** road infrastructure management, knapsack model

1 INTRODUCTION

During its lifetime, road infrastructure is exposed to significant mechanical and chemical loadings from the environment that initiate various deterioration processes in the materials and structures. As a consequence, the condition of these assets is decreasing.

In order to ensure uninterrupted transport service as well as minimum loss of value of these assets, maintenance as well as repair and rehabilitation (MR&R) actions have to be undertaken in a systematic manner throughout the planned service life of this particular infrastructure system [1-4]. Planning of these actions has to be based on rational grounds and adequate information. This first means that condition monitoring of the assets and its assessment has to be carried out on a regular basis. In addition to these data, the network managers must make decisions about maintenance and renewal alternatives based on relative risk of failure and the life cycle costs of the proposed interventions.

Since funding allocated for MR&R is always limited, it is necessary to prioritize and select the options that are best aligned with the asset managing organization's objectives. These objectives should address, in addition to cost efficiency, all three pillars of sustainable construction: environmental, social and economic one [5]. The corresponding criteria used in the decision process, however, are often uncertain, conflicting and sometimes subjective. They may include type of maintenance intervention, risk and reliability, overall network performance, life cycle costs, desired levels of service, budgetary concerns, construction costs, and social costs.

In addition, the decision makers have be aware that there are cases where demolition and replacement is preferred to the rehabilitation of a particular structure. The described situation occurs when the rehabilitation project costs are comparable to the cost of demolition and replacement of the structure; or when the execution of the rehabilitation project can not establish the required condition state (in terms of functional performance, load bearing capacity, safety, etc). In the case of road infrastructure, special attention should be paid to the road section/selected asset functional performance.

The above discussion shows that optimal selection of MR&R projects across a broad spectrum of assets presents a challenge to the decision maker. In practice, it is often carried out in a subjective manner. To overcome this current practice, a rational decision model has to be established.

The aim of the research presented in this paper is therefore development and verification of a multi criteria decision model that would help the decision maker in decisions regarding selection and priority of MR&R projects. As funding available for these projects is always limited, the model has to account for this limitation. In addition, the decision model is complemented by a separate part intended to support the decisions between MR&R and demolition and new construction project, carried out on the level of a single asset. The group of assets under consideration within this study is limited to overpasses bridging a selected motorway section.

2 MANAGEMENT OF ROAD INFRASTRUCTURE

Life cycle road asset management methodology that ensures adequate performance of the assets consists of several steps. The number of facilities/assets comprising the network is large, and they consist of road sections, bridges, tunnels, overpasses, etc. The first step is therefore creating an asset inventory system, where all assets (facilities) are identified and described by pre-defined indicators.

As the second step, field inspection, condition assessment and consequent rating of all assets has to be carried out. The rating is based on the asset's physical and functional condition as well as its structural and overall safety. Systematic condition rating has to be used to obtain objective evaluation of the assets under consideration. The final result of the rating process is a set of scores for each indicator and, if desired, an aggregated total score for each asset.

The rating phase in asset management is followed by the decision process where set of assets selected to be maintained, repaired or rehabilitated is determined. The decision process can be carried out in several ways and can be supported by various tools. Different criteria can be employed. After the MR&R projects are completed, the performance of the upgraded assets is increased. New performance level has to be recorded for each asset, and the inventory database has to be updated by the recorded data to complete the management cycle.

3 MATHEMATICAL FORMULATION OF MULTI-CRITERIA DECISION MODEL

Decision variables assigned to a given set of assets requiring maintenance, repair or rehabilitation actions are labelled as (a_1, a_2, ... a_n), where a_i equals 1 if the asset i is selected to undergo a MR&R action, and 0 when it is not selected. Utility score S_i is attained for the asset i when the action is selected to be carried out. For the selected group of assets, total utility score, S_{tot}, can be defined by using the expression

$$S_{tot} = \sum_{i=1}^{n} (a_i S_i) \quad (1)$$

All selected MR&R projects within a given time period should result in maximum total utility score, therefore the function

$$\sum_{i=1}^{n} (a_i S_i) \quad (2)$$

is the objective function of the problem to be solved. The problem solution is subject to budget constraint

$$\sum_{i=1}^{n} a_i C_i \leq C_{\max} \quad (3)$$

where C_i is the cost of the MR&R project carried out for the asset i, and C_{max} is the available overall budget for the given time period.

Depending on the criteria employed, the problem may be subject to several other constraints. The importance of different criteria used in the selection of actions can be captured by assigning criteria weights, w_j, to indicate their relative importance, where the sum of all criteria weights equals to 1. Assigning the importance to individual criteria has to be performed prior to selection process. For the purpose of this study, the relative weights were determined by the analytical hierarchy process (AHP) as proposed by Saaty [6].

Utility score assigned to the asset i, S_i, is determined with respect to m criteria. Project/asset i is assigned a utility value S_{ij} with respect to the individual criterion j used in the analysis. S_i is then determined as a weighted sum of all partial utilities S_{ij}.

$$S_i = \sum_{j=1}^{m} w_j S_{ij} \quad (4)$$

Total utility score for the set of selected projects/assets, S_{tot}, is the sum of utility values of all separate projects/assets selected (Eq.1). The objective of the decision process is to select a set of MR&R projects/assets that results in maximum total utility score (Eq.2) according to the criteria by taking into account the financial constraint (Eq.3) and compatibility constraints that depend upon the case being considered. Mathematically, the problem can be described as a knapsack problem [7] and can be solved numerically by using the SOLVER function of MS Excel software. The cost of the selected portfolio of MR&R projects, C_{tot}, is the sum of costs of selected projects.

3.1 Single facility decision support

Before the above described decision model can be implemented, the type of MR&R action has to be determined for a single asset. In most cases, the decision between rehabilitation ("R") and demolition and replacement ("D&R") has to be carried out for a single asset. This research work assumes that the decision can be based on whole life cycle cost analysis (of both alternatives).

The following costs were considered in the calculation of whole life cycle costs for the "R" option for the selected set of assets (a group of overpasses bridging the selected motorway section):

- Construction costs;

- Overpasses’ rehabilitation costs, that include costs of the closure of one side of the motorway and indirect costs;
- Overpasses’ maintenance costs over a period of 50 years after the rehabilitation is completed;
- Overpasses’ maintenance costs prior to rehabilitation/demolition project are neglected, as they are significantly smaller than the repair and rehabilitation costs.

For the “D&R” option, the following costs are taken into the account:

- Construction costs;
- Facility replacement costs, that include costs of the closure of one side of the motorway and indirect costs;
- Maintenance costs after the replacement (over a period of 50 years after the replacement is completed);
- Demolition costs of the existing facility.

It should be noted that execution of MR&R works on an overpass over the motorway section can only be performed if one side of the motorway is closed, and the vehicles travelling on the motorway in both ways have to share two lanes of the open side of the motorway. As a consequence, the average travel speed is reduced, and congestions may occur. Closure of one motorway side incurrs also costs associated with installation and surveillance of the signal signs.

Life cycle costs in case of rehabilitation (“R”), *LCCR*, and demolition and replacement (“D&R”), *LCCDR* , are calculated as

$$LCCR = C + (C_r + C_{closure} + C_{indir}) + C_{maint} \tag{5}$$

and

$$LCCDR = C + C_{demol} + (C_r + C_{closure} + C_{indir}) + C_{maint} \tag{6}$$

C is the cost of the overpass construction, C_r is the cost of rehabilitation/replacement works, $C_{closure}$ is the cost of the closure of one side of the motorway, and C_{indir} indirect costs incurred by the passengers due to increased travelling time on the motorway section under consideration. C_{maint} is the cost of maintenance after the rehabilitation/replacement., andC_{demol} is the cost of demolition works.

For the purpose of the study, the first estimate for C_{indir} was calculated as the monetary value of the additional travel time due to speed reduction from the initial (v_0) to the reduced travel speed (v_1) by using the expression

$$C_{indir} = (d_{cl} / v_1 - d_{cl} / v_0) \cdot DTV \cdot c_{aver} \cdot t_R \tag{7}$$

where d_{cl} is the length of the closed lanes (above which the overpass where the MR&R works are taking place is located). The motorway side closed for the traffic can stretch only between two consequent openings in the median barrier, which means that the location and number of existing openings may significantly affect the comfort of passengers and their lost time due to the reduced travel speed on the particular motorway section. DTV is the average daily traffic volume for the section under consideration, t_R is the average expected duration of rehabilitation works and c_{aver} is the Slovenian average net hourly wage. The decision is based on the calculated difference between rehabilitation and demolition costs, LCC_{diff}. If this value is negative, rehabilitation is preferred to the demolition and replacement scenario.

$$LCC_{diff} = LCCR - LCCDR \tag{8}$$

4 CASE STUDY

The presented model was applied to a case study of a group of 27 overpasses that enable crossing of local roads with the selected motorway section. When choosing the set of assets to be rehabilitated, the following criteria are considered:

- Asset rating (R). Rating or scores assigned to an individual asset as a whole stem for the last regular condition assessment carried out in 2004.
- Asset age (A). Implicitly, this criterion includes functional performance of the facility under consideration.
- Overpass grouping (G). When a MR&R project is carried out for a particular overpass, one side on the motorway underneath has to be closed for the traffic. This closure results in additional costs for the motorway management. Further, it can only be carried out between two consequent openings in the median barrier. If there are several overpasses between two consequent openings in the median barrier, it makes sense to carry out their MR&R actions simultaneously in order to reduce the barrier (i.e. economic and environmental) costs.
- Indirect costs (*IC*). Indirect costs of a MR&R project carried out on a overpass are incurred due to motorway traffic travel speed decrease as determined by the Eq.(7).
- MR&R project cost (*PC*). This criterion is composed of rehabilitation work costs (*RWC*) and the costs of one side closure (*LCC*),

$$PC = RWC + LCC \qquad (9)$$

The normalized utility values related to the above described criteria for the selected group of 27 overpasses (f_{ij}) are presented in Table 1. The expression

$$S_{ij} = 1 + 9 \cdot (f_{i,j} - f_{min,j}) \,/\, (f_{max,j} - f_{min,j}) \qquad (10)$$

is used to determine the normalized utility values for criteria R, A and G ($j = 1, 2, 3$), and

$$S_{ij} = 10 - 9 \cdot (f_{i,j} - f_{min,j}) \,/\, (f_{max,j} - f_{min,j}) \qquad (11)$$

for criteria IC and PC ($j=4, 5$) The weights assigned to the employed criteria are determined by AHP [6]. In this study, the values for R, A, G, IC and PC are 0,654; 0,061; 0,112; 0,061; and 0,112.

By taking into the account the budget constraint in a given time period, the decision model defined by the Eqs. (1-4) results in a set of MR&R projects/assets that yields the maximum total utility value for the following constraints: 8, and 2 million EUR for the total budget for all selected MR&R projects (Σ *PC*), and allowable indirect costs (Σ *IC*), respectively.

The resulting set is presented in Table 2 for two options: by including, and excluding the grouping of overpasses (G) due to the joint motorway side closure. The results show that taking into the account the grouping of overpasses, when possible, results in a larger number of selected assets (to be rehabilitated) and associated larger total utility score for the same budget constraints.

Table 1: Normalized utility values for the selected set of overpasses for criteria considered (shaded fields indicate the overpasses that can be grouped by joint motorway side closure).

Asset code	Utility values (S_{ij})				
			j		
i	1	2	3	4	5
	R	A	G	IC	PC
1	1,00	10,00	1,00	10,00	9,38
2	4,09	10,00	10,00	2,05	9,30
3	2,52	1,00	10,00	2,05	8,95
4	4,10	5,50	10,00	2,74	8,73
5	4,68	5,50	1,00	7,60	9,50
6	2,63	1,00	1,00	2,49	9,38
7	1,76	1,00	5,50	2,03	9,88
8	6,42	1,00	5,50	2,03	9,52
9	5,79	1,00	1,00	9,86	9,52
10	6,42	1,00	1,00	9,39	9,88
11	4,41	1,00	1,00	9,41	8,10
12	7,18	1,00	5,50	9,35	10,00
13	1,31	1,00	5,50	9,35	9,38
14	5,97	1,00	5,50	9,49	8,73
15	8,01	1,00	5,50	9,49	9,38
16	2,31	1,00	1,00	6,63	9,88
17	4,85	1,00	10,00	7,85	8,95
18	4,65	1,00	10,00	5,38	8,78
19	1,64	1,00	10,00	5,38	9,60
20	2,75	1,00	1,00	7,14	9,60
21	3,01	1,00	1,00	6,92	9,30
22	4,15	1,00	5,50	6,08	9,17
23	4,62	1,00	5,50	6,08	8,83
24	5,94	1,00	5,50	3,82	6,10
25	7,93	1,00	5,50	3,82	1,00
26	10,00	1,00	1,00	4,35	8,80
27	4,78	1,00	1,00	1,00	8,80

4.1 Single asset decision model results

Average prices for various foreseen actions are taken for all overpasses under consideration. The initial cost of construction is taken as 1120,8 EUR/ m^2, demolition cost 208,3 EUR/ m^2, cost of replacement 1620,4 EUR/ m^2, and maintenance costs 22,4 EUR/m^2/year, calculated to the m^2 area of the overpass. The assumed average time of rehabilitation or demolition and replacement works is assumed to be 120, and 180 days, respectively. The estimated cost of closing one side of the motorway is 1,618 EUR/day/m^2.

Indirect costs of a rehabilitation / demolition and replacement project carried out on an overpass, C_{indir} , are incurred due to motorway traffic travel speed decrease. For the purpose of the study, they are determined by the Eq. (7).

The differences in life cycle costs for 5 representative overpasses are presented in Table 3. Negative value indicates that rehabilitation is preferred to overpass demolition and replacement. The results show that for the case under consideration, overpass rehabilitation is always the preferred option. This result is partly a consequence of the fact that the selected group of overpasses was built within the same period, and that their condition is relatively good.

Table 2: Results (set of selected facilities) of the 5-criteria decision model.

Scenario	**A**	**B**
G included	**no**	**yes**
Facility code, i	**Decision vector, a_i**	
1	0	0
2	1	1
3	0	0
4	1	1
5	1	1
6	0	0
7	0	0
8	1	1
9	1	1
10	1	1
11	0	0
12	1	1
13	0	0
14	1	1
15	1	1
16	0	0
17	1	1
18	1	1
19	**0**	**1**
20	0	0
21	**0**	**1**
22	1	1
23	1	1
24	0	0
25	0	0
26	1	1
27	0	0
∑ (selected projects)	14	16
S_{tot}	83,5	90,7

Table 3: Results of the life cycle cost calculation for selected overpasses; negative difference implies that rehabilitation is the preferred option.

	DTV	Overpass layout area	Lane closure length	LCC_{diff}
		(m^2)	(km)	(EUR)
1	30800	460,5	1,28	-444771
2	19800	390,0	2,14	-339473
3	21000	552,6	1,99	-520006
4	26000	546,5	2,37	-565392
5	30700	1645,6	2,25	-1407410

5 CONCLUSIONS

After being built and opened for public use, road infrastructure is exposed to various deterioration processes throughout its service life. Performance of the various assets decreases at different rates and due to various causes. As a consequence, maintenance, repair and rehabilitation projects have to be carried out on several assets at various points of time.

In order to maintain the required performance of the system and consequently ensure its efficient service while not exceeding the allocated budget, management of this large asset value needs to be efficient, and it should provide the largest benefit to the users. One of the main elements of an adequate database management system that supports motorway infrastructure management is a decision support tool that facilitates the process of selection of assets where a MR&R project will be carried out. The multi criteria decision model proposed in this paper provides the solution with the largest benefit expressed by the highest cumulative utility score. The solution is expressed as a set of assets selected to undergo a MR&R project.

The presented case study deals with a group of 27 overpasses over a motorway section. 5 different criteria are taken into the account: condition rating, age, possibility of facility grouping, indirect costs and MR&R project costs. To eliminate the subjectivity of the decision makers, their relative importance is determined by the analytical hierarchy process. The results obtained show that taking into the account the grouping of underpasses selected to undego the rehabilitation due to joint closure of one side of the motorway can influence the selected set of assets. Based on the obtained results, it can be concluded that the proposed decision tool has the potential to facilitate work and rationalize the decisions of the motorway network manager.

The second part of the proposed decision model provides support to the decision maker selecting between rehabilitation and demolition (and replacement) of a single asset. The model is based on whole life cycle costs of the asset in both types of envisaged scenarios. The results show that the condition of the asset is the critical element in making the decision between rehabilitation and demolition.

References

[1] Frangopol, D. M.; Liu, M. 2007. Maintenance and management of civil infrastructure based on condition, safety, optimization and life-cycle cost, *Structure and Infrastructure Engineering,* 3 (1), pp. 29-41.

[2] Hegazy, T. 2006. Computerized system for efficient delivery of infrastructure maintenance/repair programs. *ASCE J. of Construction Engineering and Management*, 132(1), pp. 26-34.

[3] Miyamoto, A.; Kawamura, K.; Nakamura, H. 2000. Bridge management system and maintenance optimization for existing bridges, *Computer Aided Civil and Infrastructure Engineering,* 15, pp.45-55.

[4] Hallberg, D. and Racutanu, G. 2007. Development of the Swedish bridge management system by introducing a LMS concept, *Materials and Structures*, 40(6), pp.627-639.

[5] *Agenda 21 on sustainable construction,* 1999. CIB Report Publication No.237, Conseil International de Bâtiments, Rotterdam, The Netherlands.

[6] Saaty, T. L. 1990. *Multicriteria Decision Making. Vol 1, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*, 2nd edition, RWS Publications, 437 p.

[7] Čižman Anton. 2004. *Operacijske raziskave. Teorija in uporaba v organizaciji*. Kranj. Univerza v Mariboru. Fakulteta za organizacijske vede

LINEAR AND WEIGHTED GOAL PROGRAMMING IN LIVESTOCK PRODUCTION: AN EXAMPLE OF ORGANIC PIG FARMING

Jaka Žgajnar[1] and **Stane Kavčič**[2]

[1]University of Ljubljana, Biotechnical Faculty, Department of Animal Science, Groblje 3, SI-1230 Domžale, Slovenia. e-mail: jaka.zgajnar@bfro.uni-lj.si

[2]The same address as [1] e-mail: stane.kavcic@bfro.uni-lj.si

Abstract: The paper presents an optimization approach that merges linear programming and weighted goal programming. It is applied in a spreadsheet tool for organic pig ration formulation. The formulation process is based on a three-stage procedure: In the first step, common least-cost rations with different energy contents are formulated. In the second stage, a sub-model based on weighted goal programming supported by a system of penalty functions formulate nutritionally balanced and economically acceptable rations that also fulfil conditions demanded by organic farming. In the third phase, the ration formulation procedure stops when the most efficient ration is selected. The results obtained for a hypothetical farm confirmed the applicability of the implemented approach.

Keywords: linear programming, weighted goal programming, ration costs, organic farming, pork production

1 INTRODUCTION

Nowadays, livestock production demands precise management that leads to economically efficient and publicly acceptable production. This is possible through a very diverse range of measures. Due to the high share of ration cost within total production costs, ration formulation (optimization) is becoming a crucial task, especially in organic farming, which is globally characterized by higher production costs and affected by strict organic production policy constraints. However, pork production is also a livestock activity where unbalanced rations have significant negative impacts on the environment, especially if economics are taken as the most important criterion. Therefore, it is necessary to treat these kinds of problems as multiple criteria decision problems. Specifically, organic fattening that is confronted with the lack of availability of pure amino acids that results in a more unbalanced protein composition, increased feed cost, and, contrary to organic philosophy, an increased load of excessive nitrogen from manure on the environment [3]. In order to help breeders to deal with these challenges, numerous tools based on mathematical programming (MP) paradigms have been developed.

The first approach of this kind was conducted by [11], who applied the linear programming (LP) paradigm in order to formulate rations on a least-cost basis. This method was very popular in the past, especially after the rapid developments in personal computers. In the 1960s, it became a classical approach for the formulation of animal diets as well as feed mixes [2]. More recently, [5] stressed that the daily routine of ration formulation is one of the fields in which LP is most widely used.

Common to all LP problems is the concept of constraint optimization, which means that one tries to find the optimum of a single objective function. However, the exclusive reliance on just one objective (cost function) as the most important decision criteria is one of the reasons why the LP paradigm may be a deficient method in the process of ration formulation [8] and [9]. [7] stressed that, in practice, decision makers never formulate rations exclusively on the basis of a single objective, but rather on the basis of several different objectives, of which economic issues are only one of many concerns.

In common LP models for pig ration formulation, animal amino acid requirements are usually expressed in terms of minimal concentrations. Such models do not consider the total exceeded amount of protein or its quality, as long as the minimal amounts of essential amino acids are satisfied [1]. Furthermore, the same authors stressed that 'economically optimal' diets are often too rich in protein, which directly burdens the environment and does not improve animal growth. This problem could partly be solved by adding additional upper or lower constraints. However, this addition might rapidly lead into an over-constraint model that has no feasible solution. This problem is also related to the next LP drawback: the rigidity of constraints (right hand side (RHS)) [8]. This means that no constraint (e.g., the given nutrition requirements) violation is allowed at all. However, relatively small deviations in the RHS would not seriously affect animal welfare, but would result in a feasible solution [7].

Numerous methodological developments in the field of MP have eased these problems of the LP paradigm [4]. For instance, in the field of animal nutrition, [8] introduced goal programming (GP) and its improvement with a system of penalty function (PF), as well as multi-objective programming (MOP) as ways to incorporate more than one objective function. Similarly, [7] applied interactive methodologies where the optimal ration is achieved through 'computer dialog', and [5] addressed a multi-criteria fractional model.

The purpose of this paper is to present a spreadsheet tool for organic pig ration formulation, designed as a three-phase optimization approach that merges two normative MP techniques. The first part of the paper provides a brief overview of weighted goal programming (WGP) and the penalty function as the main method. This is followed by a short description of the optimization tool. Finally, the characteristics of the analysed case are presented, followed by the results and discussion.

2 MATERIALS AND METHODS

2.1 Weighted goal programming supported by a system of penalty functions

Based on the approaches reported in the literature and taking into account the primary aim of the tool presented in this paper, we decided to apply the WGP approach. In the context of ration formulation, the approach was introduced by [8].

WGP formulation is expressed as a mathematical model with a single objective (achievement) function (the weighted sum of the deviations variables). The optimal compromise solution is found through the philosophy of 'distance measure' that measures the discrepancy between the desired goal and the performance level of a goal. To consider all goals simultaneously, normalization techniques should be applied [10].

[8] introduced the PF paradigm into the WGP to keep deviations within desired limits and to distinguish between different levels of deviations. This system is coupled with the achievement function (WGP) through penalty coefficients and with additional constraints defining deviation intervals. Such an approach enables one to define allowable positive and negative deviation intervals separately for each goal. Depending on the goal's characteristics (nature and importance of 100% matching), these intervals might be different. Sensitivity is dependent on the number and size of defined intervals and the penalty scale utilized (s_i; for $i = 1$ to n).

2.2 Tool for three-phase ration formulation

The presented optimization tool for organic pig ration formulation was developed in Microsoft Excel as an add-in application. This tool is capable of formulating least-cost, nutritionally balanced and environmentally acceptable rations for 'organically' growing pigs in different production periods. In addition, it provides information about which feed mix provides the optimal energy content.

The tool is organized as a three-phase approach (Figure 1) that merges two sub-models based on MP techniques. The first sub-model is an example of a common least-cost ration formulation, based on the LP paradigm. The purpose of including this sub-model into the tool is to obtain an approximate estimate of expected ration cost. In this manner, the tool calculates the target economic goal, which is one of the goals in the second sub-model. Therefore, the first sub-model is, from the perspective of constraints, as simple as possible and is intended to exclusively measure the 'rough' cost estimation. Through cost function, it is linked to the second sub-model. The latter is based on WGP and is supported by a system of a six-sided PF. In this approach, the desired nutrition levels and ration costs are modelled as goals instead of constraints. Moreover, in the second sub-model, additional constraints are added that have an indirect influence on the environment. Consequentially, the model is much more complex and ultimately yields a better solution.

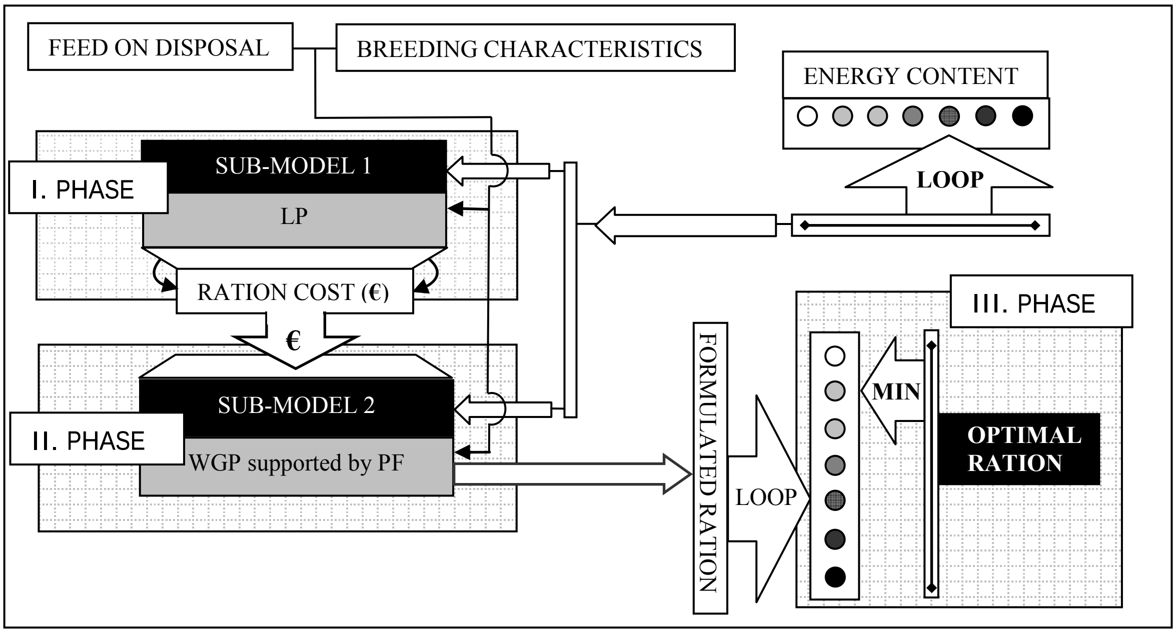


Figure 1: The scheme of the tool for three-phase organic pig ration formulation

Due to the importance of the feed mix's energy concentration and its influence on the ration structure and cost, the tool also includes a third phase (Figure 1). In this phase, a macro loop is added that runs the first and the second sub-models for n-times, and consequently it yields n–formulated rations. The number of iterations in the third phase depends on the starting/ending energy content of the feed mix and on the energy rise in each iteration step (e.g., 0.1 MJ/kg). From the n-obtained solutions, the tool selects the cheapest option and marks it as the 'optimal' feed mix structure for this given example.

2.3 Mathematical formulation of the first and the second sub-model

The first sub-model (LP) is formulated as shown in equation (1), equation (4a) and equation (7). It mostly relies on the economic (ration cost) function, C, and satisfies only the most important nutrition requirements coefficients, b_i, also known as RHS. In the first optimization phase, the desired element is the ration at the lowest possible cost (C).

$$\min C = \sum_{j=1}^{n} c_j * X_j \qquad \text{such that} \qquad (1)$$

$$\min Z = s_1 \sum_{i=1}^{k} w_i \frac{d_{i1}^- + d_{i1}^+}{g_i} + s_2 \sum_{i=1}^{k} w_i \frac{d_{i2}^- + d_{i2}^+}{g_i} \quad \text{such that} \tag{2}$$

$$\sum_{j=1}^{n} a_{ij} X_j + d_{i1}^- + d_{i2}^- - d_{i1}^+ - d_{i2}^+ = g_i \quad \text{for all } i = 1 \text{ to r and } g_i \neq 0 \tag{3}$$

$$\sum_{j=1}^{n} a_{ij} X_j \leq b_i \quad \text{for all } i=1 \text{ to m} \tag{4a}$$

$$\sum_{j=1}^{n} \left(a_{ij} - (R^{\%}) a_{kj}\right) X_j \leq 0 \quad \text{for all } i=1 \text{ to m and } i \neq k \tag{4b}$$

$$d_{i1}^- \leq g_i - p_{i1}^{\min} g_i \quad \text{for all } i=1 \text{ to r} \tag{5a}$$

$$d_{i1}^- + d_{i2}^- \leq g_i - p_{i2}^{\min} g_i \quad \text{for all } i=1 \text{ to r} \tag{5b}$$

$$d_{i1}^+ \leq p_{i1}^{\max} g_i - g_i \quad \text{for all } i=1 \text{ to r} \tag{6a}$$

$$d_{i1}^+ + d_{i2}^+ \leq p_{i2}^{\max} g_i - g_i \quad \text{for all } i=1 \text{ to r} \tag{6b}$$

$$d_{i1}^+, d_{i1}^-, d_{i2}^+, d_{i2}^-, X_j \geq 0 \tag{7}$$

The second sub-model (WGP with PF) is formulated as shown in equation (2) to equation (7). The achievement function, Z, expressed in equation (2), is defined as the weighted sum of the undesired deviation variables (d_{i1}^+, d_{i1}^-, d_{i2}^+, d_{i2}^-) from the observed goals (g_i), multiplied by the belonging penalty coefficients (s_1 and s_2) that measure the slope of the penalty function. The obtained sum-product is the subject of the minimization in equation (2). The relative importance of each goal is represented by the weights (w_i) associated with the corresponding positive or negative deviations. Penalty intervals (p_{i1}^{min}, p_{i1}^{max}, p_{i2}^{min}, p_{i2}^{max}) are in place to prevent uncontrolled deviations (equation (5a) to equation (6b)) within each goal. Because of the normalization process, only goals that have non-zero target values (equation (3)) could be relaxed with positive and negative deviations. The obtained target value, C, in the first sub-model enters into the second sub-model (WGP with PF) through the 'cost goal' ($g_i = C$). This is also the only case where negative deviation is neither penalized nor restricted to intervals. All other constraints that have no defined target values or no priority attributes are considered in equation (4a). All upper bounds for ratios (R%) are transformed into linear equations with equation (4b), and the same holds for lower bounds, which should be multiplied by –1. The non-negativity condition for both models is considered in equation (7).

2.4 Analysed example

The tool has been applied for hypothetical organic pork production, with an average genotype for less intensive fattening. In this paper, we present only the fattening period between 50 kg and 100 kg with an average daily gain of 700 g. We considered that the tool should formulate the complete ration/feed mix in relation to the nutritional requirements. It is presumed that most of the fodder is produced at the farm under organic conditions and is evaluated with the full cost approach. The rest of the feed (less than 20%) that cannot be produced at the farm is accounted for at market price. However, no synthetic substances (e.g., amino acid supplements) could be added, since they are banned by law.

The nutritional requirements (metabolizable energy (35.2 MJ/day), crude protein (399 g/day), amino acids (Lys: 19.7 g/day; Met+Cys: 11.3 g/day; Thr: 13 g/day; Trp: 3.6 g/day) and minerals (Ca: 12.88 g/day; P: 11.59 g/day; $P_{available}$: 4.89 g/day; Na: 2.58 g/day)) are taken from [3]. In order to prevent an unrealistic solution with too much of one feed in the diet, we considered recommendations for maximal feed inclusion [3] and [6], namely, through additional upper-bound constraints (Table 1). In the process of ration formulation, the tool could choose between 12 different feeds (Table 1) that could be produced at the farm (except for dehydrated alfalfa, brewer's dried yeast and potato protein concentrate that could be purchased at market price), and four mineral components (limestone, salt, monocalcium phosphate and dicalcium phosphate) that could be purchased at market price.

Table 1: Prices and nutritive values of available feed and their suggested maximal share of the ration

Feed on disposal	Price* (cents/kg)	ME MJ/kg	DM	CP	Lys	Met+Cys	Thr	Trp	Max** %
			g/kg						
Maize	18	14.1	880	85	2.5	3.5	3.0	0.8	0.6
Wheat	21	13.8	880	120	3.4	4.5	3.5	1.5	0.7
Barley	21	12.6	880	106	3.8	3.7	3.7	1.4	-
Oats	26	11.2	880	108	4.3	4.1	3.7	1.4	0.25
Wheat flour	17	12.5	880	167	7.3	5.6	6.5	2.0	0.15
Wheat bran	14	8.3	880	141	6.2	5.0	5.5	2.5	0.25
Alfalfa, dehydrated	33	6.1	910	180	8.7	4.5	7.8	2.9	-
Yeast, brewer's dried	71	13.2	900	452	32.1	11.7	21.8	5.1	0.05
Potato protein concentrate	132	15.7	930	780	56.9	20.1	45.3	10.6	0.15
Lupin seed meal	58	14.1	890	349	15.4	7.8	12.0	2.6	0.15
Faba beans	42	12.7	870	254	16.2	5.2	8.9	2.2	0.2
Pea-field	38	13.4	890	228	15.0	5.2	7.8	1.9	0.3

**Prices are estimated with model calculations—own source*
*** Suggested maximum inclusion of feedstuffs in pig diets*

The tool offers the option to switch between goals and constraints, depending on the needs and preferences of the decision maker. In the analysed case, we chose 10 goals (Table 2) that were to be met as accurately as possible.

Table 2: Importance of goals with corresponding penalty function intervals

Goal	Unit (day^{-1})	Weight (w_i)	Penalty function intervals p_{i1}^+	p_{i1}^-	p_{i2}^+	p_{i2}^-	Together p_i^+	p_i^-
			%				%	
ME	MJ	75	1	0	2	0	3	0
CP	g	60	1	0	2	0	3	0
Lys	g	80	5	1	5	3	10	4
Met + Cys	g	60	5	1	5	3	10	4
Thr and Trp	g	60	5	1	10	3	15	4
$P_{available}$	g	40	3	1	5	3	8	4
Ca and Na	g	30	3	1	5	3	8	4
Cost	cent	90	10	∞	20	∞	30	∞

p_{i1}^+, p_{i1}^-, p_{i2}^+, p_{i2}^-*: penalty intervals at the first and the second stage*

The importance of each goal is defined by belonging weights (wi) ranging from zero to 100. Relatively high values are set for amino acids, since the reduction of an unbalanced protein fraction by increased protein quality (i.e., fulfilling the amino acid ratios in relation to the energy) reduces nitrogen excretion and pollution. For each goal, deviation intervals are defined separately (Table 2) and are measured in percentage deviation from the desired level.

The cost goal is the only one that is not penalized for negative deviation; at the same time, the negative interval is unlimited.

3 RESULTS AND DISCUSSION

The main objective of the tool presented in this paper is to assist organic producers in formulating diets that are balanced and, at the same time, as cheap as possible. With a simple example, we present how the tool could be applied and the possible benefits of the utilized methods. We have presumed that the decision maker prepares a feed mix for growing pigs. For organic producers, this task is subject to numerous limitations and very complex constraints.

The range of the ration's energy content was set between 12.3 MJ and 13.7 MJ of metabolizable energy (ME) per kilogram of feed. The changing compositions of formulated rations are presented in Figure 1. In the second part of this paper, we compared ration cost obtained by the presented tool and a common least-cost approach.

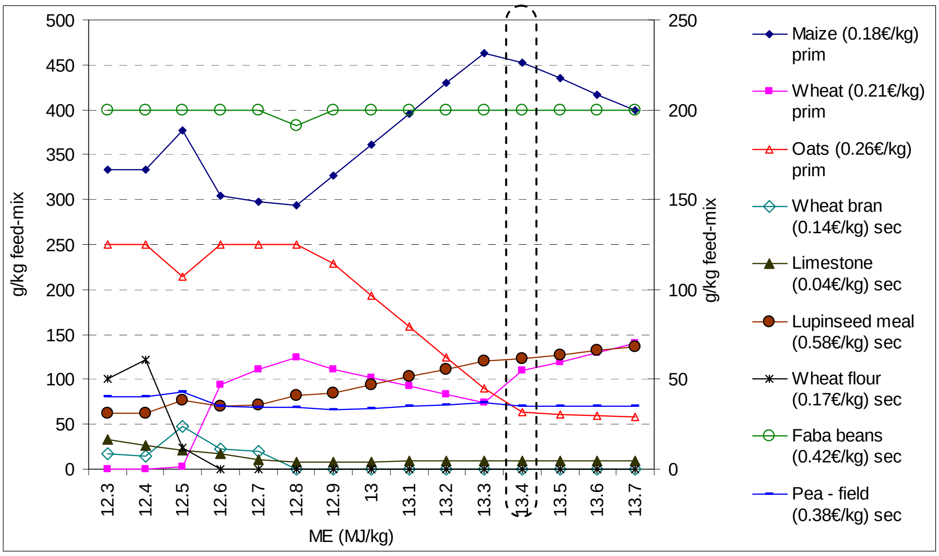


Figure 1: Formulated feed mix for organic pork production (prim = primary axis; sec = secondary axis)

One of the factors that define how much a pig is going to eat is the energy content of the feed mix. If the feed mix has a higher energy concentration, an animal will eat less, and vice versa [3]. Figure 1 presents formulated rations for the analysed fattening period. It is obvious that the energy content of the ration strongly influences the selection of the feed. With increasing energy content, the quantity of maize increases and the quantity of oats decreases. From Figure 1, it is apparent that, in spite of their high cost, faba beans enter into the solution due to their favourable amino acids structure. The same holds for peas; both are important substitutes for banned synthetic amino acid supplements.

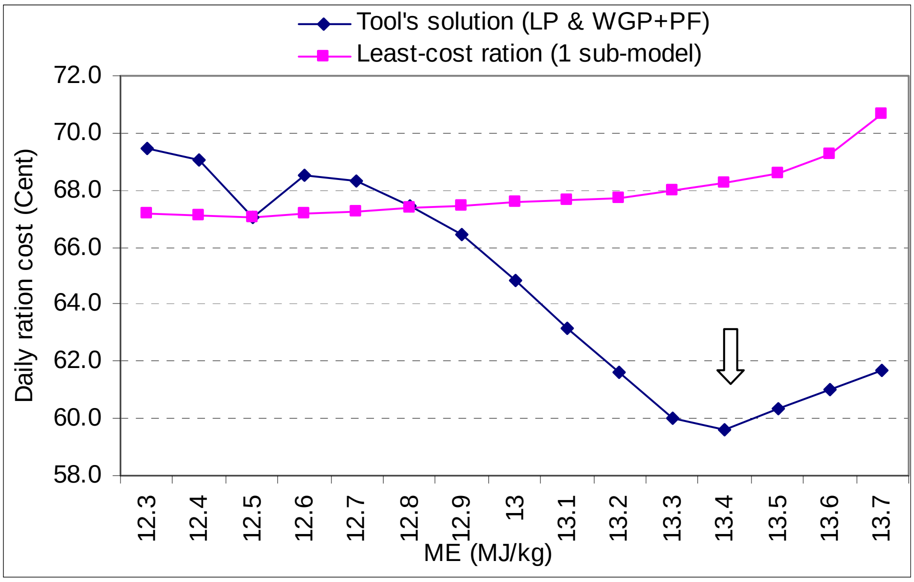


Figure 2: Daily ration costs dependent on the feed mix energy content; cost comparison of the least-cost ration and the tool solution

The difference in daily ration costs for different energy concentrations is obvious. The cost ranges from 59.61 cents up to 69.42 cents per day per pig. In any case, finding the 'optimal' energy concentration of feed mix in the daily management of organic pork production should be considered important because a feed mix with a lower energy content is harder to formulate, especially a more balanced one, which highly increases costs. Consequently, the minimal cost is achieved at a relatively high energy concentration (13.4 MJ/kg) of feed mix (Figure 2), which is unusual in organic practises that are generally less intensive. One could have legitimate doubts about the discrepancy between these results and actual practice, mainly because of the poor quality of organically produced cereals, in the sense of high nutritive value variability. However, the obtained results (Figure 2) confirm the benefits of the applied approach. Specifically, the significant discrepancy between the ration cost of the least-cost ration and the one obtained with the multiple criteria decision paradigm applied in the presented tool shows it is possible to achieve a more balanced ration (in the sense of total deviation from the target value) with a simultaneous reduction in daily ration cost of almost 13 %. One would expect the opposite situation that a more balanced ration results in increased costs. In the analysed example, cost reduction mainly occurs due to allowed negative deviations and controlled [constrained] positive deviations as result of weights and intervals of penalty functions.

4 CONCLUSIONS

The results of this study show that the three-phase optimization approach supported by mathematical programming (LP and WGP with PF) can be efficiently applied to diet formulation for organic pork production. The utilization of a multiple criteria optimization paradigm improves the quality of the obtained solution. The tool enables the formulation of efficient diets, since it supports the farmer in finding the optimal ration under various economic circumstances. With the application of this tool, problems such as unbalanced protein composition, increased feed cost, increased burdening of the environment and so forth might be mitigated. In this way, the discrepancy between the aims of organic farming and its practice may be reduced.

References

[1] Bailleul P.J., Rivest J., Dubeau F., Pomar C. 2001. Reducing nitrogen excretion in pigs by modifying the traditional least-cost formulation algorithm. Livest Prod Sci.72,199–211

[2] Black J.R. & Hlubik J. 1980. Symposium: Computer programs for dairy cattle feeding and management–past, present, and future. JDS. 63, 1366–1378.

[3] Blair, R. 2007. Nutrition and Feeding of Organic pigs. CABI. Wallingford. 322 pp.

[4] Buysse J., Huylenbroeck G.V., Lauwers L. 2007. Normative, positive and econometric mathematical programming as tools for incorporation of multifunctionality in agricultural policy modelling. Agriculture, Ecosystems & Environment, 120, 70–81.

[5] Castrodeza C., Lara P., Pena T. 2005. Multicriteria fractional model for feed formulation: economic, nutritional and environmental criteria. Agricultural Systems. 86, 76–96.

[6] Futtermittelspecifische Restriktionen. Rinder, Schafe, Ziegen, Pferde, Kaninchen, Schweine, Geflügel. 2006, 40 pp. (in German)

[7] Lara, P. & Romero, C. 1994. Relaxation of Nutrient Requirements on Livestock Rations through Interactive Multigoal Programming. Agricultural Systems 45, 443–453.

[8] Rehman, T. & Romero, C. 1984. Multiple-criteria decision-making techniques and their role in livestock ration formulation. Agricultural Systems 15, 23–49.

[9] Rehman, T. & Romero, C. 1987. Goal Programming with penalty functions and livestock ration formulation. Agricultural Systems 23, 117–132.

[10] Tamiz, M., Jones, D., Romero, C. 1998. Goal programming for decision making: An overview of the current state-of-the-art. EJOR. 111, 569–581.

[11] Waugh, F.V. 1951. The minimum–cost dairy feed. Journal of Farm Economics 33, 299-310.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section VIII:

OR Perspectives

RISK MODELING AND SIMULATIONS

Borut Jereb
University of Maribor, Faculty of Logistics
Mariborska 7, SI-3000 Celje, Slovenia
borut.jereb@fl.uni-mb.si

Abstract: As risk management is becoming a central challenge of our time, modeling and simulation of risks is in turn becoming a basic tool being employed. This paper describes an approach to risk modeling at a level that is sufficiently general to allow direct application in simulation tools.
The author's assumption is that risk is an attribute of human beings and not things or concepts. The model is based on the assumption that system processes (usually the model of business processes) and the input and output need to be divided into segments; similarly we need to make a segmentation for stakeholders whose risk we want to model and simulate.
Parameters can be used to define individual processes. Processes include functions that calculate new values of parameters and output on the basis of given input. Based on given tolerance levels for risks, impacts, and process parameters, the model determines whether these levels are acceptable. The model assumes that parameters and functions are non-deterministic, i.e. parameters and functions may change in time.
Such approach requires significantly more complex modeling and simulation. To what extent such modeling improves the quality of results, remains unanswered.

Keywords: Risk, Modeling, Simulation, Business Process

1 INTRODUCTION

Risks are an integral part of our lives and it appears that people have never devoted as much attention to the challenges of risks as we do today. Risks are addressed by numerous articles, comments, and conversations. Perhaps expectedly, there are virtually countless conceptions and definitions of the term "risk". Even if a particular community agrees upon a single definition of risk, it is still anything but certain that such community will reach uniform opinions or answers to the questions such as [2], [6], [7], [9], [10]: How to perceive risks? How to measure them? Which risks are we most exposed to in a given moment? What are the consequences of exposure to risks – what is the impact of risks? Which risks are acceptable and to which magnitude or extent? Who are the risks acceptable to and who are they not? How do risks change through time? What is their impact when observed individually and when taken together? What is their mutual effect and what are the consequences of these interactions? How should risks be managed? How to assess the amount of assets required to mitigate, or hedge the risks? The myriad of questions that have remained unanswered to this day points to the complexity of the problem imposed when one embarks on a quest to address and manage the risks in a comprehensive manner.

Risks are perhaps most easily grasped through the example of investment. Investment as the foundation of any business activity – investment enables maintenance, increase of the scope of business operations, or changing the business activity – involves risks and their management a vital part of operating activities; there are virtually no investments without risks.

Risks should be understood in order to be identified or perceived. We should be able to assess and measure their impacts, to monitor them, and after all, to manage them. In recent decades, the latter activity – i.e. risk management – increasingly employs simulations [3]. The reason for using simulations is that in practice, risks include the use of highly complex models [8] in which particular risks, in addition to their mutual interdependence, also depend on system and environment parameters.

Despite the decade-long history of contesting views on the relations between the terms risk, uncertainty, probability, risk exposure, and risk impacts, exact sciences (technical science / engineering, economics, etc.) employ a simplified approach where risk simulation models predominantly, or even exclusively, use probability distributions of risk, while failing to account for their interdependence or dependence on the environment.

In scientific literature, as well as in practice, it is quite common to address risks as something intrinsic to any object, even inanimate, although only humans have the capacity of self-awareness. In his article [4], Glyn A. Holton addresses the question of the level at which the risk is actually taken: can an organization actually be at risk, or is it in fact the individuals, i.e. the employees, who are the risk takers. In this context, they can either be regarded as individuals or as a specific community, or group, within the organization. It should be widely acceptable as a fact that in case of an undesirable event, incident, crisis, or disaster, every community generally bears its own level of risk.

