Proceedings of the 15th International Symposium on OPERATIONAL RESEARCH

SOR '19

Bled, Slovenia

September 25-27, 2019

Edited by:
L. Zadnik Stirn • M. KljajiæBorštnar • J. Žerovnik • S. Drobne • J. Povh

SOR '19 Proceedings

The 15th International Symposium on Operational Research in Slovenia

Bled, SLOVENIA, September 25 - 27, 2019

Edited by:

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

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

Proceedings of the 15th International Symposium on Operational Research
SOR'19 in Slovenia, Bled, September 25 - 27, 2019.

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

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

Co-organiser : University of Ljubljana, Faculty of Mechanical Engineering, SI-1000 Ljubljana, Aškerčeva cesta 6, Slovenia (https://www.fs.uni-lj.si/)

First published in Slovenia in 2019 by Slovenian Society Informatika – Section for Operational Research, SI 1000 Ljubljana, Litostrojska cesta 54, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

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

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

INTERNATIONAL Symposium on Operational Research in Slovenia (15 ; 2019 ; Bled)
SOR '19 proceedings / The 15th International Symposium on Operational Research in Slovenia, Bled, Slovenia, September 25-27, 2019 ; [organiser] Slovenian Society Informatika (SDI), Section for Operational Research (SOR) ; [co-organiser University of Maribor, Faculty of Organizational Sciences [and] University of Ljubljana, Faculty of Mechanical Engineering] ; edited by L. Zadnik Stirn ... [et al.]. - Ljubljana : Slovenian Society Informatika, Section for Operational Research, 2019

ISBN 978-961-6165-55-6
COBISS.SI-ID 301633536

Proceedings of the 15th International Symposium on Operational Research in Slovenia (SOR'19) is cited in: ISI (Index to Scientific & Technical Proceedings on CD-ROM and ISI/ISTP&B online database), Current Mathematical Publications, Mathematical Review, MathSci, Zentralblatt für Mathematic / Mathematics Abstracts, MATH on STN International, CompactMath, INSPEC, Journal of Economic Literature

Technical editor : Samo Drobne
Designed by : Samo Drobne
Printed by : BISTISK d.o.o., Ljubljana, Slovenia
Number of copies printed: 160

*The 15th International Symposium on Operational Research in Slovenia - **SOR '19***
Bled, SLOVENIA, September 25 - 27, 2019

Program Committee:

Organizing Committee:

Reviewers:

Chairs:

Preface

This volume, Proceedings of The 15th International Symposium on Operations Research, called SOR'19, contains papers presented at SOR'19 (http://sor19.fov.uni-mb.si/) that was organized by Slovenian Society INFORMATIKA (SDI), Section for Operations Research (SOR), University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia, and University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia, held in Bled, Slovenia, from September 25 to September 27, 2019. The volume contains blindly reviewed papers or abstracts of talks presented at the symposium.

The opening address at SOR'19 was given by Prof. Dr. Lidija Zadnik Stirn, President of the Slovenian Section of Operations Research, Mr. Niko Schlamberger, President of the Slovenian Society Informatika, Prof. Dr. Iztok Podbregar, Dean of the Faculty of Organizational Sciences, University of Maribor, Prof. Dr. Mitjan Kalin, Dean of the Faculty of Mechanical Engineering, University of Ljubljana, Prof. Dr. Immanuel Bomze, President of The Association of European Operational Research Societies (EURO),), Prof. Dr. Zrinka Lukać, President of Croatian Operational Research Society (CRORS), and presidents/representatives of some others Operations Research Societies from abroad.

SOR'19 is the scientific event in the area of operations research, another one in the traditional series of the biannual international OR conferences, organized in Slovenia by SDI-SOR. It is a continuity of fourteen previous symposia. The main objective of SOR'19 is to advance knowledge, interest and education in OR in Slovenia, in Europe and worldwide in order to build the intellectual and social capital that are essential in maintaining the identity of OR, especially at a time when interdisciplinary collaboration is proclaimed as significantly important in resolving problems facing the current challenging times. Further, by joining IFORS and EURO, the SDI-SOR agreed to work together with diverse disciplines, i.e. to balance the depth of theoretical knowledge in OR and the understanding of theory, methods and problems in other areas within and beyond OR. We believe that SOR'19 creates the advantage of these objectives, contributes to the quality and reputation of OR by presenting and exchanging new developments, opinions, experiences in the OR theory and practice.

SOR'19 was highlighted by five distinguished keynote speakers. The first part of the Proceedings SOR'19 comprises invited abstracts and papers, presented by five outstanding scientists: Acad. Prof. Dr. Ivan Bratko, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, Prof. Dr. Mirjana Čižmešija, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia, Assoc. Prof. Dr. Tibor Illés, Budapest University of Technology and Economics, Institute of Mathematics, Budapest, Hungary, Prof. Dr. Joanna Józefowska, Poznan University of Technology, Poznan, Poland (the EURO plenary), and Prof. Dr. Matej Praprotnik, Laboratory for Molecular Modeling, National Institute of Chemistry, Ljubljana, Slovenia.

Proceedings includes 106 papers or abstracts written by 203 authors. Most of the authors of the contributed papers came from Slovenia (79), then from Croatia (43), Czech Republic (13), Hungary (12), Slovak Republic (12), Poland (9), Austria (7), Spain (5), France (4), Netherlands (3), Portugal (3), Italy (2), Norway (2), Romania (2), Thailand (2), Germany (1), Indonesia (1), Ireland (1), Serbia (1), and United Kingdom (1). The papers published in the Proceedings are divided into Plenary Lectures (5 abstracts), seven special sessions: Application of Operation Research in Agriculture and Agribusiness Management (5 papers), Formal and Behavioral Issues in MCDM (6 papers and 1 abstract), Graph Theory and

Algorithms (11 papers and 1 abstract), High-Performance Computing and Big Data (4 papers), Optimization in Human Environments (7 papers), System Modelling & Soft Operational Research (5 papers), Towards Industry 4.0 (5 papers), and eight sessions: Econometric Models and Statistics (10 papers), Environment and Social Issues (5 papers and 1 abstract), Finance and Investments (11 papers), Location and Transport, Graphs and their Applications (4 papers), Mathematical Programming and Optimization (7 papers and 2 abstracts), Multi-Criteria Decision-Making (6 papers), Human Resources (4 papers), and Production and Management (6 papers).

The Proceedings of the previous fourteen International Symposia on Operations Research organized by the Slovenian Section of Operations Research, that are listed at https://www.drustvo-informatika.si/sekcije/sor/sor-publikacijepublications/, are indexed in the following secondary and tertiary publications: Current Mathematical Publications, Mathematical Review, Zentralblatt fuer Mathematik/Mathematics Abstracts, MATH on STN International and CompactMath, INSPEC. The Proceedings SOR'19 are expected to be covered by the same bibliographic databases.

The success of the scientific events at SOR'19 and the present proceedings should be seen as a result of joint effort. On behalf of the organizers we would like to express our sincere thanks to all who have supported us in preparing the event. We would not have succeeded in attracting so many distinguished speakers from all over the world without the engagement and the advice of active members of the Slovenian Section of Operations Research. Many thanks to them. Further, we would like to express our deepest gratitude to prominent keynote speakers, to the members of the Program and Organizing Committees, to the referees who raised the quality of the SOR'19 by their useful suggestions, section's chairs, and to all the numerous people - far too many to be listed here individually - who helped in carrying out The 15th International Symposium on Operations Research SOR'19 and in putting together these Proceedings. Last but not least, we appreciate the authors' efforts in preparing and presenting the papers, which made The 15th Symposium on Operations Research SOR'19 successful.

We would like to express a special gratitude to The Partnership for Advanced Computing in Europe (PRACE) for a financial support and to The Association of European Operational Research Societies (EURO) for financing the EURO plenary speaker.