If we concede that only humans have the capacity to be at risk, the ensuing question can be: "Whose risks are being managed?" [4] Perhaps all that is needed are the models of risks that would account for the specificity of particular communities or groups – given that risk is exclusively in the domain of people.

In the following step, we can ask ourselves whether the current risk models adequately account for the state of the environment which is perceived by such models, and in which past facts (as the result of past events and actions) are accumulated, which are intrinsic to the observed system and affect the state in the current moment (or in the moment being simulated). Do they at the very least account for the current effects from the environment? The models predominantly employed in scientific literature or in practice include a considerable degree of simplification and generalization. Quite expectedly so, since without simplification and generalization, there would hardly be a single practically useful model around. In this case, we are dealing with development which, if successful, always begins with simplification.

Another currently relevant area dealing with the accounting for and inclusion of "uncertainty" and "exposure" in risk models seems to open up. Namely, such inclusion becomes particularly complex as soon as one embraces the fact that risks can predominantly be taken by people or groups of people / communities which are generally specific risk takers – each person or each community are at specific risks; hence, we are dealing with a specific uncertainty and exposure in case of each individual group or community.

I shall propose a general principle for risk model, based on the proposition of segmenting the business process model into any given number of dimensions. The first four dimensions are as follows:

1. Inputs and outputs of a particular business process [1];
2. Risks and impacts as the opposite of general (data) inputs and outputs that do not take risks and do not cause effects [1];
3. Internal and external, in relation to the observed system of business processes, in such way that the inputs, outputs, risks, and impacts, are classified into internal (those that can generally be controlled to a greater extent) and external [1];
4. All internal and external business processes which are further divided into risks, impacts, general inputs, and outputs, must be defined for particular interested communities (groups) which again must be defined for the observed system of processes.

The first three of the above mentioned segmentations are described in the [1], while the segmentation according to the different communities (public) is presented in this article.

Any additionally defined dimensions depend on the requirements of the case at hand.

2 MODELING

The described model pursues the ambition to be sufficiently general to be used in various situations and in various fields where risk is encountered – perhaps as varied as suggested by Holton [4] who, in his example, takes trading natural gas, launching a new business, military adventures, …, as well as romance. Although the model described in this article can be used in a wide array of fields, the example of a business process model is taken in the following section. Depending on the particular field at hand that we wish to model, the importance of a particular model segment (various communities, internal vs. external, etc.) may differ; however, it can seldom happen that an individual model segment is completely negligible in a particular field.

2.1 Process Outline

Let us take the example of a business process A in which a clerk A in a company accepts two types of documents: document type X and document type Y. When the clerk A accepts a document, he reviews it to establish its compliance / correctness, and rejects it if required, issuing an explanation type Z which includes a request for amendment of the deficient document. If the document is correct / compliant, it is entered into the incoming mail record and forwarded to clerk B for further processing; any type X document is forwarded to clerk B, performing the business process B, and documents type Y are forwarded to clerk C, performing business process C. Figure 1 illustrates such simplified example of business processes.

In this paper, business processes are represented by process graphs *PG*, i.e. mathematical structures in which the nodes represent a particular process and the link between two nodes represents their relation.

2.2 Risks

Risks as a part of process input are of special interest in risk modeling. Risks cause some kind of (business) loss. The loss is measured by impacts (of risks) which are in turn parts of process output. Risks inevitably cause impacts; however, in addition to risks, impacts also depend on process input in general.

One example of risk in the process described above is arrival of a poorly legible document – perhaps a poor photocopy of the original document. Poor legibility of a document can pose a threat to the correctness of its further processing. The clerk A may confirm such document as being correct and forward it for further processing; however, it may turn out later in the process that an essential part of this document is illegible or not legible enough to allow any certainty as to its particular contents. This, in turn, can lead to even bigger material or non-material damage with legal consequences. Thus, processing a copy of a document (this can include a bad print due to a worn out printer cartridge or toner) always includes an increased risk of damage incurred further in the process. Similarly, damage with legal consequences may result from an unjustified rejection of a document.

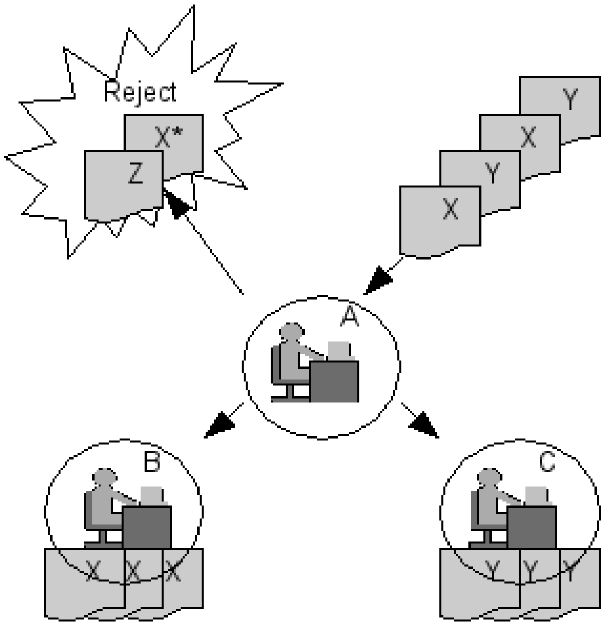


Figure 1: A simplified business process in which a clerk A reviews and sorts / classifies received documents and forwards them to business processes B (clerk B) and C (clerk C).

To remain with the above example, a detailed analysis of the example of risk may lead to a conclusion that misprocessing of type X documents may cause considerable damage while the damage due to misprocessing of type Y documents is quite negligible. Hence, in practice, the extent of potential damage would have been determined according to the share of type X documents and the expected (given the known data from the past) occurrence of misprocessing. The damage thus established represents an impact of type X document misprocessing. Meanwhile, the impact of type Y document misprocessing is negligible.

2.3 Process State

The state of each process is described by parameters. The process state depends on its specific properties which are represented by its parameters. Some examples of such parameters are: process time parameters, maturity level, sensibility to some types of risks, period within the year when its importance may be low or high, risk acceptance, impact acceptance, etc.

The proposed principle rests on the assumption that business processes include parameters whose values describe the state which a particular process is currently in. Any number of business process parameters can be accounted for. In addition, there is the function that calculates new values of process parameters (i.e. the new state) in each discrete (temporal) moment, based on the following:

1. Current values of business process parameters which define the state that the processes are currently in;
2. Business process inputs consisting of the following:
   a) general business process inputs that do not constitute risks; and
   b) risks which include the following:
      - uncertainty and
      - exposure, for each individual community / group of people.

The most important aspect of process parameters is that they allow the past life cycle of each business process to be "accumulated" within them; this accumulated information is then used to accumulate the impacts and new business process parameter values. In this way, modeling also comprises the "history" of the system being modeled. These parameters include the accumulated history of past moments and thus the past combinations of risks and other impacts relevant to the business process.

In the case of business processes of reviewing the incoming documents, process A parameter can be the number of delays resulting from untimely forwarding or rejection of a document by the clerk A (the clerk acts later than legislatively stipulated). This is an example of a parameter that records the state of the process based on an internal rule (measuring the time required for processing each document), and which has no external cause. If the clerk A never makes a mistake, type X documents are sent to clerk B. However, the clerk can be mistaken and may send a wrong document to the clerk B. A document can also be ambiguous and it may only later become evident that it is of a different type than initially believed by the clerk A. In the first or second case, the document sent to clerk B is of a wrong type. Within process B, the number of false documents received can be measured and recorded in a particular parameter of process B. This number is important because it causes extra operative costs. In this case, we are dealing with a process parameter that is changing based on the cause that is external to the process; rather, in relation to the observed process, the cause is a result of an impact from the environment.

2.4 Impacts

Impacts are again calculated with a function with similar properties than the those of the function employed to calculate parameter values. Function parameters, too, are the same; only the calculation differs.

2.5 Introducing Internal and External World

All inputs, outputs, risks, and impacts of a processes, and consequently the entire “known world” *W(PG)*, should be segmented into "internal" *W* and into "external" *W*. The observed *PG,* composed of processes with all their parameters and mutual "output-input" relationships between processes, defines the internal world, while the external world is defined by everything else. In the model, only risks as a part of external input, and impacts as part of external output of observed *PG*, are of our interest. See figure 2.

Usually, we are not able to have exact knowledge of processes from the external world with all their parameters; however, we do know the input (and risks) from external world to the observed *PG,* as well as the output (and impacts) sent to external world from observed *PG*.

In a real situation, it is difficult or even impossible to have any influence on the external risks entering the observed *PG;* on the other hand, we have the power to minimize or even to avoid internal risks (which are impacts from another process from *PG*) either by alternating the function of calculating impacts Φ_{IC} or by alternating some parameters of processes with another function, as presented later in this paper. Consequently, the ability to influence the internal risks (or impacts) is the reason why the internal world should be distinguished form external one.

In the case of clerks performing business processes based on received mail, this means that internal rules and regulations, common conduct, business practice, etc. (which are subject to our decisions) may be adopted in order to change many an internal output (including internal impacts) which defines system behavior in the future.

In our case, this means that we have the power to reduce clerk A's risks related to wrongly forwarded documents to clerks B and C. We do not, however, have any power over the quality of the print on the incoming documents; hence, image quality is an external risk.

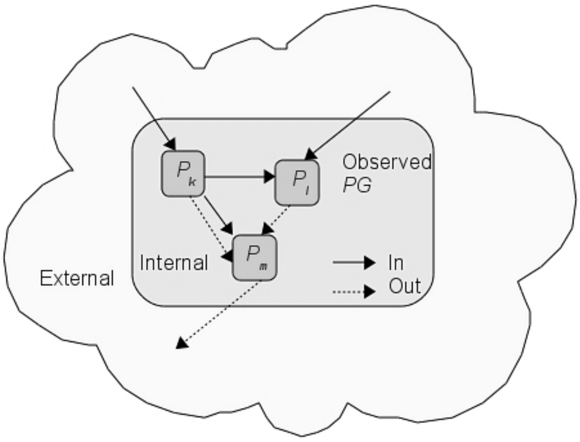


Figure 2: Segmentation of inputs and outputs of a system of processes according to whether the origins and termini of are internal or external.

Without any influence to generalization, the model should be resized from the level of processes to the level of activities or even to the level of particular processes. However, this paper will stay on the level of processes.

The approach where first knowledge is collected about internal and external input, output, risks and impacts, followed by calculation of the impacts Φ_{IC}, from the level of processes to the level of *W(TG)*, by means of parameters (which define the states) and functions, is a down-top approach. In many situations in real process modelling, we are looking to collect as much even approximate knowledge about *W(TG)* as possible, at the first step. Later, step by step, we build more precise and hopefully better models in cycles of collecting additional knowledge of inputs, outputs, risks, impacts, parameters, and knowledge about the function that calculates the impacts of processes. All new knowledge should be segmented as in the first step. Such approach is a top-down approach.

However, segmentation is obviously required to set up a model that can be used for simulations, and it does not matter which approach (top-down or down-top) is chosen. Working on both approaches together gives us the best results in the shortest time. According to our experience, the top-down approach is more effective for programming, expanding, or changing the database of knowledge (or rules) about the model of *PG*.

2.6 Introducing Time

The model should include time dimension which introduces non-determinism. In real situations, some or all processes include the time dimension in their input, output, or in the way how the following state of a process is calculated. Also, the values of parameters should be changed through time. If only one process in a *PG* has included non-determinism, then the whole system *PG* is non-deterministic. In practice, only a very generalized systems without external input are non-deterministic.

2.6.1 Cause – Effect Calculation Through Time

Each process has its input and output. Both of them should belong to the internal or external world as mentioned.

As already noted, internal risks which are caused by internal impacts have a special meaning in reality, because in many cases we have power to alter them while we do not have power to alter risks from the external world. Consequently, internal impacts also have a special meaning in the model of *PG* which is the subject of simulations.

In reality, the process of building a model of complex processes with all relations is usually too complex a problem to make the final model in just a few steps or even in one single step. Building the model means to adapt the model many times through time, making it more precise and useful by adding more knowledge about its input and rules to define calculate impacts and new stages of *PG* in the next time slot.

2.7 Introducing Public Segmentation

According to the concept and definition of risk adopted in this article, risk entails uncertainty and exposure. When understood as conditions, both can only be a condition of humans. As each human is a unique being different from all others, so our relations to a certain risk occurring with regard to a particular situation can also differ greatly. Hence, people have a different view and relation to the same risk. This may be a result of different exposure as well as different uncertainty. The problem is most commonly addressed not in relation to individuals but in relation to groups of people, i.e. individual communities that share a common stance with regard to a particular risk.

The public should also be segmented and simulation should be conducted for each community separately. In order to calculate new process states, and impacts, accounting for different communities within the public does not seem necessary. Hence, the equations in [1] assume a single uniform community (the public), or a single individual.

Given the view adopted throughout this article, however, risk must certainly be calculated for each particular community within the public. Risks can also change with time.

Accounting for the parameters of time and public segmentation, the function of risk calculating should be written as:

$$\begin{gathered}
Risk(P_k, PublicSegment_l, t) = \\
\Phi_{RC}\begin{pmatrix} Uncertainty(P_k, PublicSegment_l, t), \\ Exposure(P_k, PublicSegment_l, t) \end{pmatrix} = \\
\Phi_{RC}\begin{pmatrix} ObjectiveUncertainty(P_k, PublicSegment_l, t), \\ SubjectiveUncertainty(P_k, PublicSegment_l, t), \\ Exposure(P_k, PublicSegment_l, t) \end{pmatrix} = \\
\{R_{k_1}(PublicSegment_l, t), R_{k_2}(PublicSegment_l, t), \ldots, R_{k_m}(PublicSegment_l, t)\}
\end{gathered} \tag{1}$$

Where in (1):

1. P_k is process k.
2. *Uncertainty* (P_k, PS_l, t) is uncertainty in process P_k for the public segment (community) PS_l at time t; in the second step, uncertainty is further divided into objective (*ObjectiveUncertainty*) and subjective (*SubjectiveUncertainty*) uncertainty.

3. *Exposure* (P_k, PS_l, t) is exposure in the process P_k for the public segment (community) PS_l at time t.
4. Particular risks for process P_k are represented by a set of m risks { R_{k_1} (PS_l, t), R_{k_2} (PS_l, t), …, R_{k_m}(PS_l, t)} for a particular public segment PS_l at time t.
5. Function Φ_{RC} calculates risks.

2.7.1 Acceptance Borders

Furthermore, acceptance borders must also be defined for risks, impacts, and process states.

Each time the impacts and parameter values are calculated, the calculated values must be compared against the tolerance values for the following:

1. Risks that present specific inputs in a particular process for each community (group) respectively,
2. Calculated values of particular business process parameters, and
3. Values of calculated effects.

If any of the tolerance values is exceeded, the analysis of the causes leading to such condition must be commenced.

For risks, the acceptance border is calculated in equation (2), using the function Φ_{RAB}; acceptance border for impacts is defined with equation (3) [1] by the function Φ_{IAB}; and the acceptance border for process states is defined with the equation (4) [1] by the function Φ_{SAB}. The first of these functions has three parameters: the process, the public segment, and time; the latter two functions only include process and time.

$$RiskAcceptanceBorder\left(P_k, PublicSegment_l, t\right) = \Phi_{RAB}\left(Risk\left(P_k, PublicSegment_l, t\right)\right) = \left\{RAB_{k,l_1}(t), RAB_{k,l_2}(t), \ldots, RAB_{k,l_m}(t)\right\} \tag{2}$$

$$ImpactAcceptanceBorder\left(P_k, t\right) = \Phi_{IAB}\left(Impact\left(P_k, t\right)\right) = \left\{IAB_{k_1}(t), IAB_{k_2}(t), \ldots, IAB_{k_m}(t)\right\} \tag{3}$$

$$StateAcceptanceBorder\left(P_k, t\right) = \Phi_{SAB}\left(State\left(P_k, t\right)\right) = \left\{SAB_{k_1}(t), SAB_{k_2}(t), \ldots, SAB_{k_m}(t)\right\} \tag{4}$$

In the equations (5), (6), [1] and (7) [1], tolerable, or acceptable values for risk, impacts, and values of the process states are defined according to the given acceptance borders. IN this case, too, the first function has three parameters: the process, the public segment, and time; while the remaining two functions only include two parameters, process and time.

$$AcceptedRisks\left(P_k, PublicSegment_l, t\right) = \left\{R_{k,l_x}(t); x = 1,2,\ldots,m \wedge R_{k,l_x}(t) < RAB_{k,l_x}(t)\right\} \tag{5}$$

$$AcceptedImpacts\left(P_k, t\right) = \left\{I_{k_x}(t); x = 1,2,\ldots,m \wedge I_{k_x}(t) < IAB_{k_x}(t)\right\} \tag{6}$$

$$AcceptedStates(P_k, t) = \left\{ Param_{k_x}(t); x = 1,2,...,m \wedge Param_{k_x}(t) < SAB_{k_x}(t) \right\} \quad (7)$$

Equations (8), (9), [1] in (10), [1] define the unacceptable (intolerable) values which represent a set of values that is equal to the set of all possible values minus the set of acceptable values.

$$NotAcceptedRisks(P_k, PublicSegment_l, t) = Risk(P_k, PublicSegment_l, t) - AcceptedRisk(P_k, PublicSegment_l, t) \quad (8)$$

$$NotAcceptedImpacts(P_k, t) = Imapct(P_k, t) - AcceptedImpact(P_k, t) \quad (9)$$

$$NotAcceptedStates(P_k, t) = State(P_k, t) - AcceptedState(P_k, t) \quad (10)$$

Let us now examine the importance of public segmentation in the example of clerks receiving documents. To review, the risks related to accepting a document which is proven unacceptable later in the process due to poor quality may is that such false acceptance o the document may result in major material and non-material damage with legal consequences. This is the risk of poor document legibility, or image quality; such rusk has its impact – i.e. the consequences which are different for different people or groups of people (public segments). A considerable number of falsely processed documents may even affect the assessment of the clerk A (or several clerks performing such operations). The impact will probably be felt by the entire company or organization in which the said process is taking place – i.e. on all employees. Special impact will probably be sustained by the clerk A's superior (such as supervisor), and perhaps by some other employees as well. What about the person who submitted such document is good faith that it is correct and acceptable since it had been confirmed as such, and it was only later proven that due to the problems with legibility of the document a certain matter (such as obtaining a loan) could not be attended to in time? It this is the case of a company, it is possible that the company's bid at a tender is rejected and such tender may be of crucial importance to the company. In this case, too, various groups of people, or public segments, appear which will sustain different damage. Failed bidding may result in several layoffs (one segment of the public); or, company management may be dismissed (another segment). The media may escalate the proportions of the issue and as a result, other top-ranking officials of the company governance may be affected – which is often the case of public administration in the pre-election time. This is only an example of a single risk, a risk of poor image quality on a received document. This risk, however, includes various groups of people, or segments of the public, and each of them is affected in a different way by this risk. Hence, different stakeholders have defined different acceptance borders. Furthermore, risk acceptance borders can change for a single stakeholder through time – as in the case of the major tender or pre-election time when a different set of rules laws seems to apply.

3 CONCLUSION

Risk management is a process aimed at enhancement, maturation, and evolution of the level of security in an organization. It gives the organization a broad view of the risks, which can affect its productivity and performance, and thus enables it to make appropriate risk management decisions. The knowledge of risks to which an organization is exposed, of the reasons that caused their occurrence, and of the effects that are caused by them, is of vital importance for any organization looking to protect itself from the risks, or avoid them altogether.

Organizations need tools to handle risks; these tools should be easy to use and inexpensive. Most commonly, organizations use simulations of model(s) of their own business. This approach is relatively easy to employ and it is also rather inexpensive if the organization has already established a framework for making ongoing simulations. Establishing such framework requires relatively high investments, and it also includes investing into building a model that is the subject of the simulations. A fair model is a prerequisite for the success of the entire story of risk management which follows successful simulations.

The proposed model of risk assumes several dimensions that should be accounted for in a simulation. There are probably some instances of models in which one or another dimension was taken into account, or perhaps even some that account for all dimensions proposed in the article; however, a search through the scientific literature did not yield examples of this kind that would propose such a model in an intelligible way.

The model is complex, yet still constructed in such way that it allows omitting a particular dimension defined by a segmentation. Thus, it can be simplified to the level of models commonly in use.

The model is fairly easy to use with simulation / modeling languages such as GPSS [5]; the main problem, however, remains the definition of risks, particularly when the model is intended to be used in its entirety, i.e. including all dimensions provided. Hence, the model once again brings us to a situation when we have the tools, but we lack the real knowledge and capacity to make full use of them. The field of risk management may well have developed to a level at which it requires a special kind of experts to solve the most intricate problems. Risk management, on the other hand, requires another type of experts. At today's level of development, risk managers can probably manage the risk models as well, if they are provided with required information on risks and their properties.

The model presented herein has been used in several test runs of simulations for fictional examples; however, it has not yet been used in practice. In the test run, the input data, risks, and functions were also merely test data. Therefore, it is not possible to claim how much this approach has improved the quality and confidence in modeling and simulation. At this moment, we can only assume that the results have improved.

References

[1] Borut Jereb; Segmenting Risks in Risk Management; Logistic & Sustainable Transport; Vol 1, Issue 3

[2] Christian Bluhm, Ludger Overbeck, Christoph Wagner; An Introduction to Credit Risk Modeling; ISBN:158488326X; 2002

[3] Dan X. Houston, Gerald T. Mackulak, James S. Collofello; Stochastic simulation of risk factor potential; Journal of Systems and Software, Volume 59, Issue 3, 15 December 2001, Pages 247-257; doi:10.1016/S0164-1212(01)00066-8

[4] Glyn A. Holton; Defining Risk: Financial Analyst Journal, Volume 60; Number 6, 2004, CFA Institute

[5] GPSS; http://en.wikipedia.org/wiki/GPSS; june. 2009

[6] Jukka Hallikas, Iris Karvonenb, Urho Pulkkinenb, Veli-Matti Virolainenc, Markku Tuominena; Risk management processes in supplier networks; International Journal of Production Economics, Volume 90, Issue 1, 8 July 2004, Pages 47-58; doi:10.1016/j.ijpe.2004.02.007

[7] Lorenzo Benedetti, Davide Bixio, Filip Claeys, peter A. Vanrolleghem; Tools to support a model-based methodology for emission/immission and benefit/cost/risk analysis of wastewater systems that considers uncertainty; Environmental Modelling & Software, Volume 23, Issue 8, August 2008, Pages 1082-1091; doi:10.1016/j.envsoft.2008.01.001

[8] Matthew Pritsker ; The hidden dangers of historical simulation ; Journal of Banking & Finance; Volume 30, Issue 2, February 2006, Pages 561-582; doi:10.1016/j.jbankfin.2005.04.013

[9] Michael B. Gordy; A risk-factor model foundation for ratings-based bank capital rules; Journal of Financial Intermediation , Volume 12, Issue 3, July 2003, Pages 199-232; doi:10.1016/S1042-9573(03)00040-8

[10] Scott, Hal S.; Capital Adequacy beyond Basel, Banking, Securities, and Insurance; ISBN-13: 978-0-19-516971-3; 2005; doi:10.1093/acprof:oso/9780195169713.003.0006

DOES A SADDLE POINT OF CAPITAL-LABOR NEGOTIATION UNDER NEOLIBERAL GLOBALISATION EXIST?

Viljem Rupnik
INTERACTA, Business information processing Ltd.
Parmova 53, Ljubljana, Slovenia

Abstract: a well known dispute between capital and labor induces a temptation to answer a question on whether or not we may expect some final stabilised outcome of this seemingly never ending story. We face a double pressure, namely that of capital over labor and vice versa. The corresponding interaction may be studied in terms of neoliberal globalisation modelling in exposed to both pressures simultaneously.

Keywords: hybrid of formal and non-formal modeling, interaction analysis, the role and destiny of neoliberal globalisation goal policy.

1 INTRODUCTION

As neoliberal globalisation stretches over the world, the antagonismus between capital and labor exerts very serious conflict. On one hand, capital wants to maximize its profit and labor aims at higher wages; hence a sequence of negotiation actions on both sides takes place. Since both sides, in practice, use different arguments, it is difficult to foresee the final result and, moreover, it is hard to ensure the existence of its termination. This paper will not present the essence of feasible final result being agreed upon by both sides, although it has been proved to be realistic (see /1/). Its aim is rather focused on establishing an objective "third part" approach which might be able to furnish an answer to the second question. An objective approach should consist of one and the same decision making platform being a map of consequences of decisions of both "players", mostly being mutually interdependent. To hope that a corresponding sequence is final stepwise procedure, the architecture of the decision making platform should not be influenced by any of the players.

2 BASIC DECISION MAKING PLATFORM

Since a strict requirement imposed on a decision making platform is at the same time embeded in neoliberal globalisation, it is convenient to use primary globalisation model from /2/ which appears as labor oriented globalisation model and which serves as initial prototype of globalisation modelling. It is convenient to start from original Marxian scheme of reproduction, consisting of two sectors: S_1 sector of production of investment goods and S_2 sector of production of consumption goods. Thus, we introduce the following control theory analogues: capacity function of S_1 is $\eta = \eta(\tau, t), 0 \leq \tau \leq t \leq t_1$ measuring a capability of generating new job places in both sectors enabled by a unit of job place employed in S_1; similarly, let $\xi = \xi(t)$, $0 < t \leq t_1$, be a demand for production capacity function of sector S_1 needed by S2 per unit of its capacity; let $n = n(t), t_1 \leq t \leq t_2$ to be total number of employees in both sectors; $d = d(t), t_1 \leq t \leq t_2$ the average length of working day in both sectors; let $x(t)$ and $y(t)$ denote the growth speed in S_1 resp. S_2 sectors in terms of new job places, both defined on $0 \leq t \leq t_1$; let $\lambda = \lambda(\tau, t)$ and $\mu = \mu(\tau, t)$ be production exploatation rates in corresponding sectors, both defined on $0 \leq \tau \leq t_1, t_1 \leq t \leq t_2$. It is essential to assume that some effective demand z=$z(t)$ for products by final production sector S_2 supporting its operations.It is an active part of newly acquired labor $r(t)$ within the final control horizon $[t_1, t_2]$ considered.

To start with most simplified version of labor oriented globalisation model as described above, we may have the following formulation of its scalar-criterion version

$$\max\ r(t) = \int_{t_1}^{t_2} \left[\int_0^t \left[\lambda(\tau) x(\tau,t) + \mu(\tau) y(\tau,t) d\tau \right] \right] dt \tag{1.1.}$$

constrained by

$$x(t) + y(t) = \int_0^t \lambda(\tau,t) x(\tau) \eta(\tau,t) d\tau\,, \qquad t_1 \le t \le t_2 \tag{1.2}$$

$$z(t) = \int_0^t y(t) \mu(\tau,t)\, \xi(\tau) d\tau \qquad t_1 \le t \le t_2 \tag{1.3}$$

$$\int_0^t \left[x(\tau)\lambda(\tau,t) + y(\tau)\mu(\tau,t) \right] d\tau \le n(t) d(t) \tag{1.4.}$$

and the control domain being

$$\left. \begin{array}{l} \left. \begin{array}{l} 0 \le \lambda(\tau,t) \le 1 \\ 0 \le \mu(\tau,t) \le 1 \end{array} \right\} 0 \le \tau \le t_2, \;\; t_1 \le t \le t_2 \\ x(t) > 0, \qquad y(t) > 0, \qquad 0 \le t \le t_2 \end{array} \right\} \tag{1.5}$$

By model (1.1) to (1.5) we had made our firm decision on globalisation problem to be dealt with by means of a dynamic model. Broadly speaking, we had to define it as open ended control problem

$$r(t) = \int_{-\infty}^{t} \mathrm{T}(\tau,t,e)\, e(\tau) d\tau \tag{2.1}$$

having some response $r\ (t)$, specialised in (1.1), produced by external forces $e(t)$, specialised in $n = n(t),\ t_1 \le t \le t_2$ and $(\ d = d(t),\ t_1 \le t \le t_2$. In more general case, $r\ (t)$ and $e(t)$ are vectors, defined in linear and finite dimensional spaces. An impulse matrix T of state transition operators, being defined in nonlinear and finite dimensional space, in case of (1) reduced to (1.4.).

Furthermore, in generalised case, the behavior of external forces $e(t)$ is described by

$$\Phi(e, \frac{de}{dt}, \alpha) = 0 \tag{2.2}$$

usually appearing in the form either differential, integral or differential-integral equations, α being a system parameter. The two formulations above satisfy the additional necessary assumption on globalisation process. To illuminate the combat between capital and labor, e do not need to be involved in (2.2.).

Furthermore, for the same reason we can omit response vector $s(t)$ as a measure of performance/consequences of the globalisation process. It is urgent to assume that some secondary multidimensional effects $s(t)$ will come up from the primary response $r(t)$, say

$$s(t) = \Psi[t, \beta; r(t)] \qquad (2.3)$$

where β stands for evaluation parameter. The last constituent of general globalisation model lies in kernel of multidimensional evaluation of neoliberal globalisation consequences.

3 CONSEQUENCES OF CAPITAL PRESSURE OVER LABOR

1.When a pressure of capital over labor is taking place, a labor purchasing power tends to zero and, consequently, a S_2 selling market is decreasing; it leads therefore to $\xi(t) \to 0$ .Since both sectors are interrelated as an implication of $\eta(\tau,t)$ by $\xi(t)$ in the sense that $\xi(t) \to 0$ implies $\eta(\tau,t) \to 0$; due to (1.2) , x(t) + y(t) is limited by $\eta(\tau,t)$ and $\xi(t)$ or, simplifying

$$\lim_{\eta(\tau,t) \to 0} [x(t) + y(t)] \le 0$$

which is in an accordance with the vision of globalisation.

2. Consequently, due to $x(t) \to 0$ and $y(t) \to 0$ we also have $z(t) \to 0$, which means, when adding to the previous property that production of sector of consumer goods production is vanishing.

3. Moreover, since $x(t) \to 0$ and $y(t) \to 0$ $\Rightarrow$, the value of criterion function r(t) $\to 0$ either. Consequently, there is no employment, no new employment and no means of living. The existence triangle consisting of "natural resources – labor – capital" shrinks to a two-point active segments only; thus, it totally destroys a notion of economy in general. The two production factors, natural resources and capital become idle, or, more simply: **there will be no cheerful untertainment whatsoever!**

4.To slow down the process of $x(t) + y(t) \to 0$ we hope to increase capacity-exploatation rate $\lambda(\tau,t)$ as much as possible (but not greater than 1) provided we succeeded in building up output capacity of both sectors inside a period [t_1, t_2] large enough to offset or at least partially compensate the decline of $\xi(t)$ and $\eta(\tau,t)$. But, even when $\lambda = \lambda(\tau,t)$=1 and $\mu = \mu(\tau,t)$ = 1, the above results still hold true inasmuch the pressure of capital over labor remains in action.

5.Although $d(t)$ is increasing (but limited from above), suppose, in S_1 as well as in S_2 , the right hand side of (1.4.) is decreasing due to the active part of labor.

4 CONSEQUENCES OF LABOR PRESSURE OVER CAPITAL

Based on model (1), a labor may express its pressure through $n = n(t)$, the number of employed labor, and $d = d(t)$, the average length of working day in both sectors. In most cases, we may assume that its pressure refers to the increase of $n = n(t)$, appealing for the decrease of $d = d(t)$, but keeping the same level of wages, at least. Separate effects pf both parameters on criterion (1.1) have the opposite sign.

1. the number of employed labor becomes a live parameter increasing the capital criterion, but it does not affect directly the demand for final products; in general, it is more or less depenendent on capital's will to adjust it or not

.

2. if conditions on selling market are favourable, the positively correlated steering parameter *n=n(t),* being an instrument of labor pressure over capital acts in favour of capital and not against it.

3. if the volume of sales is fixed at existing level in the moment of bargaining , parameter *n=n(t)* loses its sense, i.e. it can never be a negatively correlated steering parameter: it is never harmful to profit as a capital criterion.

4. assuming that labor accepts an unchanged level of wages, *d=d(t)* starts acting as a steering parameter negatively correlated with capital's criterion: it may become an effective bargaining instrument in hands of labor.

5. since model (1.1 to(1.5) is labor oriented in non-value terms, we can not depict an effect of *d=d(t)* as a steering parameter negatively correlated with capital's criterion (such a case is studied in /1/).

6. most interesting is the combination of points 2. and 4.

5 INTERACTING PRESSURES OF CAPITAL AND LABOR

Any bargaining procedure has it sequence of trials, usually being effected by alternating actions of both sides (bargaining partners). From practical nature this procedure can not last for ever, which means that a corresponding sequence is finite. What is its terminal point?

The behaviour of partners is stochastic process, but each step is allowed to be influenced by variables and parameters of model (1). According to the analysis of consequences of each of separate pressures, we have two groups of attributes: $K = \{\lambda, \mu, \xi, \eta, x, y\}$ (in its limit state it reduces to $\mathrm{K} = \{\lambda, \eta, x, y\}$ only) for capital and $\mathrm{I} = \{n, d\}$ for labor. Within the context of neoliberal globalisation a labor is refused to require its criterion, the overwhelming criterion $r(t)$ belongs to the capital only, depending on $K = \{\lambda, \mu, \xi, \eta, x, y\}$ and $\mathrm{I} = \{n, d\}$. Since attributes of both groups are assumed not to be independent, but detectable as observational data reflecting on a profit as negotiation criterion. Thus, a total observational set $\Omega = \{r(t), K(t), I(t)\}$ is a domain of interaction analysis which should take place over all attributes of Ω simultaneously.

This process is generally expected in a noisy environement and therefore it is very rare to expect a perfect conditional independence. Consequently, it is not clear how to determine/recognize a dependence as insignificant: a temptation to use Bayesian approach (network) breaks down, e.i. we can not disclose the loss of reliability of the data recorded. Moreover, the Bayesean approach does not furnish any information on the importance of some particular dependencies, which might be useful to qualify connections between attributes to be important or not.Apart from additional failures of Bayesean approach we therefore try to view the negotiation process as a series of probabilistic interactions and quantifying them with entropy-based interaction information. Such a decision enables us to describe a joint probability distribution of some set of attributes on the basis of joint probability distributions of all combinations of attributes belonging to some subset of attributes (see /3/)

Having Ω defined, we should expect positive and negative interactions among its attributes. Since, in practice, they are not perfectly positive or negative, entropy-based interaction information is employed to quantify the type and magnitude of interaction; entropy as a measure of uncertainty. Such a measure will be derived from uncertainty measure which defines a dependence on shared uncertainty. It says therefore, to simplify our exposition by using Shannon's entropy, that

$$H(A) = -\sum_{a \in \Omega} P(a) \log_2 P(a) \qquad (3.1.)$$

is a measure of predictibility of attribute *A*, chosen from total observation set Ω. Smaller the entropy (3.1.), the higher reliability of our prediction of *A*. For neoliberal globalisation we have $A =$ max $r(t)$. In case of Ω consisting of two attributes, for example, we have

$$H(A \perp B) = -\sum_{a \in \Omega_a, b \in \Omega_b} P(a,b) \log_2 P(a \perp b) = H(A,B) - H(B) \qquad (3.2.)$$

as a conditional entropy for the case of 2-way interaction, whereas its mutual information is

$$I(A,B) = \sum_{a \in \Omega, b \in \Omega} P(a,b) \log_2 \tfrac{P(a,b)}{P(a)P(b)} = H(A) + H(B) - H(A,B) \qquad (3.3.)$$

thus being a measure of correlation between the two attributes; a remark on technical reason: $a \in \Omega$ runs through $\Omega \setminus a$ and similar for $b \in \Omega$. An extension to our negotiation process is evident as being observed as 9-way interaction.

We have to make a severe warning. WE have, in practice, to be aware, using model (1), that there exist some noise residual (observable or not), apart from Ω. If we suspect that some residual $\widehat{\Omega} \supset \Omega$ exist, the whole interaction analysis should be carried out on $\widehat{\Omega}$ and then to derive measures as (3.2.) or)3.3) for 9-way interaction. To simplify again, we shall assume that $\widehat{\Omega} \setminus \Omega$ is empty.

Having computed conditional entropies for 9-way interaction process on some observation time interval detect the interactional behaviour of the two partners either from interaction dendrogram or form interaction graph. As our underlying interaction analysis has been dedicated to max $r(t)$ as the dominant attribute and single criterion, we check its entropy at each observational moment in the past. A series of computed 7-way interaction entropies, say, $H\{\max r(t) \perp (\Omega - \max r(t))\} = H^{(7)}\{\max r(t)\}$, $\forall t$ through the observational period of time *T*, will, in general, fluctuate stochastically. We are looking for the term having $\min H^{(7)}\{\max r(t)\}$ for some $t \in T$. If it exists for some $t^* \in T$ both partners may stop negotiation procedure since they received most reliable prediction (minimum entrophy); in such a case, at $t^* \in T$ a **pseudo saddle point** occurs. If partners are "risk averse" from their nature, they will likely propose to stop; otherwise they continue in a hope to decrease its entropy at some later point. That is why we name it as "pseudo".

6 UNCERTAINTY CONDITIONED BEHAVIOUR OF THE TWO INTERPLAYING PARTNERS: CAPITAL AND LABOR

So far, a single economic criterion, a profit, has been considered as a measure of negotiation having been passed up to a pseudo saddle point. A profit has been forced in by capital, not by labor, unless the labor takes part in sharing of profit (as in the case of ethical economic growth). However, it quite realistic assumption that both partners will be looking at the level of uncertainty being computed for each value of economic criterion; it means therefore that they are not "risk averse"or better to say, they are sensible to enthropy as well. If they agree upon 9-way interaction analysis, e.i. they exclude any scope of additional noise, there will be two decision variables: max *r(t)* and corresponding enthropy $H^{(7)}\{\max r(t)\}$, for $\forall t$. Let the number of switches inside *T* be *N;* then we have 2-criteria 9-way interaction *N* –dimensional decision problem whose solution can not be obtained through any formal presentation. Out of existing multicriterial decision making procedures known, the RKLR method is most efficient one, for the time being (see /4/).

Let us consider this case a bit more precisely and let generalise it by assuming $p \geq 2$ criteria, q -way interaction (both being integer) and N –finite dimensional multicriterial decision making problem. At any step i =1, 2,…N the two decision variables define the **decision object**

$$S^0{}_i = \{\max r(t), H^q[\max r(t)]\} \tag{4.1}$$

Regardless of $S^0{}_i$, i =1, 2,…N being mutually dependent or not, following non-formal approach from /4/ method RKLR computes for decision object their **global measures** for $\forall i$

$$\mu(S^0{}_i) = \mu\ \{\max r(t), H^q[\max r(t)]\} \tag{4.2.}$$

which is a result of ranking of decision objects (they are ranked global measures).

To be more realistic let us assume that both negotiation partners agree upon some finite set Ξ of references for each of attributes (in our initial case we would have 9 of them), being unchanged for all *i*, we have **referenced global measures**

$$\mu(S^0{}_i, \Xi) = \mu\{\max r(t), H^q[\max r(t)], \Xi\} \tag{4.3.}$$

On one hand, operator (4.3.) represents a **relative valuing** of the *i*-th step of negotiations with respect to references available (valuing among themselves); on the other hand, this valuing is **absolute valuing**, if some ideal negotiation is conceivable. Due to RKLR properties, the ideal object $\mu(S^0{}_i, \Xi)$ is produced by (4.3.) as soon as characteristic parameters *p, q* and *N* are known. Since then their boundary values within the corresponding metric spaces are known, we should speak about **group-conditioned absolute value** of an object in question. Since arguments in (4.3.) are symmetric, to each object from (4.2.) a group-conditioned absolute value is assigned/computed as well. Thus, an ideal object is a »function« of objects involved in (4.3.).Hence, it follows that any reference implanted »from outside« (i.e. not being determined by objects in evaluation) may or may not be identical with $S^0{}_i$. The difference between the two is a measure of the distance between »internal« and »external« ideals (it may serve as an imitation/animation of negotiation partners).

References

[1] Jakulin A, Bratko, I.: Quantifying and Visualizing Attribute Interactions, Faculty of Computer and Information Science, University of Ljubljana, Slovenia, August 1, 2003

[2] Rupnik V.: An Attempt to Non-formal Modeling, Proceedings of the 4th International Symposium on OPERATIONS RESEARCH, Preddvor, Slovenia October 1-3, 1997, pp. 207-214.

[3] Rupnik V., Sundac D.: The domination of Capital = A trapping state of Mankind; in Slovenian and Croatian language, IBCC, Rijeka, 2005.

[4] Rupnik V., Research papers on globalisation modelling; unpublished; INTERACTA, Ltd, Ljubljana.

[5] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman, San Francisco, CA, USA, 1988.

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Section IX:

Statistics

TEST OF INDEPENDENCE BETWEEN VARIABLES IN THE BUSINESS SURVEY FOR THE SERVICES SECTOR – SOME EMPIRICAL EVIDENCE FROM CROATIA

Mirjana Čižmešija
Vlasta Bahovec
University of Zagreb, Faculty of Economics and Business
Trg J. F. Kennedyja 6
10000 Zagreb, Croatia
mcizmesija@efzg.hr
vbahovec@efzg.hr

Abstract: There are six questions in the services survey as a part of the business survey: business situation, turnover and total employment over the past 3 months and expected demand, expected total employment changes and the price changes over the next 3 months. The survey results in Croatia's services sector were used to find out whether there is no association between the variables in services survey questionnaire. Various statistical tests of independence for qualitative variables were applied. An analysis on a stratified sample of managers in Croatia's business survey in the services sector (n = 129) was conducted in the first quarter of 2009.

Keywords: Business Survey, Services Sector, Qualitative Variables, Pearson Chi Square Test, ML Chi Square Test, Cramer´s V

1 INTRODUCTION

Given the importance of the services sector for the overall economy, the services survey was incorporated in Croatia's business survey in the second quarter of 2008, in accordance with the Joint Harmonised EU Programme of Business and Consumer Surveys. The services survey provides information about the managers' assessment of their present and future business situation. It is a qualitative survey, like other parts of business and consumer surveys. In times of global recession, it is of interest here to find association in managers' expectations and assessments in the services sector. The EU services survey questionnaire has six monthly questions and one quarterly question. Business surveys can be used for getting answers to the questions outside harmonized methodology, which may be of particular interest for the economy under observation. In Croatia's business survey there are two additional questions: ownership and liquidity. Consequent liquidity problems, have since the socialist times been one of the major problems in business operations of companies, especially in the current global recession [6]. In the first quarter of 2009, there were 46.5 % of all companies with bad liquidity and periodical problems. Arrears mean that some of the companies, but also other buyers, such as the government, do not meet their financial obligations on time and are late for more than 360 days. They reflect a lack of financial discipline known as a part of "soft budget constraint". Such behaviour triggers a chain reaction in the economy as a whole, turning sometimes profit making companies illiquid.

2 THE SERVICES SURVEY AS A PART OF THE BUSINESS SURVEY

A business survey is a qualitative economic survey in the manufacturing industry, construction, retail trade and the services sector. This is an effective tool to monitor and predict macroeconomic developments in the national economy. Business surveys, especially services surveys and surveys in the financial services sector are a valuable tool for interpreting changes in the business climate. While the business survey results are available before the corresponding values of the macroeconomic variables, they can be used to predict their developments [3]. They are widely used to detect turning points in the economic cycle.

In Croatia, business surveys started in the second quarter of 1995 in the manufacturing industry; in the third quarter of 1995 they started in construction and in retail trade and in the services sector they started in the second quarter of 2008.[1] The surveys are being continually conducted on a quarterly basis in accordance with EU methodology (The Joint Harmonized EU Programme of Business and Consumer Surveys by the Directorate – General for Economic and Financial Affairs, in 1961 first in the manufacturing industry, construction, consumers and in retail trade [8]). The harmonized EU Programme (on the EU level) was extended to the services sector in 1996 with the aim to produce a set of comparable data for all countries [5].

The services survey covered different sectors in accordance with the NACE classification (NACE Rev. 1.1.)[2].

3 TEST OF INDEPENDENCE

The aim of this paper is to test the hypothesis of no association between the selected variables in the services survey. All variables in the services survey are qualitative. One of the most widely applied statistic tools for testing the presence of association between two categorical variables in the contingency table is the *Pearson Chi-square test*.

The *Chi-square test* requires a relatively large sample size. The *Fisher's exact test* is appropriate for small samples in the 2x2 contingency table. Both of the tests provide similar results.[3] They are useful in making conclusions whether the relationship between two variables is statistically significant at the defined significant level [4]. They do not evaluate the strength of the relationship.

Some other measures of association evaluate the existence and the strength of the association between two categorical variables. The most widely applied measures of association are *Phi a*nd *Cramer's V*[4].

The analysis of sample data in this paper was made using the SAS computer package [9]. In the SAS output, the *Pearson* and *ML chi square test* results are selected, as well as

---

[1] Privredni vjesnik – Zagreb.

[2] The sectors which are covered by a large majority of EU member States (and Croatia) in the services survey are [8]:
55 - Hotels and restaurants, 60 - Transport, 63 - Activities of travel agencies and tour operators, tourist assistance activities, 64 – Post and telecommunications, 65 - Financial intermediation, expect insurance and pension funding, 66 – Insurance and pension funding, expect compulsory social security, 67– Activities auxiliary to financial intermediation, 70 – Real estate activities, 71 - Renting of machinery and equipment without operators and renting of personal and household goods, 72 – Computer and related activities, 74 – Other business activities, 90 – Sewage and refuse disposal, sanitation and similar activities.

[3] When conducting the C*hi – square test*, the expected number of observations in each cell should be at least 5, if the preceding statistics are to be valid. If this assumption is met, the sampling distribution of the test value is approximately a C*hi-square distribution* with (r-1)(c-1) degrees of freedom [4]. Some variable modifications were applied in this paper (merging of two modalities into one), in order to meet this requirement. For 2x2 contingency tables, the *Fischer's exact test* can be computed, if a table, which does not result from missing rows or columns in a larger table, has a cell with an expected frequency less than 5.