Bled, September 25, 2019

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

Contents

Plenary Lectures 1

Special Session 1: Application of Operation Research in Agriculture and Agribusiness Management 9

Special Session 2: Formal and Behavioral Issues in MCDM 39

Special Session 5: Optimization in Human Environments 167

Special Session 6: System Modelling & Soft Operational Research 211

Special Session 7: Towards Industry 4.0 243

Session 1: Econometric Models and Statistics 275

Session 2: Environment and Social Issues 339

Session 3: Finance and Investments 373

Session 4: Location and Transport, Graphs and their Applications 441

Session 5: Mathematical Programming and Optimization 467

Session 6: Multi-Criteria Decision-Making 513

Author index

I

J

K

L

M

N

P

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Plenary Lectures

ROBOT LEARNING AND PLANNING WITH QUALITATIVE REPRESENTATIONS

Ivan Bratko
Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia
and University of Ljubljana, Faculty of Computer and Information Science,
Ljubljana, Slovenia
E-mail: bratko@fri.uni-lj.si

Abstract: To execute tasks in an unknown environment, a robot has to learn a model of the environment and use the learned model for task planning. One approach to this is reinforcement learning with deep neural networks. However, a drawback of this approach is lack of comprehensibility. In the interest of Explainable AI, the use of qualitative representations is more promising. In the presentation, an approach based on ideas of qualitative modelling and simulation are presented. The approach are illustrated on the problems of learning to fly a quadcopter, and a humanoid robot learning to walk.

ECONOMIC SENTIMENT IN QUANTITATIVE ANALYSIS

Mirjana Čižmešija
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J.F. Kennedya 6, 10 000 Zagreb, Croatia
E-mail: mcizmesija@efzg.hr

Abstract: The incorporation of the psychological sentiment in quantitative analysis, especially in macroeconomic modelling, turned out to be necessary and invaluable during and after the recent crisis in 2008. It has been shown that knowing the level and the dynamics of GDP, industrial production, stocks of finished products, employment, investment, savings etc. is not enough. Equally important are the perception and expectation of business actors and consumers about these real macroeconomic variables. One of the important sources of the economic and consumer sentiment indicators are Business and Consumer surveys (BCS). Managers' and consumers' judgements about their economic surroundings, derived from BCS results, are expressed as different, empirically confirmed, leading indicators, like economic sentiment indicator or consumer sentiment indicator. It is well known that the European Economic Sentiment Indicator (ESI) is one of the high-quality leading indicators of overall economic activity. They are based on assessments and expectations actors in five BCS sectors (industry, retail trade, services, construction and the consumer sector). Consumer Confidence Indicator (CCI) presents consumer sentiment. It is based on the consumers' perceptions about the past, and expected financial situation of households, the expected general economic situation and the intentions to make major purchases over the next 12 months.

Lately, some methodological improvements and new areas of application of economic sentiment in quantitative analyses are more present. The well-known econometric methods such as linear time series and panel data models, or simple regression and correlation analysis still exist. Nonetheless, the modern time series analysis methods such as state-space modelling, nonlinear econometrics (time-varying parameter models, threshold models and breakpoint tests) accentuate the role of economic sentiment in short-term forecasting of economic activity. Models, which include economic sentiment and all other BCS indicators, bring additional benefit to the methodological skills of economists and analysts.

Keywords: business and consumer surveys, economic sentiment indicator, consumer confidence indicator, leading indicator, time series analysis

SUFFICIENT LINEAR COMPLEMENTARITY PROBLEMS – PIVOT VERSUS INTERIOR POINT ALGORITHMS

Tibor Illés
Budapest University of Technology and Economics, Institute of Mathematics
Egry József u. 1., 1111 Budapest, Hungary
E-mail: illes@math.bme.hu

Abstract: Linear complementarity problems (LCP) generalizes some fundamental problems of mathematical optimization like linear programming (LP) problem, linearly constrained quadratic programming (LQP) problem and some other. It admits an enormous number of applications in economics, engineering, science and many other fields. The three most significant classes of algorithms for solving LCP problems are: pivot algorithms (PA), interior point algorithms (IPA) and continuation methods. Because, both PA and IPA have been developed earlier for LP and QP problems it is quite natural idea to test them on LCP problems, as well.

Concept of sufficient matrices, as generalization of positive semidefinite matrices, has been introduced 30 years ago. LCPs with sufficient matrices possess many important properties, like the solution set is convex and polyhedral; guarantees the finiteness of PAs and (pseudo) polynomial behaviour of the IPAs. Furthermore, for sufficient LCPs, it is meaningful to introduce dual LCP problem and it can be proved that from sufficient primal- and dual LCP problem, exactly one has solution that is an interesting, nice and (quite) new generalization of the old Farkas' lemma.

There are still several open questions in the area of sufficient LCPs. More importantly, solution methods developed for sufficient LCPs helps us in trying to solve LCP problems with more general matrices.

JUST-IN-TIME SCHEDULING (The EURO plenary)

Joanna Józefowska
Poznań University of Technology, Faculty of Computing
Piotrowo 3, 60-965 Poznań, Poland
E-mail: joanna.jozefowska@cs.put.poznan.pl

Abstract: Although scheduling is already a mature field of operational research it continues to inspire researchers with variety of practical applications, new models and solution approaches. One of the examples is just-in-time scheduling that has many practical applications in manufacturing as well as in computing systems. The goal of this talk is to present two approaches to just-in-time scheduling. The first one follows from the traditional Toyota system and the second one is an extension of a classical formulation of a scheduling problem. Each of them relates to slightly different production conditions. Both lead to interesting theoretical results.

SCIENTIFIC CASE FOR COMPUTING IN EUROPE 2018-2026

Matej Praprotnik
Laboratory for Molecular Modeling, National Institute of Chemistry
Hajdrihova 19, 1001 Ljubljana, Slovenia
E-mail: praprot@cmm.ki.si

Abstract: The scientific case (http://www.prace-ri.eu/third-scientific-case/) for computing in Europe 2018-2026 will be presented, which has been formulated by the PRACE Scientific Steering Committee. The scientific case addresses a number of areas of major societal relevance, and identifies both success stories and breakthroughs that will be possible with investments in next generation infrastructure.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 1:

Application of Operation Research in Agriculture and Agribusiness Management

PROJECT PLANNING FOR CATTLE STALL CONSTRUCTION USING CRITICAL PATH METHOD

Karmen Pažek
Faculty of Agriculture and Life Sciences, Chair of Agricultural Economics and Rural Development
Pivola 11, 2311 Hoče, SI
E-mail: karmen.pazek@um.si

Tina Kep
Faculty of Agriculture and Life Sciences
Pivola 11, 2311 Hoče, SI
E-mail: tina.kep@student.um.si

Abstract: The Critical Path Method (further CPM method) is widely used in time planning projects. It is applicable to all kinds of projects from various industries. The cattle stall construction projects using CPM method is presented in this paper. The data are provided in the real life practice. Results shows that critical path is 131 days and is the latest time for finishing the planning project if no delays postpone the project steps. Also, a project network is proposed to show the relationships between the activities and monitor the progress of the project. However, a delay to progress of any activity on the critical path will, without acceleration or re-sequencing, cause the overall project duration to be extended, and is therefore referred to as a 'critical delay'.

Keywords: project planning, agriculture, critical path method, stall construction

1 INTRODUCTION

The critical path method (CPM) is a fundamental concept in project planning and management. It is a useful technique used for scheduling project activities like PERT method. Although both techniques serve the same purpose, they use different methods to calculate the activity duration. Critical path of a schedule demonstrates the activities that cannot be delayed. Because any delay in the critical path causes the delay of the project.

Many different techniques and tools, e.g. Gantt chart, Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT), have been developed to support an improved project planning. These tools are used seriously by a large majority of project managers to identify critical activities and calculate the minimum time required for project completion [1-5]. Among these methods, most traditional scheduling techniques employ Gantt chart. Although this method is still a valuable tool, its application is limited for scheduling large-scale operations. In particular, the bar chart fails to delineate the complex interactions and precedence relationships existing among the project activities. Network-based procedures of PERT and CPM are well known and widely used to assist managers in planning and controlling both large and small projects of all types including construction, research, development projects and many others [6-9]. One of the clear example of CPM method application, associated with agriculture is study by Zareei (2018) that focuses on application of planning and scheduling for analysis of biogas plant construction project.