[4] If there is a 2x2 contingency table, *Phi* is the appropriate statistics. For larger contingency tables *Cramer's V* is better than *Phi*. The maximum possible value of *Phi* and *Cramer's V* is in some cases less then 1, thus creating a problem in interpretation.

measures of association between two categorical variables[5] (*Phi Coefficient, Contingency Coefficient and Cramer´s V.*) [2].

4 EMPIRICAL RESULTS

Croatia's services survey questionnaire has nine questions, i.e. nine qualitative variables[6]. All questions (except *ownership* and *liquidity*) have a similar answer scheme [8]. The sample size in Croatia's services survey, which was conducted for the year 2009 in the 1st quarter, is 129 companies in the services sector, which is representative for the entire Republic of Croatia. The source for the survey data is *Privredni vjesnik – Centre for Business Research.*

Special parts of the research in this paper were association measures between variable *liquidity* and all other variables in the questionnaire. The aim of this paper is to investigate if there is a statistically significant relationship between *liquidity,* as a special important variable in Croatia's economy in recession conditions, and all other variables in the services survey questionnaire. It is assumed that there are associations between managers' assessments of *liquidity* and their *expected demand* and the *expected business situation*. On the other hand, the assumption is that there are no associations between the *liquidity* and the *business situation* (over the past 3 months), *past demand*, *past employment, expected employment* and *expected price changing* in the next six months. With the aim to test these hypotheses, some statistical tests of no association were applied.

The measuring of association between the variable *ownership* and all other variables could not be conducted by applying the *Chi-square test*, because the requirement of the expected value in the contingency table of greater than 5 could not be met. The reason for this is that 106 of 129 companies are privately owned. The other 23 were mostly privately or mostly stated owned.

The variables are noted as follows:

Past business situation (Q1)
Past demand (Q2)
Expected demand (Q3)
Past employment (Q4)
Expected employment (Q5)
Prices change (Q6)
Expected business situation (Q7)
Ownership (Q8)
Liquidity (Q9)

---

[5] More details in: 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 Operational Research SOR '05, Zadnik Stirn, L., Drobne, S. (editors). Nova Gorica, Slovenia, 28-30 September 2005, p. 405-410. (ISBN: 961-6165-20-8)

[6] The questions in Croatia's services survey are: Q1- How has your business situation developed over the past 3 months? Q2- How has demand (turnover) for your company's services changed over the past 3 months? Q3 - How do you expect the demand (turnover) for your company's services to change over the next 3 months? Q4- How has your company's total employment changed over the past 3 months? Q5 - How do you expect your company's total employment to change over the next 3 months? Q6 - How do you expect the prices you charge to change over the next 3 months? Q7 - We expect that our business situation in the next 6 months will be in accordance with today's business situation; Q8- Ownership, Q9 – Liquidity.

The results of the tests are summarized in the SAS printout (table 1, table 2 and table 3) with the aim to test no association between two variables in the contingency table.

Table 1

Statistics for Table of Q1 by Q9

Statistic	Value	Prob
Chi-Square	3.9593	**0.0466**
Likelihood Ratio Chi-Square	3.9769	**0.0461**
Phi Coefficient	0.1752	
Contingency Coefficient	0.1726	
Cramer's V	0.1752	

Statistics for Table of Q4 by Q9

Statistic	Value	Prob
Chi-Square	4.5413	**0.0331**
Likelihood Ratio Chi-Square	4.5645	**0.0326**
Phi Coefficient	0.1876	
Contingency Coefficient	0.1844	
Cramer's V	0.1876	

Statistics for Table of Q7 by Q9

Statistic	Value	Prob
Chi-Square	5.9708	**0.0145**
Likelihood Ratio Chi-Square	5.9999	**0.0143**
Phi Coefficient	0.2151	
Contingency Coefficient	0.2103	
Cramer's V	0.2151	

Based on the sample data and no association tests in table 1, we can see that the empirical significance level (*p- value*) in the *Chi – Square test* for *liquidity* and three separated variables Q1, Q4 and Q7 are less than 0.05. The null hypotheses is not accepted, since the *p-value* is less than the theoretical significance level α=0.05. This means that at a 5% significance level, there are associations between the variable *liquidity* and the variables *past business situation (Q1), past employment* (Q4) and *expected business situation (Q7)*[7]. This means that managers' judgements about *past business situation* and *past employment* (over the past 3 months) and managers' *expectation about business situation* are associated with their judgement about *liquidity*.

One of the hypotheses was that there are associations between *liquidity* and *expected employment*. In table 2 we can be see that this assumption can be acceptable only at the 10% significance level (*p-value* = 0.0818).

Table 2

Statistics for Table of Q5 by Q9

Statistic	Value	Prob
Chi-Square	3.0297	0.0818
Likelihood Ratio Chi-Square	3.0378	0.0813
Phi Coefficient	0.1533	
Contingency Coefficient	0.1515	
Cramer's V	0.1533	

[7] p-value <0.05

Based on the results which are presented in table 3, it can be concluded that for all other variables (*past demand (Q2), expected demand (Q3)* and *price changes (Q6)*) and *liquidity* at a 5% level of significance there is no association.[8] All measures of association have small values. This means that managers' assessments and expectations in *liquidity* and in three variables (mentioned above) are not statistically significantly associated.

Table 3

Statistics for Table of Q2 by Q9

Statistic	Value	Prob
Chi-Square	0.5839	0.4448
Likelihood Ratio Chi-Square	0.5842	0.4447
Phi Coefficient	0.0673	
Contingency Coefficient	0.0671	
Cramer's V	0.0673	

Statistics for Table of Q3 by Q9

Statistic	Value	Prob
Chi-Square	0.1822	0.6695
Likelihood Ratio Chi-Square	0.1823	0.6694
Phi Coefficient	-0.0376	
Contingency Coefficient	0.0376	
Cramer's V	-0.0376	

Statistics for Table of Q6 by Q9

Statistic	Value	Prob
Chi-Square	1.1279	0.2882
Likelihood Ratio Chi-Square	1.1260	0.2886
Phi Coefficient	0.0935	
Contingency Coefficient	0.0931	
Cramer's V	0.0935	

The variables *past business situation* (*Q1), past demand (Q2)* and *expected demand (Q3*) are (in accordance with the Joint Harmonised EU Programme of Business and Consumer Survey) variables' components of the Services Confidence Indicator (SCI) as a composite indicator derived from the services survey [8]. Special attention should be paid to these variables, because time series of survey results in Croatia's services surveys will in time gain enough observations for the calculation of a composite indicator, which will be more appropriate for some more sophisticated econometrical analyses.

After the conducted test of no association between *liquidity* and all other variables, this paper also tested the connection between variable components of the Services Survey Indicator and some selected variables in the questionnaire. The survey results may be found in table 4, as follows:

[8] p-value >0.05

Table 4

Past business situation (Q1) and *Expected business situation (Q7)*

Statistic	Value	Prob
Chi-Square	17.004	0.0000
Likelihood Ratio Chi-Square	17.387	0.0000

Past business situation (Q1) and *Past employment (Q4)*

Statistic	Value	Prob
Chi-Square	17.374	0.0000
Likelihood Ratio Chi-Square	18.960	0.0000

Past demand (Q2) and Prices change (Q6)

Statistic	Value	Prob
Chi-Square	11.648	0.0030
Likelihood Ratio Chi-Square	12.050	0.0020

Expected demand (Q3) and *Expected business situation (Q7*)

Statistic	Value	Prob
Chi-Square	37.770	0.0000
Likelihood Ratio Chi-Square	43.854	0.0000

Expected demand (Q3) and *Prices change (Q6)*

Statistic	Value	Prob
Chi-Square	9.8790	0.0020
Likelihood Ratio Chi-Square	9.7580	0.0020

All examples show that the null hypotheses is rejected, since the *p-value* is smaller than the theoretical significance level α=0.05. This means that there is a statistically significant association between variable components of the Services Survey Indicator and variables such as *price changes (Q6), expected demand (Q3), expected business situation (Q1)* and *past employment (Q4).* It is interesting to see that at a 5% significance level, there is no association between *past employment (Q4)* and *price changes (Q6)* (*Chi-square* is 3.594, *p-value* is 0.058, *ML chi square* is 3.332 and *p-value* is 0.068).

5 DISCUSSION AND CONCLUSION

In various researches we are testing causality between composite indicators derived from business and consumer surveys and some selected referent macro economy variables [1]. Based on Croatia's experience it can be concluded that there is an evident existence of the relationship between the Economic Sentiment Indicator - ESI and the Gross Domestic Product - GDP, between the Industrial Confidence Indicator and the volume of industrial production, between the Retail Trade Confidence Indicator and the retail sales volume, as well as between the Construction Confidence Indicator and volume indices of construction works.

The time series of variable balances and the time series of Croatia's Services Confidence Indicator are too short for econometrical forecasting models. However, the survey results can be used to find out whether there is no association between component variables of SCI and the variables. Based on the survey results, we may conclude that there is a statistically significant association between variables components of SCI and *price changes, expected demand, expected business situation* and *past employment. Liquidity* as an additional variable in Croatia's Services Survey is a very important indicator, especially in recession. The sample data point to the conclusion that at a 5% significance level, there is an association between the variable *liquidity* and the variables *past business situation, past employment* and *expected business situation*[9]. This means that managers' judgements about *past business situation, past employment* and expected business situation are associated with their judgement about *liquidity.* All measures of association are positive, which means that if the liquidity assessment is good, past business situation is improved or remains unchanged, past employment is increased or remains unchanged, expected business situation is defined as increasing or remaining unchanged.

The mentioned results are in accordance with the economic interpretation of possible associations / non-associations between single variables, i.e. between the questions in the questionnaire, which means that the managers' answers are qualitatively good and well matched. Since the Services Survey in Croatia is being conducted since the 2nd quarter of 2008, it is not possible to apply the usual numeric procedures of summarizing the qualitative data into quantitatively expressed indicators and comparing them to reference series of the official statistics. The conclusions of the research in this paper confirm the quality of the research and the seriousness of the managers – participants in the research concerning the grading and the expectations, which guarantees the quality of the quantitative indicators, which will in time be calculated based on Croatia's Business Survey in the Services Sector.

References

[1] Bahovec, V., Čižmešija, M., Kurnoga – Živadinović, N., (2007.) *Testing for Granger Causality between Economic Sentiment Indicator and Gross Domestic Product for the Croatian Economy*. Proceedings of the 9th International Symposium on Operational Research SOR '07, Zadnik Stirn, N., Drobne, S. (editors), Nova Gorica, Slovenia, September 26-28, 2007, p. 403-408

[2] 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. (editors). Nova Gorica, Slovenia, 28-30. rujna 2005. godine., p. 405-410. (ISBN: 961-6165-20-8)

[3] Bahovec, V., Čižmešija, M., Kurnoga – Živadinović, N., (2008.) *The forecasts of some macroeconomic variables for Croatia using the information provided by business surveys.* Proceedings of the 11th International Conference on Operational Research KOI 2006,. Boljunčić, V. (editor), Pula, Croatia, September 27 - 29, 2006., str. 77 – 86. (ISBN 978-953-7498-11-5)

[4] Bodrožić, I., Jurun, E., Pivac, S. (2007.) *Chy-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. (editors), Nova Gorica, Slovenia, September 26-28, 2007, p. 415-420

[5] Čižmešija, M. (2008.). *Konjunkturni testovi Europske unije i Hrvatske*. Zagreb. Privredni vjesnik (ISBN 978-953-6488-14-8)

[9] p-value <0.05

[6] Nikić, G., Šošić, I., Čižmešija, M. (2002). *Business and investment surveys in Croatia – a case study of an economy in transition*. Proccedings (CD), 12 p. http://www.ciret.org/conferences/taipei_2002/c26_papers

(login:taipei, password: cpapers), 26th CIRET Conference in Taipei, 16-19. october 2002.

[7] Šošić, I., Čižmešija, M. (2003). *A note about forecasting accuracy of business survey in Croatia.* Bulletin of the International Statistical Institute, 54th Session, Contributed Papers, Volume LX, Book 2, p. 465 – 466., (CD). http://www.isi.de, Berlin.

[8] *The joint harmonised EU programme of business and consumer surveys*, User guide (updated 07/06/2007). European Economy. European Commission, Directorate-General for Economic and financial affairs.

http://ec.europa.eu/economy_finance/indicators/business_consumer_surveys/userguide_en.pdf

[9] Program support SAS

EVALUATING TREND MODELS IN FORECASTING EXPORTS TIME SERIES IN CROATIA

Ksenija Dumičić, Irena Čibarić and **Anita Čeh Časni**
Department of Statistics, Faculty of Economics and Business, University of Zagreb,
Trg J.F. Kennedy 6, HR-10000 Zagreb, Croatia
{kdumicic,icibaric,aceh}@efzg.hr

Abstract: The paper discusses trend models for modelling exports and relative percentage share of exports in gross domestic product in Croatia. The article focuses on examining the adequacy of four trend models: linear, quadratic, exponential and second-order exponential model. The main emphasis is on comparison of statistical quality measures of forecasts based on the fitted models. For this purpose, forecasts quality indicators as Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), Mean Squared Deviation (MSD) and information criteria (AIC ,SC, HQ) based on EViews and Minitab software output were applied.

Keywords: regression trend models forecasting, Mean Absolute Percentage Error (MAPE), Mean Square Deviation (MSD), information criteria

1 INTRODUCTION

In uncertain times of economic crisis forecasting economic time series is not an easy task to do, but still remains a very important tool for decision making in both business and economic area. It is especially important for policy makers and strategic planning of every company and every industry in the country. Forecasting economic time series is important for producers, central and commercial banks, policy makers, and all other economic actors, for people employed and those unemployed, etc. So, all the ministries, chambers of commerce, scientific and economic institutes and macroeconomic analysts in the countries all over the world put efforts to make forecasts, both subjective and objective, based on relevant macro-economic aggregate series to be well prepared for all that might happen. But, forecasts, even methodologically well done, may never replace the truth with certainty, because they are based on the past that will never repeat. If forecasts would be based on the past data which happened to be similar to the data under the time horizon in forecasting, forecasts would have a chance to be realistic and economic policy instruments might be better prepared to prevent predicted future trends, even shocks.

Many of economic actors participate in the process of creating GDP components through production, trade, collecting savings and allocation of financial assets, and some of them have got crucial role in determining the conditions for stimulating trade, and especially foreign trade, influencing GDP backwards [13]. Since exports is very important for GDP composition according to expenditure definition, the authors of the paper decided to analyse the dynamics of exports and share of exports in GDP in Croatia for the eleven years' period 1997-2007. Official time series data for exports and GDP were based on constant prices of 1997. Analysis was done with the final purpose to make statistically solid predictions based on several regression models, with all regression model assumptions tested in detail [1], [3], [10], [11], [15], and [16].

The research hypotheses was that exports and percentage share of exports in GDP depend on time variable and that some regression forecasting models are better than the others, so detailed research of past dynamics and forecasting were conducted for both time series, exports and percentage share of exports in GDP, and relevant indicators of forecasts efficiency were compared.

To better understand exports and GDP trends in Croatia, it could be said that since 2000 Croatia has got moderate but steady GDP growth between 4% and 6% led by a rebound in tourism and credit-driven consumer spending. Inflation rate over the same period has remained smooth and the currency, the kuna, stable. But, difficult problems still remain, including a growing trade deficit with uneven regional development and stubbornly high unemployment rate. The government retains a large role in the economy. In 2006 real GDP growth rate was 4.8%, in 2007 5.6% and in 2008 4.8%. The GDP per capita was (in constant 2008 US$) about $14,900 in 2006, about $15,700 in 2007, and approximately $16,100 in 2008. In 2008 exports in Croatia were at level of $12.36bn, and were based on transport equipment, machinery, textiles, chemicals, foodstuffs, fuels. The most important Croatian exporting partners in 2007 are Italy (19.3%), Bosnia and Herzegovina (13.9%), Germany (10.2%), Slovenia (8.4%), and Austria (6.2%). In the same time imports amounted $25.84bn. [15]. The EU accession process should accelerate fiscal and structural reform. While long term growth prospects for the economy remain strong, Croatia will face significant pressure as a result of the global financial crisis. Croatia's high foreign debt, weak export sector, strained state budget, and over-reliance on tourism revenue will result in higher risk to economic stability over the medium term. Several recently published analyses described the movements of GDP components in Croatia [2], [4], [5], [6].

According to Eurostat document Annual national accounts, see the GDP definition based on the expenditure method [12]: "The External balance of goods and services (ESA95, 8.68) is the difference between exports and imports of goods and services (ESA93, 3.128-3.146). It may be calculated separately for transactions in goods and in services. Exports of goods and services consist of transactions in goods and services (sales, barter, gifts or grants) from residents to non-residents (ESA95, 3.128). Imports of goods and services consist of transactions in goods and services (purchases, barter, gifts or grants) from non-residents to residents (ESA95, 3.129)." Using expenditure approach, GDP consists of private final consumption expenditure, government final consumption expenditure, gross fixed capital formation, changes in inventories, acquisition less disposal of valuables, net exports.

Some statistical descriptive technical analysis tools may help in forecasting time series, but some smoothing methods and regression analysis would help often even more. In this paper the dynamics of two time series in Croatian economy were concurrently considered with the main goal to predict the near future values with emphasis on evaluating the efficiency of several forecasting models that were applied. For methods see [3], [15], and [16].

After introductory part, the paper consists of the following parts: Data and methods, Descriptive and forecasting methods analysis (Examining the violation of regression assumptions; Measures of predictive efficiency and selection of forecasting model, and Comparing the forecasted values to the observed values), and Conclusion.

2 DATA AND METHODS

Croatian official exports and GDP data, as well as calculated % share of exports in GDP for the period 1997 – 2008 [7] are given in Table 1. Exports, GDP and relative share of exports to GDP in Croatia show growth trend in period 1997-1998.

Table 1: Exports in million kuna, at constant (1997) prices, GDP in million kuna, at constant 1997 prices, and exports-to-GDP percentage ratio for the period 1997.-2008.

Year	Time variable	Exports	GDP	% share of exports in GDP
1997	1	50873	123811	41.09
1998	2	52847	126936	41.63
1999	3	53204	125843	42.28
2000	4	59566	129438	46.02
2001	5	64397	135189	47.63
2002	6	65202	142730	45.68
2003	7	72660	150351	48.33
2004	8	76805	156758	49.00
2005	9	80320	163491	49.13
2006	10	85850	171277	50.12
2007	11	94493	180784	53.77
2008	12	(96787)	196856	(49.17)

Source: First Releases, National Accounts, Central National Bureau of Statistics

Considering methods that are going to be used for the research purpose in this paper, first, run charts and descriptive statistical indicators, such as differences, index numbers and rates were used. These tools helped to decide about possible trend analysis methods that might be used as appropriate. After that, several linear regression forecasting models with least square estimates were applied and compared. The main advantage of trend analysis is that, when model is appropriate and data exhibit a clear trend, it is possible to carry out simple analysis [1]. Because of the presence of the evident trend component in all time series, the smoothing methods did not seem to be useful [9], so they were not applied at all.

Table 2 contains first differences and chain indices for the time series of exports (in million kuna, at constant prices 1997). It is evident that the first differences are all positive, but not approximately equal. The chain indices for exports data are all higher than 100 and approximately constant at the average yearly rate of about 6%. It should be concluded that there existed absolute and relative increase of exports over all 12 years, which means that a kind of increasing trend should be clearly expected. Besides the linear, both quadratic and exponential trend models should be evaluated, at least.

Table 2 contains first differences for the time series for exports-to-GDP percentage ratio for the period 1997-2008. These differences have got opposite pre-signs, which mean that sometimes a decrease was present. Rough impression is that first differences of % share of exports in GDP were approximately constant, so a linear trend should be explored. Chain indices for the exports-to-GDP percentage ratio for data are mostly higher than 100, with the exception of two years, 2002 and 2008. It seemed to be reasonable to try modelling with quadratic, and exponential trend models, as well.

Because the goal of the research is to find the best fitted forecasting models for both time series, the natural logarithmic (ln) transformation should be introduced and evaluated.

Table 2: First differences and chain indices of exports in million kuna at constant (1997) prices, and exports-to-GDP percentage ratio for the period 1997-2008

Year	First differences of exports	Chain indices of exports	First differences of % share of exports in GDP	Chain indices of % share of exports in GDP
1997	-	-	-	-
1998	1974	103.88	0.54	101.31
1999	357	100.68	0.65	101.56
2000	6362	111.96	3.74	108.85
2001	4831	108.11	1.61	103.50
2002	805	101.25	-1.95	95.91
2003	7458	111.44	2.65	105.80
2004	4145	105.70	0.67	101.39
2005	3515	104.58	0.13	100.27
2006	5530	106.89	0.99	102.02
2007	8643	110.07	3.65	107.28
2008	2294	102.43	-4.6	91.45

3 DESCRIPTIVE AND FORECASTING METHODS ANALYSIS

Based on descriptive analysis of exports and exports-to-GDP percentage ratio in period 1997-2007, with n=11, it was decided to try modelling with four models for each of the series. The least squares estimates were calculated for linear and quadratic, both in original and in (natural) log-form. Firstly, examination of the violation of regression assumptions for all the models applied was done. After that, measures of predictive efficiency (MAPE, MAD and MSD) were compared for models in which regression assumptions were not violated. Furthermore, information criteria were used in selecting forecasting model for each of the series. On the basis of mentioned measures, appropriate forecasting model was selected for both time series. Finally, forecasted values for both series were calculated and compared with the observed values.

3.1 Examining the violation of regression assumptions

Significance tests for all the parameters as well as Durbin-Watson test of first-order autocorrelation were conducted. The results of the testing procedures are given in Table 3 and Table 4.

Table 3: Results of a time series regression modelling for exports in period 1997-2007

Trend model	Coefficient estimate	t-ratio	p-value	DW statistic
Linear	$\hat{\beta}_1 = 4310.92$	18.30	0.000	1.24
Quadratic	$\hat{\beta}_1 = 1857.93$	2.86	0.021	2.46
	$\hat{\beta}_2 = 204.42$	3.87	0.005	
Log-linear	$\ln \hat{\beta}_1 = 0.062$	25.81	0.000	2.15
Log-quadratic	$\ln \hat{\beta}_1 = 0.050$	4.81	0.000	2.39
	$\ln \hat{\beta}_2 = 0.001$	1.28	0.236	

Table 4: Results of a time series regression modelling for exports-to-GDP percentage ratio in period 1997-2007

Trend model	Coefficient estimate	t-ratio	p-value	DW statistic
Linear	$\hat{\beta}_1$=1.132	9.71	0.000	2.01
Quadratic	$\hat{\beta}_1$=1.252	2.30	0.050	2.04
	$\hat{\beta}_2$ = -0.009	-0.23	0.827	
Log-linear	ln $\hat{\beta}_1$=0.024	9.66	0.000	1.92
Log-quadratic	ln $\hat{\beta}_1$=0.031	2.70	0.027	2.06
	ln $\hat{\beta}_2$ = -0.0005	-0.61	0.561	

The parameters of models were estimated using minimum least squares approach. Table 3 shows the results of a time series regression for exports. Regression coefficients are shown to be significant at 5% level for three models: linear trend, quadratic trend and log-linear trend model. Violation of regression assumptions is examined for these models. When using regression with time series data, the assumption of independence could be violated. Thus, the error terms may be correlated over time [10]. However, the presence of autocorrelation can have adverse consequences on tests of statistical significance in a trend model. In forecasting time series, the basic requirement that the residual pattern is random is verified by examining the autocorrelation of residuals. There should be no significant autocorrelation coefficients [11]. Due to that fact, Durbin-Watson test of first-order positive autocorrelation is conducted for these three models. For models for which Durbin-Watson test was inconclusive, Breuch-Godfrey Serial Correlation LM Test was conducted. In addition, Ljung-Box autocorrelation test was conducted. For models where autocorrelation of first and second order is concluded not to be present, additionally Ljung-Box test of autocorrelation is conducted.

Durbin-Watson test is inconclusive for linear and quadratic trend model, hence Breuch-Godfrey Serial Correlation LM Test is conducted. In linear trend model, Breuch-Godfrey LM test statistic equals 1.525, with p-value equal to 0.466. Hence, there is enough evidence to accept null hypothesis of the test. Thus, Breuch-Godfrey test shows no autocorrelation of first and second order for linear trend model. Regarding quadratic trend model, Breuch-Godfrey LM test statistic equals 7.17, with p-value 0.0278. There can be concluded at 5% significance level that there exists evidence of autocorrelation of second order in quadratic trend model. Concerning log-linear trend model, Durbin-Watson test statistic leads to the acceptance of null hypothesis at 5% significance, hence it can be concluded that there does not exist the problem of autocorrelation in log-linear trend model. The autocorrelation of the error terms is in addition examined by conducting Ljung-Box test of linear and log-linear trend model. In both models, at 5% significance Ljung-Box test shows no autocorrelation for all lags $k \leq 11$.

Moreover, heteroskedasticity test is conducted due to the fact that in case of existing heteroskedasticity the least squares estimators are inefficient and estimates of variances are biased, this invalidating the tests of significance [14]. White heteroskedasticity test is conducted for models in which autocorrelation is not found to be present at 5% significance. The test shows p-values greater than 0.05 (greater than any reasonable significance level), and hence there exists enough evidence for acceptance of null hypothesis of no heteroskedasticity for both linear and log-linear trend model. Jarque-Bera test for both models shows p-values greater than 0.05 (greater than any reasonable significance level), thus there can be concluded that normal distribution of residuals exists.

Concerning exports-to-GDP percentage ratio (Table 4), regression coefficients are significant at 5% level for linear and log-linear trend. Durbin-Watson test shows no autocorrelation at 5% significance level for both models. Ljung-Box test is conducted for linear and log-linear trend model, and in both models, at 5% test shows no autocorrelation for all lags k≤11. White heteroskedasticity test shows no heteroskedasticity for both linear and log-linear trend model. Jarque-Bera test shows normal distribution of residuals.

After testing the violation of regression analysis assumptions for exports and exports-to-GDP percentage ratio, two models are selected for further analysis of predictive efficiency in forecasting exports and exports-to-GDP percentage ratio time series in Croatia: linear and log-linear (exponential) trend model.

3.2 Measures of predictive efficiency and selection of forecasting model

In order to select the model for forecasting exports and exports-to-GDP percentage ratio, the quality of the selected models is evaluated using appropriate common forecasts' quality measures such as Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), and Mean Squared Deviation (MSD). Table 5 and table 6, respectively, show the values of these indicators for selected exports and exports-to-GDP percentage ratio forecasting models. Furthermore, information criteria for selecting the most parsimonious correct model are used. Akaike, Schwarz and Hannan-Quinn criterion are useful in finding the model that best explains the data with a minimum number of parameters. These criteria are often used to measure the efficiency of the model in terms of forecasting. Table 5 and table 6, respectively, show the values of these measures for selected exports and exports-to-GDP percentage ratio forecasting models.

Suppose there are observations and forecasts for *n* time periods. Then the following statistical measures can be defined [15]:

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100 . \tag{1}$$

Mean Absolute Deviation (MAD):

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \hat{y}_t\right| . \tag{2}$$

Mean Squared Deviation (MSD):

$$MSD = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2 . \tag{3}$$

Furthermore, the basic information criteria are given by (4), (5) and (6). Let l be the value of the log of the likelihood function with the k parameters estimated using T observations. The various information criteria are all based on -2 times the average log likelihood function, adjusted by a penalty function. Information criteria used are as follows:

Akaike info criterion: $AIC = -2(I/T) + 2(k/T)$. (4)

Schwarz criterion: $SC = -2(I/T) + k\log(T)/T$. (5)

Hannan-Quinn criterion: $HQ = -2(I/T) + 2k\log(\log(T))/T$. (6)

Table 5 shows the values of all the mentioned measures calculated for linear trend model and exponential trend model of exports time series. Lower MAPE, MAD and MSD show that the deviation of predicted values from observed values is lower for exponential trend model. Additionally, information criteria are lower for exponential trend model.

Table 5: Comparison of measures of predictive efficiency for linear and exponential trend model of exports Croatia in period 1997-2007

Trend model	MAPE	MAD	MSD	AIC	SC	HQ
Linear	3	1676	4996335	18.63	18.70	18.58
Exponential	2	1224	2138072	17.73	17.81	17.70

Based on the results given in Table 5, exponential trend model is selected for forecasting exports time series in Croatia. The estimated exponential trend model for exports (natural logarithm of regression coefficient $\hat{\beta}_1$ is given in Table 3) is as follows:

$$\hat{y} = 46246.3 \times 1.0648^t \quad (7)$$

According to the estimated model, average growth of exports in period 1997-2007 was 6,48 percentage points yearly, while exports in 1996 amounted 46246.3 million kuna.

Regarding the model for forecasting exports-to-GDP percentage ratio time series in Croatia, Table 6 shows the results of calculation of mentioned measures of predictive efficiency. In this instance, lower values of MAPE, MAD and MSD, as well as of information criteria can be observed for linear trend model.

Table 6: Comparison of measures of predictive efficiency for linear and exponential trend model of exports-to-GDP percentage ratio in Croatia in period 1997-2007

Trend model	MAPE	MAD	MSD	AIC	SC	HQ
Linear	2.00	0.95	1.22	3.40	3.48	3.36
Exponential	2.05	0.97	1.26	10.83	10.90	10.78

Based on the results given in Table 6, linear trend is selected for forecasting exports-to-GDP percentage ratio time series in Croatia. The estimated linear trend model for forecasting exports-to-GDP percentage ratio is(regression coefficient $\hat{\beta}_1$ is given in Table 4):

$$\hat{y} = 39.99 + 1.13t \quad (8)$$

According to the estimated model, exports-to-GDP percentage ratio average growth in period 1997-2007 was 1.13 percentage points yearly, while exports-to-GDP percentage ratio in 1996 amounted 39.99 percent.

3.3 Comparing the forecasted values to the observed values

After the exponential trend model has been chosen for forecasting exports time series in Croatia, and linear trend model has been chosen for forecasting exports-to-GDP percentage ratio in Croatia, the comparison of forecasted values using both linear and exponential trend model for both time series is done to verify the results of the analysis conducted.

Fitted time series values for period 1997-2007 for t=1 through t=11, as well as forecasted values for a short time horizon τ=1 and t=(n+τ)=(11+1)=12, and corresponding residuals for both linear and exponential trend model were calculated. Residuals or error terms $\hat{e}_t$ represent the differences between observed and predicted values:

$$\hat{e}_t = y_t - \hat{y}_t \quad (9)$$

Table 7 shows the observed values, forecasted values and residuals for exports time series for both trend models. Residual for forecasted value of exports in 2008 is equal to 2174 million kuna when linear model is used. In other words, forecasted value underestimates the observed value of exports in 2008 by 2174 million kuna. If exponential trend model is used for forecasting exports, then the residual for forecasted value of exports in 2008 is equal to -947 million kuna for linear model. In other words, forecasted value overestimates the observed value of exports in 2008 by 947 million kuna. The difference between observed and forecasted value is smaller when using exponential trend model and that confirms previously conducted selection of exponential trend model for forecasting exports in Croatia.

Table 7: Observed and forecasted values of exports in Croatia using linear and exponential trend models calculated for period 1997-2007 and forecast for 2008

Year	**Observed values for exports, million kuna at constant (1997) prices**	**Forecasted values for exports, million kuna (Linear trend model)**	**Residuals (Linear trend model)**	**Forecasted values for exports, million kuna (Exponential trend model)**	**Residuals (Exponential trend model)**
1997	50873	47192	3681	49241	1632
1998	52847	51503	1344	52429	418
1999	53204	55814	-2610	55824	-2620
2000	59566	60125	-559	59439	127
2001	64397	64436	-39	63287	1110
2002	65202	68747	-3545	67385	-2183
2003	72660	73058	-398	71749	911
2004	76805	77369	-564	76395	410
2005	80320	81680	-1360	81341	-1021
2006	85850	85991	-141	86608	-758
2007	94493	90302	4191	92216	2277
τ=1, 2008	96787	94613	2174	97734	-947

Table 8 shows the observed values, forecasted values and residuals for exports-to-GDP percentage ratio time series for both trend models. Absolute difference between observed and forecasted value of exports-to-GDP percentage ratio in 2008 is equal to 4.41% when linear model is used, i.e. the residual equals -4.41%. Thus, forecasted value in 2008 overestimates the observed value of exports-to-GDP percentage ratio in 2008 by 4.41%. If exponential trend model is used for forecasting exports-to-GDP percentage ratio, then the residual, i.e. the difference between observed and forecasted value in 2008 equals -4.8%. Hence, the forecasted value overestimates the observed value of exports-to-GDP percentage ratio in 2008 by 4.8%. The difference between observed and forecasted value is smaller when using linear trend model and that confirms previously conducted selection of linear trend model for forecasting exports-to-GDP percentage ratio in Croatia.

Table 8: Observed and forecasted values of exports-to-GDP percentage ratio in Croatia using linear and exponential trend models calculated for period 1997-2007 and forecast for 2008

Year	Observed values for exports-to-GDP percentage ratio	Forecasted values (Linear trend model)	Residuals (Linear trend model)	Forecasted values (Exponential trend model)	Residuals (Exponential trend model)
1997	41.09	41.13	-0.04	41.29	-0.2
1998	41.63	42.26	-0.63	42.31	-0.68
1999	42.28	43.39	-1.11	43.35	-1.07
2000	46.02	44.52	1.5	44.42	1.6
2001	47.63	45.66	1.97	45.52	2.11
2002	45.68	46.79	-1.11	46.64	-0.96
2003	48.33	47.92	0.41	47.79	0.54
2004	49.00	49.05	-0.05	48.96	0.04
2005	49.13	50.19	-1.06	50.17	-1.04
2006	50.12	51.32	-1.2	51.41	-1.29
2007	53.77	52.45	1.32	52.67	1.1
τ=1, 2008	49.17	53.58	-4.41	53.97	-4.8

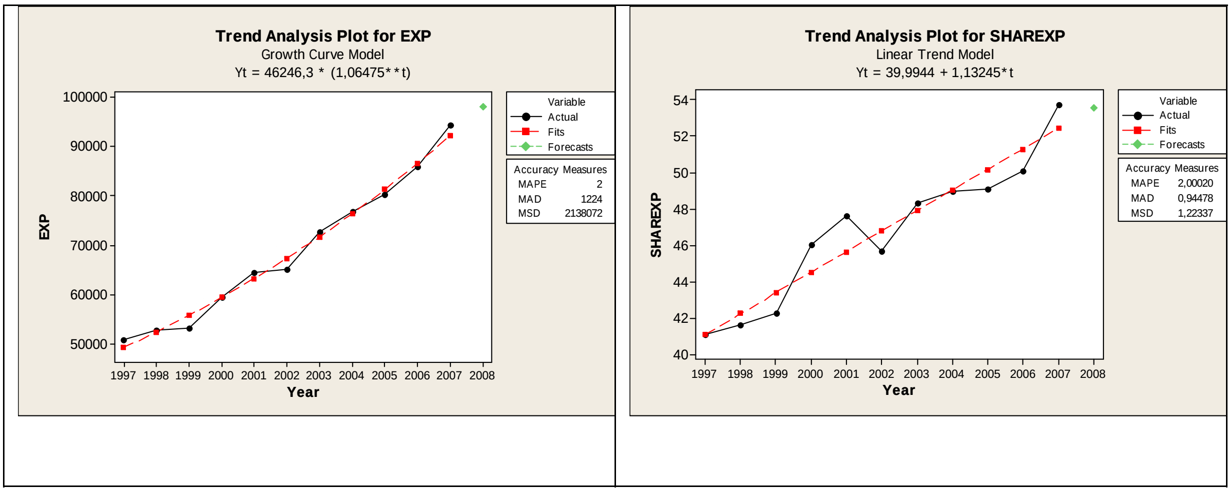


Figure 1: Minitab Trend Analysis Plot for exports and exports-to-GDP percentage ratio in Croatia in period 1997-2007 and forecast for 2008

Figure 1 shows Minitab output for exponential trend model for forecasting exports, as well as linear trend model for forecasting exports-to-GDP percentage ratio in Croatia. Only the time horizon τ=1 was applied.

4 CONCLUSIONS

Only if the future pattern would be like the past, the forecast should be reasonably accurate. In this empirical research time series regression models were used as forecasting models.

Based on descriptive analysis of exports and exports-to-GDP percentage ratio for Croatia in period 1997-2007 it was decided to try modelling with four regression models for each of the series. The forecasts were done for time horizon τ=1, i.g. only for 2008. The least squares estimates were calculated for linear and quadratic, both in original and in (natural) log-form. Firstly, examination of the violation of regression assumptions for all the models applied was done. After that, measures of predictive efficiency (MAPE, MAD and MSD) were compared for models in which regression assumptions were not violated.

Considering exports in Croatia in period 1997-2007, with the intention to make forecast for 2008, the calculated values of all the relevant measures MAPE, MAD and MSD for linear trend model and exponential trend model of exports time series indicated that exponential model was better. Simply, the deviation of predicted values from observed values was lower for exponential trend model. Additionally, information criteria were lower for exponential trend model, and this fact supports the conclusion that exponential trend model is more suitable.

Regarding exports-to-GDP percentage ratio in Croatia 1997-2007, the calculated values of all the relevant measures of deviation of predicted values from observed values, as well as information criteria were lower for the linear trend, thus linear trend model was selected for forecasting exports-to-GDP percentage ratio in Croatia.

Furthermore, when comparison of forecasted values is done, the conclusion is that forecasted value for 2008 is closer to observed value when using exponential trend model for forecasting exports, as well as when using linear trend model for forecasting exports-to-GDP percentage ratio in Croatia.

However, this research has got a limited scope, because trend models are appropriate for short run forecasts, given that circumstances under the time horizon are quite stable comparing to the past that was investigated. The restrictions of applied models refer to global financial crisis that changed the economic environment substantially. The Croatian Central Bureau of Statistics [8] has announced the country's gross domestic product (GDP) fell by 6.7 per cent in the first three months of 2009 already. It was the largest quarterly drop in GDP in the last ten years. According to the Bureau data, all categories of GDP declined with the exception of state consumption, which rose by 3.9 per cent. Household consumption declined by 9.9 per cent, the biggest decline in the last ten years. Exports of goods and services fell by 14.2 per cent, imports of them by 20 per cent. In these circumstances, it can not be assumed that neither exports nor relative percentage share of exports in gross domestic product in Croatia will move according to the selected models and the assumption that economic conditions are not changed is violated.

References:

[1] Aczel, A. D., Sounderpandian, J. (2009). Complete Business Statistics. 7th Edt., New York: McGraw Hill.

[2] Babić, Z., Čondić-Jurkić, I., Mervar, A., Nestić, D., Pavuna, D. Slijepčević, S., Smilaj, D., Švaljek, S., Vizek, M. (2008). Usporavanje u drugom tromjesečju?, Privredna kretanja i ekonomska politika, pp. 7-22. http://www.eizg.hr/AdminLite/FCKeditor/UserFiles/File/PKIEP115-current-trends.pdf

[3] Bowerman, B. L., O'Connell, R. T., Koehler, A. B. (2005). Forecasting, Time Series and Regression. 4th Edt., Belmont, California: Duxbury Press.

[4] Broz, T., Buturac, G., Pavuna, D., Rašić Bakarić, I., Slijepčević, S., Smilaj, D. (2008a). Najava recesije. Privredna kretanja i ekonomska politika, 117, zima, pp. 7-26. http://www.eizg.hr/AdminLite/FCKeditor/UserFiles/File/PKIEP117-current-trends.pdf

[5] Broz, T., Buturac, G., Pavuna, D., Rašić Bakarić, I., Slijepčević, S., Smilaj, D. (2008b). Usporavanje. Privredna kretanja i ekonomska politika, pp. 7-25. http://www.eizg.hr/AdminLite/FCKeditor/UserFiles/File/PKIEP116-current-trends.pdf

[6] Broz, T., Buturac, G., Pavuna, D., Rašić Bakarić, I., Slijepčević, S., Smilaj, D. (2009). Nepovoljni trendovi, Privredna kretanja i ekonomska politika, 118, pp. 9-29. http://www.eizg.hr/AdminLite/FCKeditor/UserFiles/File/PKIEP_118_Current_economic_trends.pdf

[7] Central Bureau of Statistics, Republic of Croatia, First Releases, National Accounts, http://www.dzs.hr (July 25 2009)

[8] Central Bureau of Statistics, Republic of Croatia, Quarterly Gross Domestic Product Estimate, 2009, http://www.dzs.hr (July 30 2009)

[9] Dumičić, K., Čeh Časni, A., Gogala, Z. (2008). Evaluating Holt's Double Exponential Smoothing and Linear Trend Forecasting of Basic Tourism Time Series in Croatia. 4th International Conference «An Enterprise Odyssey: Tourism – Governance and Enterpreneurship. University of Economics and Business – Zagreb, Cavtat, Croatia.

[10] Groebner, D. F., Shannon, P. W., Fry, P. C., Smith, K. D. (2008). Business Statistics: A Decision Making Approach. 7th Edt., Upper Saddle River, NJ: Prentice Hall.

[11] Hanke, J. E., Reitsch, A. G. (2008). Business Forecasting. 9th Edt. Needheim Heights: Allyn and Bacon.

[12] http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/en/nama_esms.htm (June 27, 2009)

[13] http://www.theodora.com/wfbcurrent/croatia/croatia_economy.html (July 30, 2009)

[14] Maddala, G.S. (2001). Introduction to Econometrics. 3rd Edt., New York: Wiley.

[15] Makridakis, S., Wheelwright, S.C., Hyndman, J. (1998). Forecasting, Methods and Applications. 3rd Edt., New York: Wiley.

[16] Montgomery, D.C, Jennings, L.J., Kulahci, M. (2008). Introduction to Time Series Analysis and Forecasting. New York: Wiley.

MULTIPLE MONTE CARLO SIMULATIONS – CASE STUDY OF CROATIAN COUNTIES

Elza Jurun and **Snježana Pivac**
University of Split, Faculty of Economics
Matice hrvatske 31, 21000 Split, Croatia
{elza,spivac}@efst.hr

Abstract: This paper is a part of an extensive research about Croatian economic challenges within global recession environment. Its focus is estimation of integral Croatian Counties model by multiple Monte Carlo simulations. The main intention is to quantify each partial influence of a set of independent regional variables (employment, gross investment, production of more important agricultural products, GVA per person employed, value of construction works, exports, imports, foreign tourists arrivals, foreign tourists nights, ecology…) to regional GDP per capita. Integral part of this paper is the comparative analysis with the results given by classical approach of econometric multiple regression.

Keywords: Monte Carlo simulations, multiple regression model, Gauss Markov conditions, regional GDP per capita, comparative analysis.

1 INTRODUCTION

At the end of the period of transition economy, Croatia confronts with challenges and problems as modern East European societies. Nowadays in global crises environment Croatia, as regional economic leader, pretends to become a member of developed European countries community.

Although as EU accession country Croatia is divided into three NUTS 2 (from French language - Nomenclature des unites territoriales statistiques) regions, twenty one Croatian Counties show significant economic and social disproportions. These disproportions are in the focus of extended on-going scientific research. In this research phase integral Croatian Counties model is estimated by multiple Monte Carlo simulations. In multiple regression model it is estimated how regional GDP per capita as regresand variable depends about a set of independent regional variables (employment, gross investment, production of more important agricultural products, GVA per person employed, value of construction works, exports, imports, foreign tourists arrivals, foreign tourists nights, ecology…). Dependence between variables is estimated by the standard statistical way and subsequently parameters are evaluated by Monte Carlo simulations. The comparative analysis of the results confirms the advantage of Monte Carlo simulation in evaluation of interdependence especially regarded on Gauss-Markov conditions on stochastic part of econometric model. For this regional analysis multiple Monte Carlo simulation of regression model is used for the first time to measure individual average influence of each mentioned independent variable. As one of the main methodological assumptions in modeling is normal residuals distribution, Monte Carlo simulation is used also to fulfill this basic assumption [9].

It is necessary to emphasize that the whole methodological procedure is done upon data base of real regional indicators. Unfortunately, the data base does not consist of recent indicators because of Croatian official statistics data time lag. That is why the latest year with complete data for each Croatian county was 2005 and these indicators are taken as inputs for this research.

In spite of division into three NUTS 2 regions of Croatia as EU accession country, for the purpose of this phase of research Croatia is divided into twenty one counties: County of Zagreb, County of Krapina-Zagorje, County of Sisak-Moslavina, County of Karlovac, County of Varaždin, County of Koprivnica-Križevci, County of Bjelovar-Bilogora, County of Primorje-Gorski kotar, County of Lika-Senj, County of Virovitica-Podravina, County of Požega-Slavonia, County of Slavonski Brod-Posavina, County of Zadar, County of Osijek-Baranja, County of Šibenik-Knin, County of Vukovar-Sirmium, County of Split-Dalmatia, County of Istria, County of Dubrovnik-Neretva, County of Međimurje, City of Zagreb.

The structure of the paper is organized as follows. After introduction part in the second section methodology approach and data used in analysis are described. In the third section classical econometric and then multiple Monte Carlo simulations approach are applied on case study of Croatian counties. The fourth section summarizes the results of the research and provides conclusions.

2 METHODOLOGICAL BACKGROUND

For analyzing, planning and managing with regional GDP per capita in each Croatian county at the beginning of this research, the intention is to define direction and intensity of various partial influences on it. Between regional economic indicators which are measured continuity by official statistics for the purpose of this paper are chosen these ones whose values are published for each of twenty one Croatian county for 2005.

First of all the classic econometric approach has been used and regional GDP per capita has been defined as a product of chosen regressor variables by multiple regression model. Later these results will be used for comparative analysis with corresponding results given by multiple Monte Carlo simulations [9].

Like regression analysis, Monte Carlo simulation is a general term that has many meanings. The word “simulation” signifies that we build an artificial model of a real system to study and understand the system. The Monte Carlo part of the name alludes to the randomness inherent in the analysis. The name Monte Carlo Was coined during Manhattan Project of Word War II, because of the similarity of statistical simulation to games of chance, and because the capital of Monaco was a center for gambling and similar pursuits [2]. Monte Carlo is now used routinely in many diverse fields for the simulation of various complex phenomena. Monte Carlo simulation is a method of analysis based on artificially recreating a chance process (usually with a computer), running it many times and directly observing the results [7]. Because Monte Carlo simulation is based on repeatedly sampling from a chance process, it stands to reason that random numbers are a crucial part of the procedure.

Therefore, from all above mentioned the conclusion arises that experimenting with different values of $\beta_0, \beta_1, ..., \beta_k$ parameters results with different sums of square residuals. This means that estimated regression parameter values are expected values of their probability distributions [1]. In this case from linear variables dependences:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + ... + \beta_k \cdot X_{ki} + e_i \qquad (1)$$

outcomes that parameters distributions depend of random error distribution, i.e. depend of real regresand variable Y_i values distribution [13].

In the case of classical regression model it is assumed that random variable e_i is normal distributed with zero mean and standard deviation equal to one [6]. Whereas, very often in practice it is not fulfilled. Namely, random variable distribution may be distributed by Student distribution, uniform distribution, lognormal distribution etc…, or random variable distribution is not defined [14]. In these cases instead of appropriate transformation and in order to fulfill the normality assumption, Monte Carlo simulation method is suggested for parameter estimation. Monte Carlo method enables parameter estimations even when random variable distribution is not normally. The guideline in the process of discovering appropriate residual distribution is empirical variable Y_i values distribution.

If empirical variable Y_i values are normally distributed, it can be assumed that residuals are normal distributed with zero mean and standard deviation equal to one [16].

3 CASE STUDY OF CROATIAN COUNTIES

Official indicator for numerous macroeconomic analyses is GDP per capita. Croatia as accession EU country has to fulfill a lot of social, legal and economic presumptions in which GDP per capita is also very important criterion. Moreover, this criterion was superior in the process of classifying Croatian territory into three NUTS 2 regions: North West Croatia, Central East (Pannonian) and Adriatic Croatia.

In this research official national division of Croatian territory into twenty one above mentioned counties is accepted. The main intention is to quantify each partial influence of a set of independent regional variables (employment, gross investment, production of more important agricultural products, GVA per person employed, value of construction works, exports, imports, foreign tourists arrivals, foreign tourists nights, ecology…) to regional GDP per capita [10]. These independent regional variables are chosen because it is possible to form complete data base about them. Namely, only observations of these variables are continual published in the official statistical publications. It is necessary to emphasize either at this point that the latest year with complete data for each Croatian county was 2005 and these indicators are taken as inputs for this research.

3.1 Multiple regression model

It is wise to construct complete matrix of correlation coefficients between all observed variables as groundwork for process of estimating multiple model [3]. First interpretations of correlation coefficients from Table 1 lead to essential conclusion that same variables as for example *Production of more important agricultural products* and *Foreign tourists nights* have no important influence on *GDP per capita*. Moreover, the same variables have no important influence neither on all others regressors variables.

The most intensive effect on *GDP per capita* scope has *Exports* with coefficient correlation of 0,850. Excluding this effect of *Exports, Foreign tourist arrivals* has the major influence expressed through partial correlation coefficient.

It is necessary to mention that *Exports* has strong correlation with all other regressors variables while *Foreign tourist arrivals* has no multicolinearity effect.

Multicolinearity effects are the main reason why Stepwise method of selecting regressors variables in the phase of specifying the multiple regression model has selected only two variables as statistical significant for defining *GDP per capita* scope.

In Table 2 there are the previous results of multiple regression model with noted chosen regressors variables: *Exports* and *Foreign tourist arrivals.*

Although there are only two regressors variables R-square shows high percentage of regression sum of squares.

Regarding to Durbin–Watson coefficient there is no autocorrelation problem (at the level of 0,05 Durbin-Watson coefficient is in the inconclusive interval, but at the level of 0,01 it declines to no autocorrelation interval) [5].

Table 1: Correlation matrix of regional variables for Croatian counties in 2005.

Correlations

		GDP per capita, kn	Employment	Gross investment (000 kn)	Production of more important agricultural products (t)	GVA per person employed (kn)	Value of construction works (000 kn)	Exports (000 kn)	Imports (000 kn)	Foreign tourists arrivals	Foreign tourists nights	Ecology (000 kn)
GDP per capita, kn	Pearson Correlation	1,000	,825**	,835**	-,378	,741**	,875**	,850**	,827**	,481*	,419	,814**
	Sig. (2-tailed)		,000	,000	,091	,000	,000	,000	,000	,027	,058	,000
	N	21,000	21	21	21	21	21	21	21	21	21	21
Employment	Pearson Correlation	,825**	1,000	,963**	-,128	,647**	,950**	,966**	,965**	,211	,127	,964**
	Sig. (2-tailed)	,000		,000	,581	,002	,000	,000	,000	,358	,583	,000
	N	21	21,000	21	21	21	21	21	21	21	21	21
Gross investment (000 kn)	Pearson Correlation	,835**	,963**	1,000	-,172	,679**	,909**	,967**	,997**	,094	,015	,994**
	Sig. (2-tailed)	,000	,000		,455	,001	,000	,000	,000	,687	,948	,000
	N	21	21	21,000	21	21	21	21	21	21	21	21
Production of more important agricultural products (t)	Pearson Correlation	-,378	-,128	-,172	1,000	-,146	-,221	-,141	-,177	-,414	-,363	-,150
	Sig. (2-tailed)	,091	,581	,455		,527	,335	,542	,442	,062	,106	,517
	N	21	21	21	21,000	21	21	21	21	21	21	21
GVA per person employed (kn)	Pearson Correlation	,741**	,647**	,679**	-,146	1,000	,723**	,676**	,684**	,320	,304	,638**
	Sig. (2-tailed)	,000	,002	,001	,527		,000	,001	,001	,157	,181	,002
	N	21	21	21	21	21,000	21	21	21	21	21	21
Value of construction works (000 kn)	Pearson Correlation	,875**	,950**	,909**	-,221	,723**	1,000	,917**	,906**	,426	,341	,904**
	Sig. (2-tailed)	,000	,000	,000	,335	,000		,000	,000	,054	,130	,000
	N	21	21	21	21	21	21,000	21	21	21	21	21
Exports (000 kn)	Pearson Correlation	,850**	,966**	,967**	-,141	,676**	,917**	1,000	,970**	,184	,125	,969**
	Sig. (2-tailed)	,000	,000	,000	,542	,001	,000		,000	,423	,589	,000
	N	21	21	21	21	21	21	21,000	21	21	21	21
Imports (000 kn)	Pearson Correlation	,827**	,965**	,997**	-,177	,684**	,906**	,970**	1,000	,082	,007	,992**
	Sig. (2-tailed)	,000	,000	,000	,442	,001	,000	,000		,723	,977	,000
	N	21	21	21	21	21	21	21	21,000	21	21	21
Foreign tourists arrivals	Pearson Correlation	,481*	,211	,094	-,414	,320	,426	,184	,082	1,000	,987**	,078
	Sig. (2-tailed)	,027	,358	,687	,062	,157	,054	,423	,723		,000	,737
	N	21	21	21	21	21	21	21	21	21,000	21	21
Foreign tourists nights	Pearson Correlation	,419	,127	,015	-,363	,304	,341	,125	,007	,987**	1,000	-,004
	Sig. (2-tailed)	,058	,583	,948	,106	,181	,130	,589	,977	,000		,987
	N	21	21	21	21	21	21	21	21	21	21,000	21
Ecology (000 kn)	Pearson Correlation	,814**	,964**	,994**	-,150	,638**	,904**	,969**	,992**	,078	-,004	1,000
	Sig. (2-tailed)	,000	,000	,000	,517	,002	,000	,000	,000	,737	,987	
	N	21	21	21	21	21	21	21	21	21	21	21,000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Source: www.dzs.hr

Table 2: Stepwise method variables selection.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,850[a]	,722	,708	7992,445	
2	,912[b]	,831	,812	6407,022	1,217

a. Predictors: (Constant), Exports (000 000 kn)

b. Predictors: (Constant), Exports (000 000 kn), Foreign tourists arrivals (u 000)

c. Dependent Variable: GDP per capita, kn

Source: www.dzs.hr

In Table 3 there are results of variance analysis. According to all usual classical statistic indicators this multiple regression model of *GDP per capita* is significant at each significance level.

Table 3: Variance analysis results.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,157E9	1	3,157E9	49,426	,000[a]
	Residual	1,214E9	19	6,388E7		
	Total	4,371E9	20			
2	Regression	3,632E9	2	1,816E9	44,240	,000[b]
	Residual	7,389E8	18	4,105E7		
	Total	4,371E9	20			

a. Predictors: (Constant), Exports (000 000 kn)

b. Predictors: (Constant), Exports (000 000 kn), Foreign tourists arrivals (u 000)

c. Dependent Variable: GDP per capita, kn

Source: www.dzs.hr

Estimated parameters in Table 4 shows positive effects of *Exports* and *Foreign tourist arrivals* on *GDP per capita*. Although both regression parameters are statistical significant, according to standardized coefficient it can be concluded that *Exports* has stronger influence on regressand variable. Taking into account collinearity statistics it is obvious that model has no multicolinearity problem.

Table 4: Estimated parameters, confidence intervals and collinearity statistics.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	37080,594	2049,477		18,093	,000	32790,988	41370,199		
	Exports (000 000 kn)	3,077	,438	,850	7,030	,000	2,161	3,993	1,000	1,000
2	(Constant)	34583,930	1799,482		19,219	,000	30803,357	38364,502		
	Exports (000 000 kn)	2,853	,357	,788	7,992	,000	2,103	3,603	,966	1,035
	Foreign tourists arrivals (u 000)	7,558	2,222	,335	3,401	,003	2,889	12,227	,966	1,035

a. Dependent Variable: GDP per capita, kn

Source: www.dzs.hr

Heteroscedasticity problem is tested by parametric Spearman test [4]. Results in Table 5 shows that there is no absolute residual rang correlation with anyone regressor variable.

Whereas normal distribution of standard error is one of the basic assumptions for using parameters estimation method and numerous validity tests, after parameters estimation in this case study residuals normal distribution tests have been done.

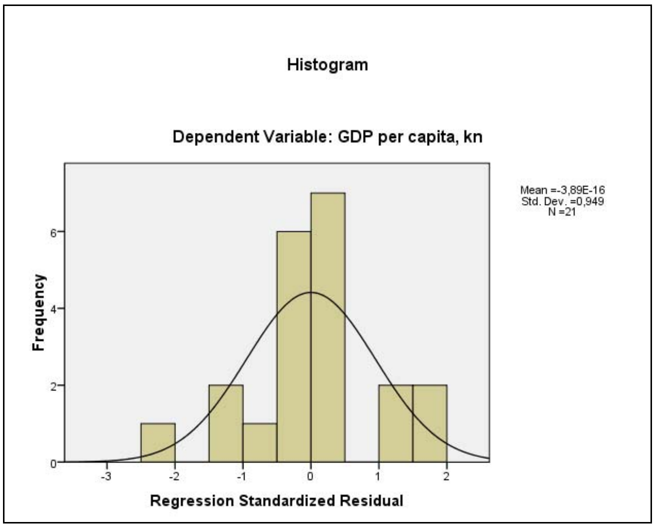
Table 5: Estimated parameters, confidence intervals and collinearity statistics.

Correlations

			Absolut res. 2	Exports (000 kn)	Foreign tourists arrivals
Spearman's rho	Absolut res. 2	Correlation Coefficient	1,000	-,173	,235
		Sig. (2-tailed)	.	,454	,305
		N	21	21	21
	Exports (000 kn)	Correlation Coefficient	-,173	1,000	,352
		Sig. (2-tailed)	,454	.	,118
		N	21	21	21
	Foreign tourists arrivals	Correlation Coefficient	,235	,352	1,000
		Sig. (2-tailed)	,305	,118	.
		N	21	21	21

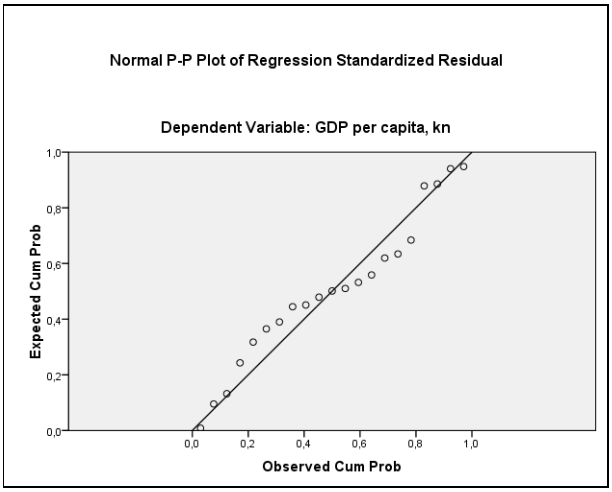
Source: www.dzs.hr

Results of residuals normal distribution testing by Kolmogorov-Smirnov test in this case study are presented in Table 6. Previous at Figure 1 regression standardized residuals are presented and it is obvious that their mean is almost zero with standard deviation equal to unity [14].



Source: www.dzs.hr

Figure 1: Regression standardized residual histogram.



Source: www.dzs.hr

Figure 2: Normal P-P Plot of regression standardized residual.

Table 6: One-Sample Kolmogorov-Smirnov Test of Unstandardized Residual.

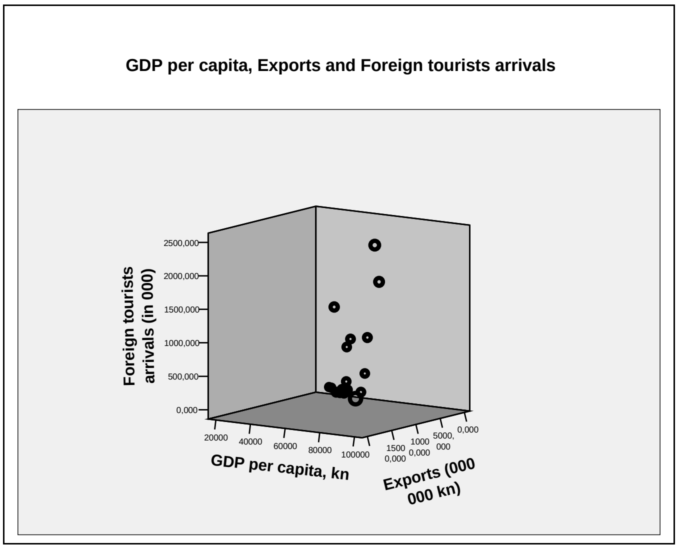
One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual (Finall model)
N		21
Normal Parameters[a]	Mean	,0000000
	Std. Deviation	6,07823454 E3
Most Extreme Differences	Absolute	,121
	Positive	,121
	Negative	-,120
Kolmogorov-Smirnov Z		,554
Asymp. Sig. (2-tailed)		,919

a. Test distribution is Normal.

Source: www.dzs.hr

To fulfill multiple regression validity tests on Figure 2 normal P-P plot of regression standardized residual is shown to confirm that estimated model is significant according to all standard statistical and econometric tests. On Figure 3 there is multidimensional visual presentation of interdependence between variables in final multiple model.



Source: www.dzs.hr

Figure 3: Multidimensional visual presentation of interdependence between GDP per capita, Exports and Foreign tourists arrival.

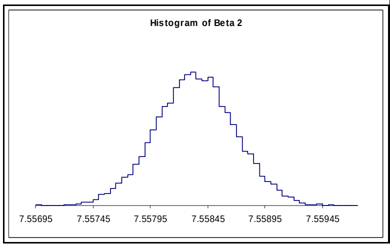
3.2 Multiple Monte Carlo simulations

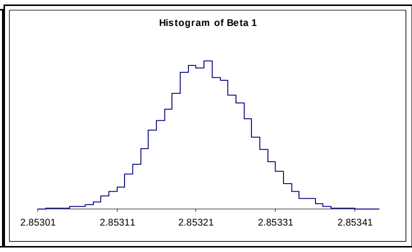
Monte Carlo simulations is based on repeatedly sampling from a chance process. It stands to reason that random numbers are a crucial part of the procedure [2]. Simple but important claim is that Excel, like all other computer software can not draw a true sequence of random numbers. At best, Excel’s random draws can mimic the behavior of truly random draws, but true randomness is unattainable. The inability of computer software to generate truly random numbers results from a computer program’s having to follow a deterministic algorithm to produce its output. If the previous number and the algorithm are known, so is the next number. Because the essence of randomness is that it is not known what is going to happen next, numbers produced by computer software are not genuinely random. Thus, Monte Carlo simulations with Excel are based on pseudorandom number generation [12]. For multiple Monte Carlo simulations in order to estimate multiple econometric model array formulas are necessary instruments. An array formula can perform multiple calculations and then return either a single result or multiple results.

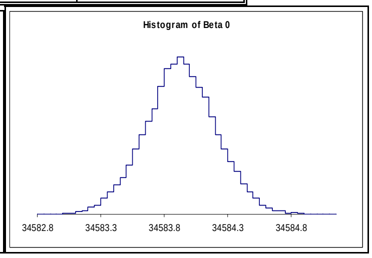
In economic sense multiple Monte Carlo simulations are precious algorithm when data base of observed phenomena is not reach enough and desired mean level with associated dispersion level of the phenomena is defined [9]. According to case study of Croatian counties very often desired level of GDP per capita is known for achieving defined development or similar level. Nowadays it is especially important because of Croatian ambition to become legal EU member.

Table 7: Parameters estimations by multiple Monte Carlo simulations.

Summary Statistics for β_2		Summary Statistics for β_1		Summary Statistics for β_0	
	7,558		2,851		34583,929
SD	0,0003	SD	0,0001	SD	0,2784
Max	7,560	Max	2,853	Max	34585,121
Min	7,557	Min	2,853	Min	34582,804
10000	repetitions	10000	repetitions	10000	repetitions
8	seconds	6	seconds	6	seconds





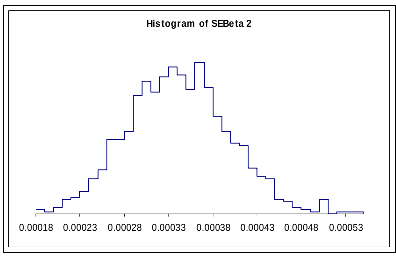


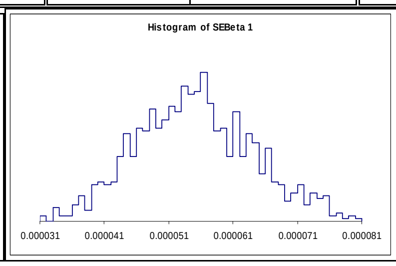
Source: *www.dzs.hr*

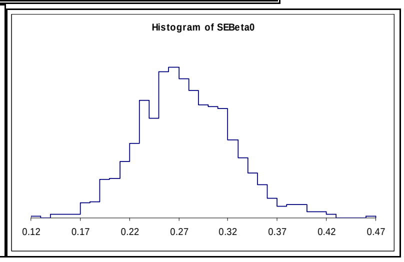
In Table 7 there are parameters estimations of final GDP per capita model by multiple Monte Carlo simulations with 10000 repetitions. Comparative analysis with corresponding estimates given by classic econometric approach shows that both sets of parameter estimations are very similar. Multiple Monte Carlo simulations approach in this case study confirms, like some kind of corrective factor, results given by classic econometric approach. Namely, both methodologies confirm that *GDP per capita* in Croatian counties is explained very well only by *Exports* and *Foreign tourists arrivals* according classic econometric criteria.

Table 8: Parameters standard errors estimations by Monte Carlo simulations.

Summary Statistics for SE β_2		Summary Statistics for SE β_1		Summary Statistics for SE β_0	
	0,00034		0,000055		0,277
SD	0,000057	SD	9,087E-06	SD	0,0479
Max	0,00054	Max	0,00008	Max	0,464
Min	0,00018	Min	0,000031	Min	0,125
1000	repetitions	1000	repetitions	1000	repetitions
1	seconds	1	seconds	1	seconds





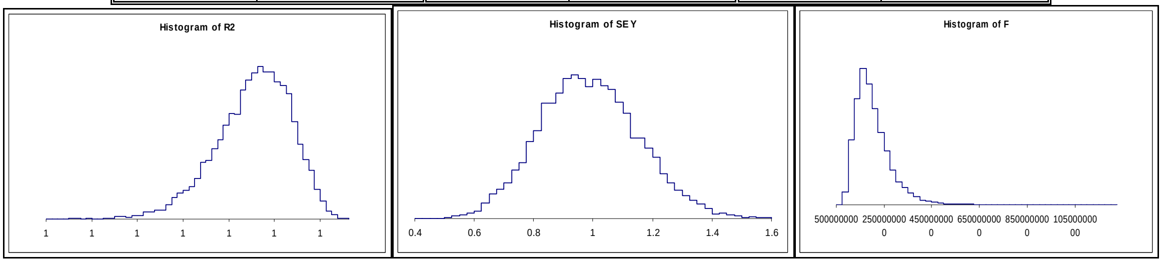


Source: *www.dzs.hr*

This is obvious also from parameters standard errors estimations given in Table 8 with 1000 repetitions, as well as from Table 9 in which there are representative model indicators and F-test results with 10000 repetitions.

Table 9: Representative model indicators and F-test results by Monte Carlo simulations.

Summary Statistics for R^2		Summary Statistics for SEY		Summary Statistics for F	
	0,9999		0,987		2,048E 09
SD	1,6559E-09	SD	1,663	SD	785E 06
Max	0,9999	Max	1,589	Max	12E 09
Min	0,9999	Min	0,415	Min	571E 06
10000	repetitions	10000	repetitions	10000	repetitions
6	seconds	6	seconds	6	seconds

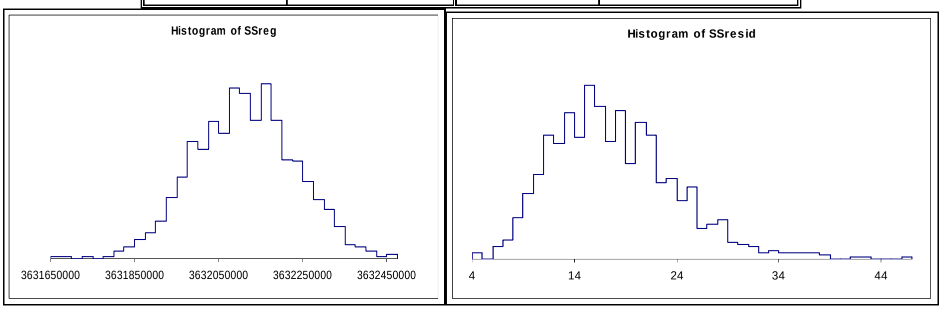


Source: *www.dzs.hr*

Comparative analysis results are confirmed also by sum of squared residuals (SSresid) and by sum of squared regression (SSreg) results given by multiple Monte Carlo simulations in Table 10 with 1000 repetitions.

Table 10: SSresid and SSreg results by Monte Carlo simulations.

Summary Statistics for $SSreg$		Summary Statistics for $SSresid$	
	3,6E 09		17,991
SD	120566,54	SD	6,1086
Max	3,63E 09	Max	46,260
Min	0,9999	Min	4,785
1000	repetitions	1000	repetitions
0	seconds	0	seconds



Source: *www.dzs.hr*

4 CONCLUSION REMARKS

This paper as a part of en extensive research about Croatian economic challenges within global recession environment has estimation of integral Croatian counties GDP per capita model in its focus. The main intention is to quantify each partial influence of a set of independent regional variables (employment, gross investment, production of more important agricultural products, GVA per person employed, value of construction works, exports, imports, foreign tourists arrivals, foreign tourists nights, ecology…) to regional GDP per capita. Stepwise variable selection method points out that regional GDP per capita is very well explained only by two regional variables: Exports and Foreign tourist arrival. Namely, correlation analysis shows strong multicolinearity effects between Exports and all other regressors variables. Regional Croatian counties GDP per capita model is estimated by classical multiple regression approach and then by multiple Monte Carlo simulations.

The comparative analysis of the results confirms the advantage of Monte Carlo simulation in evaluation of interdependence especially regarded on Gauss-Markov conditions on stochastic part of econometric model. For this regional analysis multiple Monte Carlo simulation of regression model is used for the first time to measure individual average influence of each regressor variable on the GDP per capita.

Multivariate approach of multiple Monte Carlo simulations is essentially incorporate in the logic of Central limit theorem because of huge number of repetitions involved in its computing. This enables multiple Monte Carlo simulations algorithm to be close as possible to real values of parameters estimated.

References

[1] Anderson, T. W., 2003. An Introduction to Multivariate Statistical Analysis (Wiley Series in Probability and Statistics), Wiley-Interscience, 711 p.

[2] Barreto, H., Howland, F. M., 2006. Introductory Econometrics, Using Monte Carlo Simulation with Microsoft Excel. Cambridge, University Press, 774 p.

[3] Enders, W., 2003. Applied Econometrics Time Series. Second edition, New York, John Wiley & Sons, 467 p.

[4] Fang, K. T., Yao-Ting, Z., 1990. Generalized Multivariate Analysis. Springer-Verlag, New York, 483 p.

[5] Harrell, F. H., 2001. Regression Modeling Strategies. With Application to Linear Models, Logistic Regression, and Survival Analysis. Springer, 568 p.

[6] Hill, G., Lim, 2008. Principles of Econometrics. Third edition, New York, John Wiley & Sons, 513 p.

[7] Hull, J. C., 2006. Options, futures and other derivatives. Forth edition. Prentice Hall, 683 p.

[8] Johnson, R. A., Wichern, D. W., 2002. Applied Multivariate Statistical Analysis. Prentice Hall, 497 p.

[9] Jurun, E., Pivac, S., Arnerić, J., 2006. Primijenjena ekonometrija 1, Kvantitativne financije. University of Split, Faculty of economics, 138 p.

[10] Jurun, E., Arnerić J., Pivac S., 2008. Multivariate Risk-Return Decision Making within Dynamic Estimation, Economic Analysis Working Papers, Vol.7, No.11, pp.1-11.

[11] Jurun, E., Arnerić J., Pivac S., 2007. Stock Prices Tehnical Analysis; In: Zadnik Stirn, L., Drobne, S. (eds.), Proceedings of the 9th International Symposium on Operational Research SOR'07, Nova Gorica, Slovenia, pp.397-402.

[12] de Levie, R. 2004. Advanced Excel for Scientific Data Analysis, Oxford University Press, 615 p.

[13] Knusel, L. 2004. On the Accuracy of Statistical Distributions in Microsoft Excel 2003, Computational Statistics & Data Analysis, Vol. 48, No 3, pp. 445-119.

[14] McCloskey, N., Ziliak, S. T., 1996. The Standard Error of Regressions, Journal of Economic Literature, Vol. 34, No 1, pp. 97-114.

[15]Verbeek, M., 2005. A Guide to Modern Econometrics. Second edition, London, John Wiley & Sons, 429 p.

[16]Wooldrige, J. M., 2000. Introductory Econometrics: A Modern Approach. Second edition, South-Western: Cincinnati, 805 p.

SOME SECOND GENERATION PANEL UNIT ROOT TESTS FOR INVESTIGATING THE VALIDITY OF PURCHASING POWER PARITY

Alenka Kavkler, **Darja Boršič** and **Jani Bekő**
Faculty of Economics and Business, University of Maribor
Razlagova 14, 2000 Maribor, Slovenia
e-mail : {alenka.kavkler, darja.borsic, jani.beko}@uni-mb.si

Abstract: The aim of this paper is to present and discuss three of the second generation panel unit root tests, namely Moon and Perron test, Choi test and Pesaran test. Second generation panel unit root tests alleviate the cross-sectional independency assumption that is rather restrictive and somewhat unrealistic in the majority of macroeconomic applications. As an illustration of the described theoretical approach, we investigate the validity of purchasing power parity for a group of 12 transition countries with respect to dollar and euro by covering the period of 1994–2008.

Keywords: panel unit root tests, purchasing power parity, real exchange rates.

1 INTRODUCTION

The study of unit roots has played an increasingly important role in empirical analysis of panel data. The investigation of integrated series in panel data has reached a great development and panel unit root tests have been applied to various fields of economics: analysis of the purchasing power parity (PPP) hypothesis, growth and convergence issues, saving and investment dynamics, international R&D spillovers.

The first generation of panel unit root tests is based on the cross-sectional independency hypothesis: Levin et al. [16], Breitung [2], Im et al. [14], Maddala and Wu [17], Choi [5], Hadri [11]. Within this context, correlations across units constitute nuisance parameters. The cross-sectional independency hypothesis is rather restrictive and somewhat unrealistic in the majority of macroeconomic applications of unit root tests, like the studies of convergence or the analysis of PPP where co-movements of data series are often observed.

The second generation of panel unit root tests is characterized by the rejection of the cross-sectional independence hypothesis. Within this second generation of tests, two main approaches are distinguished. The first one relies on the factor structure approach and includes the contributions of Moon and Perron [18], Choi [6], Pesaran [21] and Bai and Ng [1]. The second approach consists of imposing few or none restrictions on the residuals covariance matrix and has been adopted notably by Chang [3] and [4], who proposed the use of nonlinear instrumental variables methods or the use of bootstrap approaches to solve the nuisance parameter problem due to cross-sectional dependency. A comprehensive overview of second generation unit root tests is provided by Hurlin and Mignon [13].

In this study, we discuss and employ three of the second generation panel unit root tests for inspecting the validity of purchasing power parity in a heterogeneous group of 12 transition countries with respect to US dollar and euro. The paper proceeds as follows. Section 2 examines second generation panel unit root tests. Three of the tests are described in detail, namely Moon and Perron test, Choi test and Pesaran test. In Section 3, the general model of PPP is presented and the testing procedure is elaborated. Results of the empirical analysis are discussed and interpreted in Section 4. Concluding remarks of our study are given in the final section.

2 SECOND GENERATION PANEL UNIT ROOT TESTS

The general setting for panel unit root testing takes into account the following first order autoregressive process (AR(1)):

$$y_{i,t} = \rho_i y_{i,t-1} + X_{i,t}\delta_i + \varepsilon_{i,t}\,, \tag{1}$$

where i represents one of the N cross-section units observed over periods $t=1, 2, ..., T$, $X_{i,t}$ are exogenous variables in the model (any fixed effects or individual trends), ρ_i are autoregressive coefficients, while $\varepsilon_{i,t}$ are idiosyncratic disturbances. Under the cross-sectional independence hypothesis of first generation unit root tests, the disturbances are assumed to be independent across the units. If absolute value of the autoregressive coefficient ρ_i is less then 1, y_i is said to be weakly stationary. If the absolute value of the autoregressive coefficient is 1, y_i contains a unit root. Two different assumptions about the autoregressive coefficients can be made: (1) the coefficients are common across cross-sections ($\rho_i=\rho$ for all i), and (2) ρ_i vary across cross-sections. The first assumption implies a common unit root process, while in the second case individual unit root processes are assumed.

The second generation of panel unit root tests is characterized by the rejection of the cross-sectional independence hypothesis. The cross-sectional dependencies can be specified by using a factor structure model (Moon and Perron [18], Choi [6], Pesaran [21], Bai and Ng [1]). Alternatively, cross-sectional dependencies may be achieved by imposing restrictions on the covariance matrix of residuals (Chang [4]). Hurlin and Mignon [13] provide a discussion on the recent developments relating to panel unit root tests. In our paper, Moon and Perron test, Choi test and Pesaran test are described in more detail. We also attempted to compute the Chang test statistics, but the results are not reported here due to problems with near-singular matrices. Bai and Ng test is complicated, performed in several steps and sensitive to different assumptions and is therefore omitted.

2.1 Moon and Perron test

Moon and Perron [18] consider a dynamic panel model of the form

$$y_{i,t} = \alpha_i + y^0_{i,t} \tag{2}$$
$$y^0_{i,t} = \rho_i y^0_{i,t-1} + \varepsilon_{i,t}$$

that allows for fixed effects. We adopted some of the notation from Hurlin [12]. The null hypothesis of unit root $H_0 : \rho_i = 1, \forall i$ is tested against the alternative that at least one cross-sectional unit is stationary, i.e. $H_1 : \rho_i < 1$ for some i. To model the cross-sectional correlation, Moon and Perron assume that the error terms in equation (2) are generated by an unknown number of common factors:

$$\varepsilon_{i,t} = \beta_i' F_t + e_{i,t}\,, \tag{3}$$

where F_t is an r-dimensional vector of common factors, β_i is the vector of factor loadings and $e_{i,t}$ are idiosyncratic shocks. Let us assume that $e_{i,t} = \sum_{j=0}^{\infty} d_{ij} v_{i,t-j}$, where $v_{i,t}$ are iid(0,1) across i and over t, have a finite eighth moment. We define $\sigma^2_{e,i} = \sum_{j=0}^{\infty} d^2_{ij}$, $\omega^2_{e,i} = \left(\sum_{j=0}^{\infty} d_{ij}\right)^2$ and $\lambda_{e,i} = \sum_{l=0}^{\infty}\sum_{j=0}^{\infty} d_{ij} d_{ij+l}$. $\sigma^2_{e,i}$ is the variance of $e_{i,t}$, $\omega^2_{e,i}$ is the long-run variance and $\lambda_{e,i}$ the one-sided long-run variance of $e_{i,t}$. The following four limits are supposed to be well defined as $N \to \infty$: $\omega_e^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \omega_{e,i}^2$, $\Phi_e^4 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \omega_{e,i}^4$, $\sigma_e^2 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \sigma_{e,i}^2$ and $\lambda_e = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N \lambda_{e,i}$. It is easier to work with matrix notatation:

$$Z_i = \left(y_{i,1},\dots,y_{i,T}\right)',\ Z_{-1,i} = \left(y_{i,0},\dots,y_{i,T-1}\right)',\ Z = \left(Z_1,\dots,Z_N\right),\ Z_{-1} = \left(Z_{-1,1},\dots,Z_{-1,N}\right). \quad (4)$$

The pooled autoregressive estimator can now be computed as

$$\hat{\rho}_{pool} = \frac{tr\left(Z'_{-1}Z\right)}{tr\left(Z'_{-1}Z_{-1}\right)}. \quad (5)$$

Because of the cross-sectional correlation due to common factors the conventional central limit theorem cannot be applied. Moon and Perron propose to eliminate the common factors with the help of the projection matrix

$$Q_\beta = I - \beta\left(\beta'\beta\right)^{-1}\beta'. \quad (6)$$

The i-th row of the matrix β is the vector β_i'. The data is thus projected onto the space orthogonal to the factor loadings. The modified pooled OLS estimator on the de-factored data is defined as

$$\hat{\rho}_{pool}^{+} = \frac{tr\left(Z_{-1}Q_\beta Z'\right) - NT\lambda_e^N}{tr\left(Z_{-1}Q_\beta Z'_{-1}\right)} \quad (7)$$

with $\lambda_e^N = \frac{1}{N}\sum_{i=1}^N \lambda_{e,i}$. Moon and Perron calculate two modified t-statistics

$$t_a = \frac{\sqrt{N}T\left(\hat{\rho}_{pool}^{+} - 1\right)}{\sqrt{2\Phi_e^4 / \omega_e^4}} \quad (8)$$

$$t_b = \sqrt{N}T\left(\hat{\rho}_{pool}^{+} - 1\right)\sqrt{\frac{1}{NT^2} tr\left(Z_{-1}Q_\beta Z'_{-1}\right)\frac{\omega_e^2}{\Phi_e^4}}$$

that have a standard normal distribution under the null hypothesis. t_a and t_b are not feasible as they depend on unknown parameters. Consistent estimators of λ_e^N, ω_e^2 and Φ_e^4 can be obtained as follows:

$$\hat{\lambda}_{e,i} = \sum_{j=1}^{T-1} w\left(\frac{j}{h}\right)\hat{\Gamma}_i(j), \quad \hat{\omega}_{e,i}^2 = \sum_{j=-T+1}^{T-1} w\left(\frac{j}{h}\right)\hat{\Gamma}_i(j). \quad (9)$$

w denotes a kernel function and h stands for bandwidth parameter, while the sample covariance $\hat{\Gamma}_i(j)$ is computed as $\hat{\Gamma}_i(j) = \frac{1}{T}\sum_t \hat{e}_{i,t}\hat{e}_{i,t+j}$. The final estimators are of the form

$$\hat{\lambda}_e^N = \frac{1}{N}\sum_{i=1}^N \hat{\lambda}_{e,i},\ \hat{\omega}_e^2 = \frac{1}{N}\sum_{i=1}^N \hat{\omega}_{e,i}^2 \text{ and } \hat{\Phi}_e^4 = \frac{1}{N}\sum_{i=1}^N \hat{\omega}_{e,i}^4. \quad (10)$$

Quadratic spectral kernel is used in our computations and optimal bandwidth is chosen according to the Newey and West [19] procedure. We determine the number of common factors (r) with the help of the Schwarz information criterion, while the maximal possible number of common factors is set to 12. Moon and Perron [18] also derive the asymptotic normal distribution of the two tests under the null hypothesis and study their asymptotic power. Several assumptions of a technical nature have to hold for the derivation of the asymptotic distribution (see Moon and Perron [18] for details).

2.2 Choi test

Choi [6] models the cross-sectional correlation with the help of the error-component models. The two-way error-component model can be expressed as

$$y_{i,t} = \alpha_0 + x_{i,t} \qquad (11)$$
$$x_{i,t} = \alpha_i + \lambda_t + v_{i,t}$$
$$v_{i,t} = \sum_{l=1}^{q_i} a_{il} v_{i,t-l} + e_{i,t} \, ,$$

where α_0 is the common mean for all i, α_i denotes the individual effect, λ_t is the time effect and $v_{i,t}$ is the random component modeled as autoregressive process of order q_i. The null hypothesis of unit root, $H_o : \sum_{l=1}^{q_i} a_{il} = 1, \forall i$, is tested against the alternative hypothesis $H_1 : \sum_{l=1}^{q_i} a_{il} < 1$ for at least one i. The constant term, the nonstochastic trend component and the cross-sectional correlations are first eliminated by Elliott et al. [9] GLS-based detrending and the conventional cross-sectional demeaning for panel data. Let us assume that the largest root of $v_{i,t}$ is equal to $1 + \frac{c}{T}$ for all i. If we regress $\left(y_{i,1}, y_{i,2} - \left(1 + \frac{c}{T}\right) y_{i,1}, \ldots, y_{i,T} - \left(1 + \frac{c}{T}\right) y_{i,T-1} \right)'$ on $\left(1, 1 - \left(1 + \frac{c}{T}\right), \ldots, 1 - \left(1 + \frac{c}{T}\right) \right)'$, we derive the GLS estimator of parameter α_0 in model (11) and denote it $\hat{\alpha}_{0i}$. Following Elliott et al. [9], Choi [6] also sets $c = -7$ for all i. After demeaning $y_{i,t} - \hat{\alpha}_{0i}$, one obtains

$$z_{i,t} = y_{i,t} - \hat{\alpha}_{0i} - \frac{1}{N} \sum_{i=1}^{N} \left(y_{i,t} - \hat{\alpha}_{0i} \right). \qquad (12)$$

The deterministic components α_i and λ_t are eliminated from the series $y_{i,t}$. As pointed out by Hurlin [12], the transformed variables $z_{i,t}$ are thus independent across i for large T and N.

The main idea employed by Choi [6] is to combine p-values from a unit root test of each time series. The augmented Dickey-Fuller (ADF) test is applied to detrended series and the p-values are calculated. From regression

$$\Delta z_{i,t} = \theta_{i0} z_{i,t-1} + \sum_{l=1}^{q_i - 1} \theta_{il} \Delta z_{i,t-l} + u_{i,t} \qquad (13)$$

the t-ratio statistic for coefficient estimate $\hat{\theta}_{i0}$ is derived. This test is called the Dickey-Fuller GLS test and follows the Dickey-Fuller distribution. The proposed pooled unit root test statistics P_m, Z and L^* combine the p-values of the ADF test applied to individual time series:

$$P_m = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\ln(p_i) + 1 \right) \qquad (14)$$
$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Phi^{-1}(p_i)$$
$$L^* = \frac{1}{\sqrt{\pi^2 N/3}} \sum_{i=1}^{N} \ln\left(\frac{p_i}{1 - p_i} \right).$$

p_i denotes asymptotic p-value of the Dickey-Fuller GLS test for cross-section i. $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. Under the null hypothesis, all three statistics are normally distributed.

2.3 Pesaran test

Pesaran [21] starts with a simple dynamic panel model of the form

$$y_{i,t} = (1-\rho_i)\mu_i + \rho_i y_{i,t-1} + \varepsilon_{i,t} . \quad (15)$$

The cross-section dependence is imposed by assuming a single-factor structure of the error term

$$\varepsilon_{i,t} = \beta_i f_t + e_{i,t} , \quad (16)$$

where f_t is the unobserved common effect and $e_{i,t}$ is the idiosyncratic error. The last two equations can be rewritten as

$$\Delta y_{i,t} = \alpha_i + \theta_i y_{i,t-1} + \beta_i f_t + e_{i,t} \quad (17)$$

with $\alpha_i = (1-\rho_i)\mu_i$ and $\theta_i = \rho_i - 1$. The unit root hypothesis $\rho_i = 1, \forall i$, is equivalent to $H_o : \theta_i = 0, \forall i$, which is tested against the alternative that at least one cross-sectional unit is stationary ($H_1 : \theta_i < 0$ for some i). Pesaran [21] argues that the common factor f_t can be proxied by the cross-section mean of $y_{i,t}$, namely $\bar{y}_t = \frac{1}{N}\sum_{j=1}^{N} y_{j,t}$, and its lagged values $\bar{y}_{t-1}, \bar{y}_{t-2}, \ldots$ for N sufficiently large. If the residuals $\varepsilon_{i,t}$ are not serially correlated, the effects of the common factor can be filtered out using only $\bar{y}_t$ and $\bar{y}_{t-1}$, or equivalently $\bar{y}_{t-1}$ and $\Delta\bar{y}_t$. The following cross-sectionally augmented Dickey-Fuller (CADF) regression is obtained:

$$\Delta y_{i,t} = a_i + h_i y_{i,t-1} + c_i \bar{y}_{t-1} + d_i \Delta\bar{y}_t + e_{i,t} . \quad (18)$$

Let us denote the t-statistic of the OLS estimate of h_i by $t_i(N,T)$. This t-ratio is the so-called cross-sectionally augmented Dickey-Fuller (CADF) statistic for the i-th cross section. Following Im et al. [14], Pesaran [21] defines the *CIPS* panel unit root statistic as the average of individual CADF t-type statistics:

$$CIPS = \frac{1}{N}\sum_{i=1}^{N} t_i(N,T) . \quad (19)$$

To avoid large influence of extreme outcomes for samples with small *T*, Pesaran also consideres a truncated version of the proposed panel unit root statistic calculated as

$$CIPS^* = \frac{1}{N}\sum_{i=1}^{N} t_i^*(N,T) . \quad (20)$$

$t_i^*(N,T)$ is a truncated CADF statistic given by

$$t_i^*(N,T) = \begin{cases} t_i\,(N,T), & \text{if } -K_1 < t_i(N,T) < K_2 \\ -K_1, & \text{if } t_i(N,T) \le -K_1 \\ K_2, & \text{if } t_i(N,T) \ge K_2 \end{cases} \quad (21)$$

K_1 and K_2 are positive constants such that the probability $P\left(-K_1 < t_i(N,T) < K_2\right)$ is sufficiently high, usually more than 0.9999. The generalization of this approach to serially correlated residuals is straightforward and requires introduction of the lagged cross-section mean values into equation (18).

The critical values of the *CIPS* and *CIPS** statistics are tabulated. If more than one lag is taken into account in auxiliary regression (18), the *CIPS* and *CIPS** statistics coincide. In our empirical study, we set the maximal lag length of auxiliary regression to 12 since we employ monthly data. Pesaran [21] also derives asymptotic results for both CADF and CADF* statistics.

3 PURCHASING POWER PARITY

The theory of purchasing power parity (PPP) states that in case of absolute PPP the exchange rate equals the relative price level between countries, whereas in case of relative PPP the exchange rate movements equal the difference between the relative price level shifts. Although the majority of empirical tests has produced rather mixed outcomes, researchers agree on two issues related to this exchange rate theory (Rogoff [22]): first, real exchange rates tend to converge on levels predicted by PPP in the long-run; and second, short-run deviations from the PPP relationship are substantial and variable.

The basic model of testing for relative PPP can be derived in the following form (Froot and Rogoff [10]):

$$e_t = \alpha_0 + \alpha_1 p_t + \alpha_2 p_t^* + \xi_t, \tag{22}$$

where e_t stands for nominal exchange rates, defined as the price of foreign currency in the units of domestic currency; p_t denotes domestic price index and p_t^* foreign price index; while ξ_t stands for the error term showing deviations from PPP. All the variables are given in logarithmic form. The strict version of PPP contains two types of restrictions imposed on the parameters. Under $\alpha_0=0$, the symmetry restriction applies such that α_1 and α_2 are equal in absolute terms, whereas the limitation of α_1 and α_2 being equal to 1 and -1, respectively, is called the proportionality restriction.

4 EMPIRICAL ANALYSIS

In the present study we relied on relevant monthly data frequency covering the period of January 1994–December 2008 for the following countries: Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Macedonia, Poland, Romania, Slovakia and Slovenia. Primary data included monthly averages of nominal exchange rates and consumer price indices gathered from the central banks of individual countries, from the European Central Bank, Eurostat, and from national statistical offices of individual countries. Each of the exchange rates has been defined as the number of units of domestic currency for the dollar and for the euro. Consumer price indices used in this study for all countries refer to January 1994.

The empirical analysis usually starts off with testing the characteristics of real exchange rates (strict version of equation (22)). Following relative PPP, the movements in nominal exchange rates are expected to compensate for price level shifts. Thus, real exchange rates should be constant over the long-run and their time series should be stationary (Parikh and Wakerly [20]). The real exchange rates are a function of nominal exchange rates and relative price indices in two observed economies. They are calculated from the nominal exchange rates using the consumer price indices:

$$RE_t = E_t\,(P_t^*/P_t), \tag{23}$$

where RE_t stands for the real exchange rate, E_t is the price of a foreign currency in units of the domestic currency, and P_t^* and P_t represent the foreign price index and the domestic price index, respectively. Taking the logarithms of equation (23), the real exchange rates are defined as:

$$re_t = e_t + p_t{}^* - p. \tag{24}$$

Thus, the validity of PPP can be investigated by testing for stationarity of real exchange rates. The growing popularity of panel unit root tests in empirical PPP studies is due to higher power of the panel tests.