2 METHODOLOGY

Network analysis can help identify the interrelationships between tasks that make up complex processes and establish the most appropriate moment for their execution. They help in preparing the project program and determining critical paths. Programs describe the sequence in which tasks must be carried out so that a project (or part of a project) can be completed on time. A typical network represents a set of different "arrow diagrams" which go from the origin node to the destination node. In this sense, the path is defined as a sequence of connected events which flow from the start of the project to the end. The time necessary in covering any of these paths is the sum of the time corresponding to each of the tasks involved. The critical path is the one that requires the longest period of time to progress from start to completion and indicates the minimum timeframe necessary to complete the whole project. The reduction of the total execution timeframe will only be possible if the activities on this path can be shortened, since the time necessary to execute non-critical activities does not affect the project's total duration. Decreasing the duration of one or more critical activities, can reduce the project's total timeframe; but it may also change the critical path so that activities which were not previously critical become so. The CPM was developed by James E. Kelley and Morgan R. Walker as a scheduling technique in the 1950's [11, 12]. The graphic representation of a project is called a network and consists of a list of activities and priorities. Activity dependencies represent the relationship between the activities. The time required to follow one of these paths is the sum of the times corresponding to each of the activities. The diagram below is a graphic representation using the arrow diagram technique. Note that a fictitious activity (F) of no duration had to be entered so that activities F and G do not have the same start and finish node; only in this way is it possible to differentiate both activities and calculate the float for one of them, in particular that of F (Figure 1).

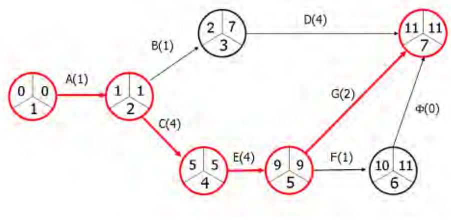


Figure 1: Graphic representation using CPM method

An algorithm is used to calculate the critical path which includes two phases: forward step and backward step. The forward step starts at the origin node and ends at the destination node. At each node a number is calculated which represents the earliest start time for the corresponding action. These numbers are represented in the figure above in the upper left sector. For the backwards step the calculations are taken from the destination node and flow towards the origin node. The number calculated at each node (shown within the upper right sector) represents the latest finish time for the corresponding activity. The early start time for each activity is the earliest date on which this activity may begin, assuming that the priority activities have started on their respective earliest start dates. The forward step begins at the origin node, where Dij is the duration of the activity (i,j) and Ei the early start time for all the activities which share origin node i; if i=1, therefore it is the origin node and by convention E1=0 for all critical path calculations. The calculations for this forward step are obtained using the following expression:

$$E_j = \max_i \{ E_i + D_{ij} \} \quad (1)$$

If the time necessary to carry out an activity is Dij, the early finish time can similarly be determined as Tij=Ei+Dij. The last finish Li, is the latest an activity can finish without delaying the project beyond the deadline. In the same way, the last start Iij, is the latest an activity can start without delaying the project's completion date; it is defined as Iij=Lj-Dij.

The second phase, the backwards step, begins at the destination node. Its objective is to calculate the last finish (Li) for all activities which conclude at node i. If i=n, then this is the destination node, Ln=En, and corresponds to the start of the backwards step. In general, for any node i:

$$L_i = \min_j \{ L_j - D_{ij} \} \tag{2}$$

Once the two phases are complete, the activities which comprise the critical path can be identified; they are those which satisfy the following conditions:

$$\left.\begin{array}{ll} \text{(i)} & E_i = L_i \\ \text{(ii)} & E_j = L_j \\ \text{(iii)} & E_j - E_i = L_j - L_i = D_{ij} \end{array}\right\} \tag{3}$$

There are two important types of floats: total and free. The total float, Hij for activity (i,j), is the difference between the maximum time available to carry out the activity (Lj-Ei) and its duration(Dij); it represents the maximum amount of time the start date for the activity can be delayed, in relation to the early start without delaying the completion of the whole project:

$$Hij = Lj - Ei - Dij = Iij - Ei = Lj – Tij \tag{4}$$

In terms of the free float, it is assumed that all activities start as early as possible. In this case the free float, Fij for activity (i,j), is the excess of time available (Ej-Ei) over its duration (Dij); it represents the delay allowed for an activity without holding up the early start date for the initiation of another activity:

$$Fij = Ej - Ei – Dij \tag{5}$$

In the example represented it can be seen that the activities which make up the critical pathare A, C, E and G, turning out a project with an 11 unit duration. The activities B, D and F have floats, both total and free, of 3, 2, and 1 month respectively [11, 12]. Assuming that the number of activities involved in the critical path is sufficiently high, that the duration distribution functions for each task are of the same type and are statistically independent, it is possible to estimate the project time frame based on the sum of the activity durations which comprise the critical path. In addition, the variances in the project's duration are evaluated using the sum of the variances in activities which form the critical path, using the Gauss curve distribution function. These conditions are sometimes not fully followed on projects; nonetheless, the PERT technique has offered good practical results. Graphically, the nodes represent the activities and the arrows their interrelationships. These methods allow a more complete understanding than a Gantt chart for the different tasks involved in a project. They provide information on the dependency relationships and the decisions to be made to reach the proposed aim [11].

3 RESULTS AND DISCUSSION

Activities can be determined by using the work breakdown structure and project scope. The scope of work defines the deliverables required to complete the project. The work breakdown structure divides the project scope into meaningful work packages (Table 1).

Table 1: Activity chart for new construction of cattle stable.

No.	Activity	Activity description	Duration	Start	End	Previous activities
1	1,1	Ordering of armature, concrete, gravel, grates, equipment, wood	1 day	Mon 1.4.19	Mon 1.4.19	0
2	1,2	Demolition of the existing construction and mark of new construction	2 days	Tue 2.4.19	Wed 3.4.19	1
3	1,3	Armature delivering	1 day	Tue 2.4.19	Tue 2.4.19	1
4	1,4	Roof delivering	4 days	Tue 2.4.19	Fri 5.4.19	1
5	1,5	Grates delivering	40 days	Tue 2.4.19	Mon 27.5.19	1
6	1,6	Equipment delivering	21 days	Wed 3.4.19	Wed 1.5.19	1
7	1,7	Earth excavation and earth straighten	5 days	Thurs 4.4.19	Wed 10.4.19	2
8	1,8	Gravel delivery	1 day	Thurs 11.4.19	Thurs 11.4.19	7
9	1,9	Gravel straighten	2 days	Fri 12.4.19	Mon 15.4.19	8
10	2,0	Reinforcement and other material	8 days	Tue 16.4.19	Thurs 25.4.19	9;3
11	2,1	Concrete preparation and delivery	1 day	Fri 26.4.19	Fri 26.4.19	10
12	2,2	Walls concreting and in concreting of manure cave base	2 days	Mon 29.4.19	Tue 30.4.19	11
13	2,3	Concrete drying	14 days	Wed 1.5.19	Mon 20.5.19	12
14	2,4	Preparation of project land, reinforcement, panels	2 days	Wed 1.5.19	Thurs 2.5.19	12;1
15	2,5	Wood construction delivery	5 days	Thurs 2.5.19	Wed 8.5.19	1
16	2,6	Concrete preparation and delivery	1 day	Fri 3.5.19	Fri 3.5.19	14
17	2,7	Base concreting	1 day	Mon 6.5.19	Mon 6.5.19	16
18	2,8	Concrete drying	14 days	Tue 7.5.19	Fri 24.5.19	17
19	2,9	Construction wood coloring	10 days	Thurs 9.5.19	Wed 22.5.19	15
20	3,0	Panels and reinforcing of walls, other stable equipment installation	10 days	Mon 27.5.19	Fri 7.6.19	17;1;13;18
21	3,1	Concrete preparation and delivery	1 day	Mon 10.6.19	Mon 10.6.19	20
22	3,2	Concreting of walls and corresponding equipment (pillars, feedstock table,..)	2 days	Tue 11.6.19	Wed 12.6.19	21
23	3,3	Concrete drying	14 days	Thurs 13.6.19	Tue 2.7.19	22
24	3,4	Panels and reinforcement	5 days	Thurs 13.6.19	Wed 19.6.19	22;1
25	3,5	Concrete preparation and delivery	1 day	Thurs 20.6.19	Thurs 20.6.19	24
26	3,6	Further concreting of the existing floorboard, panels and pillars	1 day	Fri 21.6.19	Fri 21.6.19	25
27	3,7	Concrete drying	14 days	Mon 24.6.19	Thurs 11.7.19	26
28	3,8	Gratis installation	1 day	Wed 3.7.19	Wed 3.7.19	23;5
29	3,9	Sinking walls of manure with earth	3 days	Wed 3.7.19	Fri 5.7.19	23
30	4,0	Mechanical smoothing of concrete and grinding	2 days	Fri 12.7.19	Mon 15.7.19	23;27
31	4,1	Installation of a work platform for the roof	1 day	Fri 12.7.19	Fri 12.7.19	27;29
32	4,2	Wood construction assembly	5 days	Mon 15.7.19	Fri 19.7.19	19;31;28
33	4,3	Roofing installation	5 days	Mon 22.7.19	Fri 26.7.19	32;4
34	4,4	Epoxy coating in feeding area	5 days	Mon 29.7.19	Fri 2.8.19	30;33