Figure 1 and Figure 2 present the graphs of real exchange rates of the selected 12 Central and Eastern European economies in comparison to the US dollar and the euro, respectively. As can be seen from Figure 1 and Figure 2, the real exchange rates of selected countries with respect to the US dollar and as well as to the euro appear to be marked by appreciation trends, although phases of relative stability of real exchange rates are not uncommon in the sample observed (see, e.g., Estonia, Latvia, and Lithuania against the US dollar and the euro; Macedonia, Romania, and Slovenia against the euro). The pattern of relative (in)stability in real exchange rate movements is explained in the literature by a range of factors, including inherited macroeconomic imbalances in transition countries, mixed performance of chosen exchange rate arrangements, monetary difficulties arising from huge capital inflows, the inflationary impact of wage and price adjustments, and real exchange rate appreciation due to the catching-up process (Égert et al. [8]).

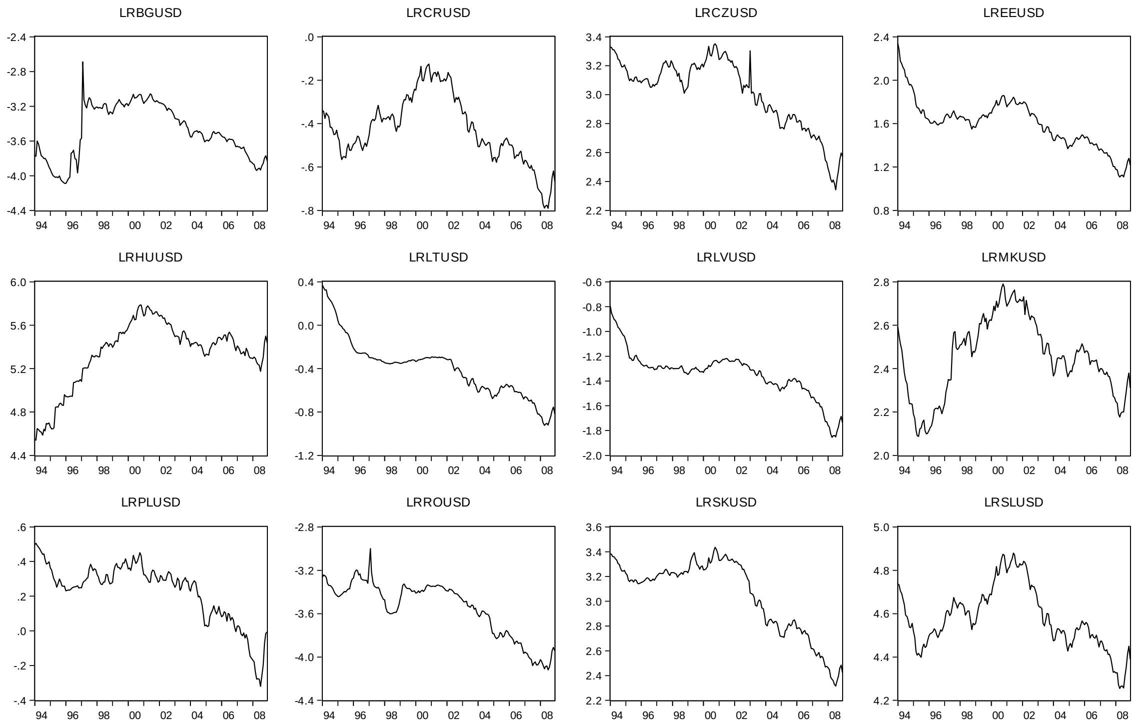


Figure 1: Real Exchange Rates of 12 Central and Eastern European Economies in Comparison to the US Dollar

Notes: L stands for logarithm, R for real; the next two letters represent the abbreviation of the name of selected Central and Eastern European countries; and the last three letters denote the US dollar.

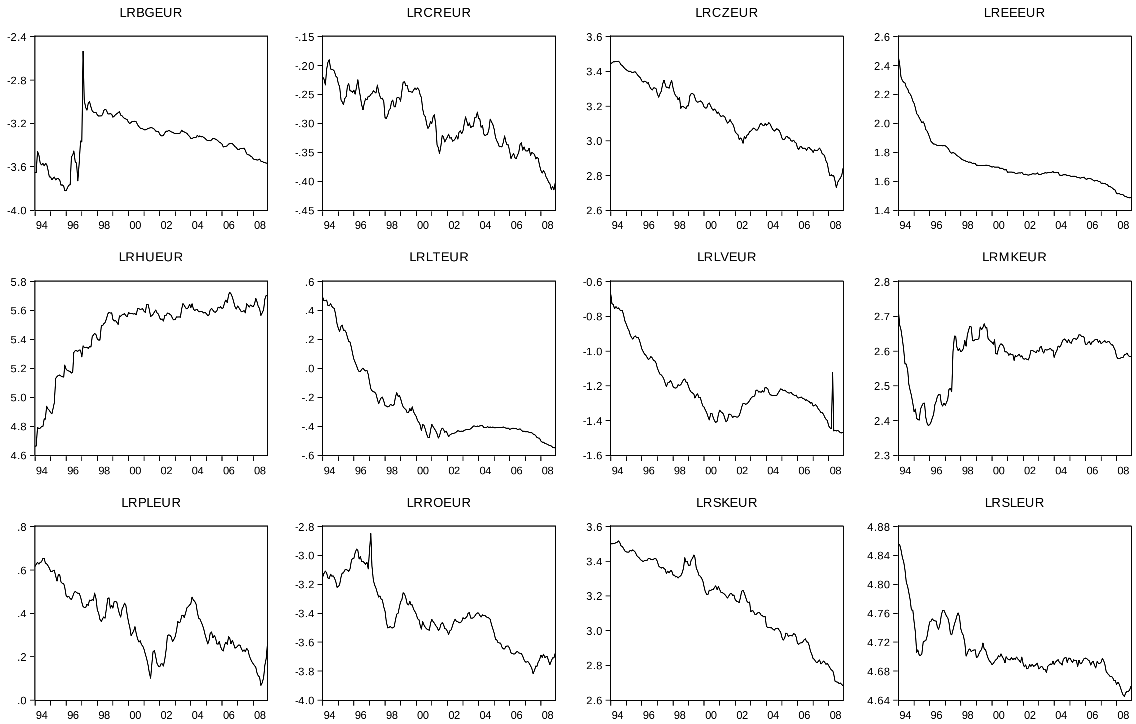


Figure 2: Real Exchange Rates of 12 Central and Eastern European Economies in Comparison to the Euro

Notes: L stands for logarithm, R for real; the next two letters represent the abbreviation of the name of selected Central and Eastern European countries; and the last three letters denote the euro.

The results of Moon and Perron test are given in Table 4, with p-values in brackets. The t_a^* and t_b^* statistics reject the null hypothesis of unit root for both reference currencies.
The results of all three Choi-type panel unit root tests in Table 5 are similar and also in favor of PPP. Maximal lag order of individual ADF regressions is set to 12 due to monthly data. The null hypothesis of unit root cannot be rejected only in case of Pesaran test for the US dollar based panel of real exchange rates (Table 6).

Table 4: Results of Moon and Perron test

Reference currency	t_a^*	t_b^*
USD	-9.5543 (0.0000)	-13.2711 (0.0002)
EUR	-10.9973 (0.0000)	-11.8394 (0.0000)

Table 5: Results of Choi test

Reference currency	P_m	Z	L^*
USD	2.0917 (0.0182)	-2.3458 (0.0095)	-2.2702 (0.0116)
EUR	4.9393 (0.0000)	-3.2344 (0.0006)	-3.6799 (0.0001)

Table 6: Results of Pesaran test

Reference currency	*CIPS*
USD	-2.2307 (0.0950)
EUR	-2.8100 (0.0100)

5 CONCLUSION

In the present study we discussed three of the second generation panel unit root tests and employed them to check the validity of PPP in a sample of 12 transition countries. Testing for the stationarity of real exchange rates of selected countries showed firm evidence in favor of PPP in comparison to the euro, whereas the stationarity of real exchange rates in panel against the US dollar was not suggested by Pesaran test.

In contrast to rather weak empirical evidence in favor of PPP for transition countries found in the early papers, results from recent studies seem to provide more support for PPP. This can be attributed to at least three factors. First, availability and extensive employment of longer data sets for transition countries by researchers. Second, the growing popularity of panel data approach that uses both time series and cross-section observations increasing thereby the power of unit root tests (Taylor [24]). And third, the authors pay special attention to the possibility of nonlinearities in the formation of real exchange rates. Ignoring nonlinearities in the behavior of real exchange rates might bias the testing results towards the failure of PPP (Cuestas [7]).

Since the results found in this article complement the empirical evidence reported in recent studies on PPP in reforming economies (see, e.g., Solakoglu [23] and Koukouritakis [15]), we plan to expand our analysis in two directions. The first reasonable extension is to re-examine the PPP theory with a panel unit root test in a SUR framework by keeping the period length and monthly frequency of data consistent with the present study. The second direction for research involves searching for a long-run linear relationship among chosen nominal exchange rates and individual time series of consumer prices by a panel cointegration approach.

References

[1] Bai, J. and S. Ng. 2001. A PANIC Attack on Unit Roots and Cointegration, Mimeo, Boston College, Department of Economics.

[2] Breitung, J. 2000. The Local Power of Some Unit Root Tests for Panel Data. In: B. Baltagi, ed. Advances in Econometrics, Vol. 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels. Amsterdam: JAI Press, 161-178.

[3] Chang, Y. 2002. Nonlinear IV Unit Root Tests in Panels with Cross-Sectional Dependency. – *Journal of Econometrics*, 110, 261-292.

[4] Chang, Y. 2004. Bootstrap Unit Root Tests in Panels with Cross-Sectional Dependency. – *Journal of Econometrics*, 120, 249-272.

[5] Choi, I. 2001. Unit Root Tests for Panel Data. – *Journal for International Money and Finance*, 20, 249-272.

[6] Choi, I. 2002. Combination Unit Root Tests for Cross-Sectionally Correlated Panels, Mimeo, Hong Kong University of Science and Technology.

[7] Cuestas, J. C. 2009. Purchasing Power Parity in Central and Eastern European Countries: An Analysis of Unit Roots and Nonlinearities. – *Applied Economics Letters*, 16, 1, 87–94.

[8] Égert, B., Halpern, L. and R. MacDonald. 2006. Equilibrium Exchange Rates in Transition Economies: Taking Stock of the Issues. – *Journal of Economic Surveys*, 20, 2, 257–324.

[9] Elliott, G., Rothenberg, T. and J. Stock. 1996. Efficient Tests for an Autoregressive Unit Root. – *Econometrica*, 64, 813-836.

[10] Froot, K. A. and K. Rogoff. 1995. Perspectives on PPP and Long-Run Real Exchange Rates. In Grossman, G. and K. Rogoff, eds. Handbook of International Economics Vol. III, Elsevier Science, 1647–1688.

[11] Hadri, K. 2000. Testing for Stationarity in Heterogenerous Panel Data. – *Econometric Journal*, 3, 148-161.

[12] Hurlin, C. 2007. What would Nelson and Plosser find had they used panel unit root tests? *HAL Working Papers*, No. 00156685 v1.

[13] Hurlin, C. and V. Mignon. 2007. Second Generation Panel Unit Root Tests. *HAL Working Papers*, No. 00159842 v1.

[14] Im, K. S., Pesaran, M. H. and Y. Shin. 2003. Testing for Unit Roots in Heterogenerous Panels. –*Journal of Econometrics*, 115, 53-74.

[15] Koukouritakis, M. 2009. Testing the Purchasing Power Parity: Evidence from the New EU Countries. – *Applied Economics Letters*, 16, 1, 39-44.

[16] Levin, A., Lin, C. F. and C. Chu. 2002. Unit Root Testing in Panel Data: Asymptotic and Finite-Sample Properties. – *Journal of Econometrics*, 54, 159-178.

[17] Maddala, G. S. and S. Wu. 1999. A Comparative Study of Unit Root Tests with Panel and a New Simple Test. – *Oxford Bulletin of Economics and Statistics*, 61, 631-653.

[18] Moon, H. R. and B. Perron. 2004. Testing for Unit Root in Panels with Dynamic Factors. – *Journal of Econometrics*, 122, 81-126.

[19] Newey, W. and K. West. 1994. Automatic Lag Selection and Covariance Matrix Estimation. – *Review of Economic Studies*, 61, 631-653.

[20] Parikh, A. and E. Wakerly. 2000. Real Exchange Rates and Unit Root Tests. – *Weltwirtschaftliches Archiv*, 136, 3, 478–490.

[21] Pesaran, M. H. 2007. A Simple Panel Unit Root Test in the Presence of Cross-Section Dependece. – *Journal of Applied Econometrics*, 22, 2, 265-312.

[22] Rogoff, K. 1996. The Purchasing Power Parity Puzzle. – *Journal of Economic Literature*, 34, 2, 647–668.

[23] Solakoglu, E. G. 2006. Testing Purchasing Power Parity Hypothesis for Transition Economies. – *Applied Financial Economics*, 16, 7, 561–568.

[24] Taylor, M. P. 2006. Real Exchange Rates and Purchasing Power Parity: Mean-Reversion in Economic Thought. – *Applied Financial Economics*, 16, 1–2, 1–17.

MULTIVARIATE ANALYSIS OF ENTREPRENEURIAL PRODUCTIVITY: CASE OF CROATIA

Nataša Kurnoga Živadinović
Petar Sorić
University of Zagreb - Faculty of Economics & Business
Trg J. F. Kennedyja 6
10000 Zagreb, Croatia
nkurnoga@efzg.hr
psoric@efzg.hr

Abstract: The aim of this paper was to analyze the relationship between Croatian enterprises' size and productivity. A thousand top Croatian companies were analyzed regarding two indicators: revenue per employee and profit per employee. Using K-means clustering method enterprises were grouped into three clusters. It was found that with a rise in employment level enterprises tend to get less productive and vice versa. Hence, the observed relationship is found to be reversely proportional. Cluster analysis results were corroborated using discriminant analysis. Dummy variable regression models were also formed to give answers about the strength and magnitude of the relationship of interest. It was confirmed that small Croatian top enterprises have significantly higher revenue per employee and profit per employee then large corporations.

Keywords: Cluster Analysis, Discriminant Analysis, Regression Analysis, Entrepreneurship, Productivity

1 INTRODUCTION

Entrepreneurship is recognized worldwide as the key factor of a country's both economic and social development. It has a central role in recuperating, transforming and boosting economic growth through setting up new ventures, opening new vacancies and generally introducing modern, innovative products and services [1]. The importance of entrepreneurship is particularly emphasized in Croatia, where stimulating entrepreneurship could be an effective regional policy instrument for solving high unemployment problems, especially in areas devastated by war [2]. The mentioned priority is explicitly stated in Croatian Government's Strategic Development Framework for 2006-2013 [3].

Hence it is very important to closely examine the structure of Croatian entrepreneurial sector. In accordance with the European Union legislation, current and relevant classification of economic subjects in Croatia emphasizes the distinction between small, medium and large enterprises [4]. The main criteria for the mentioned segregation are yearly financial results and the average yearly number of employees. Precisely the latest criteria will be employed here in order to analyze top 1000 Croatian enterprises and question the relationship between Croatian firms' size and their productivity.

Croatian transition process can be divided into two diametrically opposed periods vis-a-vis enterprises' productivity. From the sole beginning of Croatian independence until the year 1999 small enterprises dominated over medium and large companies regarding productivity performance. In that breakpoint year small enterprises started to fall behind the medium and large ones, and the mentioned divergence held up until today [5]. That brings us to the main hypothesis of this paper: The existent lagging of small enterprises regarding productivity in the economy as a whole is not pronounced in the case of top Croatian companies.

2 CONCEPTUAL FRAMEWORK

The term *productivity* as a measure of efficiency can generally be defined as a ratio of output volume and input volume. There exists a wide spectrum of productivity measures within this widely accepted definition, depending on the sole purpose of their calculation and the type of input(s) or output(s) that are being analyzed. First of all, production inputs are usually divided into several categories (labor, capital and intermediate goods). In line with that, it is usually stated in the literature that productivity indicators are categorized in the following way [6]:

- Single-factor productivity measures (relating output to only one production factor)
- Multi-factor productivity measures (relating output to several production factors)

In the first case productivity is calculated by putting in relation the volume of output and the volume of labor or capital. On the other hand, second type productivity measures aim at relating output to both mentioned production factors in addition to intermediate goods such as energy, production material or services.

As far as the purpose of productivity analysis is concerned, literature survey crystallized its 5 main objectives (functions): tracing technology changes, monitoring efficiency shifts, identifying cost savings in production, benchmarking different production processes and understanding the transition of living standard [7].

Here the emphasis will be on two specific functions of productivity analysis. First of all, productivity will be monitored as a measure of enterprises' production process efficiency. The concept of full efficiency is here understood in a sense that the production process has achieved the maximum amount of output using a given amount of production factors [8]. At the same time, technology level and organizational structure are not being analyzed, meaning that possible determinants of (in)efficiency are treated as exogenous variables (not examined in this particular analysis).

Having in mind the classification of Croatian entrepreneurial sector according to the level of employment (small, medium and large companies), another productivity measurement function arises from this research. Namely, the conducted analysis is thought out to identify Croatian entrepreneurial subsectors that are characterized by certain inefficiency. Such research strategy is undoubtedly directing the analysis towards the last mentioned function, benchmarking different production processes.

3 DATA AND METHODS

The aim of this paper was to analyze the relationship between Croatian firms' size and their productivity. As every registered company in Croatia has an obligation to submit their annual financial reports, it was possible to compare enterprises' operating results. With that in mind, a large scale database was formed of the top 1000 Croatian companies in the year 2007 by the criterion of added value[1] [9].

The first step of the research was to identify the needed productivity indicators. Since different productivity measures are not independent of each other, it is advisable to use all available indicators at the same time to draw conclusions about the effectiveness of Croatian enterprises [6]. Hence two labor productivity indicators were formed: revenue per employee (REV) and profit per employee (PROF).

In doing so, profit was defined as the difference between revenues and expenditures in a financial year, expressed in the amount before taxation. Financial data (revenue and profit) was expressed in thousands of euros, and calculated according to the exchange rate valid on

[1] Non-profit organizations, public institutions and associations are not involved in this analysis.

31st December of 2007. Since productivity indicators were expressed per employee, a total of 16 enterprises were excluded from the analysis because they reported an average yearly number of employees equal to zero. That led us to the total of 984 enterprises examined in this research. In accordance with the hypothesis about the reversely proportional relationship between Croatian firms' size and productivity, first phase of the research was to classify firms with respect to their productivity, and secondly clusters formed in that way were analyzed vis-a-vis object size.

Multivariate methods were applied to test the hypothesis about the reversely proportional relationship between Croatian firms' size and productivity. Firstly, cluster analysis was used to classify enterprises into homogenous groups. Discriminant analysis was used to identify the enterprises that did not fit perfectly in the groups given by the cluster analysis. Finally, regression models with dichotomous (dummy) variables were provided to quantify the differences in productivity of small, medium and large enterprises.

Cluster analysis classifies the objects into relatively homogenous groups on the basis of their characteristics. These groups are clusters suggested by the data and their structure is not defined a priori. The objects in each cluster are similar to each other and dissimilar to objects in other clusters [10].

The non-hierarchical cluster analysis was applied for the classification of the enterprises. The K-means method provides non-hierarchical cluster analysis. It requires that the number of clusters is specified in advance. Another characteristic of the K-means method is handling a large number of cases. For the reason that 1000 enterprises were analyzed the K-means method were applied.

Multicollinearity among the variables affects the results. The presence of multicollinearity[2] had to be examined for two selected variables. TOL values were found to be greater than 0.2 which denotes that there is no high multicollinearity.

Cluster analysis is quite sensitive to differing scales among the variables and the cluster solution is influenced by measurement differences [11]. Consequently the clustering variables were standardized.

Discriminant analysis classifies the objects within defined groups according to the selected independent variables. The dependent variable consists of two or more groups which are defined a priori. It is appropriate when the dependent variable is a categorical variable and the independent variables are numerical variables [12]. The discriminant analysis was used to classify the enterprises within defined groups by the K-means method.

The regression analysis with dummy variables was also provided. Since the main idea is that regression equations for different categories of enterprises (small, medium and large) will differ only in the constant term (regression lines represent parallel shifts of each other) it is appropriate to use intercept dummy variables [13]. The first model used profit per employee as a dependant variable, while revenue per employee was set as the dependent variable in the second model. Two categorical variables (D1 and D2) were modelled to investigate the productivity discrepancy between enterprises categories and dummy D3 was set as the reference variable:

[2] Multicollinearity is the fact that an independent variable is a linear combination of other independent variables. The most common measure of multicollinearity is tolerance (TOL). TOL is the proportion of variation of an independent variable that is not explained by all of the other independent variables. It is defined as 1 minus the squared multiple correlation of this variable with all other independent variables in the regression equation. A small value of TOL denotes high multicollinearity.

$$D_1 = \begin{cases} 1 \text{ for small enterprise} \\ 0 \text{ otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 \text{ for medium enterprise} \\ 0 \text{ otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 \text{ for large enterprise} \\ 0 \text{ otherwise} \end{cases}$$

To ensure results comparability, regression analysis was also conducted using standardized variables.

4 EMPIRICAL RESULTS

As previously stated, productivity indicators are dependant of each other, what is empirically confirmed here for the case of Croatian enterprises. The obtained correlation coefficients are presented in the following table.

Table 1: Correlation matrix of productivity indicators

	PROF	REV
PROF	1.000000	0.849274
REV	0.849274	1.000000

Hereby it is confirmed that different productivity measures are linearly dependant of each other. Nevertheless, as noted before, the analysis will be continued with variables PROF and REV.

The non-hierarchical cluster analysis was used to group the enterprises. The K-means method of non-hierarchical cluster analysis with Euclidean distances as the distance measure was applied. In the K-means method standardized variables were used. The given number of clusters was three. The first cluster comprises 6 small enterprises, the second 20 (18 small and 2 medium) enterprises and the third 958 (140 small, 490 medium and 328 large) enterprises.

Figure 1 shows the plot of means for the three clusters obtained by the K-means method. It can be seen that the clusters differ according to their productivity. The first cluster (6 small enterprises) is the most productive; the revenue and profit are well above the average. Those variables for the third cluster (mostly medium and large enterprises) are below the average.

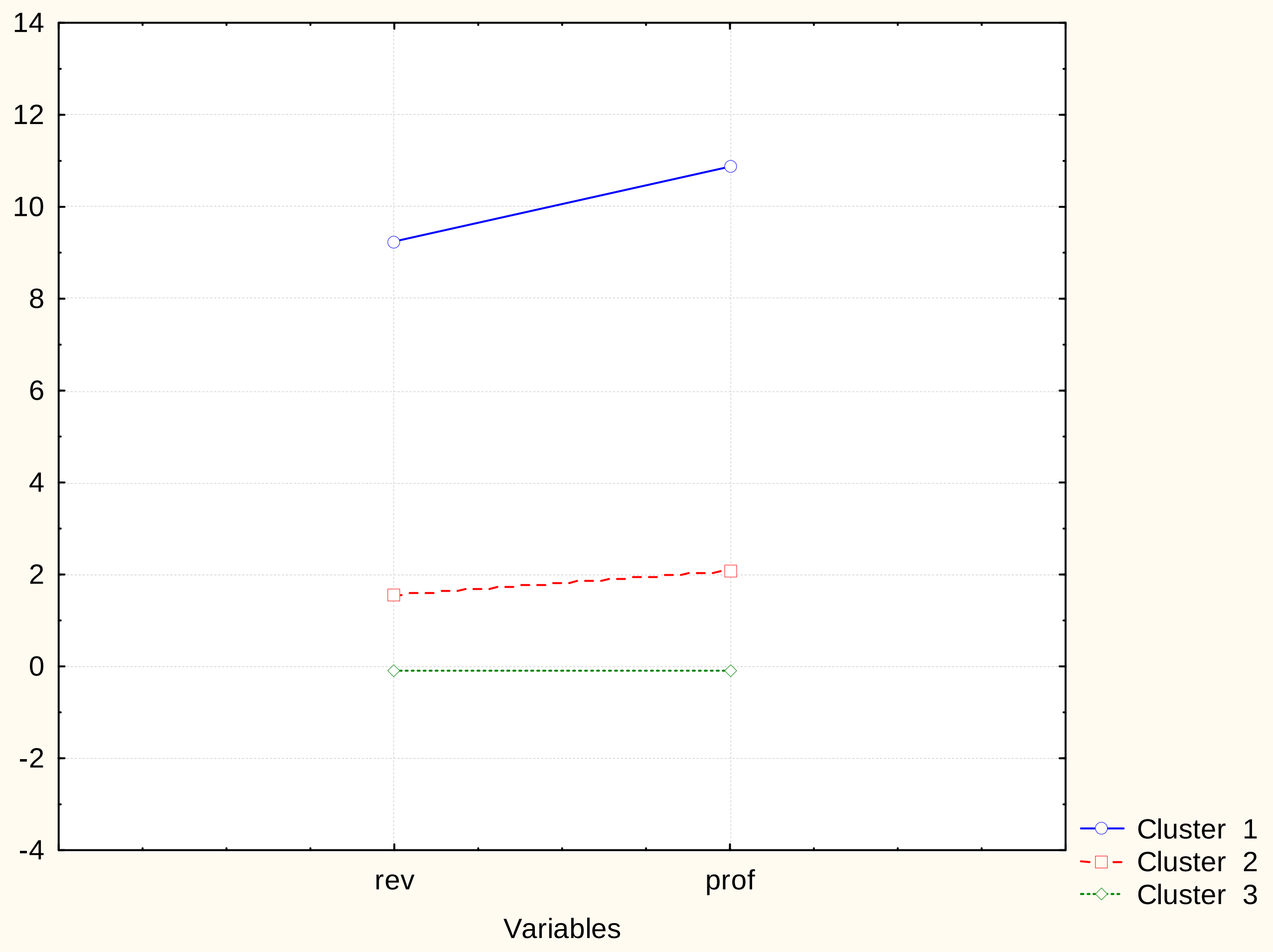


Figure 1: Plot of means for each cluster

After 3 clusters were extracted, it was essential for the testing of this paper's main hypothesis to analyze in what way the enterprises were grouped vis-à-vis their size. The results are given in the following table.

Table 2: Affiliation analysis results

	Absolute frequency			Percentage		
	Small	Medium	Large	Small	Medium	Large
1.cluster	6	0	0	100,00%	0,00%	0,00%
2.cluster	18	2	0	90,00%	10,00%	0,00%
3.cluster	140	490	328	14,61%	51,15%	34,24%

Just a glance at the reported results reveals a certain pattern. Namely, it is obvious that as clusters tend to be less productive, enterprises grouped in them become more labor intensive (larger in employment level). To be precise, first (and the most productive) cluster incorporates only small enterprises. The following cluster is also mostly consisted of small enterprises (90%), while small enterprises are present in the last cluster with a rather neglectable share of 14%. This interpretation can also be stated in the opposite direction: there is a tendency of medium and large enterprises to magnify their share in clusters as they diminish in productivity.

The discriminant analysis was applied to see if it can classify objects successfully into three clusters which are obtained by the K-means method. The cluster membership for three clusters obtained by the K-means method was used as the grouping variable. Revenue per employee and profit per employee were used as independent variables. The Wilks lambda was applied to develop the discriminant function.

Table 3 shows the classification matrix where rows represent the observed classifications while columns represent the predicted classifications. It can be seen that all enterprises were correctly classified only in the third cluster. In the first cluster only one enterprise was incorrectly classified; it was classified in the second cluster. In the second cluster four enterprises were incorrectly classified; those enterprises were classified in the third cluster.

Table 3: Classification matrix
(rows: observed classifications; columns: predicted classifications)

Group	Percent Correct	G_1:1	G_2:2	G_3:3
G_1:1	83,3333	5	1	0
G_2:2	80,0000	0	16	4
G_3:3	100,0000	0	0	958
Total	99,4919	5	17	962

Table 4 shows five misclassified enterprises and accompanying squared Mahalanobis distances. According to the squared Mahalanobis Distances four enterprises were classified in the third cluster due to smallest distances. One enterprise was classified in the second cluster.

Table 4: Squared Mahalanobis distances from group centroids - incorrect classifications

Case	Observed Classif.	G_1:1 p=,00610	G_2:2 p=,02033	G_3:3 p=,97358
* 5	G_2:2	613,3506	24,401	8,484
* 184	G_1:1	387,3781	201,981	275,537
* 313	G_2:2	692,0386	54,631	24,493
* 513	G_2:2	711,8455	80,878	50,207
* 652	G_2:2	556,2625	15,856	12,027

So far it seems that indeed small Croatian enterprises score significantly higher productivity than medium and large sized ones. At this point regression analysis was conducted using dichotomous (dummy) variables in order to quantify the mentioned productivity gap.

The first examined model treats profit per employee as a dependant variable, while two categorical variables are modelled to investigate the productivity discrepancy between enterprises categories. Dummy D3 (indicating large sized companies) was set as the reference variable on purpose so the obtained coefficients could be interpreted as the absolute difference in productivity between small (D1), medium (D2) and large corporations.

Table 5: Dummy variable regression results (PROF as dependant variable)

Dependent Variable: PROF				
Variable	Coefficient	Std. Error	t-statistic	p-value
C	-0.144032	0.052478	-2.744620	0.0062
D1	0.835689	0.090987	9.184698	0.0000
D2	0.009521	0.067818	0.140395	0.8884
F-statistic	51.96441			
p-value	0.000000			

The analysis showed that the only statistically significant dummy variable was D1 (at 1% significance level), meaning that, according to the regression model, profit per employee of small Croatian enterprises is, in average, 0.836 standard deviations higher than the one of a large Croatian enterprise.

Identical methodological steps were taken for the model with revenue per employee set as the dependant variable.

Table 6: Dummy variable regression results (REV as dependant variable)

Dependent Variable: REV				
Variable	Coefficient	Std. Error	t-Statistic	p-value
C	-0.154980	0.053063	-2.920705	0.0036
D1	0.766819	0.092000	8.334945	0.0000
D2	0.054465	0.068573	0.794257	0.4272
F-statistic	40.07996			
p-value	0.000000			

Here the obtained results confirm those of the previous model to a large extent. Again at 1% significance level it is possible to conclude that small enterprises have 0.767 standard deviations higher revenue per employee than large companies.

5 CONCLUSION

The aim of this paper was to closely examine Croatian entrepreneurial sector and the relationship between companies' size and labor productivity. For that purpose two productivity indicators were used: profit per employee and revenue per employee. One thousand top Croatian enterprises were grouped into clusters with respect to the two mentioned variables using K-means clustering method.

It was found that the progression in productivity is quite intensively followed by a reduction in firms' size. The obtained cluster classification was submitted to discriminant analysis, which confirmed previously stated results. At the end, in order to corroborate the mentioned results, dummy variable regression analysis was pursued. It also unambiguously confirmed the inversely proportional relationship between top Croatian enterprises' size and productivity. Such conclusions draw back some important policy implications. Although general economic structure does not confirm such conclusions, small enterprises most definitely have the potential to be the driver of Croatian economic and social development. Economic intuition behind that result indicates that such phenomenon is present because the observed small enterprises are extremely capital intensive, which enables them to head the whole Croatian entrepreneurial sector in terms of added value.

It would also be worthwhile to remark some caveats of the approach used here and to give some guidance for further research. First of all, this research concentrated on top Croatian enterprises, so the drawn conclusions cannot be generalized on the total Croatian economy to any extent. Further on, due to data availability, we concentrated only on labor productivity. Perhaps it would be interesting to see if the obtained results would diverge from those gathered here.

References

[1] Kružić, D., 2007. Poduzetništvo i ekonomski rast: Reaktualiziranje uloge poduzetništva u globalnoj ekonomiji, *Ekonomska misao i praksa*, No. 2, pp. 167-192

[2] Maleković, S., 2000. Croatian experience in regional policy, Institute for economic research Ljubljana, Occasional paper No.2.

[3] Government of the Republic of Croatia, 2006. Strategic Development Framework for 2006-2013

[4] European Commission, 2003. Commission Recommendation concerning the definition of micro, small and medium-sized enterprises, available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:124:0036:0041:EN:PDF [20 May 2009.]

[5] Škrtić, M., Mikić, M., 2006. Economic importance of Croatian entrepreneurship- possibilities and pitfalls, *Zbornik Ekonomskog fakulteta u Zagrebu*, pp. 191-204

[6] Nestić, D., 2004. Bilješka o proizvodnosti-definicija, mjerenje i povezanost s politikom plaća, *Privredna kretanja i ekonomska politika*, 14, No.101, pp. 54-74.

[7] OECD, 2001. Measuring productivity- measurement of aggregate and industry-level productivity growth, OECD Manual, available at: http://www.oecd.org/dataoecd/59/29/2352458.pdf [22 May 2009.]

[8] Diewert, E.W. and Lawrence,D., 1999. Measuring New Zealand's Productivity, *Treasury Working Paper* 99/5

[9] Institute for Business Intelligence Zagreb

[10] Everitt, B. S., Landau, S., Leese, M., 2001. Cluster Analysis. 4th Ed. London, Arnold.

[11] Johnson, R.A., Wichern, D. 2007. *Applied Multivariate Statistical Analysis*. Pearson International Education.

[12] Hair, J. F., Black, W.C., Babin, B. J., Anderson, R. E., Tatham, R. L., 2006. Multivariate Data Analysis. 6th Ed. Prentice Hall. Upper Saddle River.

[13] Bahovec, V. and Erjavec, N., 2009. Uvod u ekonometrijsku analizu, Element Zagreb

DIFFERENCES TESTING BETWEEN CLUSTERS: EVIDENCE FROM MARKET ORIENTATION OF CROATIAN FACULTIES

Zoran Mihanović, Josip Arnerić and **Mario Pepur**
University of Split, Faculty of Economics
Matice hrvatske 31, 21000 Split, Croatia
{zoran.mihanovic,jarneric,mario.pepur}@efst.hr

Abstract: This work investigates the impact of a particular managerial component and that is the organizational size on the development of market orientation towards multiple stakeholders on faculties in Croatia. Hierarchical cluster procedure will be used to group individuals into clusters, with goal to maximize the homogeneity of individuals within the clusters while also maximize the heterogeneity between the clusters. Ward method will be performed in which the similarity used to join clusters is calculated as the sum of squares between the clusters summed over all variables. Graphical representation (tree graph) of the results of a hierarchical procedure will be shown by dendrogram. Afterwards differences testing between clusters will be performed according to organizational size on the development of market orientation towards multiple stakeholders on faculties in Croatia.

Keywords: Differences testing, Cluster analysis, Ward method, Euclidean distance metric, Market orientation, Organizational size, Higher education.

1 INTRODUCTION

In theoretical as well as practical sense, there is a notable lack of work on marketing management and market orientation in education in general, and particularly on the impact of organizational size on market orientation of educational institutions, regardless of the fact whether the participants of marketing implementation in education come from profit or non-profit sector. Due to the particularities of educational institutions, one should ask the question whether marketing concept is an adequate philosophy in these institutions. The mission and goals of educational institutions differ from the goals of profit organizations. Therefore, a special and adapted marketing approach within educational institutions is essential. Marketing, whose standard concept is not applicable in these situations, is often neglected in such institutions whereas it ought to be quite the opposite. Taking account of organizational size of educational institution, adapted concept would alleviate the achievement of positive performance for managers of these institutions, and the satisfaction of education customers and other target groups.

2 THEORETICAL CONSTRUCT

2.1 Marketing concept and market orientation

Interest for developing marketing concept has been constant as it can be seen in the work of many authors and many researches from the 1960s onward, [10]. Later, in the 1990s, much more attention has been paid to the implementation of marketing concept. Most researches define market orientation as acquisition and implementation of marketing concept and emphasize that marketing should not be considered simply as a function of a specific department but, more importantly, as the leading philosophy for the entire organization [12]. There are three essential backbones underlying most marketing concept definitions and they are: (1) customer orientation; (2) coordinated marketing and (3) profitability. It is reasonable to conclude from the literature that market oriented organization is the one where the three mentioned foundations of marketing conception are operatively manifested [16].

Taking into consideration the research of prominent authors in the late 80s and from the 90s onward, who were mostly preoccupied with market orientation in the profit sector, there have been many attempts to conceptualize market orientation. One of the first conceptualizations was the approach of *organizational decision making process* where the main author Shapiro [27] states that the term market orientation represents a set of procedures influencing all aspects of the firm with the firm being market oriented if it emphasizes the following three aspects: (1) the firm has to fully understand its market and its customers and all important intelligence, gathered in various ways, have to spread and reach all organizational functions; (2) suggests that strategic and tactical decisions have to be taken interfunctionally and interdepartmentally; (3) with a strong and internal connection, different departments and functions take well coordinated decisions and zealously execute them. Unlike Shapiro, Narver and Slater [22] in their approach of *culturally based behaviour* suppose that market orientation consists of three behavioural components (customer orientation, competition orientation, interfunctional coordination) and two decision criteria (long-term direction and profitability). Market orientation has been defined as "organizational culture which most efficiently and successfully generates behaviour necessary to create better value for the customer thus continuing with improved performances for the firm "[22]. One of the most accepted market orientation conceptualizations is the approach of *market information* where its main authors Kohli and Jaworski state a slightly different definition of market orientation. Their research shows profitability as the result of market orientation, and the other two determinants of marketing concept (customer orientation and coordinated marketing) constitute the basis of market orientation. Market orientation requires: *(1) active participation of one or more departments in activities organised for the purpose of developing the understanding of customer's current and future needs and factors which influence them (environmental influence); (2) effective communication and distribution of these findings throughout departments and (3) active participation of different departments in activities created to satisfy certain current and future customer needs* [16]. Other authors have tried to define market orientation and test its measure looking back on and broadening the research of earlier authors. All these studies make it clear to what extent organizational effects, such as *organizational size, the number of departments, employees and customers*, can influence dissemination of intelligence within the organization and the level of organization's market orientation.

Most studies in the profit sector are oriented towards elements which constitute market orientation and the existence of potential relations between market orientation and organizational performance. However, not all studies developed the measure/scale for market orientation simultaneously, that is, there was a problem with the method of measuring market orientation. Most of the chief and mostly used measures developed from the way in which certain authors view the concept of market orientation. The first scale with fifteen items was developed and tested by Narver and Slater [7]. Researches in many countries and many different industries have proven it to be valid. The most used measure of market orientation and the most tested one in various studies is the measure by Kohli and Jaworski, MARKOR, with twenty items [15; 17]. The third significant measure was developed by Desphande, Farly and Webster [7]. Other authors prominent of the profit sector have mainly upgraded the existing measures relating them to their own constructs of market orientation and its important and constituting components [5; 6; 9; 31].

2.2 Market orientation of educational institutions

Croatian educational system offers educational services on the level of preschool education, elementary, secondary and higher education and the system of adult education. Educational institutions have more goals and a much broader mission than the institutions in the profit sector, and have relationships with a far greater number of subjects. They have to weigh customer needs and tendencies while at the same time maintaining institution's academic reputation and its other goals and commitments. The mission of education is to satisfy expected future needs in the function of development, and one is often mistaken when thinking that education is directed to existing needs [14]. Exactly these needs raise the importance of marketing and marketing concept implementation in education because marketing is the one that can play a significant role in joining long-term needs of the entire society and satisfying customer needs with quality presentation of that which is not wanted in short-term, but which will affect its development in long-term and the development of society directing these needs in a socially acceptable manner, thus eventually fulfilling institution's mission and goals.

The implementation of marketing concept in the non-profit sector is not as applicable as in the profit sector. For non-profit organizations, a goal corresponding to profitability is survival which means earning enough income to provide for long-term expenses and/or long-term satisfaction of all important clients [18]. Non-profit business can be placed within the context of a stakeholder model. This concept views the organization as an open socioeconomic system, consisting of numerous subjects (besides the organization's founder/owner as well as other insiders), with legitimate demands towards the organization and a part in its functioning [13]. Managing the components of marketing mix of non-profit organizations will probably differ from the profit ones, thus, activity accomplishment is at times measurable only in quality terms relating customer satisfaction to the firm's goals [32]. Strategy of non-profit organizations is not based on money nor is it the centre of their planning. They start with their performances, goals and mission [8]. Taking into consideration ever moving social and economic changes, there is a need for improving the flexibility of educational system and its porosity for individuals to have better options of changing education/qualification, in accordance with shifting needs of labour market and the concept of lifelong learning [24]. This implies a significant shift towards customer-oriented educational systems with loose borders between different sectors [20]. Kotler and Levy in their seminal article, first proposed that the marketing philosophy could be extended to non-profit organizations [19]. Nevertheless, no matter the importance which was recognised in the latter, most empirical work in the non-profit marketing has been focused on the use or direct transfer of specific marketing techniques to non-profit organizations [11; 26]. Certain authors suppose that, bearing in mind the particularities of service sector and education and the results of market orientation research in profit and non-profit sector, there are three behaviouristic components necessary to shape the framework of market orientation in education: customer orientation, internal orientation (orientation towards its employees) and integrated marketing efforts (organizational coordination) with the condition of long-term survival as the determining criterion [29]. Even though they consider it to be the vital element of changeable environment, customer orientation is excluded from conceptual framework since their research has shown that it was not a crucial element of market orientation in educational institutions. Based on this, market orientation in education has been defined as follows: *organization follows market orientation in so much that its structure, culture, systems and procedures are established and developed in a way that ensures long-term relations with customers (both clients and employees) within limited resources and conditioned long-term survival of that organisation, and is therefore, on the basis of the study results, suggested that marketing needs set forth from the organization itself due to serviceable educational nature, and top management's understanding and commitment to market orientation is of utmost importance* [29].

Many studies which have been measuring market orientation and the relation between market orientation and organizational performance in profit and non-profit sector have been using this behaviourist approach by Kohli and Jaworski [4; 28] and adapted measure of Kohli, Jaworski and Kumar [17], or a combination of the approaches and the approach of multiple market orientation taking into consideration multiple clients. All these studies are encouraging because many of them support the argument that a higher level of market orientation will lead to better organizational performance.

Market orientation in non-profit organizations is inhibited by ideological or attitudinal barriers coming from the organization management. The control over non-profit organizations and decision making is spread and distributed among a large number of internal and external publics making organization-wide co-ordination, communication and responsiveness problematical thus inhibiting market orientation development [21]. Market orientation and its emphasis on organizational awareness of citizen needs and demands are proposed as the critical element of better management in non-profit organizations. Market orientation could be used to enhance managerial and financial efficiency in the non-profit sector [28]. Most non-profit organizations lack customer orientation since they are mostly sales-oriented. This was believed to be for the reasons related to management of the organizational system. *From everything stated, it seems that many obstacles in adopting marketing concept are again related to the field of management. Thus, this work investigates the impact of a particular managerial component and that is the organizational size on the development of market orientation towards multiple stakeholders on faculties in Croatia*. For example, Narver and Slater [22] have initially conceptualized relative organizational size as control factor in relation between performance and market orientation. Their results suggest that larger organizations are more reluctant in becoming market-oriented. Looking at market orientation from a behaviourist approach, Kohli and Jaworski [16] suggest that a larger number of departments within the organization can act as a communication barrier and inhibit the dissemination of intelligence between departments and employees. Similarly, efficiency measures seem to be related more to organizational systems and management related factors [3]. It seems that efficiency in many non-profit organizations is strongly mediated by organizational and management related factors. It is the development of appropriate organizational and management system as a result of stakeholders' market orientation (or stakeholders' demands) that will increase efficiency and stakeholder market orientation itself.

3 METHODOLOGY AND EMPIRICAL RESULTS

In the empirical part of the research, a multivariate data analysis was applied on a representative sample with the use of a questionnaire as the instrument of research. The questionnaire was constructed on the basis of the research by Kohli, Jaworski and Kumar [16; 17] where they created MARKOR - Measure of Market Orientation with three basic market orientation components – *gathering and dissemination of intelligence and appropriate responsiveness* – whose reliability and validity has been proven in the later works of other authors. The questionnaire has been altered and adapted to the non-profit sector, education and Croatian higher education institutions. Population gathered all faculties in the Republic of Croatia (59 of them). The framework of sample choice was the list and categorization of faculties according to the Ministry of science, education and sports [23]. Public school of professional higher education and Polytechnics were excluded from the research considering the size and different organizational hierarchy/structure. The questionnaire was answered by 46 faculties which makes 78 % of the population. Managers of faculties were used as research subjects, because they are the instigators of certain behaviour and architects of organizational culture and philosophy and their familiarity with overall situation within the organizations themselves make them the most competent subject for providing certain answers on the organizational activities. This research included five stakeholders (students, potential students, economic sector, competent Ministry and faculty) who faculties establish or should establish effective relationships with. Education is a labour intensive activity. Hence, organizational size is defined by the number of employees/staff the faculty employs. The research was conducted by the end of 2006 and the beginning of 2007.