35	4,5	Coloring of interior and exterior concrete surfaces (walls)	3 days	Mon 5.8.19	Wed 7.8.19	27;34
36	4,6	Fencing	3 days	Thurs 8.8.19	Mon 12.8.19	35;6
37	4,7	Machine installation	5 days	Tue 13.8.19	Mon 19.8.19	36
38	4,8	Electricity installation	5 days	Tue 20.8.19	Mon 26.8.19	37
39	4,9	Front size wood coloring	4 days	Tue 27.8.19	Fri 30.8.19	38
40	5,0	Color drying	4 days	Mon 2.9.19	Thurs 5.9.19	39
41	5,1	Installation of an anti-wind network	1 day	Mon 2.9.19	Mon 2.9.19	39
42	5,2	Placing the rubber on the floor	1 day	Tue 3.9.19	Tue 3.9.19	41
43	5,3	Placing of doors	1 day	Wed 4.9.19	Wed 4.9.19	42
44	5,4	Placing of wood on the construction	2 days	Fri 6.9.19	Mon 9.9.19	40;43
45	5,5	Landscaping	1 day	Tue 10.9.19	Tue 10.9.19	44

Table 1 shows all the activities needed for the completion of the analyzed project and their duration, respectively. Previous activities are also shown in the last column of the table.

After the identification of activities and establishing their dependencies, a network diagram can be drawn. Below figure illustrates a presented network diagram for analyzed case study (Figure 2). The diagram was developed using Microsoft Office Excel 2016.

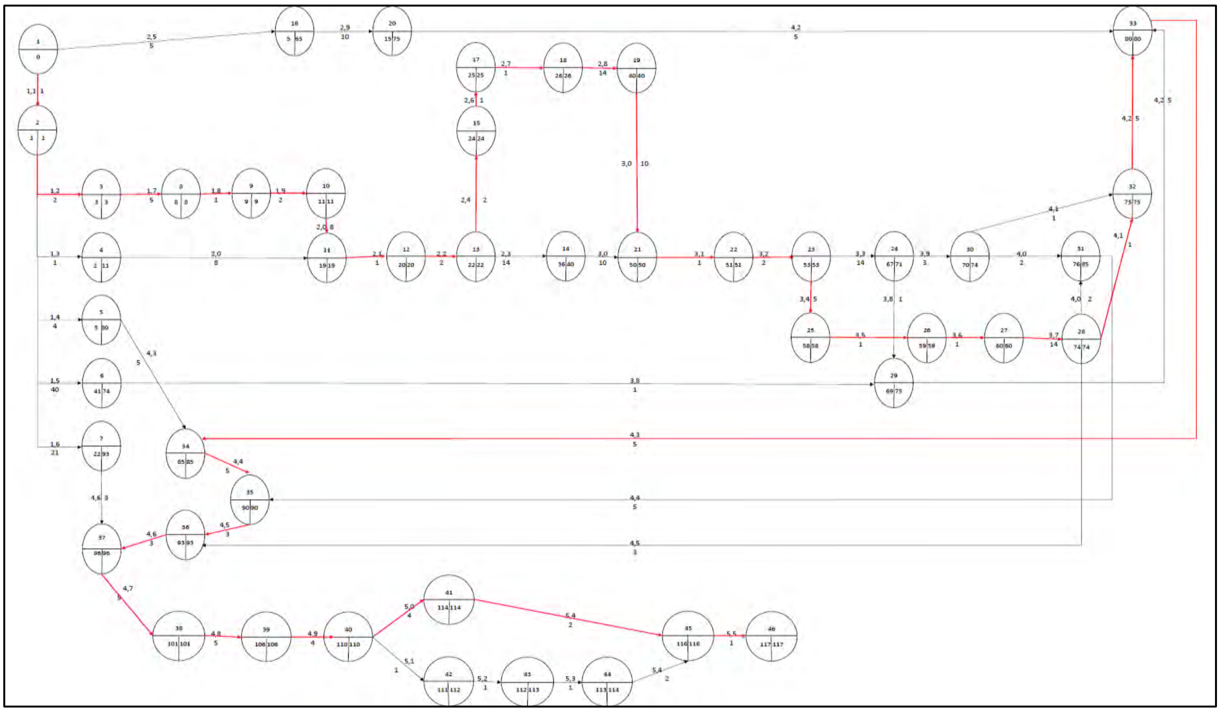


Figure 2: Network diagram for the cattle construction project.

Forward and backward pass calculation is used to determine the critical path and total float. By the help of forward pass and backward pass calculations, the earliest start and finish dates, and the latest start and finish dates for each activity can be identified and the Critical Path of the network diagram can be determined. The Critical Path of the analyzed project is 131 days (S21=1+2+3+8+9+10+11+12+13+15+17+18+19+21+22+23+25+26+27+28+32+33+34+35+36+37+38+39+40+41+45+46 days) and is the maximum time available to provide out the whole project. According to the calculated results the early project finish time is 41 days. The red color shows the critical path of the project, as mentioned before, 131 days (in presented case starting on 1st of April, 2019 and finishing on 10th of September, 2019). However, there advantages of CPM methods as follows; it improves decision making within the project team, it is a visual technique which enables to show activities, activity dependencies and durations in the same diagram, further the method enables the project team to make time optimizations, it enables to manage and organize large and complex projects to. Very important is the calculation of The Earliest Start/Finish and The Latest Start/Finish dates in order to manage activities and procurement tasks. Although the process of CPM estimation allows decision maker to break down complex tasks into simpler, the levels to which the process can break down these tasks can make the critical path diagram for the entire project much more complex than necessary. For large and complex projects, there'll be thousands of activities and dependency relationships. Without software it can be mighty difficult managing this. To make matters worse, if the plan changes during project execution then the precedence diagram will have to be redrawn. Fortunately, we do have relatively cheap software that can handle this with ease.

Conclusion

The CPM helps to determine which activities can be delayed without delaying the project and at last it is a Schedule compression methods such as fast tracking and crashing rely on the critical path method. On the other side there are some limitations of CPM too, like is hard to manage activities in large and complex projects without software, it does not consider resource allocations, activity durations should be determined correctly otherwise, the critical path of the project will be wrong and it will be hard to determine the critical path if there are many other similar duration paths in the project. The method gives the project management teams correct completion dates for their projects. However, it is not easy to apply this method to large scaled projects that have thousands of activities without support of a software. Correspondingly it can be used to determine the critical path of a project easily.

References

[1] Awani, A.O. 1983. Project management techniques. Petrocelli Books, p. 208.
[2] Chizari, A.H, Amirnejad, H. 1998. Management of corn drying factory construction project using CPM and PERT methods. Agric. Econ. Dev. Vol (29), 257-73.
[3] Fahimifard, S.M., Kehkha, A.A. 2009. Application of project scheduling in agriculture (case study: grape garden stabilization). Am-Eurasia Jour. Agric. Environ. Sci., Vol (5,3): 313-21.
[4] Foulds, L.R., Wilson, J.M. 2005. Scheduling operations for the harvesting of renewable resources. J. Food Eng. Vol (70), 281–92.
[5] Hillier, F.S., Lieberman, G.J. 2005. Chapter 22: project management with PERT/CPM. Introduction to operations research. 8th Ed. McGraw-Hill, Boston, MA.
[6] Kelley, J., Walker, M. 1959. Critical-Path Planning and Scheduling. Proceedings of the Eastern Joint Computer Conference.
[7] Monjezi, N., Sheikhdavoodi, M.J., Basirzadeh, H., Zakidizaji, H. 2012. Analysis and evaluation of mechanized greenhouse construction project using CPM methods. Res. Appl. Sci. Eng. Technol. Vol (4, 18), 3267–73.
[8] Pollack, J.B. 1998. Project management software usage patterns and suggested research directions for future development. Proj. Manag. J., Vol (29), 19-29.
[9] Zaree, S. 2018. Project scheduling for constructing biogas plant using critical path method. Renewable and Sustainable Energy Reviews, Vol (81,1): 756-759.
[10] Yang, J.B. 2007. How the critical chain scheduling method is working for construction. Cost. Eng., Vol (49,4), 25-32.
[11] https://www.projectcubicle.com/critical-path-method/ [04/06/2019].
[12] https://www.scl.org.uk/resources/delay-disruption-protocol/ [04/06/2019].

USING DATA ENVELOPMENT ANALYSIS AND ANALYTIC HIERARCHY PROCESS TO MEASURE EFFICIENCY OF TOURISM FARMS: CASE OF SLOVENIA

Boris Prevolšek
University of Maribor, Faculty of tourism
Cesta prvih borcev 36,8250 Brežice, Slovenia
E-mail: boris.prevolsek@um.si

Karmen Pažek
University of Maribor, Faculty of agriculture and life sciences
Pivola 10, 2311 Hoče, Slovenia
E-mail:karmen.pazek@um.si

Maja Žibert
University of Maribor, Faculty of agriculture and life sciences
Pivola 10, 2311 Hoče, Slovenia
E-mail: maja.zibert@um.si

Črtomir Rozman
University of Maribor, Faculty of agriculture and life sciences
Pivola 10, 2311 Hoče, Slovenia
E-mail:crt.rozman@um.si

Abstract: Nowadays, many rural areas face the difficulty how to motivate farmers to undertake diversified activities, such as farm tourism, while raising their efficiency [15]. Though economic viability and low productivity of small-scale tourist farms have been extensively dealt with, there is no information neither about the economic analysis nor efficiency of the tourism sphere [6]. The aim of this paper is to examine the efficiency of tourist farms in Slovenia by adopting the approach using a framework of non-parametric programming – Data Envelopment Analysis (DEA) and Analytic Hierarchy Process (AHP). The findings of this paper can help the tourist farm managers to improve efficiency of their tourist farm. It can help managers to get important insights for their strategic and operational decisions to improve performance of their business, as well.

Keywords: farm tourism, efficiency, Data Envelopment Analysis (DEA), Analytic Hierarchy Process (AHP).

1 INTRODUCTION

In spite of increase of the number of tourist farms and inrease of the demand for their services and in spite of their sustainable potential [14, 16, 23], they often fail on the market [4]. Accordingly, the challenge appears how to preserve efficient and competitive operation of tourist farms [6]. Efficiency of tourist farms is a crucial factor when planning economic successfulness. Though economic viability and low productivity of tourist farms have been extensively dealt with, there is no information neither about the economic analysis nor efficiency of the tourism sphere [6]. Efficiency is understood as a measure of operational excellence in rational utilization of resources and refers to decision-making, possibility of improvements and benchmarks of resource allocation [8].

2 EFFICIENCY ASSESSMENT APPROACHES

A measure of efficiency determines the ability of an organization to attain the output(s) with the minimum inputs. Efficiency is not a measure of successfulness on the market but a measure of operational excellence in the rational utilization of resources. Efficiency refers to decision-making, possibility of improvements and benchmarks of resource allocation [8]. For measuring efficiency which is associated with 'doing things right' two approaches are available, DEA and AHP.

2.1 DEA

The Data Envelopment Analysis (DEA) was moslty used as a non-parametric method introduced by [10]. This approach is used in assessment of relative efficiency for evaluating decision making units (DMUs). Each DMU i.e. tourist farm selects its best set of corresponding weights to consider inputs and outputs and the values of weights may thus vary from one DMU to another. Furthermore the DEA calculate each DMU's performance score ranging between 0 and 1. This result represents its relative degree of efficiency [24].

As far as the efficiency measuring literature is concerned, DEA is quite popular in efficiency measuring in general, but it is not so long ago that DEA started to be used in the tourism and hospitality industry [25]. As far as the DEA application papers are concerned the share of tourism is just estimated to only 1.34% [13]. In the sphere of farm tourism only one research is traceable, [6] applying DEA to tourist farms in South Korea.

DEA is known as CCR (Charnes-Cooper-Rhodes) model, which is built on the assumption of constan returns to scale of activities. That's mean, if an activity (x, y) is feasible, then, for every positive scalar t, the activity (tx, ty) is also feasible. The efficient production frontiers have constant returns-to-scalecharacteristics for the single-input and single-outputcase [5]. While BCC (Banker-Charnes-Cooper) model accepts the convex combinations of the decision-making units as the production possibility set [6]. The frontiers have piece-wise linear and concave characteristics which,leads tovariable returns-to-scale characterizations [5].

2.2 AHP

The Analytic Hierarchy Process (AHP) is a commonly used multi-criteria decision making method [18]. AHP was first proposed by the author Thomas L. Saaty in 1971 [17, 18, 19]. AHP performs pairwise comparisons between factors to make priorities among them by means of the eigen-value calculation [11].

The AHP method is used in many spheres, such as activity planning, alternative choosing, optimization, resource allocation, conflict resolution etc. [1], but it has been used also in business, energy, health, resource management, and transportation. However, tourism is seldom discussed in scientific documents based on the Expert Choice 2000 [7]. In the field of farm tourism there is no research using AHP.

2.3 USE OF COMBINED MODELS

There is a limited number of papers on combinations of DEA and AHP approaches [2], Irrespective of the constructive efforts and in many ways positive results of combing DEA and AHP, most existing studies use DEA and AHP models separately.

Among other things, combinations of both models can be used for selection of a flexible manufacturing system [22], evaluation of quality management activities [27], proposed facility layout design [26], measuring the relative efficiency of non-homogeneous decision-making units (DMUs) [20], facility layout design (FLD) in a manufacturing system [9] evaluation of supplier selection [21], improvement and optimization of railway system [3] evaluation of efficiency performance of international airports [12].

Papers combining DEA and AHP methods for tourism sector as well as for the farm tourim have not yet been published.

In our study discussing efficiency of tourist farms the DEA and AHP models were used separately and the results were compared.

3 CASE STUDY

The combined DEA and AHP model has been used on tourist farms in Slovenia, where the tourist farm efficiency was tested on 45 samples. For assessing efficiency of tourist farms the following input variables were used:

- the number of full-time employees in the basic agricultural activity,
- the number of full-time employees contributed, in total, by other family members to basic agricultural activity,
- the number of rooms,
- the number of beds,
- the number of seats and
- the number of full-time employees in tourist activity on the farm.