Hierarchical cluster analysis was used to determine which subsets of faculties can be classified by their size. For size indicator number of faculty employees is selected. Before clustering procedure all variables were standardized. This procedure has created 2 clusters from the 46 observations supplied using Ward's method with Euclidean distance metric. The clusters are groups of observations with similar characteristics. To form the clusters, the analysis began with each observation in a separate group. It then combined the two observations which were closest together to form a new group. After recomposing the distance between the groups, the two groups then closest together were combined. This process was repeated until one group remained. Ward's method has the tendency to result in clusters of approximately equal size due to its minimization of within-group variation. A dendrogram shows how each of the clusters was formed. A horizontal line connecting two groups shows that the groups were combined at the distance shown on the vertical axis.

```
                         Rescaled Distance Cluster Combine

  C A S E      0         5        10        15        20        25
Label     Num  +---------+---------+---------+---------+---------+

Case 23    23   -+
Case 45    45   -+
Case 31    31   -+---+
Case 21    21   -+   |
Case 22    22   -+   +---------------------+
Case 41    41   -+   |                     |
Case 24    24   -+   |                     |
Case 25    25   -+---+                     |
Case 10    10   -+                         +---------------------+
Case 46    46   -+                         |                     |
Case 16    16   -+                         |                     |
Case 36    36   -+                         |                     |
Case 43    43   -+---------+               |                     |
Case 40    40   -+         +---------------+                     |
Case 26    26   -+---------+                                     |
Case 44    44   -+                                               |
Case 2      2   -+                                               |
Case 38    38   -+-+                                             |
Case 5      5   -+ |                                             |
Case 17    17   -+ |                                             |
Case 4      4   -+ +---------+                                   |
Case 12    12   -+ |         |                                   |
Case 42    42   -+ |         |                                   |
Case 20    20   -+ |         |                                   |
Case 3      3   -+-+         |                                   |
Case 33    33   -+           |                                   |
Case 37    37   -+           |                                   |
Case 1      1   -+           +-----------------------------------+
Case 13    13   -+           |
Case 28    28   -+           |
Case 8      8   -+---+       |
Case 19    19   -+   |       |
Case 7      7   -+   |       |
Case 30    30   -+   |       |
Case 6      6   -+   +-------+
Case 11    11   -+   |
Case 9      9   -+   |
Case 27    27   -+   |
Case 15    15   -+   |
Case 29    29   -+---+
Case 14    14   -+
Case 34    34   -+
Case 18    18   -+
Case 35    35   -+
Case 39    39   -+
Case 32    32   -+
```

Figure 1: Dendrogram computed according cluster analysis using Ward's method with Euclidean metric.

From dendrogram it can be notice that optimal number of clusters is 2, because the greater distance is in transition from two clusters to one. The first cluster contains 30 faculties which are small by their size (67 faculty employees on average), while the second cluster contains 16 faculties which can be classified as large by their size (218 faculty employees on average).

However, the faculties within each cluster shows distribution differences according total market orientation. From Figure 2 it can be concluded that total market orientation of faculties is proportional to theirs size, i.e. faculties with greater number of employees are more market oriented in comparison to the "small" faculties.

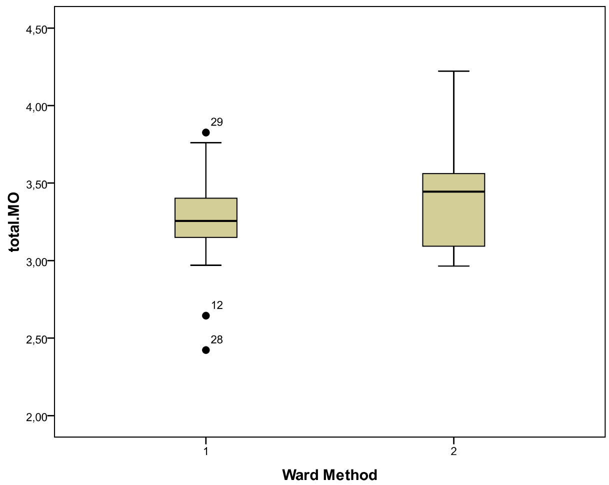


Figure 2: Box and Whisker plot of market orientation between two clusters obtained using Ward method.

Figure 2 shows two Box and Whisker plots, one for each cluster. The rectangular part of the plot extends from the lower quartile to the upper quartile, covering the center half of each sample. The center lines within each box show the location of the sample medians. The whiskers extend from the box to the minimum and maximum values in each sample, except for any outside or far outside points, which are plotted separately. Outside points are points which lay more than 1,5 times the interquartile range above or below the box and are shown as small circles. In this case, there are 3 outside points only in the first cluster. The presence of outside points may indicate outliers.

To test the differences in means between clusters the F-test from ANOVA is used. Namely, differences between clusters are tested in three basic market orientation components – *gathering and dissemination of intelligence and appropriate responsiveness*. Since the p-values of the F-ratios are less than 0.05 for all variables (except *gathering*) the conclusion is that there is a statistically significant difference between the means of the clusters (Table 1).

Table 1: Differences testing in means between clusters in three basic market orientation components

Variables	Clusters		ANOVA results of differences testing	
	Mean value for "small" faculties	Mean value for "large" faculties	F-ratio	p-value
Gathering	2.750	2.882	0.152	0.698
Dissemination	2.905	3.402	3.595	0.047
Responsiveness	3.154	3.674	10.312	0.002

Differences between clusters are also tested towards multiple stakeholders: students, potential students, economic sector, competent Ministry and faculty. The p-values of the F-ratios are less than 5% only for two stakeholders: students and faculty. For deeper analysis the differences between clusters are also tested towards three basic market orientation components within two stakeholders, which are shown to be statistically significant. From Table 2 it can be concluded that a significant difference exists in the second (dissemination) and the third (responsiveness) component of the market orientation degree towards students between the two clusters. "Small" faculties are distributing the data on students inside their own institution in a lesser degree than faculties with greater number of employees, and they respond on collected and distributed information in a significantly lesser degree. Moreover, "small" faculties are significantly less collecting information from faculty and responding on collected data.

Table 2: Differences testing in means between clusters towards three basic market orientation components within two stakeholders: students and faculty

Stakeholders/market orientation components		Clusters		ANOVA results of differences testing	
		Mean value for "small" faculties	Mean value for "large" faculties	F-ratio	p-value
Students	Gathering	3.49	3.56	0.228	0.636
	Dissemination	3.47	3.75	3.793	0.048
	Responsiveness	3.78	4.14	7.330	0.010
Faculty	Gathering	3.44	3.77	4.229	0.046
	Dissemination	3.56	3.67	0.554	0.460
	Responsiveness	3.87	4.11	5.130	0.028

Cluster scatter-plot (Figure 3) shows the members of each cluster within (1) number of faculty employees, (2) total market orientation and (3) market orientation toward students.

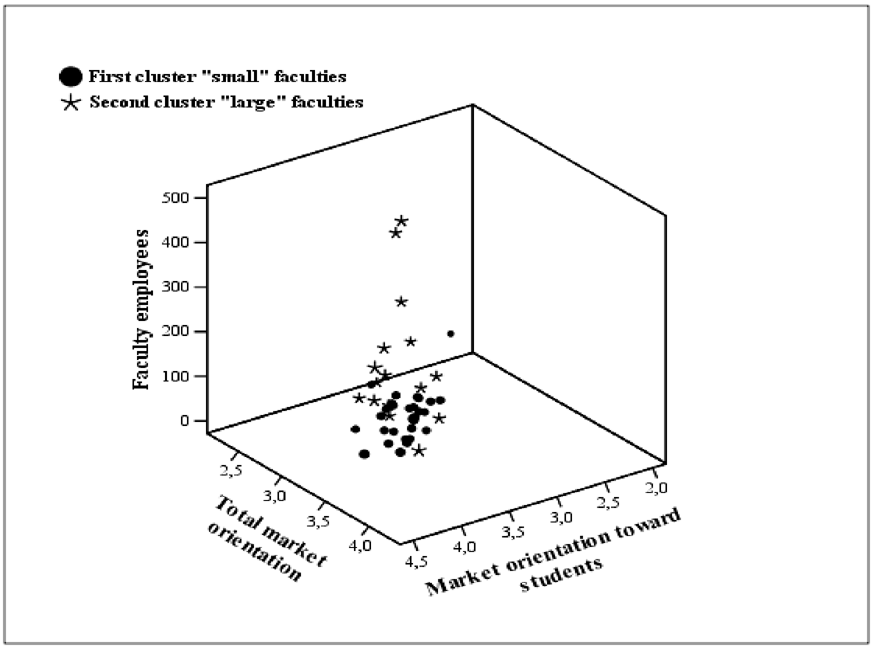


Figure 3: Cluster-scatter plot of faculties membership according number of faculty employees, total market orientation and market orientation toward students.

From Fiqure 3 it can be noticed that there are significant differences between faculties according to their size within total market orientation and particularly within market orientation toward students.

4 CONCLUSION

From the presented research it can be concluded that total market orientation of faculties is proportional to their size, i.e. faculties with greater number of employees (“large” faculties) are more market oriented in comparison to the “small” faculties. Differences between clusters are also tested towards multiple stakeholders: students, potential students, economic sector, competent Ministry and faculty. The p-values of the F-ratios are less than 5% only for two stakeholders: students and faculty. There are significant differences between faculties according to their size within total market orientation and particularly within market orientation toward students. For deeper analysis the differences between clusters are also tested towards three basic market orientation components (gathering and dissemination of intelligence and appropriate responsiveness) within two stakeholders, which are shown to be statistically significant. It can be concluded that a significant difference exists in the second (dissemination) and the third (responsiveness) component of the market orientation degree towards students between the two clusters. “Small” faculties are distributing the data on students inside their own institution in a lesser degree than faculties with greater number of employees, and they respond on collected and distributed information in a significantly lesser degree. Moreover, “small” faculties are significantly less collecting information from faculty and responding on collected data. These results are especially important since students are the most important stakeholder of the faculties and primary target group as well, and since the faculty is one of the most important stakeholders in the education process, having direct exchange of knowledge with the students. Therefore, it is very important that “small” faculties recognize the significance of market orientation especially towards the students and the faculty.

5 DIRECTION FOR FURTHER RESEARCH

Future research should include the following: (1) internal information within the organization should be studied, also the opinions of various client/stakeholders who institutions establish relationships with, who it depends on, especially actual customers and not just subjective opinion of managers in order to evaluate the level of organizational market orientation; (2) in the future, performance of higher education institutions should be studied and related to market orientation and also the impact of organizational size on these two components; (3) other segments of education (secondary and elementary education) should also be studied and then one should determine the role that marketing plays or can play within them.

References

[1] Balser, D., McClusky, J., 2005. Managing stakeholder relationships and nonprofit organization effectiveness, Nonprofit Management & Leadership, Vol. 15, No. 3, pp. 295-315

[2] Bennett, R.C.; Cooper, R.G., 1979. Beyond the marketing concept, Business Horizons, Vol. 22, No, 3, pp. 76-83

[3] Callen, J.L. and Falk, H., 1993. Agency and efficiency in nonprofit organizations – the case of specific health focus charities, Accounting Review, Vol. 68 No. 1, pp. 48-65.

[4] Caruana, A., Ramaseshan, B., Ewing, M.T., 1998. Do universities that are more market orientated perform better?", International Journal of Public Sector Management, Vol. 11, No. 1, pp. 55-69

[5] Chen, S.C., Pascale G.Q., 2005. Developing a Value-Based Measure of Market Orientation in an Interactive Service Relationship, Journal of Marketing Management, Vol. 21, No. 7/8, pp. 779-808

[6] Deng, S., Dart, J., 1994. Measuring Market Orientation: A Multi-factor, Multi-item Approach", Journal of Marketing Management, Vol. 10, No. 8, pp. 725-742

[7] Deshpandé, R., Farley, J.U., 1998. Measuring market orientation: generalization and synthesis, Journal of Market-focused Management, Vol. 2, No. 3, pp. 213-232

[8] Drucker, P.E., 1989. What business can learn from non-profits, Harvard Business Review, July-August, pp. 88-93.

[9] Ellis, P. D., 2005. Market orientation and marketing practice in a developing economy, Euroeopean Journal of Marketing, Vol. 39. No. 5/6, pp. 629-645

[10] Felton, A.P., 1959. Making the Marketing Concept Work, Harward Business Review, Vol. 37 No. 3, pp. 55-65

[11] Giunipero, L.C., Crittenden, W. and Crittenden, V., 1990. Industrial marketing in nonprofit organizations, Industrial Marketing Management, Vol. 19 No. 3, pp. 279-85.

[12] Hooley G.J., Lynch, J.E., Shepherd, J., 1990. The Marketing Concept: Putting the Theory into Practice European Journal of Marketing; Vol. 24, No. 9, pp. 7-24

[13] Hinterhuber, H.H., Krauthammer, E., 1998. The leadership wheel: the tasks entrepreneurs and senior executives cannot delegate, Strategic Change, Vol. 7, No. 3, pp. 149-162

[14] Hrvatska u 21. stoljeću, 2002. ODGOJ I OBRAZOVANJE, Bijeli dokument o hrvatskom obrazovanju, Koncepcija promjena odgojno-obrazovnog sustava u Republici Hrvatskoj, Ured za strategiju Razvitka Republike Hrvatske.

[15] Jaworski, B.J., Kohli, A.K., 1993. Market orientation: Antecedents and consequences, Journal of Marketing, Vol. 57, No. 3, pp. 53-70

[16] Kohli, A.K.; Jaworski, B.J., 1990. Market Orientation: The Construct, Research Propositions, and Managerial Implications, Journal of Marketing, Vol. 54, No. 2, pp. 1-18

[17] Kohli, A.K., Jaworski, B.J., Kumar, A., 1993. MARKOR: A Measure of Market Orientation, Journal of Marketing Research, Vol. 30, pp. 467-477

[18] Kotler, P., Andreasen A. R., 1996. Strategic Marketing for Nonprofit Organizations, Prentice Hall, 536 p.

[19] Kotler, P., Levy, S.J., 1969. Broadening the Concept of Marketing, Journal of Marketing, Vol. 33, pp. 10-15

[20] Mihanović, Z., 2007. Uloga korisnika u visokom obrazovanju: jesu li studenti aktivni dionici ? (The role of customers in higher education: are students active stakeholders?). Tržište, Vol. 19, No. 1, pp. 115-132.

[21] Mokwa, M.P., 1990. The policy characteristics and organizational dynamics of social marketing, in Fine, S.H. (Ed.), Social Marketing: Promoting the Causes of Public and Nonprofit Agencies, Allyn & Bacon, Needham Heights, MA, pp. 43-55.

[22] Narver, J.C.; Slater, S.F., 1990. The effect of a market orientation on business profitability, Journal of Marketing, Vol. 54, No. 4, pp. 20-35

[23] Obrazovanje 2005./2006. (Statističke informacije), Republika Hrvatska - Državni zavod za statistiku 2006.

[24] Plan razvoja sustava odgoja i obrazovanja 2005.-2010., 2005. Ministarstvo znanosti, obrazovanja i športa.

[25] Riggs, H.E., 1986. Fund-raising lessons from high-tech marketing, Harvard Business Review, Vol. 64 No. 6, pp. 64-69

[26] Schwenk, C.R., 1985. The use of participant recollection in the modelling of organizational decision process, Academy of Management Review, Vol. 10 No. 3, pp. 496-503.

[27] Shapiro, B.P., What the Hell Is Market oriented, Harvard Business Review, Vol. 66, No. 6, 1988. pp. 119-125

[28] Shoham, A., et. al., 2006. Market Orientations in the Nonprofit and Voluntary Sector: A Meta-Analysis of Their Relationships With Organizational Performance, Nonprofit and Voluntary Sector Quarterly, Vol. 35, No. 3, pp. 453-476

[29] Siu Noel, Y.M., Wilson, R.M.S., 1998. Modelling Market Orientation: An Application in the Education Sector, Journal of Marketing Management, Vol. 14 No. 4/5, pp. 296-313

[30] Thomas, J.B., Clark, S.M. and Gioia, D.A., 1993. Strategic sense making and organizational performance – linkages among scanning, interpretation, action, and outcomes", Academy of Management Journal, Vol. 36 No. 2, pp. 239-70.

[31] Tse, A. C. B., 2003. Market orientation and business performance in a Chinese business environment, Journal of Business Research, 56, pp. 1331-1340

[32] Yorke, D.A., 1984. Marketing and Non-Profit-Making Organisations, European Journal of Marketing, Vol. 18 No. 2, pp. 17-22

MODELING WEALTH DISTRIBUTION FOR 100 REACHEST PEOPLE IN SOUTH-EASTERN EUROPE

Jelena Minović[1], **Miroslav Minović**[2] and **Miloš Milovanović**[3]
[1]Belgrade Banking Academy, Faculty for banking, Insurance and Finance, Zmaj Jovina 12, Union University, 11000 Belgrade, Serbia, jelena.minovic@gmail.com; jminovic@ien.bg.ac.rs
[2,3]Faculty of Organizational Science, Jove Ilića 154, University of Belgrade, 11000 Belgrade, Serbia, mminovic@fon.rs, milovanovicm@fon.rs

Abstract: This paper presents modeling of wealth for 100 richest people in South-Eastern Europe using data from article Vprost monthly. For modeling we applied generalized Lotka-Volterra (GLV) model. This paper initially provides theoretical background of this model, following presentation results of numerical simulation of generalized Lotka-Volterra (GLV) model. In addition, we developed software for numerical simulation using programming language C#. In our work we tried to determine that GLV model is financial market model which reflect power law (Pareto) for probability distribution of wealth very well. Thus, we find that wealth for 100 richest people in South-Eastern Europe is distributed according to Pareto distribution.

Keywords: wealth distribution, power-law, Generalized Lotka-Volterra model, numerical simulation

1 INTRODUCTION

Polish magazine “Vprost” issued a list of 100 wealthiest people in south-east Europe for 2008. Six people from Serbia found themselves on this list. “Vprost” analysis indicates that out of 40 wealthiest people in the region 30 are from Russia while observing the entire list we can notice 55 Russians. List also contains 14 Polish, seven Ukrainian and Romanian and three Lithuanian and Bulgarian citizen. Czech Republic and Croatia have two representatives on this list [14]. In south-east Europe region current distribution of wealth and “fresh capital” depends on fluctuation of political parties in the government, pace of reforms and privatization as well as the degree of adjusting the legislation to western standards [15].

The Generalized Lotka Voltera (GLV) formalism has been introduced in order to explain the power law distributions in the individual wealth. The wealth distribution is examined using Vprost’s list. In this paper we study a stochastic dynamical model, based on the Lotka-Volterra system that gives rise to the power-law distribution. The model consists of coupled dynamic equation which describe the discrete time evolution of the basic system components ω_i, $i = 1,\ldots,N$. The structure of these equations resembles the logistic map and they are coupled though the average value $\overline{\omega}(t)$. We find that the system components spontaneously evolve into a power-law distribution [7]. GLV model was applied for simulation of individual wealth for 100 richest people in South-Eastern Europe listed by „Vprost“ [13]. Numeric simulation shown that distribution individual wealth poses properties of power law. Program was written in C# [10], [12]. In the paper, Section 2, we present theoretical aspects of generalized Lotka-Volterra model. Definition of power-law is presented in Section 3. In Section 4 we present results of our numerical simulations. Section 5 concludes.

2 THEORETICAL ASPECTS OF GLV MODEL

2.1 Definition of model

In this Section we present theoretical framework of generalized Lotka-Volterra (GLV, Lotka, 1925, and Volterra, 1926) model. This model describes the time evolution of the wealth distribution of individuals in a society [8], [7]. The individual wealth of investors *i* are given by set of time dependent variables ω_i, $i = 1,..., N$. The average wealth $\bar{\omega}(t)$ is:

$$\bar{\omega}(t) = \frac{1}{N}\sum_i \omega_i(t) \tag{2.1}$$

The microscopic representation method models the market index $\bar{\omega}(t)$ as a collective quantity emerging from the interactions of a macroscopic number of traders. More precisely, "microscopic representation" of the stock market is composed of many investors $i = 1,....,N$ each having a certain personal wealth ω_i(t). The investors buy and sell stocks, using more or less complicated strategies designed to maximize their wealth. The resulting index is then proportional to the total sum of the individual investor's wealth [11], [8].

In this paper, we will often use the term „investors" instead of microscopic elements and talk about the „wealth" ω_i of each investor and the „average wealth" $\bar{\omega}(t)$ [11].

The generalized Lotka-Volterra equation system describes the evolution in discrete time of *N* dynamic variables ω_i. In this paper it represents the wealth of an individual investor [7]. The GLV is an interactive multi-agent model. We have *N* investors, each having wealth ω_i, and each ω_i is evolving in time according to following equation:

$$\omega_i(t+1) = \lambda(t)\omega_i(t) + a(t)\bar{\omega}(t) - c(t)\bar{\omega}(t)\omega_i(t) \tag{2.2}$$

where *i* is in 1, 2, ..., *N* [11]. Here $\bar{\omega}$ is the average value of the dynamical variables at time *t*, or the average wealth [7]:

$$\bar{\omega} = \frac{\omega_1 + \omega_2 + ... + \omega_N}{N} \tag{2.3}$$

λ is a positive random variable with a probability distribution $\Pi(\lambda)$. The first term on the right hand side of equation (2.2) is a stochastic autocatalytic term (representing investments, $\lambda(t)\omega_i(t)$) [7], [3]. In a stock market system it represents a growth (or decrease) of investors capital determined by random multiplicative factor $\lambda(t)$ [3]. The coefficients *a* and *c* are in general functions of time, reflecting the changing conditions in the environment [11]. Value $a(t)\bar{\omega}(t)$ is drift term (representing social security payments) [7]. It represents the wealth the individuals receive as members of the society in subsidies, services and social benefits. That is the reason it is proportional to the average wealth [11]. Value $c(t)\bar{\omega}(t)\omega_i(t)$ is a time dependent saturation term (due to the finite size of the economy) [7]. It has the effect of limiting the growth of $\bar{\omega}$ to values sustainable for the current conditions and resources [11], [3], [8].

In recent years there has been considerable interest in the collection and analysis of large volumes of economic data. Such data includes the distribution of the income and wealth of individuals, the market values of publicly traded companies as well as their short term and long term fluctuations. A common observation is that distributions of economic data exhibit a power-law behavior [7].

2.2 GLV- financial interpretation and its advantages

In financial interpretation of GLV model, there is no inflation. We first discuss the assumption that the individual investments/gains/losses are proportional to the individual wealth:

$$\omega(t+1) = \lambda\omega(t). \tag{2.4}$$

This is actually not applicable for the low income/wealth individuals whom incomes do not originate from the stock market. In fact the additive term $a(t)\bar{\omega}(t)$ tries to account for the departures related to additional amounts originating in subsidies, salaries and other fixed incomes. However, for the range of wealth where one expects power laws to hold ($\omega > \bar{\omega}$) it is well documented that the investment policies, the investment decisions and the income, measured yearly, are in fact proportional to the wealth itself [11].

The statistic uniformity of the relative gains and losses of the market participants is a weak form to express the fairness of the market and the lack arbitrage opportunities (opportunities to obtain systematically higher gains λ - 1 than the market average without assuming higher risks): for instance, if the distribution of λ would be systematically larger for small-ω-investors, then the large-ω-investors would only have to split their wealth in independently managed parts to mimic that low-ω superior performance. This would lead to an equalization of their λ to the λ of the low-ω investors. Therefore in the end, the distribution $\Pi(\lambda)$ will end up ω-independent as we assumed it to be in GLV from the beginning [11].

The model described by equation (2.4) is of non-interactive multiplicative processes with fixed lower bound (eq. (2.5)). For the purpose of creating a power law (eq. 3.2) with the exponent of $\alpha > 0$ multiplicative equation process (2.4) must be modified in such a matter that variation $\omega(t)$ in equation (2.4) is limited with lower bound $\omega_{\min}$ (minimum capital):

$$\omega(t) > \omega_{\min}. \tag{2.5}$$

Relation between minimum capital $\omega_{\min}$ and average capital is presented with equation (2.6):

$$q = \frac{\omega_{\min}}{\bar{\omega}} \tag{2.6}$$

In inflation circumstances q should be limited instead of $\omega_{\min}$, since locking the minimum capital would not be effective for a longer period of time. In order to formalize the sophistic application after several trivial analytical steps, relation is reached, that binds exponent α from power law (eq. (3.2)) with the coefficient q (eq. (2.6)). Exponent of stepped law is reached as:

$$\alpha = \frac{1}{1-q} \text{ [11], [8].} \tag{2.7}$$

The dependence of α on q given by Equation (2.7) indeed confirms $\alpha > 1$ for $q < 1$ [6]. This result is quite satisfying since for lower bound $0 < q < 1/2$ predicts that exponent of power law is $1 < \alpha < 2$, which represents the value of exponent noticed in real world. Main issue with mechanism of generating power law based on single-agent dynamics is that it occurs that $\omega(t+1)$ (except for surroundings of $\omega_{\min}$) is lower than $\omega(t)$. This is not the case in reality where economies and population are expanding. Exponent α in power law is rather instable when influenced with fluctuation in system parameters. Modeling changes in $\bar{\omega}$ leads to significant fluctuation of q, and that implies substantial variation of exponent α [11], [8]. This contradicts the extreme stability of exponent α noticed in real world. Key element in this problem is determining interaction between individual elements ω_i through involvement of values (or lower bounds) proportional $\bar{\omega}$ [8].

The advantages of GLV model:

- The GLV model insures a stable exponent α of the power-law (described by equation (3.2)) even in the presence of large fluctuations of the parameters.
- The average value of $\langle \lambda \rangle$ and the average $\bar{\omega}$ can vary during the run (and between the runs) without affecting the exponent α of the power-law distribution.
- can take typical values both larger and smaller than 1 [11].

For further economic and financial interpretation of GLV model you can see reference [11]. Additionally, for detail theoretical framework of GLV systems and models that came before it you can see the same reference [11], and [8].

3 POWER-LAW DISTRIBUTION

Multiplicative processes have been well studied in different contexts and widely applied to various research fields, such as the biological, social, and economic systems. One of the major interests in these processes is the generation of power laws, which have been observed in many natural domains and indicate the properties of scale invariance and universality. In fact, the investigation of power law behaviors in systems with stochastic multiplicative dynamics can be traced back to decades ago, and the underlying physical mechanisms are still to be understood [5].

For the application in financial problems, the assumption of multiplicative property for the individual capital investments ω_i (the index i = 1,..., N may correspond to various investors/traders or companies (stocks) in the market) leads to

$$\omega_i(t+1) \sim \lambda(t)\omega_i(t) \tag{3.1}$$

appearing in the dynamics of wealth/capital evolution. That is, the individual capital ω_i at time t + 1 is proportional to the invested capital itself. The random factor $\lambda(t)$ reflects the relative gain/loss incurred by individuals between time t and t + 1, and is chosen from a probability distribution $\Pi(\lambda)$ [5].

The model described by equation (3.1) is multiplicative processes coupled through the lower bound (eq. (2.5)). However, we consider a system of N degrees of freedom ω_i which is governed by the following dynamics: at each time t, one of the ω_i's is chosen randomly to be updated according to the formula (3.1), while all the other ω_i's are left unchanged [11].

The Pareto power-law distribution of individual wealth ω_i is [5]

$$P(\omega) \sim \omega^{-1-\alpha}\,. \tag{3.2}$$

Power-law states that appearance of small values is quite common, while large values appear rarely. This law is literately regarded as Zipf or Pareto law [1]. If $N_t(\omega)$ represents the number of investors that hold the capital in the moment t in interval (ω, $\omega + d\omega$). In border situation when $N \to \infty$ this defines distribution of individual capital $P_t(\omega)$ by relation:

$$P_t(\omega)\,d\omega = \lim_{N\to\infty} \frac{N_t(\omega)}{N}\,. \tag{3.3}$$

This distribution has properties:

- $P(\omega) = 0$, for $\omega < 0$, that corresponds to fact that capital is always positive; $P(0) = 0$;
- $\lim_{\omega \to \infty} P(\omega) = 0$
- There is a maximum between $\omega = 0$ and $\omega = \infty$.

For many systems that poses large ω it is determined that distribution act according to power low presented in equation (3.2) [11]. Power laws are common in systems that are constituted out of pieces that have no specific size and in systems that are constituted out of self-catalyzing elements [2]. They represent the link between simple microscopic basic laws on individual level and macroscopic phenomenon's that occur collectively [11], [8].

It was observed that fluctuations in financial markets exhibit non-Gaussian "fat tailed" distributions [6].

The value of the exponent α depends on the value of the lower wealth bound. Indeed, empirical studies show that the value of α changes across different countries, and is typically in the range $1 < \alpha < 2$. Theoretical analysis shows that for N not too small, the exponent is determined by the parameter q, according to equation (2.7) [6].

4 RESULTS OF NUMERICAL SIMULATION

To examine the theoretical predictions presented in Sections 2 and 3, we have performed computer simulations of generalized Lotka-Volterra model described by Equation (2.2) [7].

Simulations [4] of the model were performed for $N = 100$. For simulation purposes [4] the computer program was written using C# programming language. Program provides selection of all the required parameters and model values, starting the simulation and finally graphically presenting the results [9]. Data values are taken from the list published by Polish magazine "Vprost" [13].

The value of the parameter q was obtained directly from the Vprost data [13], as the ratio $\omega_{\min} / \bar{\omega}$ (eq. (2.6)), averaged for 2008. year, where $\omega_{\min}$ is the wealth of the least wealthy investor on the list. We find a value of $q = 0.18$ (680 mln USD/3690.20 mln USD). Starting from a homogeneous distribution of the wealth, the wealth distribution spontaneously evolved towards a power-law distribution (shown in Figure 4.2), with an exponent $\alpha = 1.22$ [6].

The results are of much relevance to empirical studies. They show that the distribution of the individual wealth in different countries during different periods in the 20th century has followed a power-law distribution with $1 < \alpha < 2$ [7].

It provides a strong connection between the lower-cutoff, and the exponent α affects the wealth distribution of the entire population, including those at the top. The relationship (2.7) holds precisely for the simulation using the Vprost 100 lists - the value of $q = 0.18$ corresponds to a value of $\alpha = 1/(1 - 0.18) = 1.22$ [6].

The model predicts that the power-law distribution of wealth extends well beyond the top 100 investors. To examine this feature we performed simulations with $N = 10{,}000$ investors and the same value of q. We obtained a power-law distribution of wealth with the same value of α [6].

Consider the average wealth

$$\bar{\omega}(t) = \sum_{i=1}^{100} \omega_i(t) / 100 \tag{4.1}$$

of the 100 investors at year t [6].

Factor of value change $\lambda(t)$ is taken by chance, according to uniform distribution, from the given interval. Parameters and variables of GLV model in our case have following meaning [9]:

- $\omega_i(t)$ - wealth of investor *i* in moment *t*
- $\lambda(t)$ – coefficient of wealth growth in moment *t*
- *a* – parameter that describes the influence of society wealth and state, and in increase investors wealth
- *c* – parameter that describes the influence of competition and cooperation between investors on increasing their wealth
- $\bar{\omega}(t)$ – average wealth of all investors in moment *t* (3690.20 mln USD according to Vprost list for 2008.)
- $\omega_{\min}$ – the wealth of the least wealthy investor on the list (680 mln USD according to Vprost list for 2008.))
- $\omega_{\max}$ – the wealth of the wealthiest investor on the list (24500 mln according to Vprost list for 2008.))

Parameters of model are determined by experience and the selected values are [3]:

- $\lambda(t) = 1.3$
- $a = 0.000001$
- $c = 0.00001$.

Regarding that the main cause is realizing how model acts, instead of acquiring as precise economic indicators as possible, selected values will prove suited for this simulation. Because of limited computational power as well as easier statistical analysis of acquired results, the number of iterations is limited at 500000. Simulation results are given on images 4.1, and 4.2. Horizontal axis presents individual starting wealth and individual final wealth, while vertical axis presents the occurrence frequency, respectively [9]. Figure 4.1 provides starting wealth distribution of investors that is actually uniform.

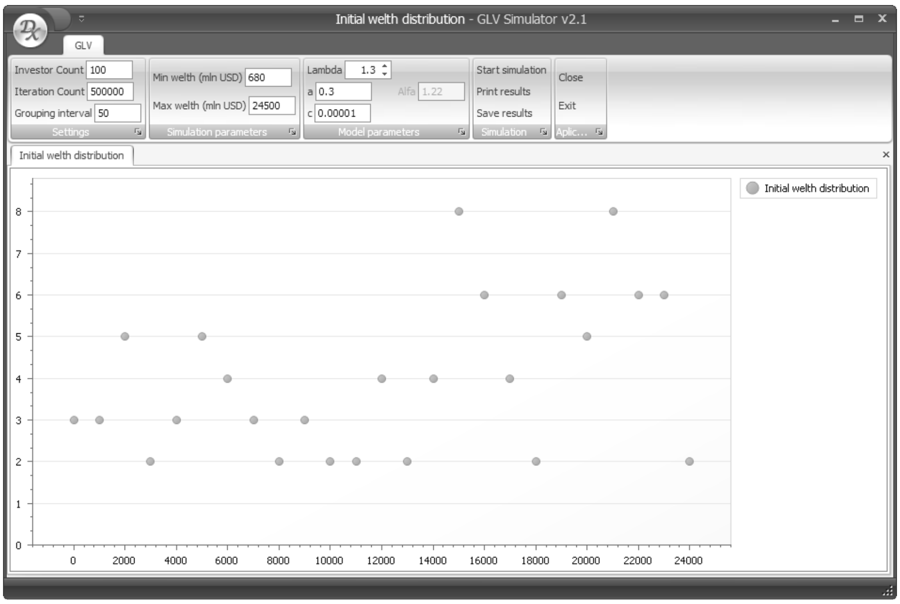


Figure 4.1: Starting uniform distribution of wealth.

Upon ending the simulation the program provides the results of reached individual wealth of 100 wealthiest people in South-east Europe. When those data is processed by program, given distribution can be seen on image 4.2 [9].

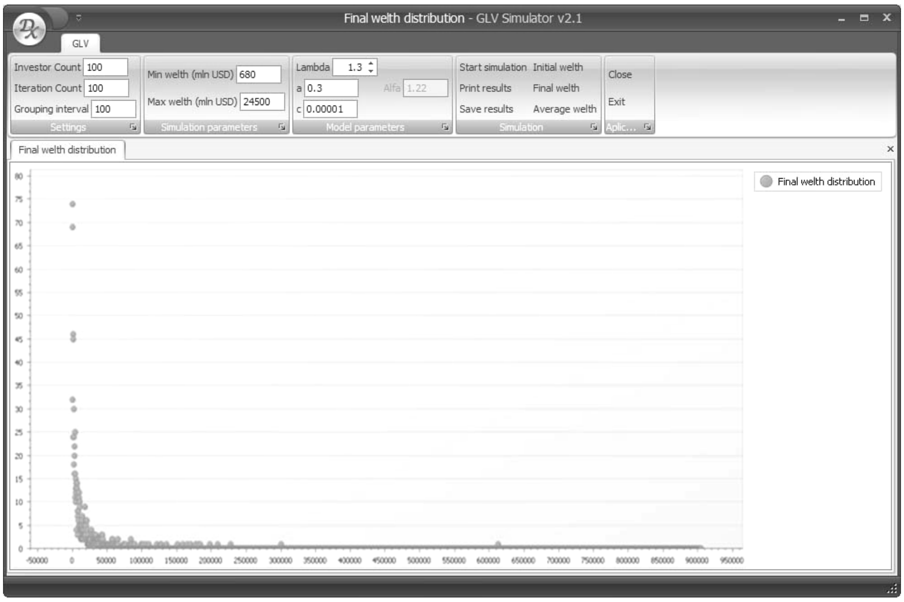


Figure 4.2: Final distribution of wealth

Final distribution of individual wealth is actually a power law as can be seen on 4.2. This represents a final conclusion of simulation, since starting from uniform distribution of wealth we reached a distribution that small amount of investors possesses great amount of wealth a while majority possesses small amount of wealth. This law is empirically known as overwhelming in many natural systems (from stock markets to animal world population), is now reached mathematically, through GLV model [9].

Interesting fact is that power law will occur in most cases, in spite of changing values in model parameters. In addition it will not be dependent on average incomes. Decrease of parameter $\lambda(t)$ can lead to slowing down the occurrence of power law but it won't cancel it out [3].

However, a crucial assumption in the model is that the same distribution $p(\lambda)$ is used for all the investors. It was found that simulations in which this assumption is violated do not give rise to a homogeneous power-law distribution of the wealth. Thus, according to the model, the power-law wealth distribution in the Vprost data indicates that the *ex-ante* return distribution is similar for all the investors. The multiplicative dynamics greatly magnifies the differences between more successful and less successful investors, and builds up the power-law distribution of wealth [6].

The assumption that the return distribution $p(\lambda)$ is the same for all the investors is an implementation of the efficient market hypothesis, which states that no investor can consistently obtain a return distribution better (adjusted for risk) than the return distributions of the other investors. Thus, our model provides a connection between the efficient market hypothesis and the Pareto distribution of wealth [6].

The main goal of this simulation was to show the commodity of applying GLV model in practice, as well as confirm the claim on social wealth distribution acting according to power. We also tried to indicate the influence of changing some of the models on entire model act. Simulation results cannot be used in purpose of predicting the real situation on the market because the model is simplified. Building an applicable model would require a much more detail analysis of parameters as well as simulation conditions [9].

We can conclude that for this class of models, in order to obtain a power-low distribution, it is sufficient that relative returns of the agents are stochastically equivalent. Therefore, the presence of a power-low distribution may be a sign of „market efficiency“[7].

5 CONCLUSION

In this paper we modeled wealth for 100 richest people in South-Eastern Europe using data from article Vprost monthly. The Vprost 100 lists reveal a striking agreement with the Pareto power-law distribution of wealth. For modeling wealth distribution we used generalized Lotka-Volterra (GLV) model. Results of numerical simulation in this paper presented that generalized Lotka-Volterra (GLV) model reflects Pareto power law for probability wealth distribution of individual investors. Downside of the Lotka-Volterra is that it does not provide a satisfactory level for real application and for that purpose it should be improved. GLV actually presents a general method that can be used for simulation, analysis and understanding of vast class of phenomenon where power law occurs.

References

[1] Adamic, L. A., Retrieved on 2007. Zipf, Power-lows, and Pareto-a ranking tutorial, Xerox Palo Alto Researches Center, Palo Alto, CA 94304, http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html

[2] Anderson, P.W., 1995, in The Economy as an Evolving Complex System II (Redwood City, Calif: Addison-Wesley.

[3] Biham, O., Malcai, O., Levy, M., Solomon, S. 1998. Generic emergence of power law distributions and Levy-stable intermittent fluctuations in discrete logistic systems, Phys. Rev. E 58, p. 1352. http://xxx.lanl.gov/abs/adap-org/9804001.

[4] Gogal, M. 2007. Computer-Based Numerical & Statistical Techniques, Infinity Science Press.

[5] Huang, Z. F., Solomon, S. 2002. Stochastic multiplicative processes for financial markets. Physica A, 306, 412–422.

[6] Klass, O. S., Biham, O., Levy, M., Malcai, O., Solomon, S. 2004. The Forbes 400, the Pareto Wealth Distribution and Efficient Markets. The European Physical Journal B, 55, pp. 143-147.

[7] Malcai, O., Biham, O., Richmond P., Solomon, S. 2002. Theoretical Analysis and Simulations of the Generalized Lotka-Volterra Model. http://arxiv.org/PS_cache/cond-mat/pdf/0208/0208514v1.pdf

[8] Minović, J. 2008. Prikaz generalisanog Lotka-Volterra modela. SYM-OP-IS 2008, zbornik radova sa međunarodne konferencije, Soko Banja, Srbija

[9] Minović, M., Minović, J. 2008. Softver za simulaciju tržišta kapitala Srbije. InfoFest, zbornik radova sa međunarodne konferencije, Budva, Crna Gora.

[10] Rattz, J. C. 2007. Pro LINQ Language Integrated Query in C# 2008, Apress.

[11] Solomon, S. 1999. Generalized Lotka-Volterra Models, in “Aplications of Simulations to Social Sciences”, eds. G. Balot and G. Weisbuch; Hemes Science Publications. http://xxx.lanl.gov/PS_cache/cond-mat/pdf/9901/9901250v1.pdf

[12] Troelsen, A., 2007. Pro C# 2008 and the .NET 3.5 Platform, fourth edition, Apress.

[13] Official web site of „Vprost“ magazine: http://najbogatsieuropejczycy.wprost.pl/

[14] Official web site of „Srpska Dijaspora“ magazine: http://www.srpskadijaspora.info/vest.asp?id=6276

[15] Official web site of „Vreme“ magazine: http://www.vreme.com/arhiva_html/495/22.html

MODELING THE IMPACT OF TRANSIT SERVICE UNRELIABILITY ON DEMAND

Dejan Paliska*, Daša Fabjan**
*University of Ljubljana, Faculty of Maritime Studies and Transport, Portorož, Slovenia
**University of Primorska, Faculty of Tourism Studies, Portorož, Slovenia
e-mails: dejan.paliska@fpp.uni-lj.si; dasa.fabjan@turistica.si

Abstract: This paper discusses the effect that bus transit service unreliability has on demand and presents the results of its study. In the research a set of regression models that estimate route-level demand were developed using data collected with Automatic Passenger Counters and Automatic Vehicle Location systems installed on buses, and demographic, socio-economic and land use information from other sources. Eight regression models were introduced according to time of the day and whether the route was inbound or outbound. Focusing on service related variables, especially bus departure delay variable, the regression results showed this turned out statistically significant in all four outbound routes. The estimation of the coefficients in this case showed a negative value, which in turn proved that in some cases the unreliability does result in a decreased demand.

Keywords: statistical model, transit demand, transit service, level of service, reliability.

1 INTRODUCTION

Reliable service is important for both transit passengers and transit providers. Many surveys have shown that reliability is strongly related to passenger satisfaction and perception of service quality [1], as well some preference experiments have found that passengers implicitly value reliability [2] and consider it in their mode choice decisions [3]. Unreliable service results in additional waiting time for passengers [4], [5] [6], [7], the unit cost of which has been estimated to exceed the cost of in-vehicle travel time by a factor of three [8].

Unreliable service has negative economic consequences also for transit providers. Effective service capacity is diminished when vehicles become unevenly spaced and platooning, or “bus bunching”, occurs. Bus bunching results in more frequent passenger overloads, which necessitates provision of additional service. Such service expansions would not be required if vehicles were more regularly spaced and passenger loads were more evenly distributed [9]. Financial investments in the vehicle fleet are affected because reliability problems are most acute during peak service periods [9].

There has been considerable research on the underlying causes of unreliable service [3], [9], [11], [12]. Primary causes of unreliability have been attributed to route characteristics (e.g. length, the number of signalized intersections, the extent of on-street parking, stop spacing), operating conditions (e.g. traffic volume, service frequency, passenger activity), and vehicle operators behavior (e.g. departure delays, operator-specific behavior differences).

Considerable attention has also been devoted to identifying operations control actions to improve reliability [3], [9], [11], [12], [13]. Examples of control actions include vehicle holding, stop-skipping, leap-frogging and short-turning. Many of the *Level of Service* (LOS) factors affecting transit use however cannot be easily quantified and there is always a problem of generally not having data available. The literature generally supports the ability of transit system with high-quality service to attract more users, as well as for poor service to encourage more automobile use [14]. Public opinion indicates increases in LOS as an important factor.

While much has been done to understand the causes of unreliable service and to find the corrective actions that can be taken, research on this subject has been hampered by the costs of manual data collection. However, the introduction and deployment of the *Advanced Public Transit System* (APTS) technologies, particularly *Automatic Vehicle Location* (AVL) and

Automatic Passenger Counter (APC) systems, have transformed the data environment for transit providers. Comprehensive data on vehicle operations and passenger activity are now being recovered and archived at very low cost. The new data environment is facilitating more extensive and detailed analysis of transit operations, the benefits of which are reflected in service planning, scheduling, dispatching and operations control improvements [9] , [15].

A number of econometric models have been developed analyzing the determinants of bus transit demand. The models differ according to method used, variable selection and level of aggregating data. Most previous studies seeking to explain the determinants of transit demand have been conducted at either route level [16], [17], [18], [10] or route-segment level [20], [21]. Stop-level transit demand has been discussed in the literature as being the most appropriate level of analysis, but there are very few actually applied models at that level [21], [22]. Historically, regression models have been popular because they are relatively easy to use, well established, comparable with other available procedures, and well suited for parameter estimation problems. Nonetheless, regression is not the only possible estimation approach, and other methods such as time series analysis have been explored [23].

2 TRANSIT SERVICE RELIABILITY

Transit service reliability is a multidimensional phenomenon and there is no single measure that can adequately address service quality. The most common measures of transit service reliability typically relate to schedule adherence, running times, and headways. The usefulness of each service reliability measure is largely determined by service frequency, whether or not timed transfers must be met, and functional needs. Important distinctions exist between passengers and operators in their perceptions of service.

Departure delay (actual departure time minus scheduled departure time) effectively measures schedule adherence for a given bus at a particular location. Schedule adherence is an important reliability measure for infrequent users, timed transfers, and service characterized by large headways. Transit providers have employed a number of service measures. The indicator that is most widely recognized, and the one that probably has the greatest intuitive appeal, measures on-time performance (OTP) [9]. In practice, on-time performance is probably most relevant in situations of infrequent service, where bus riders tend to time their arrivals in relation to the schedule, or in trips which involve transfers.

Our work focused on measuring on-time performance, which requires precise location positioning of the vehicle.

3 STUDY AREA AND DATA COLLECTION

The study area was part of the bus transit system in the city of Koper. The city is situated in the southwestern part of Slovenia, at the northern edge of the Adriatic Sea. With 25,000 inhabitants, the city of Koper is a regional center of the Slovenian coastal area. The settlement pattern is distinctively longitudinal, stretching along the coast. This makes it difficult to set up an effective transit service. In addition, rapidly increasing car ownership in the last 10 years has had a negative impact on transit use.

In the past, there have been different attempts to revitalize the transit public system; one of the latest was the introduction of a small bus transit system. The system runs on fixed routes with a fixed schedule, connecting the old city core with newer, high density suburban areas. The aim of the new system was to offer to commuters a more attractive transit service. The response of the passengers and the level of use did not meet the operator’s expectations, and to motivate the passengers to use those new routes, the community of Koper began subsidizing this service.

This study operated with the data collected within four of the mentioned bus routes. The route lengths were between 6.4 and 7.8 km, with 9 to 12 bus stops in each direction. There are two common bus stops with higher passenger demand, and routes cross 4 to 5 signalized intersections. All the routes start near the old city core, passing near port, shopping malls, city center, and run through high density suburban areas. The scheduled headway differs during the day with a minimum value of 7 minutes during morning and evening peak periods, and increasing to 30 during off-peak. The evening headway varies from 30 minutes up to 130 minutes during late evening.