The outputs used were:

- the number of tourist arrivals in 2017
- the number of tourist nights in 2017
- total revenue from the basic agricultural activity in 2017
- total revenue from tourism in 2017

In excess of the said inputs and outputs, included in DEA, several additional variables have been used with AHP:

- additional possibilities on the farm, such as: location near wine road, pets are welcome, house with tradition, beekeeping on the farm, access by bus, camper parking lot, ecological farming, etc.
- additional activities on the farm, such as: hiking, biking, swimming in pool, river or lake, sauna, horse-riding, playground for children, hunting, fishing, etc.
- food and drink services, such as: bed and breakfast, half board, full board, all inclusive, a la carte food service, domestic and local speciality, domestic and local wines, etc.
- specialized offer, such as: ecological tourist farm, family friendly tourist farm, bikers friendly tourist farm, tourist farm offering healthy vacation, etc.

4 RESULTS

The figure 1 herebelow shows the results of both models ranking the farms in terms of efficiency. Within DEA and its sub-models CCR and BCC it was found, that 50 % of tourist farms were efficient, supposing that they reached the efficiency degree 1. The rusulting efficiency degree 1 means that the tourist farm is efficient. The remaininig tourist farms, which have lower efficiency degree than 1 were evaluated as parly efficient implying that the lower the value the less efficient the tourist farm. As far as the AHP method and inclusion of the additional variables are concerned, it was found that up to unit 11 the AHP ranked in the same order as the DEA. In the continuation, the values of efficiency assessment follow intermittently. It is evident that AHP model, involving additional variables, ranked differently than DEA model, implying that in some cases those additional variables had significitant impact.

DMU	AHP assessment	Ranking	DEA ASSESSMENT							
			BCC-I	Ranking	BCC-O	Ranking	CCR-I	Ranking	CCR-O	Ranking
DMU1	0,9747	2	1	1	1	1	1	1	1	1
DMU2	0,6643	19	1	1	1	1	1	1	1	1
DMU3	0,7944	7	1	1	1	1	1	1	1	1
DMU4	0,7185	14	1	1	1	1	1	1	1	1
DMU5	0,9017	4	1	1	1	1	1	1	1	1
DMU6	0,7226	13	0,71882	6	0,67128	5	0,67032	6	0,67032	6
DMU7	0,6616	20	1	1	1	1	1	1	1	1
DMU8	1	1	1	1	1	1	1	1	1	1
DMU9	0,6436	21	0,57343	13	0,28889	15	0,25195	16	0,25195	16
DMU10	0,6345	24	1	1	1	1	1	1	1	1
DMU11	0,5448	36	0,54575	14	0,19012	19	0,18942	19	0,18942	19
DMU12	0,9208	3	1	1	1	1	1	1	1	1
DMU13	0,5638	34	0,58384	11	0,38674	11	0,371	10	0,371	10
DMU14	0,4879	44	0,86957	3	0,90323	2	0,8284	4	0,8284	4
DMU15	0,7404	11	0,75192	5	0,17108	20	0,12884	21	0,12884	21
DMU16	0,7088	16	1	1	1	1	1	1	1	1
DMU17	0,5425	38	0,2627	22	0,25243	16	0,17596	20	0,17596	20
DMU18	0,4958	42	1	1	1	1	1	1	1	1
DMU19	0,5867	32	0,31447	21	0,12336	21	0,10701	23	0,10701	23
DMU20	0,6398	22	1	1	1	1	1	1	1	1
DMU21	0,5448	35	0,58182	12	0,32521	14	0,28961	15	0,28961	15
DMU22	0,5966	31	1	1	1	1	1	1	1	1
DMU23	0,61	27	0,48595	16	0,48867	10	0,43889	9	0,43889	9
DMU24	0,7226	12	0,71882	6	0,67128	5	0,67032	6	0,67032	6
DMU25	0,4959	41	1	1	1	1	0,93517	2	0,93517	2
DMU26	0,6044	29	1	1	1	1	1	1	1	1
DMU27	0,7137	15	0,42706	18	0,35011	13	0,30031	14	0,30031	14
DMU28	0,8362	5	1	1	1	1	1	1	1	1
DMU29	0,7057	17	1	1	1	1	1	1	1	1
DMU30	0,5702	33	1	1	1	1	1	1	1	1
DMU31	0,7733	8	1	1	1	1	1	1	1	1
DMU32	0,5075	39	0,8042	4	0,74976	4	0,73286	5	0,73286	5
DMU33	0,5075	40	0,88758	2	0,88028	3	0,87264	3	0,87264	3
DMU34	0,6957	18	0,61743	9	0,36306	12	0,36055	11	0,36055	11
DMU35	0,5426	37	0,65603	8	0,23767	17	0,21991	18	0,21991	18
DMU36	0,6042	30	0,4755	17	0,10892	22	0,10819	22	0,10819	22
DMU37	0,7599	10	1	1	1	1	1	1	1	1
DMU38	0,6374	23	0,538	15	0,66754	7	0,53682	8	0,53682	8
DMU39	0,4205	45	0,41513	19	0,496	9	0,35219	12	0,35219	12
DMU40	0,6145	26	1	1	1	1	1	1	1	1
DMU41	0,6329	25	1	1	1	1	1	1	1	1
DMU42	0,6075	28	0,35016	20	0,59862	8	0,34946	13	0,34946	13
DMU43	0,8156	6	1	1	1	1	1	1	1	1
DMU44	0,4934	43	0,58737	10	0,2273	18	0,22716	17	0,22716	17
DMU45	0,7686	9	1	1	1	1	1	1	1	1

Figure 1: The results of DEA and AHP models

5 CONCLUSION

A tourist farm can be considered to be efficient, when the relation between inputs and outputs is approoriate. For example, this means that the number of rooms and beds results in creation of a profitable number of tourist arrivals and nights. This refers also to other relations of inputs in outputs.

The results show that the use of combined models DEA and AHP is justified, since, by inclusion of additional variables, the AHP model gives more precise values and different ranking on certain units (tourist farms) than the DEA model.

The results of the comparison of the combined use of the DEA and AHP models for tourist farms might be useful also for other sectors within tourism and wider.

References

[1] Ahmad, N., Berg, D., & Simons, G. R. (2006). The integration of analytical hierarchy process and data envelopment analysis in a multi-criteria decision-making problem. *International Journal of Information Technology & Decision Making*, *5*(02), 263-276.

[2] Assaf, A. G., & Gillen, D. (2012). Measuring the joint impact of governance form and economic regulation on airport efficiency. *European journal of operational research*, *220*(1), 187-198.

[3] Azadeh, A., Ghaderi, S. F., & Izadbakhsh, H. (2008). Integration of DEA and AHP with computer simulation for railway system improvement and optimization. *Applied Mathematics and Computation*, *195*(2), 775-785.

[4] Busby, G., & Rendle, S. (2000). The transition from tourism on farms to farm tourism. *Tourism management*, *21*(6), 635-642.

[5] Cooper, W. W., Seiford, L. M., & Tone, K. (2000). Data envelopment analysis. *Handbook on Data Envelopment Analysis, 1st ed.; Cooper, WW, Seiford, LM, Zhu, J., Eds*, 1-40

[6] Choo, H., Ahn, Y. H., & Park, D. B. (2018). Using the Data Envelopment Analysis to Measure and Benchmark the Efficiency of Small-scale Tourism Farms in South Korea. *Journal of Rural and Community Development*, *13*(2).

[7] Deng, J., King, B., & Bauer, T. (2002). Evaluating natural attractions for tourism. *Annals of tourism research*, *29*(2), 422-438.

[8] Drucker, P. F. (1977). *An introductory view of management: instructor's manual*. Harper and Row.

[9] Ertay, T., Ruan, D., & Tuzkaya, U. R. (2006). Integrating data envelopment analysis and analytic hierarchy for the facility layout design in manufacturing systems. *Information Sciences*, *176*(3), 237-262.

[10] Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, *120*(3), 253-290.

[11] Kurttila, M., Pesonen, M., Kangas, J., & Kajanus, M. (2000). Utilizing the analytic hierarchy process (AHP) in SWOT analysis—a hybrid method and its application to a forest-certification case. *Forest policy and economics*, *1*(1), 41-52.

[12] Lai, P. L., Potter, A., Beynon, M., & Beresford, A. (2015). Evaluating the efficiency performance of airports using an integrated AHP/DEA-AR technique. *Transport Policy*, *42*, 75-85.

[13] Liu, J. S., Lu, L. Y., Lu, W. M., & Lin, B. J. (2013). A survey of DEA applications. *Omega*, *41*(5), 893-902.

[14] Meraner, M., Heijman, W., Kuhlman, T., & Finger, R. (2015). Determinants of farm diversification in the Netherlands. *Land Use Policy*, *42*, 767-780.

[15] Ohe, Y. (2017). Assessing Managerial Efficiency of Educational Tourism in Agriculture: Case of Dairy Farms in Japan. *Sustainability*, *9*(11), 1931.

[16] Park, D. B., Doh, K. R., & Kim, K. H. (2014). Successful managerial behaviour for farm-based tourism: A functional approach. *Tourism Management*, *45*, 201-210.

[17] Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, *15*(3), 234-281.

[18] Saaty, T. L. (1980). The analytic hierarchy process: planning, priority setting, resources allocation. *New York: McGraw*, *281*.

[19] Saaty, T. L. (1982). The analytic hierarchy process: A new approach to deal with fuzziness in architecture. *Architectural Science Review*, *25*(3), 64-69.

[20] Saen, R. F., Memariani, A., & Lotfi, F. H. (2005). Determining relative efficiency of slightly non-homogeneous decision making units by data envelopment analysis: a case study in IROST. *Applied Mathematics and Computation*, *165*(2), 313-328.

[21] Sevkli, M., Lenny Koh, S. C., Zaim, S., Demirbag, M., & Tatoglu, E. (2007). An application of data envelopment analytic hierarchy process for supplier selection: a case study of BEKO in Turkey. *International Journal of Production Research*, *45*(9), 1973-2003.

[22] Shang, J., & Sueyoshi, T. (1995). A unified framework for the selection of a flexible manufacturing system. *European Journal of Operational Research*, *85*(2), 297-315.

[23] Võsu, E., & Kaaristo, M. (2009). An ecological approach to contemporary rural identities: The case of tourism farms in south-east Estonia. *Journal of Ethnology and Folkloristics*, *3*(1), 73-94.

[24] Wei, Q., & Chang, T. S. (2011). Optimal profit-maximizing system design data envelopment analysis models. *Computers & Industrial Engineering*, *61*(4), 1275-1284.

[25] Wöber, K. W. (2007). Data envelopment analysis. *Journal of Travel & Tourism Marketing*, *21*(4), 91-108.

[26] Yang, T., & Kuo, C. (2003). A hierarchical AHP/DEA methodology for the facilities layout design problem. *European Journal of Operational Research*, *147*(1), 128-136.

[27] Yoo, H. (2003). A study on the efficiency evaluation of total quality management activities in Korean companies. *Total Quality Management & Business Excellence*, *14*(1), 119-128.

ASSESSING THE NUTRIENT CYCLING POTENTIAL IN AGRICULTURAL SOILS USING DECISION MODELLING

Aneta Trajanov[1], Jaap Schröder[2], David Wall[3], Antonio Delgado[4], Rogier Schulte[5], Marko Debeljak[1]

[1]Jozef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
E-mails: aneta.trajanov@ijs.si, marko.debeljak@ijs.si

[2]Wageningen University and Reasearch, Plant Science Group, Wageningen, Netherlands
E-mail: jaap.schroder@wur.nl

[3]Teagasc - Crops, Environment and Land Use Programme, Johnstown Castle, Wexford, Ireland
E-mail: david.wall@teagasc.ie

[4]University of Sevilla, Departamento de Ciencias Agroforestales, Sevilla, Spain
E-mail: adelgado@us.es

[5]Wageningen University and Reasearch, Department of Plant Sciences, Wageningen, Netherlands
E-mail: rogier.schulte@wur.nl

Abstract: One of the essential functions that agricultural soils provide is nutrient cycling. The capacity of soils to provide this function is influenced by the interactions between soil properties, climate and management. Understanding these interactions can help in assessing the soil nutrient cycling potential on a field and in identifying best management options. To optimize this process, we developed a multi-attribute decision model using the DEXi modelling tool. The outputs from this model may be used to obtain recommendations for farmers and other stakeholders and assist them with the selection of management practices fostering the nutrient cycling potential of soils.

Keywords: soil functions, nutrient cycling, decision model, DEXi, recommendations.

1 INTRODUCTION

Soils provide ecosystem services such as primary productivity, water regulation and purification, habitats for biodiversity, climate regulation and nutrient cycling [12]. Soils differ in their capacity to deliver these functions in response to climate, weather, intrinsic soil properties and also management. Compaction, erosion, desertification, salinization, sealing and contamination may have a negative effect on the capacity of soils to deliver these services. Therefore, these threats may ask for adjustments of the management of soils [6, 7, 13].

Societies are in need of nutrient cycling to minimize the use of finite resources and to avoid the accumulation of 'wastes', from here on referred to as by-products. Examples of these by-products are crop residues, manures and industrial and municipal refusals. Nutrient cycling encompasses the capacity of a soil to accommodate the reception of by-products, to provide nutrients to crops from by-products and from intrinsically present resources, to support the acquisition of nutrients by these crops, and to effectively carry over these nutrients into the harvested parts of crops [10]. If this function does not perform well, there will be a greater need for putting in additional nutrients to compensate for the nutrients that are fixed or lost to water and air. This will increase the depletion of finite resources of nutrients, such as mined phosphorus (P) rock and of fossil fuels needed for manufacturing mineral nitrogen (N) fertilizers [5, 8, 14].

Assessing and optimizing the nutrient cycling soil function is a complex decision process that depends on interactions between the soil properties, climate and weather and soil management practices. To capture these interactions and assess the capacity of soils to recycle nutrients, we developed a multi-attribute decision model.

2 MATERIALS AND METHODS

The nutrient cycling decision support model was developed together with experts from the H2020 project LANDMARK using Multi-Criteria Decision Analysis, in particular using the DEX (Decision Expert) integrative methodology [2, 3, 4] for qualitative decision modelling. Using this methodology, the main decision problem (concept, in our case nutrient cycling) is decomposed into smaller, less complex sub-problems (sub-concepts) in a hierarchical way. The attributes at the lowest level of the tree are the basic attributes, representing the main drivers of nutrient cycling: attributes describing the soil, environmental and management properties. The attributes in the intermediate levels represent aggregated attributes. The values of the basic attributes are represented in a qualitative way and the values of the intermediate attributes are obtained using decision rules integrating the combined effect of lower-level attributes. The decision rules are a tabular representation of a mapping of values of lower-level attributes to higher-level attributes. From the above points of view, the capacity of a soil to recycle nutrients is reflected by the ratio of the amounts of nutrients that are applied (input) and nutrients that are harvested (output). That ratio (output/input) is determined by the summed product of one or more (n = i) types of inputs and their respective nutrient fertilizer replacement values (NFRV, 'mineralizability', i.e. the extent to which nutrient availability in by-products is equivalent to that in mineral fertilizers), the apparent nutrient recovery fraction (ANR, the extent to which nutrients are effectively taken up by crops) and the nutrient harvest index (NHI, fraction of the nutrients in crops that eventually leaves the field in harvests), divided by the inputs:

$$Nutrient\ cycling = \frac{output}{input} = \frac{\sum_i (input_i \times NFRV_i) \times ANR \times NHI}{\sum_i input_i}$$

The construction of the model is further based on the premise that each of these three factors is in turn determined by combinations of underlying factors that are ultimately ruled by basic attributes. As opposed to non-agricultural environments, mineralization (NFRV) is not seen in the model as the major limiting factor for nutrient cycling, apart from relatively rare situations where it is too dry, too cold, too acid or too alkaline for biological decomposition. Nitrogen, as opposed to P, is very mobile and therefore gets easily lost between 'field, food, fork and fire'. As a result of these losses (leaching, denitrification, ammonia volatilization) most by-products (e.g. crop residues, livestock manures, composts, incineration-ashes, etc.) contain less N per unit P than what most crops need when these by-products are used as a fertilizer [9]. Considering that, the presence or provision of sufficient N is deemed key for effective nutrient cycling, in particular via the apparent nutrient recovery (ANR). The model further assumes that an effective export of nutrients from the field (NHI) is mainly determined by the absence of crop failures and the decision to remove crop residues.