The data sample was collected between January 1 and February 15 of 2005. A GPS based bus positioning system was used together with an *Automatic Passenger Counter* to record the data at the bus stop level. A GPS receiver recorded the location each time the bus stopped and opened the doors. When the door opened, the APC unit started to count passengers boarding and alighting; when the doors closed the data were recorded. During the observation period, a sample of 2,996 stops was recorded.

4 MODELLING AND METHODOLOGY

As previously said, most of the transit demand models described in the literature are based on route level, mostly due to costs and time required for manual data collection. Previous studies on transit demand mostly used *ordinary least squares* (OLS) regression to relate transit demand with different socio-economic, demographic, land use and system performance related variables [20], [21], [26], [27], [14].

4.1 Defining independent variables

Zhao [14] classified the factors affecting transit demand into four categories; transit *level of service* (LOS), accessibility, land use and urban design, and transit users' socioeconomic and demographic characteristics. Although tremendous efforts have been devoted to exploring the factors that significantly affect transit use, contradictory findings were cited by different researchers and practitioners in different study areas, indicating that some of the factors may not be transferable from one urban area to the other [14]. In some studies some accessibility variables were demonstrated to significantly affect transit demand [28], [29], while in others not. Land use and urban design variables may include population density, employment density, land use mix, land use balance, etc. Although some of the previous researches consider the land use and urban design factor to be an important component affecting travel demand, its effects are not as significant as individual characteristics such as gender, ethnicity, and age [30]. Socioeconomic and demographic characteristics of transit users result in rather contradictory conclusions. In some previous studies income was found not to significantly affect transit use [21], [29], [30], while other researchers concluded otherwise [31], [32]. Conflicting findings can also be observed for age, gender, and vehicle availability variables. In addition, a household's or an individual's socioeconomic and demographic characteristics are usually highly correlated, such as in the case of vehicle ownership and income [14].

An attempt was made to include as many variables as possible in our analysis. Unfortunately, not all data were available. As a dependent variable in the models, an actual number of boarded passengers (PSPP) was selected. The independent variables selected to develop the route level demand model were: mean scheduled headway (CPVR), headway variation at previous bus station (ZCPP), departure delay - OTP (ZCP), job-housing balance

(JOBS_HH), competition - number of buses in from other transit agencies (PPK), average number of workers in household (POVZAPG), percentage of households without children (BREZOT), percentage of households without car (BAVTO), average number of cars per household (PAVTO), total population (POP), number of workers in bus station service area (POPDEL), number of bus stations (PAP) and percentage of population under 16 years of age (POPV16).

Service related variables (CPVR, ZCPP, ZCP, PPK, PAP, PSPP) were either collected with AVL and APC systems or provided by service operators. Data for demographic related variables (POPV16, POP, POPDEL, BREZOT, BAVTO, PAVTO, POVZAPG, JOBS_HH) were extracted from the Census 2002 database. Other socioeconomic variables used in the model were collected by a stratified sample of 15% of population in the area covered in our study.

The jobs housing indicator (JOBS_HH) was calculated using the Ewing formula and the GIS approach was used as well [34]:

$$JOBS - HH_i = \frac{|E_i - P_i|}{E_i + P_i} \qquad (1)$$

where:

$JOBS\text{-}HH_i$ = job housing balance index for zone i;
E_i = employment size in zone i;
P_i = population size in zone i.

For the allocation of socioeconomic, demographic and other land use information the transit routes service areas were defined using GIS. First, walking distances between each household and transit stations were calculated. For the calculation, the corrected Zhao's method with distance decay functions was used [14], [33]. Since the study area has significant variations in altitude, corrections that consider the effect of the slope on walking were used. A combined vector/raster approach on digital elevation model and street network was then applied to calculate a uniform service area. Afterwards each household was associated with only one, the nearest, bus station. Since all four routes have the same starting and ending points, and there are no significant differences in service frequency, the assumption was that the riders would most likely walk to the nearest station. This way the problem of service areas overlapping was avoided. At the end, the variables allocated according to bus station service areas were either summed or averaged in routes service areas.

4.2 Model estimation

Multiple regression analysis was performed to determine the most significant factors that affect transit demand. In *Minitab 14* the best subset and the stepwise (backward and forward) procedure was employed to select independent variables. Variables having significance level values higher than 0.1 were considered to be insignificant and were further not included in the models.

The decision whether a model was reasonable or not was based on the positive or negative value of the coefficients (+, -), R-square value, and analysis of the residuals. The presence of autocorrelation was controlled with Durbin-Watson statistics and multicolinearity was detected by examining the variance inflation (VIF) score. If the VIF factor exceeded a value of 10, the correlation factors between independent variables were examined, and the model was corrected either by eliminating or by joining the variables.

The statistical correlation matrix between different independent variables indicated a strong multicolinearity between certain demographic variables. Thus, the independent variables that caused most multicolinearity problems, such as percentage of population under 16 years of age (POPV16), were removed from further analysis. The summary statistics for selected variables in each model are presented in Table 1.

Table 1: Descriptive statistics for all selected variables

Variable	Mean	StDev	Minimum	Maximum	Skewness	Kurtosis
CPVR	69.98	22.59	0.00	150.00	0.28	0.46
ZCP	***1.26***	***1.89***	***-7.00***	***7.00***	***0.02***	***1.14***
ZCPP	0.01	2.26	-10.00	12.00	0.27	1.74
PAP	5.38	3.49	0.00	11.00	0.11	-1.31
POP	845.89	349.80	163.00	1264.00	-0.86	-0.66
JOBS_HH	0.54	0.13	0.25	0.89	0.72	0.95
PPK	15.15	14.97	0.00	58.00	0.64	-0.32
POPDEL	1248.40	649.70	204.00	2171.00	-0.43	-1.51
POVZAPG	1.71	0.09	1.45	1.83	-1.45	2.42
BREZOT	0.14	0.06	0.06	0.33	1.80	4.52
BAVTO	8.33	5.82	0.00	20.00	-0.14	-1.09
PAVTO	1.56	0.17	1.25	1.83	-0.64	-0.30
PSPP	0.72	0.46	0.00	2.15	1.03	1.24

Our aim was to analyze the effect that the departure delay has on demand. Thus the stress is on the ZCP variable, for which we can read the distribution from the table above. It can be noted that delays have a rather normal distribution with a mean delay time of +1.26 minutes and almost 2 minutes standard deviation.

A deeper analysis showed that departure delay times vary among different bus stations. There were 16 included in our study and the Figure 1 shows the distributions of departure delay times for each station. The same figure also shows how frequently buses actually stop on each bus station, and which stations tend to be skipped. Checking the skewness on the graphs we can notice almost all bus stations most of the times experience delays.

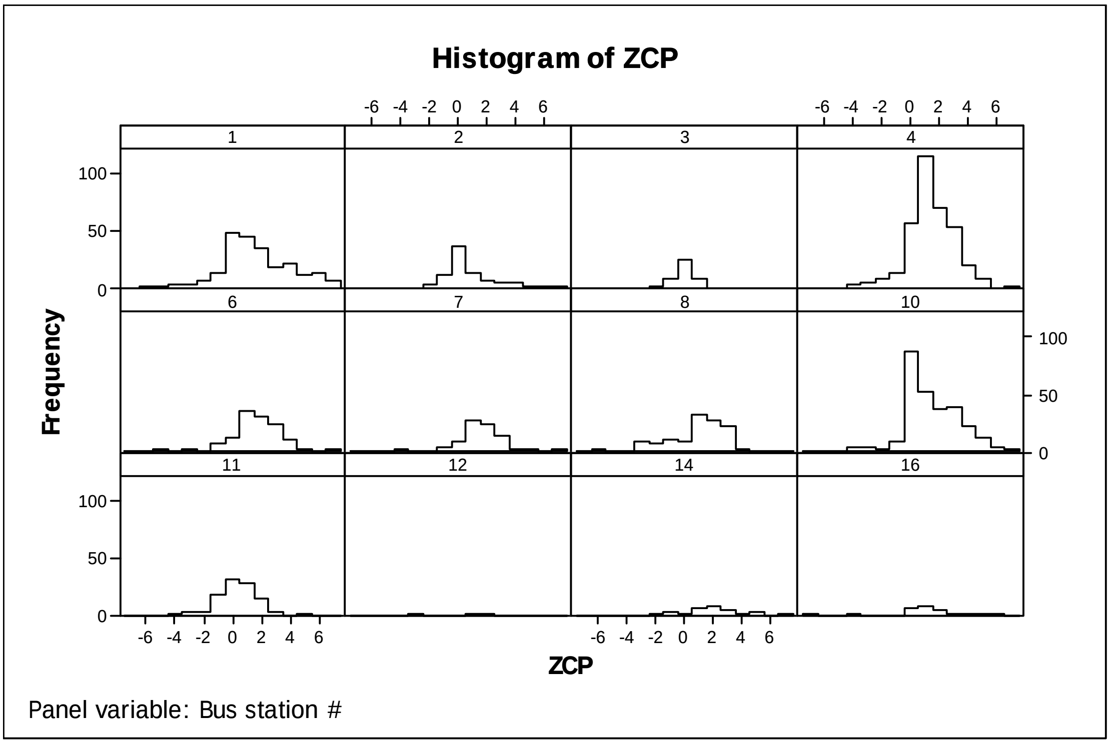


Figure 1: Distributions of the departure delay times for 16 bus stations.

The distances analysis also resulted in frequencies of the stops that buses actually make on routes. In Figure 2 we can see a combined graph showing the variation in distribution of stops with or without delays and the frequency of stops at stations located at different lengths of the route.

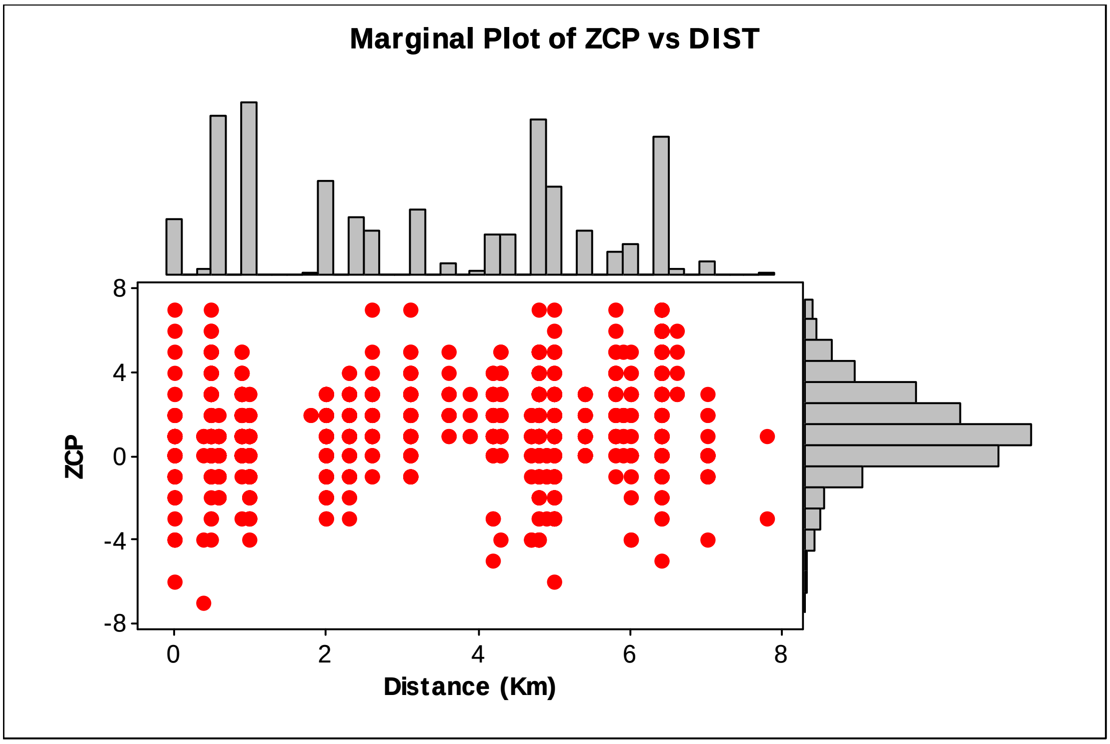


Figure 2: Total bus departure delay times distribution and their occurrence on the route.

4.3 Demand models parameters estimation

The analysis of numerous regression model results further indicated that it was best to develop a separate model for each time period as follows: AM (6:00-9:59); Mid-day (10:00-14:59); PM (15:00-18:59); and Evening (19:00-00:00).

Considering the above eight models (four for inbound and four for outbound routes) were developed using the acquired data, with the R-square values obtained ranging from 0.73 up to 0.98.

Focusing on one of the most important variables among service related variables for measuring the level of reliability, i.e. the departure delay – OTP (ZCP), the tables below show only the four outbound models were ZCP variable turned out to be statistically significant.

Table 2: Parameter estimation for AM demand model

Variable	Coefficient	T-Value	P-Value	Elasticity
CPVR	0.0059	2.2270	0.0290	0.4411
ZCP	***-0.1217***	***-8.1230***	***0.0000***	***-0.0831***
ZCPP	-0.0309	-2.1280	0.0360	-0.0225
POVZAPG	0.2171	0.2465	0.8060	0.5090
PAVTO	-13.1790	-21.2300	0.0000	-29.2338
JOBS_HH	-5.1629	-7.5890	0.0000	-4.2951
POPDEL	-0.0024	-14.8100	0.0000	-4.4231
SSE	3.73			
R-Sq	0.9092			
R-Sq(adj)	0.9011			
F statistic	112.95			

Table 3: Parameter estimation for Mid-day demand model

Variable	Coefficient	T-Value	P-Value	Elasticity
CPVR	0.0000	0.1639	0.8700	0.0014
ZCP	***-0.0065***	***-3.5190***	***0.0010***	***-0.0078***
ZCPP	0.0022	1.6440	0.1010	-0.0001
PAP	-0.0624	-44.6600	0.0000	-0.3563
POPDEL	0.0001	16.6100	0.0000	0.1520
POVZAPG	-0.9568	-34.0200	0.0000	-2.0895
BAVTO	0.0027	5.0580	0.0000	0.0338
PAVTO	-0.2389	-13.2900	0.0000	-0.4722
BREZOT	5.1932	65.7100	0.0000	0.9129
JOBS_HH	-3.6510	-81.7900	0.0000	-2.5012
SSE	0.52			
R-Sq	0.985			
R-Sq(adj)	0.984			
F statistic	963.26			

As shown in Tables 2 to 5, the highest value of R-square was obtained in the Mid-day demand model and the smallest, although still very high, in the Evening demand model. One can see that the variables with positive coefficient cause demand to increase, while those with negative coefficient do the opposite.

A negative value of the ZCP variable coefficient was expected, meaning that the unreliability does result in a decreased demand. Coefficients values are small in all models, the biggest importance of the variable however is noticed in the Evening outbound model (coefficient elasticity accounts for -0.1074).

Table 4: Parameter estimation for PM demand model

Variable	Coefficient	T-Value	P-Value	Elasticity
CPVR	0.0000	-0.1061	0.9160	-0.0011
ZCP	***-0.0074***	***-3.6240***	***0.0000***	***-0.0087***
ZCPP	0.0025	1.6160	0.1070	0.0000
POVZAPG	-0.9734	-29.4400	0.0000	-2.1113
BREZOT	5.3071	50.1900	0.0000	0.9030
BAVTO	0.0019	2.4890	0.0140	0.0240
JOBS_HH	-3.7148	-64.7000	0.0000	-2.4890
POPDEL	0.0001	14.2700	0.0000	0.1509
PPK	0.0004	2.0730	0.0390	0.0102
PAP	-0.0634	-38.7700	0.0000	-0.3685
PAVTO	-0.2494	-11.8200	0.0000	-0.4870
SSE	0.20			
R-Sq	0.984			
R-Sq(adj)	0.983			
F statistic	1018.77			

Table 5: Parameter estimation for Evening demand model

Variable	Coefficient	T-Value	P-Value	Elasticity
POPDEL	0.0000	-0.0610	0.9520	-0.0052
ZCP	***-0.0205***	***-2.6070***	***0.0110***	***-0.1074***
ZCPP	0.0084	1.3370	0.1850	-0.0005
PPK	-0.0100	-3.1710	0.0020	-0.2192
BREZOT	-0.4576	-1.8440	0.0690	-0.1652
BAVTO	0.0318	9.2800	0.0000	0.8151
CPVR	0.0036	7.1280	0.0000	0.9029
POVZAPG	1.0028	4.3130	0.0000	4.3247
SSE	0.522			
R-Sq	0.813			
R-Sq(adj)	0.792			
F statistic	39.29			

The departure delay variable turned out to be statistically significant, showing that unreliability causes decrease in demand, especially in the outbound models, which can be attributed to the fact that it is easy to find alternative transport in the city center, thus the passengers are less prepared to wait in uncertainty for the bus.

Among the four demand models presented the coefficient value of departure delay variable ZCP is the highest in the morning outbound model. This gives a proof that it is much more important to passengers to be on time for their jobs, school or other morning activities, while their sense of being on time loses its importance during the day, and it gains importance again in the evening, when the scheduled headway is increased.

5 CONCLUSION

This research has aimed to investigate the variables that cause unreliable public transport service and their relationship with transit demand. A statistical model was developed for estimating the effect of several factors on increasing or decreasing of transit demand on a case study. The set of variables included different socio-economic, demographic, land use and bus performance factors, among these also a variable of service reliability (departure delay) was taken in consideration.

After either corrections or omitting of some variables, that showed autocorrelation or multicolinearity, the results of regression analysis indicated it was more reasonable to introduce separate models for morning, midday, afternoon and evening as well as for inbound and outbound bus routes. In our case these models explained from 73% to up to 98% of the variability in the number of passengers.

The coefficient of departure delay variable turned out to be statistically significant in all four outbound models. Its value was negative implying that bigger delays have a rather decreasing effect on transit demand. Quite interesting was also the varying value of coefficient and its elasticity, showing the varying importance of the variable throughout the day. In the morning people tend to give more importance to being on-time according to their morning activities (e.g. job, school, doctor’s appointment), less during the day and again more importance in the evening, which could be correlated to the bigger scheduled headway.

These models and their results can serve as a useful information or even tool for transit service providers and transport planners to define the appropriate timings and schedules according to the transit demand. Of course it would be appropriate when new sets of sample data were collected throughout the year, some data updated, especially those that had to be interpolated in our study. And it would be a big improvement when also data on bus driver behavior were available. The results of the models would definitely be more consistent, however given the circumstances the models present good foundations to understand the impact of analyzed variables.

References

[1] Transit Cooperative Research Program (TCRP) (1999): *A Handbook for Measuring Customer Satisfaction and Service Quality*. TCRP Report 47. Washington, DC: Transportation Research Board, National Research Council.

[2] Bates J., Polak P, Jones J. and A. Cook (2001): *The Valuation of Reliability for Personal Travel. Transportation Research*. Part E, No. 37, pp. 191-229.

[3] Prioni P. and D. Hensher (2000): *Measuring Service Quality in Scheduled Bus Services.* Journal of Public Transportation, vol. 3, pp. 51-74.

[4] Welding P. (1957): *The Instability of Close Interval Service*. Operational Research Quarterly, No. 8, pp. 133-148.

[5] Turnquist M.(1978): *A Model for Investigating the Effects of Service Frequency and Reliability on Bus Passenger Waiting Times.* Transportation Research Record, 663, pp. 70-73.

[6] Bowman L. and M. Turnquist (1981): *Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops*. Transportation Research, Part A, vol. 15, pp. 465-471.

[7] Wilson N., Nelson D., Palmere A., Grayson T. and C. Cederquist, (1992): *Service Quality Monitoring for High Frequency Transit Lines.* Paper presented at the 71 Annual Meeting of the Transportation Research Board, Washington, DC.

[8] Mohring H., Schroeter J.and P. Wiboonchutikula (1987): *The Values of Waiting Time, Travel Time, and a Seat on the Bus.* Rand Journal of Economics, No.18 (1), pp. 40-56.

[9] Strathman J., Kimpel T., Dueker K., Gerhart R.and S. Callas (2002): *Evaluation of Transit Operations: Data Applications of Tri-Met's Automated Bus Dispatching System*. Transportation, vol. 29, pp. 321-345.

[10] Strathman J., Dueker K., Kimpel T., Gerhart R., Turner K., Taylor P., Callas S. and D. Griffin (2000): *Service Reliability Impacts of Computer-Aided Dispatching and Automatic Vehicle Location Technology: a Tri-Met Case Study.* Transportation Quarterly , vol. 54, pp. 85-102.

[11] Sterman B. and J. Schofer (1976): *Factors Affecting Reliability of Urban Bus Services.* Transportation Engineering Journal, vol. 102, pp. 147-159.

[12] Abkowitz M. and I. Engelstein (1984*): Methods for Maintaining Transit Service Regularity.* Transportation Research Record 961, , pp. 1-8.

[13] Levinson H. (1991): Supervision *Strategies for Improved Reliability of Bus Routes.* NCTRP Synthesis of Transit Practice 15. Washington, DC: Transportation Research Board, National Research Council.

[14] Zhao F., Min-Tang L., Chow L. F., Gan A. and D. Shen (2002): *FSUTMS Mode Choice Modeling: Factors Affecting Transit Use and Access*. Final Report, National Center For Transit Research (NCTR), University of South Florida, Tampa.

[15] Casey R. (2000): *What Have we Learned about Advanced Public Transportation Systems? In What Have We Learned About Intelligent Transportation Systems?*. Volpe Transportation Systems Center, Federal Highway Administration, US Department of Transportation, Chapter 5.

[16] Kemp, M. A. (1981): *A Simultaneous Equations Analysis of Route Demand and Supply, and its Application to the San Diego Bus System.* Washington, DC: UMTA, Report DTUM-60-80-71001.

[17] Horowitz, A. J. (1984): *Simplifications for Single-Route Transit Ridership Forecasts.* Transportation 12, pp. 261-275.

[18] Azar, K. T., and J. F. Ferreira (1994): *Integrating Geographic Information Systems into Transit Ridership Forecast Models.* Journal of Advanced Transportation 29(3), pp. 263-279.

[19] Hartgen, D., and M. W. Horner (1997): *A Route-Level Transit Ridership Forecasting Model for Lane Transit District.* Oregon. NC, Center for Interdisciplinary Transportation Studies, Report No. 170.

[20] Peng, Z. (1994): *A Simultaneous Route-Level Transit Patronage Model: Demand, Supply, and Inter-Route Relationships*. Unpublished doctoral dissertation, Portland State University, Portland, OR.

[21] Kimpel, T. J.(2001): *Time Point-Level Analysis of Transit Service Reliability and Passenger Demand.* Portland, OR: Unpublished Doctor of Philosophy in Urban Studies, Portland State University.

[22] Furth, P. G., B. Hemily, T. H. Mueller, and J. G. Strathman (2003): *Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Review and Potentia.,* Washington DC: Transportation Research Board, TRCP Report No. 23 (Project H-28), http://gulliver.trb.org/publications/tcrp/tcrp_webdoc_23.pdf.

[23] Vaziri, M., Hutchinson, J., and M. Kermansha (1990): *Short-Term Demand for Specialized Transportation: Time-Series Model.* Journal of Transportation Engineering 11, pp. 6105-121.

[24] Arrowhead Space & Telecommunications, Inc (1999): *Bus Driver Fatigue and Stress Issues Study*. Final Report No. DTGH61-99-Z-00027, Washington DC. http://ntl.bts.gov/lib/7000/7800/7883/busfatigue.pdf.

[25] Peng, Z., and K. J. Dueker (1995): S*patial Data Integration in Route-Level Transit Demand Modeling*. Journal of the Urban and Regional Information Systems Association 7, pp. 26-37.

[26] Pendyala, R.M.(1999): *Integrated Transit Demand and Supply Model User Manual*. ITSUP Version 0.50, Public Transit Office, Florida Department of Transportation, Tallahassee.

[27] Zhao, F (1998): *GIS Analysis of the Impact of Community Design on Transit Accessibility*. Proceedings of the ASCE South Florida Section 1998 Annual Meeting, Sanibel Island, FL., pp. 1-12.

[28] Brinckerhoff, P. (2000): *Tour and Trip Mode Choice Models.* Draft Technical Report, Prepared for Sacramento Area Council of Governments, Sacramento, California.

[29] Sun, X., Wilmont, C.G., and T. Kasturi: H*ousehold Travel, Household Characteristics, and Land Use: An Empirical Study from the 1994 Portland Travel Survey*, Transportation Research Record 1617, Transportation Research Board, National Research Council, Washington, D.C. ,1998.

[30] Loutzenheiser, D.R. (1997*): Pedestrian Access to Transit: Model of Walk Trips and Their Design and Urban Form Determinants around Bay Area Rapid Transit Stations.* Transportation Research Record 1604, Transportation Research Board, National Research Council, Washington, D.C., pp. 40-49.

[31] Dargay, J., and M. Hanly (2002): *The Demand for Local Bus Service in England*. Journal of Transport Economics and Policy 36, pp. 73-91.

[32] Preslar, D.A. (1998): *Transit Ridership Forecasting Using a GIS.* Proceedings of the ASCE Conference on Transportation, Land Use, and Air Quality - Making the Connection pp. 595-605.

[33] Zhao, F., L. Chow, M. Li, I. Ubaka, and A. Gan (2003): *Forecasting Transit Walk Accessibility: Regression Model Alternative to Buffer*. Transportation Research Record 1835, pp. 34-41.

[34] Ewing, R. (1995): *Measuring Transportation Performance*. Transportation Quarterly, 49 (1), pp. 91-104.

APLICATION OF ASSOCIATION RULES METHOD IN TOURISM PRODUCT DEVELOPMENT

Pejić Bach Mirjana, Ph.D and **Dumičić Ksenija, Ph.D**
Faculty of Economics and Business, University of Zagreb
Trg J. F. Kennedya 6, 10 000 Zagreb, Croatia
{mpejic , kdumicic }@efzg.hr

Zrinka Marušić, univ.spec.
Bartolići 33, 10000 Zagreb, Croatia
zrinka.marusic@iztzg.hr

Abstract: The association rules methods is used in the analysis of the consumer basket, whereby the analysis of the frequency and the importance of specific rules results in the choice of the activity in a tourist centre which attracts the largest number of guests. The association rules method was applied to the data on activities performed by guests who had participated in TOMAS research on attitudes and consumption habits of tourists visiting the Republic of Croatia. A large number of rules were generated in order to enhance the quality of the insight to the guest activities in tourist destinations.

Keywords: data mining, tourism demand, association rules, sample design

1 INTRODUCTION

Tourism, as one of the main Croatian economic branches should not be left to unsystematic development. In order to gather data which could be used in the decision-making process in the field of tourism in Croatia, the Institute for Tourism organised a research on attitudes and consumption habits of tourists visiting the Republic of Croatia called TOMAS [14].

The tourism demand depends on numerous factors, among which an important role is played by the offer in tourist destinations. A large number of tourists do not seek only for natural beauties, the sun and the see, but they have additional interests such as enjoying cultural contents, adventure activities and similar events. Thus, the tourism product development can be regarded as the design of the market basket in a tourist destination. A potential tourist decides to visit a particular destination, among other things, depending on the offer of the contents from which he chooses the most attractive ones for him. Moreover, tourists rarely choose only one content and they are more often interested in various contents, combining those which are incompatible at first glance. For example, a tourist can combine enjoying wellness offer with adventure trips in national parks. In order to model the consumer basket in an optimal way and thus develop a tourism product, we propose to apply the data mining approach and, in particular, the association rules method.

The data mining has only recently been used in the field of tourism, and the articles on the application of association rules in the tourism demand are rare [12]. To our knowledge, the only reference is the one from the year 2005 where association rules have been applied in order to perform the segmentation of hotel guests in Bursa [1]. The analysis was aimed at the segmentation of the domestic tourism demand in Bursa, although the authors have not applied association rules but the decision tree, which resulted in classification rules such as the following: "if a guest stays in a hotel and values the quality of accommodation and food, he will stay in the hotel six or more days" [9]. In that way, the authors have defined two sub segments of the domestic demand related to guests preferring a longer stay (defined as the stay of six and more days): the segment of guests preferring tourist villages to whom the price of additional services is very important and the segment of guests preferring hotel accommodation and the high quality of accommodation and gastronomic offer. One of the rare applications of data mining, i.e. the application of neuron networks in the tourism research, is the segmentation of the international tourist market in Cape Town, Republic of South Africa [2]. Moreover, some authors in their works analyse the slowness and unpreparedness of "tourism" to accept the knowledge management approach, which can be one of the reasons why the techniques of knowledge discovery from data bases as well as data mining approach have not been more frequently applied in the field of tourism [7, 17].

The aim of this paper is to explore the possibilities of the application of association rules to the development of the tourist product based on the data from TOMAS research on attitudes and consumption habits of tourists visiting the Republic of Croatia. The paper is divided into several sections: after an introductory section, the second section of the article is devoted to characteristics of the association rules method and the third one explores vacation activities of the guests, i.e. at their destination, as one of the important research themes in the field of tourism demand. The fourth section presents the results of a research related to tourist activities during their stay in Croatia. The fifth section shows the model application of association rules in the analysis of the features of the tourism demand while in the sixth section we present the application of association rules in the analysis of guest activities.

2 ASSOCIATION RULES

Association rules [3] are defined in the same way, and described as „simple credible phrases on common appearance of certain events in the data base“, which can be shown as follows (1):

$$\textit{If } A = 1 \textit{ and } B = 1 \textit{ then } C = 1 \textit{ with probability } p, \tag{1}$$

whereby *A*, *B* and *C* are binary variables, and *p* is conditional probability that the variable *C* will gain the value 1 under the condition that variables *A* and *B* gain the value 1, i.e. $p = p(C = 1 \mid A = 1, B = 1)$.

The line of an association rule is the number of variables integrated in the rule. In the phrase mentioned above, the rule line is k + 1. In the association rule, shortly presented as $A \rightarrow B$, A is the antecedent or the body of the rule. B is the consequent of the head of the rule.

The main goal of the association rules method is to detect an „interesting“ set of association rules from the chosen data set, i.e. the rules that can shed a new light onto the analysed data.

In order to determine the interestingness of the rules, various criteria are applied. There are three main measures of statistical significance, or measures of the lift, i.e. the quality of association rules, used in detecting the set of interesting rules. They are support, confidence and lift. If $N_{A \rightarrow B}$ is the number of observations, transactions, in which the rule $A \rightarrow B$ has been satisfied, the significance of the association rule $A \rightarrow B$ is defined in the following way (2):

$$\text{significance}\{A \rightarrow B\} = \frac{N_{A \rightarrow B}}{N}, \tag{2}$$

N is the total number of observations or transactions. The significance of the rule is the number of repeated observations in which the rule $A \rightarrow B$ has been satisfied.

The confidence of the association rule $A \rightarrow B$ is calculated as a proportion of the number of repetitions of the rule $A \rightarrow B$ and the body of the rule A, i.e. as the ratio between the significance of the rule $A \rightarrow B$ and the significance of the body rule A (3):

$$\text{conficende}\{\mathrm{A} \rightarrow \mathrm{B}\} = \frac{N_{A \rightarrow B}}{N_A} = \frac{\frac{N_{A \rightarrow B}}{N}}{\frac{N_A}{N}} = \frac{\text{significance}\{\mathrm{A} \rightarrow \mathrm{B}\}}{\text{significance}\{\mathrm{A}\}}. \tag{3}$$

The confidence expresses the relative number of repetitions of the head of the rule B in the rules containing the body of the rule A. The confidence is the most commonly used measure of interestingness of association rules. The goal of applying the confidence is to measure the strength between the body and the head of the rule. Higher the confidence, higher the probability of appearance of the head of the rule A in case of the body of the rule B.

The correlation of the rule is calculated as a ratio of the rule A → B and the root of the product of the significance of the head of the rule A and the body of the rule B (4):

$$\text{corellation}\{A \rightarrow B\} = \frac{\text{significance}\{A \rightarrow B\}}{\sqrt{\text{significance}\{A\}\text{significance}\{B\}}}. \tag{4}$$

The association rules are calculated by means of several algorithms, whereby the algorithm *A priori* is the oldest and the most commonly used one. They are most frequently applied in the market basket analysis in food industry [10], but in the field of on-line purchase of computers and computer parts as well [16]. Text analysis and text mining are also the fields where association rules are applied. The method of fuzzy association rules[8] is applied in text mining with the aim of better alignment of demands on the basis of extracted rules. Association rules are applied [5] in order to determine the user's basic education depending on which literature he consults. Recently, association rules have been applied in the analysis of web pages search [10]. The goal of such an analysis is to predict the behaviour of internet visitors, i.e. to predict which pages they will visit on the basis of previously visited ones. Besides the assessment of the probability that the user will visit a certain web page, the analysis enables the assessment of the probability of the visit to a certain page from any other web page. Association rules [13] are applied in the analysis of the election survey data in the United States (2003). The author researches the correlation between demographic characteristics of the voters, voting for a particular political party and influence of the media.

3 VACATION ACTIVITIES

In the last two decades, the field of tourism has been characterized by a growing need for adapting the offer to the demand, which shows a great interest in the participation in different sporting and recreational, entertainment, cultural as well as educational activities. Regardless of the main motive for the arrival to a certain destination, the consumer tries to „cease“ every moment of his vacation, expand his knowledge, rise the level of his psychical and physical condition and get acquainted with the culture and other habits of the environment. The tourists want to harmonize their vacation with their living style and enable themselves a creative, physical or spiritual development.

A large portion of the research on tourist activities conducted so far has underlined the correlation of recreational activities and tourism, i.e. has emphasized the major role of recreational activities in the formation of the tourist product [6]. Some authors even consider that it is difficult to distinguish recreation from tourism. There are very few studies on the correlation [3] between daily activities and vacation activities, and that such an information, if a correlation exists, would significantly influence the management and the offer of tourist activities [15]. The authors conducted a research whereby they confirmed a positive correlation between the participation in vacation activities with the participation in the same daily activities. The correlation between daily recreation and recreation during vacation is explained by means of Participation Theory [4].

In order to become even more competitive, the existing destinations, therefore, carefully research and closely follow how potential tourists spend their leisure time, which are their regular interests and hobbies, activities and attitudes. Their success largely depends on the guest satisfaction, but not only in the field of accommodation and gastronomy, but also related to all the contents which will fulfil their need for diverse activities. Tourist market segmentation of successful destinations heavily relies on the research results regarding guest activities and it is reflected in the „tailoring“ of the tourist product as well as in its promotion. Thus, the research on guests' vacation activities, i.e. activities at their tourist destination, is one of the most important research themes in the field of tourism demand.

4 RESEARCH ON TOURIST ACTIVITIES DURING VACATION IN CROATIA

The data which will be analysed by means of association rules relate to the features of tourism demand in Croatia and are gathered in the framework of the research on tourism demand *Attitudes and consumption habits of tourists in Croatia – TOMAS – Summer 2007* [14].

By the allocation of the sample to the destinations with the highest number of tourist arrivals in 2006, a total number of 86 destinations along the coast and on the islands in seven coastal counties were included in the survey. In these destinations, from 71% to 89% of the tourist traffic in 2007 was realized in each of the analysed counties. The sample was allocated proportionally to coastal destinations with regard to realized tourist traffic in 2006. For each chosen destination and type of accommodation (hotel, tourist village, camp and private accommodation) one or more facilities where the survey took place were chosen. In total, there were 130 hotels, 50 tourist villages and 70 camps. The choice of the facility depended on the presence of different categories of facilities and the presence of the size of the facility (measured by the number of permanent beds). In order to perform the survey in private accommodation, the interviewers had to contact tourist boards in selected destinations.

Since the sample has been defined based on the data on the tourist traffic in 2006, and having in mind the used sample design (stratified design but not entirely proportional in individual strata), the sample is not self-pondering. Therefore, the differences in strata structure between the sample and the population are balanced by means of weight coefficients – ponders. The weight coefficient for the stratum i is defined as a ratio between the realized traffic in 2007 in the analysed period and the obtained sample, i.e. ponder of the stratum i which represents tourists from a country n, in an accommodation s and in a county z (5):

$$w_{z,s,n} = \frac{N_{z,s,n}}{n_{z,s,n}}, \qquad (5)$$

whereby

$N_{z,s,n}$ is the tourist traffic realized by tourists from a country n in an accommodation s and in a z in 2007[1], and

$n_{z,s,n}$ is the number of respondents from a country n, in an accommodation s and in a county z.

[1] Central Bureau of Statistics of the Republic of Croatia (2008*). Monthly briefings on the tourist traffic in Croatia in 2007*.

5 MODEL APPLICATION OF ASSOCIATION RULES IN THE ANALYSIS OF TOURISM DEMAND FEATURES

Having in mind the sample design and the use of weights or ponders in order to obtain representative results for the tourism demand in 2007, weights defined for each observation (respondent) at the level of counties, accommodation type and country of origin were used in association rules analysis as well. The weighted (pondered) sample mean is used as an estimate of the population mean, as in case of the estimate of the proportion which also comes down to an estimate of the population mean (6):

$$\overline{X_w} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \tag{6}$$

whereby for every stratum i, N_i is the size of the stratum in the population, n_i is the size of the stratum, and w_i is defined as (7):

$$w_i = \frac{N_i}{n_i}. \tag{7}$$

In case of association rules, it means that the significance of the rule, defined as (8):

$$\text{significance}\{\text{A} \rightarrow \text{B}\} = \frac{N_{A \rightarrow B}}{N}, \tag{8}$$

will be calculated as weighted, pondered significance (9):

$$\overline{\text{significance}\{\text{A} \rightarrow \text{B}\}} = \frac{\sum_{i=1}^{n} w_i A_i B_i}{\sum_{i=1}^{n} w_i}, \tag{9}$$

whereby A_i and B_i are the values of the variables A and B, and w_i is calculated weight.

For the purpose of association rules modelling, the two measures of interestingness will be taken into account: significance and confidence. Due to the fact that the majority of analyzed variables do not appear frequently, the threshold (minimal value) of significance and confidence will be set relatively low and separately for each model. Association rules were applied by means of the programme package *STATISTICA 7*, module *Data Mining*.

6 APPLICATION OF ASSOCIATION RULES IN THE ANALYSIS OF GUEST ACTIVITIES

The guests visiting tourist destinations participate in a large number of activities and it is highly important to set limitations in order not to generate too many rules. If the level of minimal significance is set at 30% and if the activities performed by more than 80% of guests are left out (swimming and bathing, visiting coffee shops and restaurants), more than 800 rules are generated which renders finding causality or correlation between different activities very difficult.

Therefore, the analysis did not include the activities performed by more than 66% of the guests; the level of minimal significance was set to 30%, and the level of minimal confidence to 60%. The results are shown in the Table 1.

The eight rule (*organized trips → trips to national parks*) and the twelfth rule (*trips to national parks → organized trips*) suggest that there is a correlation between going to organized trips and going on trips to national parks. About 36% of guests go to both national parks and on organized trips. Based on that data, it is not clear whether the guests implied only one trip, i. e. organized trip to a national park. However, based on the eight rule (*organized trips → trips to national parks*), it can be concluded that 68% of guests participating in organized trips also go on trips to national parks. On the other hand, 67% of guests going to national parks participate in organized trips as well. Although we cannot asses what is the portion of guests which go on organized trips in national parks, we can assume that all the trips to national parks are not organized and that all organized trips are not trips to national parks.

Diving is practiced by 43% of guests (Table 1). Based on the first rule (*diving → boat riding*) it can be deducted that about 30% of guests go for a boat ride and practice diving, whereby 71% of guests practicing diving go for a boat ride as well. The second rule (*diving → dance/disco*) shows that an equal portion of the guests (30%) go diving and go dancing, whereby 71% of guests practicing diving go dancing as well. The high confidence of this rule is probably due to the fact that younger guests practice diving and therefore go dancing as well.

The highest confidence is encountered in the twenty-fifth rule (*visits to theatre end events → visits to museums and exhibitions*) showing that almost one third of all the guests (significance of the rule 32%) visit theatres, events, museums and exhibitions. However, 85% of the guests visiting theatres and events visit museums and exhibitions as well. Unlike the twenty-fifth rule (*visits to theatre and events → visits to museums and exhibitions*), the twenty-third rule (*visits to museums and exhibitions → visits to theatre and events*) suggests that about two-thirds of those visiting museums and exhibitions will also go to the theatre or an event. As it is the case with research results related to guests' motivation for arrival to a certain destination, the results of this analysis suggest that there is a probability for the guests to participate in a certain combination of activities, which can enable tourism product creators to „pack" certain groups of products depending on the preferences of a particular target group.

Table 1:. Association rules for the set of 21 activities with minimal significance of 30% and minimal confidence of 60%

	Body of the rule	**>**	**Head of the rule**	**Significance (%)**	**Confidence (%)**	**Correlation (%)**
1	Diving	>	Boat riding	30,33	70,52	65,49
2	Diving	>	Dance/disco	30,47	70,83	62,26
3	Boat riding	>	Diving	30,33	60,82	65,49
4	Boat riding	>	Organized trips	32,82	65,81	63,80
5	Boat riding	>	Trips to national parks	33,98	68,14	65,36
6	Boat riding	>	Dance/disco	32,78	65,73	62,21
7	Organized trips	>	Boat riding	32,82	61,85	63,80
8	Organized trips	>	Trips to national parks	36,14	68,11	67,40
9	Organized trips	>	Visits to museums and exhibitions	31,94	60,20	63,56
10	Organized trips	>	Dance/disco	33,74	63,58	62,07
11	Trips to national parks	>	Boat riding	33,98	62,70	65,36
12	Trips to national parks	>	Organized trips	36,14	66,69	67,40
13	Trips to national parks	>	Visits to concerts	34,38	63,43	67,86
14	Trips to national parks	>	Visits to museums and exhibitions	35,20	64,95	69,31
15	Trips to national parks	>	Dance/disco	36,20	66,80	65,91
16	Visits to concerts	>	Organized trips	30,84	65,14	61,53
17	Visits to concerts	>	Trips to national parks	34,38	72,60	67,86
18	Visits to concerts	>	Visits to museums and exhibitions	32,54	68,72	68,54
19	Visits to concerts	>	Dance/disco	35,88	75,78	69,89
20	Visits to museums and exhibitions	>	Organized trips	31,94	67,12	63,56
21	Visits to museums and exhibitions	>	Trips to national parks	35,20	73,97	69,31
22	Visits to museums and exhibitions	>	Visits to concerts	32,54	68,37	68,54
23	Visits to mus. and exhib.	>	Visits to theatre&events	31,82	66,86	75,26
24	Visits to mus. and exhib	>	Dance/disco	33,01	69,35	64,12
25	Visits to theatre and events	>	Visits to museums and exhibitions	31,82	84,73	75,26
26	Dance/disco	>	Organized trips	33,74	60,60	62,07
27	Dance/disco	>	Trips to national parks	36,20	65,03	65,91
28	Dance/disco	>	Visits to concerts	35,88	64,45	69,89

7 CLOSING REMARKS

The paper presents the analysis of guest activities during tourist trips by means of association rules method, whereby the data was gathered in the framework of TOMAS – Summer 2007 research. The application of association rules resulted in already familiar and logic findings on guest activities in tourist destinations, which proved the credibility of this method. Some new insights related to analysed features have been gained as well, which has contributed to the existing interpretation of the results of TOMAS – Summer 2007 research. The significance of these new insights is, however, lower than expected, which is probably the result of the value of features themselves and the ways in which they have been defined. Namely, a great difference existed in the frequency of appearance of analysed variables, which, on the one hand, when the significance threshold was set relatively low, resulted in a great number of rules and difficult interpretation of results, and on the other hand, when the significance threshold was set high, resulted only in appearance of „obvious“ rules. However, the method revealed to be entirely applicable for this set of variables, which opens new possibilities of its application in future research in the field of tourism. We can, therefore, conclude that the aims of the research have been entirely fulfilled. Based on the acquired knowledge on the application of this method to tourism demand data, the future data analysis in the field of tourism by means of association rules should be expanded by including additional features of tourism demand which could contribute to finding new information and knowledge in the data (for example, country of origin, type of accommodation, age and similar).

References

[1] Akat, O., Emel, G. G., Taskin, C. (2005). The Use Of Association Rule Mining For Hotel Customers Profiling: An Application In Bursa. *First International Conference on Business, Management and Economics*, Yaşar University, Çeşme-İzmir, Turska, str. 16–19, lipanj 2005.

[2] Bloom, J. Z. (2005). Market segmentation: A Neural Network Application. *Annals of Tourism Research*, 32 (1), str. 93–111.

[3] Brey, E. T., Lehto, X. Y. (2007). The Relationship Between Daily and Vacation Activities. *Annals of Tourism Research*, 34 (1), str. 160–180.

[4] Broderic, A. i Mueller, R. (1999). A Theoretical and Empirical Exegesis of the Consumer Involvement Construct: The Psychology of the Food Shopper. *Journal of Marketing Theory and Practice*, 7(4), str. 97–108.

[5] Chen, X. i Wu, Y. B. (2006). Personalized Knowledge Discovery: Mining Novel Association Rules from Text. *Proceedings of SIAM SDM'06*, 20. – 22. travnja, 2006, Bethesda, MD, SAD, str. 589–593.

[6] Chubb, M. i Chubb, H. (1981). *One Third of Our Time? An Introduction to Recreation Behavior and Resources.* John Wiley & Sons, Chichester, Engleska.

[7] Cooper, C. (2006). Knowledge management and tourism. *Annals of Tourism Research*, 33 (1), str. 47–64.

[8] Delgado, M., Martín-Bautista, M. J., Sánchez, D., Serrano, J. M. i Miranda, M. A. V. (2003). Association rules and fuzzy association rules to find new query terms. EUSFLAT Conference, str. 49˜53.

[9] Emel, G. G. i Taskin, C. (2005). Identifying Segments of a Domestic Tourism Market by Means of Data Mining. *Operations Research Proceedings*, vol. 2005, str. 653–658.

[10] Giudici, P. (2003). Applied *Data Mining: Statistical Methods for Business and Industry*. John Wiley & Sons, Chichester, Engleska.

[11] Hand, D., Mannila, H. i Smyth, P. (2001). *Principles of Data Mining*. The MIT Press, Cambridge, Massachusetts, SAD.

[12] Law, R., Mok, H. i Goh, C. (2007). Data Mining in Tourism Demand Analysis: A Retrospective Analysis. *Lecture Notes in Computer Science*, vol. 4632, str. 508–515.

[13] MacDougall, M. (2003). Shopping for Voters: Using Association Rules to Discover Relationships in Election Survey Data. *Proceedings of the Twenty-Eighth Annual SAS® Users Group International Conference*. Dostupno na: www2.sas.com/proceedings/sugi28/122-28.pdf. [1. prosinca 2008.]

[14] Marušić, Z., Čorak, S., Hendija, Z. i Ivandić, N. (2008). *Stavovi i potrošnja turista u Hrvatskoj – TOMAS Ljeto 2007*. Institut za turizam, Zagreb.

[15] McKercher, B. (1996). Differences between Tourism and Recreation in Parks. *Annals of Tourism Research*, 23, str. 563–575.

[16] Tan, P., Steinbach, M. i Kumar V. (2006). Introduction to Data Mining. Addison-Wesley, Reading, Massachusetts, SAD.

[17] Xiao, H. i Smith, S. L. J. (2007). The use of Tourism knowledge: Research Propositions. *Annals of Tourism Research*, 34 (2), str. 310–331.

AVERAGE WEIGHTS OF THE COMPOSITE PRICE INDICES

Snježana Pivac and **Anita Udiljak**
Faculty of Economics Split, University of Split
Matice hrvatske 31, 21000 Split, Croatia
spivac@efst.hr, anita.udiljak@efst.hr

Abstract: The composite price indices are frequently used in official statistics and in economic analysis. In this paper the price indices are considered in the term of their weights. In literature there were unweighted indices which in case of quantity changes do not give the real situation about price changes. In practice are more frequently used weighted indices. Many of them are calculated according to the modified Laspeyres formula, which means with weights from the base period. That is the basic indices disadvantage. The suggestion is average weights from the base period and the period for which the index is computed. The comparative analysis is made on the Toyota's cars prices indices in Split - Dalmatian County in Croatia.

Keywords: composite indices, price indices, estimating weights, average weights, harmonic mean weights

1 INTRODUCTION

Price indices have several potential uses, specifically they can help producers with business plans, pricing and can be useful in helping to guide investment. From the macroeconomics point of view they are used as measure of inflation and as deflators of value indicators in economies [3]. In this paper the price indices are considered in the term of their weights and that is the main problem of their calculation and interpretation in practice [9]. The suggestion is average weights from the base period and the period for which the index is computed.

The paper is constructed in four parts. After this introduction part, it is given the overview of the composite price indices. Firstly unweigthed indices are listed and then indices with different weights. The emphasis is on the price indices with weights as arithmetic, geometric and a new harmonic mean. The experiences of the price indices in official statistics in some countries are presented. In the third part of this paper the comparative analysis of indices is made on the Toyota's cars prices in Split - Dalmatian County in Croatia. At the end some conclusion remarks are given.

2 THE COMPOSITE PRICE INDICES

The price indices are typically a weighted average of prices for a given class of goods or services in a given region, during a given period of time. The index is usually computed yearly or quarterly in some countries [1]. They are statistically designed to help to compare how researched prices, taken as a whole, differ between time periods or geographical locations. Various price indices have been constructed in order for better economical analysis and interpretations [11]. Here some most frequently used and familiar price indices are addressed and new price indices with harmonic means weights are involved.

2.1 Unweighted indices

Unweighted price indices or elementary price indices only compare prices between two periods. They do not make any use of quantities or expenditure weights [7]. These indices are called elementary because they are often used at the lower levels of aggregation for more comprehensive price indices. At these lower levels, weights do not matter since only one type of good is being aggregated. That assumption in practice it is not fulfilled and this indices are not represented in complex economic analysis.

Carli index was developed in 1764 by Carli, an Italian economist. Its formula is the arithmetic average of the price relative between a period *t* and a base period *0* [1]:

$$P_C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{p_{ti}}{p_{0i}}\right)\cdot 100 \tag{1}$$

where:
p_{ti} - prices from the period for which the index is computed,
p_{0i} - prices from base period,
n - number of aggregated one type of good.

Dutot index was proposed in 1738 by French economist Dutot. The index is calculated by dividing the average price in period *t* by the average price in period *0* [10]:

$$P_D = \frac{\frac{1}{n}\sum_{i=1}^{n} p_{ti}}{\frac{1}{n}\sum_{i=1}^{n} p_{0i}}\cdot 100 = \frac{\sum_{i=1}^{n} p_{ti}}{\sum_{i=1}^{n} p_{0i}}\cdot 100\,. \tag{2}$$

Jevons index was proposed in 1863 by English economist Jevons. It is calculated taking the geometric average of the price relative of period *t* and base period *0*. When used as an elementary aggregate, the Jevons index is considered a constant elasticity of substitution index since it allows for product substitution between times periods what is the basic disadvantage in practice [1]:

$$P_J = \sqrt[n]{\prod_{i=1}^{n}\frac{p_{ti}}{p_{0i}}}\,. \tag{3}$$

Harmonic mean of price relative index is the harmonic average counterpart to the Carli index. The index was proposed by Jevons in 1865 and by Coggeshall in 1887 [1]:

$$P_{HR} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{p_{0i}}{p_{ti}}}\cdot 100\,. \tag{4}$$

Ratio of harmonic means index or harmonic means price index is the harmonic average counterpart to the Dutot index [10]:

$$P_{RH} = \frac{\sum_{i=1}^{n}\frac{n}{p_{0i}}}{\sum_{i=1}^{n}\frac{n}{p_{ti}}}\cdot 100\,. \tag{5}$$

2.2 Paasche and Laspeyres price indices

The two most basic formulas used to calculate composite price indices are the Paasche index and the Laspeyres index [8].

The Laspeyres price index is computed as the weighted arithmetic mean of individual price indices where the weights are quantities of goods from base period [7]. The base-period index number is thus 100, and periods with higher price levels have index numbers greater than 100 [4]. The simple formula is:

$$P_{0t}(q_0) = \frac{\sum_{i=1}^{n} p_{ti} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} \cdot 100 \quad (6)$$

with the abbreviations standing for:
q_{0i} - quantities from base period,
n - number of aggregated goods.

The Paasche index is computed as the weighted arithmetic mean of individual price indices where the weights are quantities of goods from the period for which the index is computed [7]. The simple formula is:

$$P_{0t}(q_t) = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{0i} q_{ti}} \cdot 100 \quad (7)$$

with the abbreviations standing for:
q_{ti} - quantities from given period,
n - number of aggregated goods.

Laspeyres takes into consideration quantities from base period and Paasche counts with quantities from given period. It is clear that quantities are different from period to period and these indices are not perfect. In practice most frequently used is modified Lasperyers index formula because quantities from base period are more available.

2.3 Marshall-Edgeworth index and Fisher index

The Marshall-Edgeworth index is price index that takes in consideration both quantities from the base period and from the given period, combining the Laspeyres index and the Paasche index. It is created by Marshall (1887) and Edgeworth (1925). The index actually uses the arithmetic average of the current and based period quantities for weighting. The use of the Marshall-Edgeworth index can be problematic in cases such as a comparison of the price level of a large country to a small one. In such instances, the set of quantities of the large country will overwhelm those of the small one [1]:

$$P_{0t} = \frac{\sum_{i=1}^{n} p_{ti} \cdot \frac{1}{2} \cdot (q_{0i} + q_{ti})}{\sum_{i=1}^{n} p_{0i} \cdot \frac{1}{2} (q_{0i} + q_{ti})} . \quad (8)$$

The Fisher index is a price index, computed for a given period by taking the square root of the product of the Paasche index value and the Laspeyres index value, i.e. their geometric mean.

$$P_F = \sqrt{P_{0t}(q_0) \cdot P_{0t}(q_t)} \quad (9)$$

where:
$P_{0t}(q_0)$ - Laspeyres price index,
$P_{0t}(q_t)$ - Paasche price index.

The basic disadvantage of this index is the problem with economic interpretation.

2.4 Hedonic price index

Hedonic price index is any price index, which uses information from hedonic regression [8]. Hedonic regression describes how product price could be explained by the product's characteristics. Hedonic price indices proved to be very useful when applied for information and communication products (e.g. personal computers) to calculate price indices, because they can successfully mitigate such problems as new goods and rapid quality change.

2.5 Walsh price index

The Walsh price index is the weighted sum of the current period prices divided by the weighted sum of the base period prices with the geometric average of both period quantities serving as the weighting mechanism [1]:

$$P_W = \frac{\sum_{i=1}^{n}\left[p_{ti}\cdot\left(q_{0i}\cdot q_{ti}\right)^{\frac{1}{2}}\right]}{\sum_{i=1}^{n}\left[p_{0i}\cdot\left(q_{0i}\cdot q_{ti}\right)^{\frac{1}{2}}\right]}. \qquad (10)$$

2.6 The price indices with average weight

Until now, familiar price indices are Marshall-Egdeworth index in equation (8) and Walsh price index in equation (10). Marshall-Edgeworth index presents composite price index which is based on arithmetic mean of quantities. Walsh price index is composite price index with geometric mean of quantities as weights. It can be taken as presumption that values of these indices are very similar because both of them are taking in consideration some average weights.

It is introduced new composite price Index with harmonic mean of quantities as weights:

$$P_H = \frac{\sum_{i=1}^{n} p_{ti}\left[\frac{2}{\frac{1}{q_{0i}}+\frac{1}{q_{ti}}}\right]}{\sum_{i=1}^{n} p_{0i}\left[\frac{2}{\frac{1}{q_{0i}}+\frac{1}{q_{ti}}}\right]}, \qquad (11)$$

what can be written as:

$$P_H = \frac{\sum_{i=1}^{n} p_{ti}\left[\frac{2q_{0i}q_{ti}}{q_{0i}+q_{ti}}\right]}{\sum_{i=1}^{n} p_{0i}\left[\frac{2q_{0i}q_{ti}}{q_{0i}+q_{ti}}\right]}. \qquad (12)$$

When, taken in comparison, Marshall-Edgeworth and Walsh indices with Index with harmonic mean, it is expected that results will give similar prices trends. Namely, these already familiar indices and this new index with harmonic mean weights are taking into account averages of quantities from base and given periods. Arithmetic, geometric and harmonic means are average values which are used in economic practice and they give similar results [5]. Therefore these average values as indices weights are causing almost identical values.

There is no practical experience about which average value is most reliable. According to that, there is no suggestion which index is most authentic. The real observed price trends are in between of these indices.

2.7 The price indices in official statistics - experience of some countries

In Croatia, Central Bureau of Statistics, from 2004 calculates (before it was cost of living indices) Consumer Price Indices (CPI) on the basis of index lists of representative products and services as well as related weights (or structures), according to the modified Laspeyres formula:

$$I = \frac{\sum_{i=1}^{n} \frac{p_{ti}}{p_{0i}} \cdot w_{0i}}{\sum_{i=1}^{n} w_{0i}} \cdot 100 \qquad (13)$$

where:
w_{0i} - relative structure of the sales value in the base period.

Individual indices for each representative product are calculated from data on prices by dividing prices of a current month by prices of a reference month. Aggregate indices and the total index are calculated by weighted arithmetic mean from individual indices according to the modified Laspeyres formula [13]-[29].

The Consumer Prices Index (CPI) and Retail Prices Index (RPI) are the main UK measure of inflation for macroeconomic purposes and the Government's inflation target are formed on these indices. They are also used for international comparisons. CPI and RPI both measure the average changes month-to-month in prices of consumer goods and services purchased in the UK, although there are differences in coverage and methodology.

Internationally, the CPI is known as the Harmonized Index of Consumer Prices (HICP), although the two indices remain one and the same. The First Release contains HICP data for countries across the European Union; where each country computes some 80 prescribed sub-indices, their weighted average constituting their national Harmonized Index. The classification was developed in the national accounting context.

On some stock markets, indices have been very useful for stock exchange traffic analysis and it helps the investors diversify their risk. For example, the NASDAQ Composite index is a market capitalization-weighted grouping of approximately 5000 stocks listed on the NASDAQ market. In Zagreb Stock Exchange Crobex index is weight composite index for most frequently rated stocks. These indices are useful tools for measuring and tracking price level changes to an entire stock market or sector. Therefore, they provide a useful benchmark against which to measure an investor's portfolio. The goal of a well diversified portfolio is usually to outperform the main composite indices. So, composite price indices have great affect in every important economic analysis on macroeconomic and microeconomic levels.

3 THE PRICES CHANGES OF THE TOYOTA CARS

Toyota Motor Corporation is a multinational corporation headquartered in Japan, and is the world's largest automaker. Toyota employs approximately 316 000 people worldwide.

In consideration were taken prices and quantities for 6 Toyota car models: Yaris, Auris, Rav4, Corolla, Avensis and Land Cruiser. Data were taken from Toyota Condic-Kadmenovic branch office in Croatian, Split - Dalmatian County in period from 2005 to 2009 year. It must be mentioned that in 2009 year were taken proximal values based on first four months of the year.

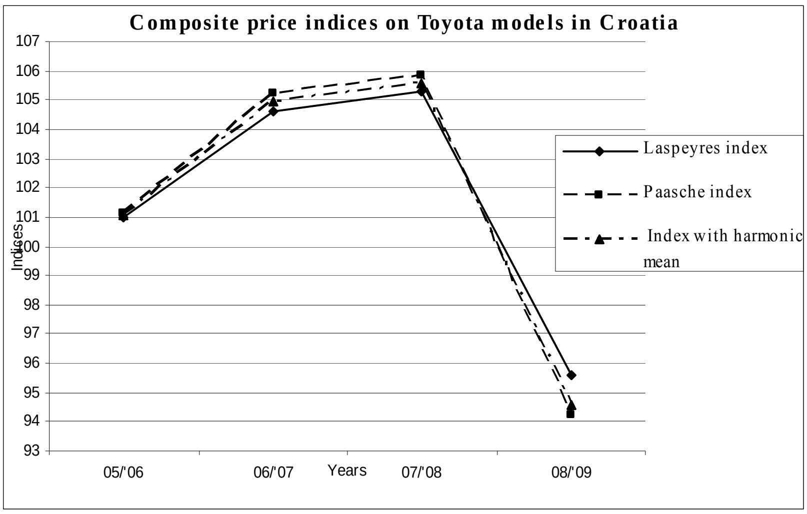
In Table 1, there are Laspeyres, Paasche [2], [12] and Index with harmonic mean of chosen Toyota cars prices in Split - Dalmatian County. From results in Table 1 it can be concluded that there are differences between Laspeyres and Paasche indices. That was expected because Laspeyres index is calculated according to cars quantities from base period and Paasche index according to given period. Index with harmonic mean has values between these two before mentioned indices.

Table 1: Composite price indices on Toyota cars in Split - Dalmatian County, Croatia.

Years	***Laspeyres index***	***Paasche index***	***Index with harmonic mean***
05/'06	100.974	101.151	101.061
06/'07	104.618	105.191	104.930
07/'08	105.295	105.842	105.553
08/'09	95.593	94.200	94.571

Source: Calculated according to Toyota Condic-Kadmenovic office database

The same conclusion is obvious from Figure 1. Index with harmonic mean [6], weights is between Laspeyres and Paasche indices for all observed years. In last year, there is evident downfall of Toyota cars prices. The reasons for this downfall can be found in global recession environment, which has not bypass Croatian consumption market.



Source: Calculated according to Toyota Condic-Kadmenovic office database

Figure 1: Laspeyres and Paasche price indices and Index with harmonic mean.

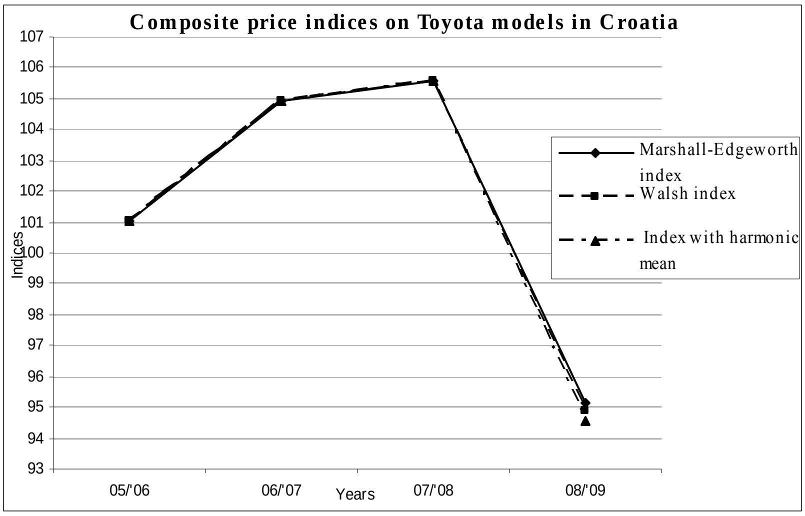
In Table 2, the price indices with average weights are calculated. Marshall-Edgeworth price index has arithmetic mean of car quantities from base and given period as weights and Walsh price index is with geometric mean of car quantities as weights. The new Index with harmonic mean is shown in the same table. Comparative analysis of results shows the same values of price indices for appropriate period. These indices take in consideration quantities of sold Toyota cars in base and current periods as weights, in Split - Dalmatian County.

Table 2: Composite price indices on Toyota cars in Split - Dalmatian County, Croatia.

Years	***Marshall-Edgeworth index***	***Walsh index***	***Index with harmonic mean***
05/'06	101.066	101.064	101.061
06/'07	104.910	104.920	104.930
07/'08	105.557	105.555	105.553
08/'09	95.168	94.860	94.571

Source: Calculated according to Toyota Condic-Kadmenovic office database

The price indices from Table 2 are shown on Figure 2. It can be seen that values of these indices are overlapping. The real composite price changes are somewhere in between of presented indices values.



Source: Calculated according to Toyota Condic-Kadmenovic office database

Figure 2: Marshall-Edgeworth and Walsh price indices and Index with harmonic mean.

Indices with average weights are showing actual price trends. While in the case of large quantites changes in base and current periods, there is great difference from Laspeyres index to Pasche index.

4 CONCLUSION REMARKS

In this paper the price indices are considered in the term of their weights. In practice are more frequently used weighted indices. Many of them are calculated according to the modified Laspeyres formula, which means with weights from the base period. It is presumption that it is used because of easier data availability in previous period. Although, it is known that quantities are changeable from period to period and that is the basic indices disadvantage.

The new Index with harmonic mean as weight is presented. It shows similar price trends as other known composite price indices with average weights. Because of great importance of these indices in economic analysis, it is advised when calculating price changes to accept weights from both observed periods.

In regard to, composite price indices with average weights are showing similar values, in practice price movement should be contemplate with mention indices.

The comparative analysis of different price indices is made on the Toyota's cars prices in Split - Dalmatian County in Croatia. It is indicated that prices of Toyota car models were increased from 2005 to 2008, but in 2009 is registered downfall of prices due to global economic recession environment.

References

[1] Diewert, W. E., 1993. Chapter 5: Index Numbers in Essays in Index Number Theory. Eds: W.E. Diewert and A.O. Nakamura. Vol 1. Elsevier Science Publishers.

[2] De Levie, R., 2004. Advanced Excel for Scientific Data Analysis. Oxford University Press, Oxford, 615 p.

[3] Fanning, S. F., 2005. Market Analysis for Real Estate: Concepts and Applications in Valuation and Highest and Best Use. Appraisal Institute, Chicago, 543 p.

[4] Pavlić, I., 1971. Statistička teorija i primjena. Tehnička knjiga, Zagreb, p 343.

[5] Petz, B., 2004. Osnovne statističke metode za nematematičare. Naklada Slap, Zagreb, p. 384.

[6] Pivac, S, Jurun, E., 2005. Parameter Estimation in Excel. Proceedings of the 28th International Convention MIPRO 2005. Computers in Education, Opatija, pp. 168.-173.

[7] Pivac, S., Šego, B., 2005. Statistika, udžbenik sa zbirkom zadataka za IV. razred srednje ekonomske škole. Alkascript, Zagreb, p. 194.

[8] Rozga, A., Grčić, B., 2003. Poslovna statistika. Ekonomski fakultet Split, p. 271.

[9] Siegel, A. F., 1994. Practical Business Statistics. IRVIN Publishing, Boston, Massachusetts, p. 820.

[10] Silver, M., Heravi, S., 2006. Why elementary price index number formulas differ: Evidence on price dispersion. Journal of Econometrics, Vol. 140, Issue 2, pp. 874.-883.

[11] Šošić, I., 1983. Metode statističke analize. Sveučilišna naklada Liber, Zagreb, p 474.

[12] Whigham, D., 1998. Quantitative Business Methods Using Excel. Oxford University Press, Oxford, p. 467.

[13] http://www.abs.gov.au/ausstats/abs%40.nsf/mf/6401.0; (Consumer Price Index, Australia, Australian Bureau of Statistics)

[14] http://www.rba.gov.au/Statistics/Bulletin/G02hist.xls; (Consumer Price Index, Reserve Bank of Australia)

[15] http://www.stats.gov.cn/english/statisticaldata/index.htm; (Consumer Price Index, National Bureau of Statistics of China)

[16] http://www.banrep.gov.co/series-estadisticas/see_precios.htm; (Banco de la República Colombia, Series estadísticas, Precios- IPC -IPP)

[17] http://www.dzs.hr; (The Central Bureau of Statistics, Croatia)

[18] http://www.stat.fi/til/khi/index_en.html; (Consumer Price Index, Statistics Finland)

[19] http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Statistiken/Zeitreihen/WirtschaftAktuell/Basisdaten/Content100/vpi101a,templateId=renderPrint.psml; (Verbraucherpreisindex für Deutschland, Federal Statistical Office of Germany)

[20] http://www.rediff.com/money/2007/jun/07infla.htm; (Central Statistical Organisation India)

[21] http://www.cso.ie/statistics/consumpriceindex.htm; (Consumer Price Index, Central Statistics Office Ireland)

[22] http://www.istat.it/salastampa/comunicati/in_calendario/precon/20081014_00/consumer%20price%20092008.pdf; (Italian national statistical institute)

[23] http://www.censtatd.gov.hk/hong_kong_statistics/statistical_tables/; (Census and Statistics Department of Hong Kong)

[24] http://www.banxico.org.mx/sitioIngles/index.html); (Banco de Mexico)

[25] http://www.stats.govt.nz/datasets/economic-indicators/consumers-price-index-cpi.htm; (Statistics New Zealand)

[26] http://www.jos.nu/; (Statistics Sweden)

[27] http://www.bfs.admin.ch/bfs/portal/en/index/themen/systemes_d_indicateurs/economic_and_financial/data.html; (Swiss Federal Statistical Office)

[28] http://www.statistics.gov.uk/cci/nugget.asp?id=181; (Office for National Statistics U.K.)

[29] http://www.bls.gov/bls/glossary.htm; (United States Department of Labour, Bureau of Labour Statistics)

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Appendix

Authors' addresses

Addresses of SOR'09 Authors

(The 10th International Symposium on OR in Slovenia, Nova Gorica, SLOVENIA, September 23 – 25, 2009)

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
1.	Assil	Al-Ibrahim	University of Amsterdam, Faculty of Economics and Business	Roetersstraat 11	1018WB	Amsterdam	the Netherlands	a.alibrahim@uva.nl
2.	Zdravka	Aljinović	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	zdravka.aljinovic@ efst.hr
3.	Josip	Arnerić	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	jarneric@efst.hr
4.	Jan	Babič	Jožef Štefan Institute	Jamova 39	1000	Ljubljana	Slovenia	jan.babic@ijs.si
5.	Zoran	Babić	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	babic@efst.hr
6.	Vlasta	Bahovec	University of Zagreb, Faculty of Economics and Business	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	bahovec@efzg.hr
7.	Darko	Bečaj	Snaga javno podjetje d.o.o.	Nasipna 64	2000	Maribor	Slovenia	darko.becaj@ snaga-mb.si
8.	Jani	Bekő	University of Maribor, Faculty of Economics and Business Maribor	Razlagova 14	2000	Maribor	Slovenia	jani.beko@uni-mb.si
9.	Ludvik	Bogataj	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1000	Ljubljana	Slovenia	ludvik.bogataj@ ef.uni-lj.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
10.	Marija	Bogataj	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1000	Ljubljana	Slovenia	marija.bogataj@ ef.uni-lj.si
11.	Darja	Boršič	University of Maribor, Faculty of Economics and Business Maribor	Razlagova 14	2000	Maribor	Slovenia	darja.borsic@uni-mb.si
12.	Andrej	Bregar	Informatika d.d.	Vetrinjska ulica 2	2000	Maribor	Slovenia	andrej.bregar@ informatika.si
13.	Vincenc	Butala	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	vincenc.butala@ fs.uni-lj.si
14.	Kristijan	Cafuta	University in Ljubljana, Faculty of Electrical Engineering		1000	Ljubljana	Slovenia	kristijan.cafuta@ fe.uni-lj.si
15.	Tibor	Csendes	University of Szeged, Institute of Informatics	Árpád tér 2	6720	Szeged	Hungary	csendes@ inf.u-szeged.hu
16.	Vesna	Čančer	University of Maribor, Faculty of Economics and Business Maribor	Razlagova 14	2000	Maribor	Slovenia	vesna.cancer@uni-mb.si
17.	Anita	Čeh Časni	University of Zagreb, Graduate School of Economics and Business, Department of Statistics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	aceh@efzg.hr
18.	Irena	Čibarić	University of Zagreb, Graduate School of Economics and Business, Department of Statistics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	icibaric@efzg.hr
19.	Mirjana	Čižmešija	University of Zagreb, Faculty Economics and Business,	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	mcizmesija@efzg.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
20.	Andreja	Čurin	University of Maribor, Faculty of Logistics	Mariborska cesta 7	3000	Celje	Slovenia	andreja.curin@ fl.uni-mb.si
21.	Louis	Dorard	University College London, Department of Computer Science	WC1E 6BT		London	United Kingdom	l.dorard@cs.ucl.ac.uk
22.	Samo	Drobne	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	samo.drobne@ fgg.uni-lj.si
23.	Ksenija	Dumičić	University of Zagreb, Graduate School of Economics and Business, Department of Statistics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	kdumicic@efzg.hr
24.	Nataša	Erjavec	University of Zagreb, Faculty of Economics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	nerjavec@efzg.hr
25.	Daša	Fabjan	University of Primorska, Faculty of Tourism Studies			Portorož	Slovenia	dasa.fabjan@ turistica.si
26.	M. Victoria	de la Fuente	Universidad Politécnica de Cartagena	Campus Muralla del Mar	30202	Cartagena	Spain	
27.	Dorota	Glowacka	University College London, Department of Computer Science	WC1E 6BT		London	United Kingdom	d.glowacka@ cs.ucl.ac.uk
28.	Petra	Grošelj	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	petra.groselj@bf.uni-lj.si
29.	Robert W.	Grubbström	Linköping Institute of Technology	83 Linköping	SE-581	Linköping	Sweden	robert@grubbstrom.com
30.	Gregor	Guna	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	gunagregor@yahoo.com

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
31.	René	Haijema	Wageningen University and Research Centre, Logistics, Decision and Information sciences	Hollandseweg 1	6706 KN	Wageningen	the Netherlands	r.haijema@wur.nl
32.	Karsten	Hentsch	Chemnitz University of Technology, Department of Computer Science	Straße der Nationen 62	09107	Chemnitz	Germany	karsten.hentsch@ informatik.tu-chemnitz.de
33.	Dušan	Hvalica	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1001	Ljubljana	Slovenia	dusan.hvalica@ ef.uni-lj.si
34.	Dmitry	Ivanov	Chemnitz University of Technology, Faculty of Economics and Business Administration		09107	Chemnitz	Germany	dmitry.ivanov@ wirtschaft.tu-chemnitz.de
35.	Marko	Jakšič	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1001	Ljubljana	Slovenia	marko.jaksic@ ef.uni-lj.si
36.	Jaroslav	Janaček	University of Žilina, Faculty of Management and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republik	jaroslav.janacek@ fri.uniza.sk
37.	Marta	Janáčková	University of Žilina, Faculty of Science and Informatics	Univerzitná 2	010 26	Žilina	Slovak Republik	marta.janackova@ fstroj.uniza.sk
38.	Borut	Jereb	University of Maribor, Faculty of Logistics	Mariborska 7	3000	Celje	Slovenia	borut.jereb@fl.uni-mb.si
39.	Elza	Jurun	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	elza@efst.hr
40.	Joachim	Kaeschel	Chemnitz University of Technology, Faculty of Economics and Business Administration		09107	Chemnitz	Germany	

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
41.	Stane	Kavčič	University of Ljubljana, Biotechnical Faculty, Department of Animal Science	Groblje 3	1230	Domžale	Slovenia	stane.kavcic@ bfro.uni-lj.si
42.	Alenka	Kavkler	University of Maribor, Faculty of Economics and Business Maribor	Razlagova 14	2000	Maribor	Slovenia	alenka.kavkler@ uni-mb.si
43.	Igor	Klep	University of Maribor, Faculty of Natural Science and Mathematics, Universtity of Ljubljana, Faculty of Mathematics and Physics		2000 1000	Maribor, Ljubljana	Slovenia	igor.klep@ fmf.uni-lj.si
44.	Peter	Köchel	Chemnitz University of Technology, Faculty of Informatics, Chair of Modelling & Simulation	Straße der Nationen 62	09107	Chemnitz	Germany	peter.koechel@ informatik.tu-chemnitz.de
45.	Michal	Koháni	Faculty of Management and Informatics, University of Žilina		010 26		Slovak Republic	michal.kohani@ fri.uniza.sk
46.	Miha	Konjar					Slovenia	mkonjar@yahoo.com
47.	Danijel	Kovačić	University of Ljubljana, Faculty of Economics	Kardeljeva ploščad 17	1001	Ljubljana	Slovenia	danijel.kovacic@ ef.uni-lj.si
48.	Uroš	Kramar	University of Maribor, Faculty of Logistics	Mariborska cesta 7	3000	Celje	Slovenia	uros.kramar@ fl.uni-mb.si
49.	Tomaž	Kramberger	University of Maribor, Faculty of Logistics	Mariborska cesta 7	3000	Celje	Slovenia	tomaz.kramberger@ uni-mb.si
50.	Markus	Kraus	Institute for Advanced Studies, Department of Economics and Finance	Stumpergasse 56	1060	Vienna	Austria	kraus@ihs.ac.at

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
51.	Janez	Krč	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	janez.krc@bf.uni-lj.si
52.	Miran	Kuhar	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	mkuhar@fgg.uni-lj.si
53.	Nataša	Kurnoga Živadinović	University of Zagreb, Faculty Economics and Business,	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	nkurnoga@efzg.hr
54.	Ali Osman	Kusakci	Faculty of Engineering and Natural Sciences, International University of Sarajevo		71000	Sarajevo	Bosnia and Herzegovina	akusakci@ius.edu.ba
55.	Janez	Kušar	University of Ljubljana, Faculty of Mechanical Engineering	Aškrčeva 2	1000	Ljubljana	Slovenia	janez.kusar@fs.uni-lj.si
56.	Lado	Lenart	Jožef Štefan Institute	Jamova 39	1000	Ljubljana	Slovenia	lado.lenart@ijs.si
57.	Anka	Lisec	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	anka.lisec@fgg.uni-lj.si
58.	Zrinka	Marušić		Bartolići 33	10000	Zagreb	Croatia	zrinka.marusic@ iztzg.hr
59.	Miklavž	Mastinšek	University of Maribor, Faculty of Economics and Business	Razlagova 14	2000	Maribor	Slovenia	mastinsek@uni-mb.si
60.	Peter	Matis	Faculty of Management and Informatics, University of Žilina		010 26		Slovak Republic	peter.matis@ fri.uniza.sk
61.	Zoran	Mihanović	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	zoran.mihanovic@ efst.hr
62.	Miloš	Milovanović	Faculty of Organizational Science, University of Belgrade	Jove Ilića 154	11000	Belgrade	Serbia	milovanovicm@fon.rs

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
63.	Jelena	Minović	Belgrade Banking Academy, Faculty for Banking, Insurance and Finance	Zmaj Jovina 12	11000	Belgrade	Serbia	jminovic@ien.bg.ac.rs
64.	Miroslav	Minović	Faculty of Organizational Science, University of Belgrade	Jove Ilića 154	11000	Belgrade	Serbia	mminovic@fon.rs
65.	Dubravko	Mojsinović	Consule d.o.o.	Franje Tuđmana 8	10434	Strmec Samoborski	Croatia	dmojsinovic@ consule.hr
66.	Dejan	Paliska	University of Ljubljana, Faculty of Maritime Studies and Transport	Pot pomorščakov 4	6320	Portorož	Slovenia	dejan.paliska@ fpp uni-lj.si
67.	Polona	Pavlovčič Prešeren	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	ppavlovc@fgg.uni-lj.si
68.	Mirjana	Pejić Bach	University of Zagreb, Graduate School of Economics and Business, Department of Statistics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	mpejic@efzg.hr
69.	Mario	Pepur	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	mario.pepur@efst.hr
70.	Tunjo	Perić	Bakeries Sunce	Rakitska cesta 98	10437	Bestovje	Croatia	tunjo.peric1@ zg.t-com.hr
71.	Igor	Pesek	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	igor.pesek@ uni-mb.si
72.	Špera	Pezdevšek Malovrh	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	spela.malovrh@ bf.uni-lj.si
73.	Snježana	Pivac	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	spivac@efst.hr

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
74.	Janez	Povh	Faculty of Information Studies	Na Loko 2	8000	Novo mesto	Slovenia	janez.povh@ fis.unm.si
75.	Marion	Rauner	University of Vienna, Faculty of Business, Economics and Statistics	Bruenner Str. 72	1210	Vienna	Austria	marion.rauner@ univie.ac.at
76.	Lorenzo	Ros	Universidad Politécnica de Cartagena	Campus Muralla del Mar	30202	Cartagena	Spain	
77.	Viljem	Rupnik	INTERACTA, LTD, Business Information Processing	Parmova 53	1000	Ljubljana	Slovenia	viljem.rupnik@siol.net
78.	John	Shawe-Taylor	University College London, Department of Computer Science	WC1E 6BT		London	United Kingdom	jst@cs.ucl.ac.uk
79.	Ivana	Simeunović	Institute of Economic Sciences	Zmaj Jovina 12	11000	Belgrade	Serbia	isimeunovic@ ien.bg.ac.rs
80.	Aleš	Slak	Iskra ISD-Strugarstvo d.o.o.	Savska loka 4	4000	Kranj	Slovenia	ales.slak@iskra-isd.si
81.	Moshe	Sniedovich	University of Melbourne, Department of Mathematics and Statistics		VIC 3010		Australia	moshe@ unimelb.edu.au
82.	Boris	Sokolov	Russian Academy of Science, St. Petersburg Institute of Informatics and Automation	39, 14 Linia, VO	199178	St.Petersburg	Russia	sokol@iias.spb.su
83.	Petar	Sorić	University of Zagreb, Faculty Economics and Business,	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	psoric@efzg.hr
84.	Aleksander	Srdić	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	asrdic@fgg.uni-lj.si

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
85.	Oskar	Sterle	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	osterle@fgg.uni-lj.si
86.	Bojan	Stopar	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	bstopar@fgg.uni-lj.si
87.	Alžbeta	Szendreyová	University of Žilina, Faculty of Science and Informatics	Univerzitná 2	010 26	Žilina	Slovak Republik	alzbeta.szendreyova@ fstroj.uniza.sk
88.	Nataša	Šarlija	University of Osijek, Faculty of Economics	Gajev trg 7	31000	Osijek	Croatia	natasa@efos.hr
89.	Jana	Šelih	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova 2	1000	Ljubljana	Slovenia	jselih@fgg.uni-lj.si
90.	Blanka	Škrabić	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	bskrabic@efst.hr
91.	Kristina	Šorić	University of Zagreb, Faculty of Economics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	ksoric@efzg.hr
92.	Ivana	Tadić	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	itadic@efst.hr
93.	Jože	Tavčar	Iskra Mehanizmi	Lipnica 8	4245	Kropa	Slovenia	joze.tavcar@ iskra-mehanizmi.si
94.	Ivana	Tomas	Facilty of Economics, University of Rijeka	Ivana Filipovića 4	51000	Rijeka	Croatia	ivana.tomas@efri.hr
95.	Tadeusz	Trzaskalik	The Karol Adamiecki University of Economics in Katowice, Department of Operations Research	Ul. Bogucicka 14	40-587	Katowice	Poland	ttrzaska @ae.katowice.pl

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
96.	Danijela	Tuljak Suban	University of Ljubljana, Faculty of Maritime Studies and Transport			Portorož	Slovenia	danijela.tuljak@ fpp.uni-lj.si
97.	Anita	Udiljak	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	anita.udiljak@efst.hr
98.	Erik	Van der Sluis	University of Amsterdam, Faculty of Economics and Business	Roetersstraat 11	1018WB	Amsterdam	the Netherlands	h.j.vandersluis@uva.nl
99.	Jan	Van der Wal	University of Amsterdam, Faculty of Economics and Business	Roetersstraat 11	1018WB	Amsterdam	the Netherlands	j.vanderwal@uva.nl
100.	Nico M.	Van Dijk	University of Amsterdam, Faculty of Economics and Business	Roetersstraat 11	1018WB	Amsterdam	the Netherlands	n.m.vandijk@uva.nl
101.	Jelena	Vidović	University Centre for Professional Studies	Livanjska 5/III	21000	Split	Croatia	jvidovic@oss.unist.hr
102.	Josipa	Višić	University of Split, Faculty of Economics	Matice Hrvatske 31	21000	Split	Croatia	josipa.visic@efst.hr
103.	Silvija	Vlah	University of Zagreb, Faculty of Economics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	svlah@efzg.hr
104.	Višnja	Vojvodić Rosenzweig	University of Zagreb, Faculty of Economics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	vvojvodic@efzg.hr
105.	Tomasz	Wachowicz	The Karol Adamiecki University of Economics in Katowice	ul. Bogucicka 14	40-227	Katowice	Poland	tomasz.wachowicz@ ae.katowice.pl
106.	Pawel	Wieszala	The Karol Adamiecki University of Economics in Katowice	ul. Bogucicka 14	40-227	Katowice	Poland	pawel.wieszala @ ae.katowice.pl
107.	Di	Yuan	Department of Science and Tehnology, Linköping University		601	Norrköping	Sweden	diyuan@itn.liu.se

ID	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
108.	Lidija	Zadnik Stirn	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	lidija.zadnik@bf.uni-lj.si
109.	Aleksandar	Zdravković	Union Universtity of Belgrade, Belgrade Banking Academy	Zmaj Jovina 12	11000	Belgrade	Serbia	aleksandar.zdravkovic@ ien.bg.ac.rs
110.	Gašper	Žerovnik	Jozef Stefan Institute	Jamova 39	1000	Ljubljana	Slovenia	gasper.zerovnik@ ijs.si
111.	Janez	Žerovnik	IMFM	Jadranska 19	1000	Ljubljana	Slovenia	janez.zerovnik@ imfm.uni-lj.si
112.	Jaka	Žgajnar	University of Ljubljana, Biotechnical Faculty, Department of Animal Science	Groblje 3	1230	Domžale	Slovenia	jaka.zgajnar@ bfro.uni-lj.si

The 10th International Symposium on
Operational Research in Slovenia

SOR '09

Nova Gorica, SLOVENIA
September 23 - 25, 2009

Appendix

Sponsors' Notices

ASO LJUBLJANA

AUSTRIAN SCIENCE AND RESEARCH LIAISON OFFICE LJUBLJANA

Austrian Science and Research Liaison Office (ASO) Ljubljana

The Austrian Science and Research Liaison Office Ljubljana has been established in October 1990 as branch office of the Vienna based Austrian Institute for East and Southeast European Studies to foster scientific co-operation between Austria and Slovenia. ASO Ljubljana has been reorganised in March 2004 and is since that time part of the Centre for Social Innovation (ZSI) in Vienna. ASO Ljubljana receives its funding mainly from Austrian Federal Ministry of Science and Research (bm:wf) and partially also from Ministry of Higher Education, Science and Technology of Republic of Slovenia.

ZSI is in charge of coordination of activities of ASO Ljubljana with bm:wf as well as of coordination with regard to national, bilateral and international initiatives and programmes. The Austrian Science and Research Liaison Office in Ljubljana supports the science policy of Austria in South Eastern Europe which is co-ordinated on European level with projects and initiatives like SEE-ERA.net www.see-era.net, RTD polcy dialogue platform Western Balkan INCO.net http://wbc-inco.net/ , etc.

Some highlights of ASO Ljubljana work:

ASO Ljubljana has been co-organiser of the **Bled Forum on Europe SEE Foresight training seminars** since 2006.
In September 2006 ASO Ljubljana organised together with UNESCO Office in Venice and Slovenian Ministry of Higher Education, Science and Technology the International conference and Ministerial Roundtable "**Why invest in science in SEE countries?**" http://investsciencesee.info/

ASO Ljubljana has been partner of The **Academic Council on the United Nations System (ACUNS)** www.acuns.org in organizing the eighteenth ACUNS-ASIL Summer Workshop on International Organization Studies; the topic has been "**Building the Knowledge Base for Global Governance**" in summer 2008 (Mr. Polzer has been academic co-director of the summer workshop) http://www.aso.zsi.at/sl/veranstaltung/2995.html

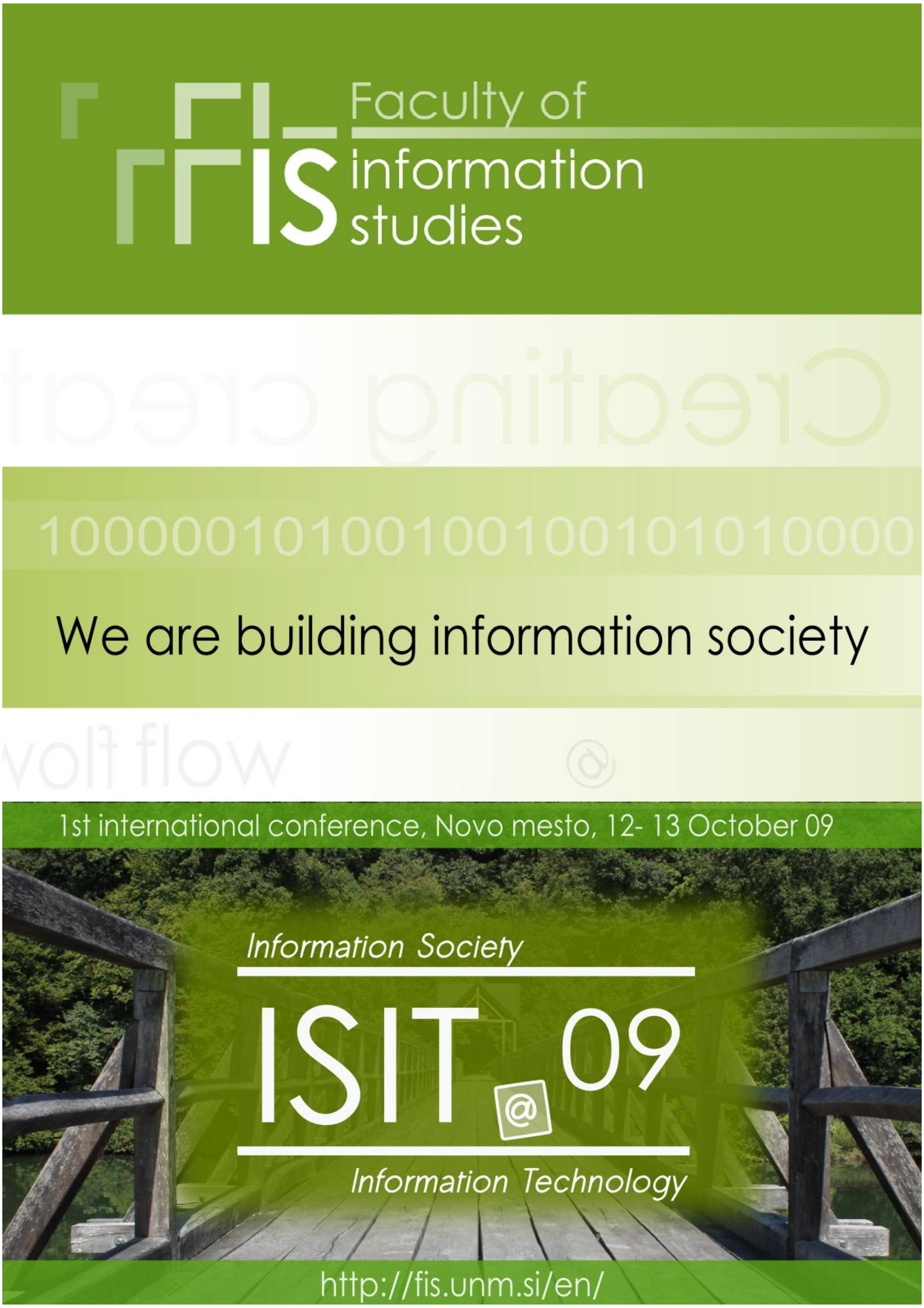
On behalf of Austrian Federal Ministry of Science and Research ASO Ljubljana in cooperation with colleagues at ZSI administer on an biannual basis "**Calls for Proposals of the ASO Ljubljana on Research Cooperation and Networking between Austria, Slovenia, South Eastern Europe**" (for more details see: http://www.aso.zsi.at/ausschreibung/list ; next call is expected to be launched end of 2009).

Contact:

Austrian Science and Research Liaison Office Ljubljana (ASO) / Avstrijski znanstveni institut v Ljubljani/ Österreichisches Wissenschaftsbüro Ljubljana,
Dunajska 104; SI-1000 Ljubljana; Slovenija
e-mail: aso-ljubljana@zsi.at; homepage: www.aso.zsi.at
tel: 00 386 (0) 1 5684 168 fax: 00 386 (0) 1 5684 169


Miroslav Polzer, Head of ASO Ljubljana


Gorazd Weiss, Programme Manager at ASO Ljubljana


FIS
Faculty of information studies
1000001010010010010101000
We are building information society
1st international conference, Novo mesto, 12- 13 October 09
Information Society
ISIT @ 09
Information Technology
http://fis.unm.si/en/