When the initial version of the model was completed, a sensitivity analysis was carried out. The goal of the sensitivity analysis is to find input attributes whose values have negligible impact on the outputs of the model. Because different attributes had different scales of values, the weights were normalized to the same unit interval, i.e., we used global normalized weights, which take into account the relative importance of sub-models to the overall model [4]. If the weights of the basic attributes were less than 1%, they were removed from the model and the corresponding decision rules were modified accordingly. This allowed us simplify the model to its present form.

3 RESULTS

The final decision model obtained after the sensitivity analysis is presented in Figure 1. It consists of three sub-models: the mineralization model (NFRV), the nutrient recovery model (ANR) and the harvest index model (NHI). These three rule the top concept– the capacity of a soil to provide and cycle nutrients.

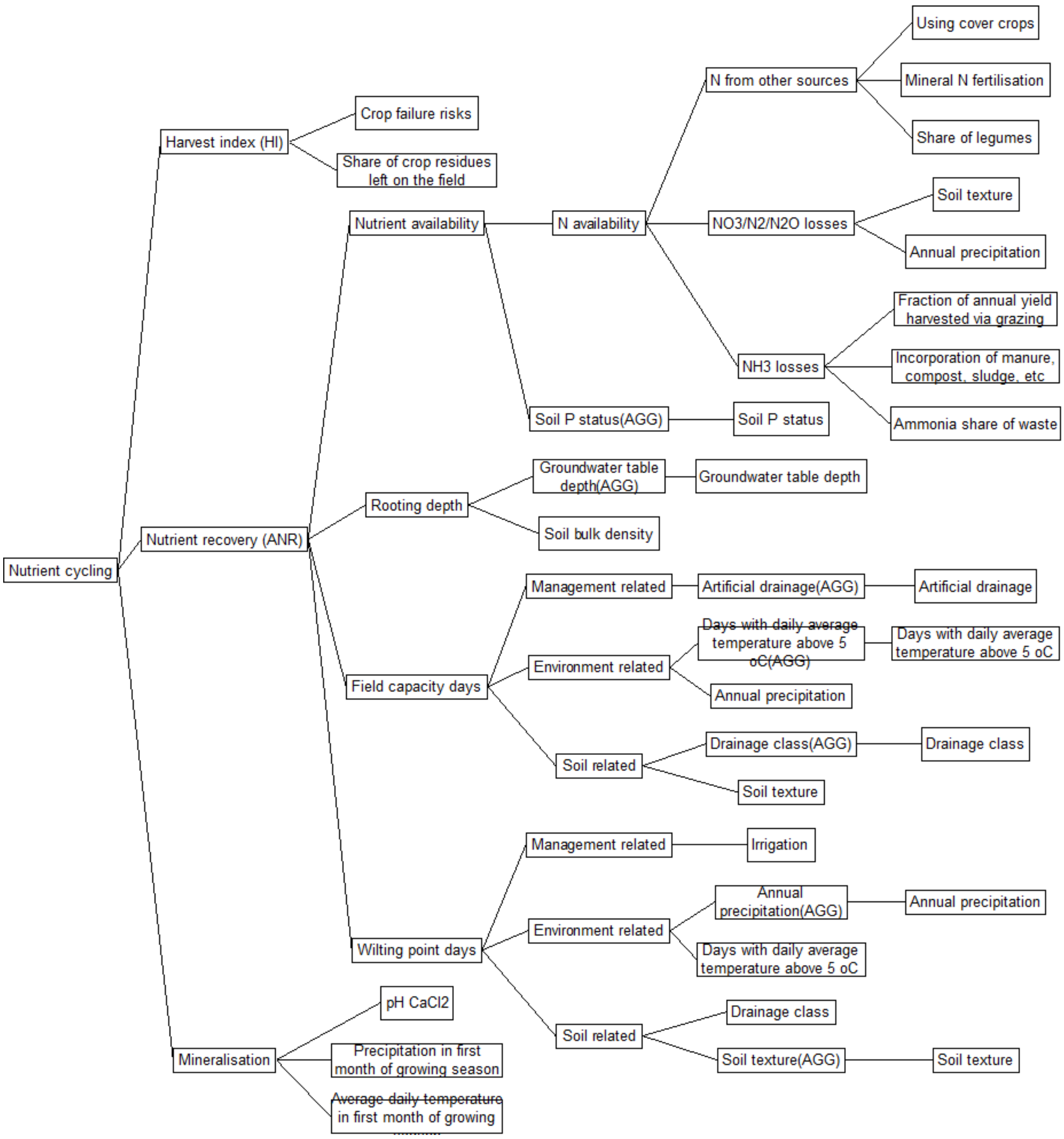


Figure 1. Decision model for assessment of the ability of a soil to provide and cycle nutrients.

The model was validated using a French national dataset collected within the French Soil Monitoring Network (RMQS) [1]. The dataset comprised soil, climatic and management data for 534 site-years across France using wheat grain yields as a proxy of harvested nutrients

[11]. In hindsight the dataset was of limited value for validation, because most of the site-years pertained to commercial fields from mainly one environmental zone with relatively mild weather conditions, where P deficiency nor soil compaction occurred. Moreover, ample amounts of mineral fertilizer N had been and hardly any by-products. Consequently, only 5% of the sites exhibited a nutrient cycling score 'low' and 47% a score 'medium' and grain yields that were only slightly less than when the score was 'high'. Unfortunately, it turns out to be extremely difficult to find a complete data set with sufficient variation, with focus on the cycling of by-products and a long history to make sure that annual nutrient inputs and outputs are more or less in equilibrium.

Nevertheless, the decision model appears to give a reasonable basis for deriving recommendations for farmers and policy makers to improve the capacity of soils to provide and cycle nutrients. These recommendations pertain to attention for a sufficiently high pH of soils, drainage or irrigation wherever one of the two is needed, avoiding compaction, and tuning the provision of both available N and P as precisely as possible to what a crop rotation requires.

4 CONCLUSIONS

In this study, we presented a decision model for assessment of the potential of the soil to provide and cycle nutrients. The model was derived from expert knowledge, sensitivity analyses were carried out and validation was performed using a French national dataset. The proposed approach enabled us to obtain recommendations for farmers and policy-makers that could improve the management practices in order to improve the nutrient cycling soil function. Finally, this model is being integrated in the LANDMARK H2020 project Soil Navigator DSS tool that integrates decision models for four more soil functions besides the nutrient cycling: primary productivity, soil biodiversity and habitat provision, water purification and regulation, and climate regulation and carbon sequestration. The tool will provide an overall assessment of the soil status at a field level in terms of the provision of the five main soil functions. The integration of five functions into one tool enables users to identify synergies and trade-offs of these functions and to make better informed decisions.

Acknowledgement

This research was conducted as part of the LANDMARK (LAND Management: Assessment, Research, Knowledge Base) project. This project has received funding the project LANDMARK within the European Union's Horizon 2020 research and innovation program under the Grant Agreement 635201.

References

[1] Arrouays, D., Saby, N. P. A., Thioulouse, J., Jolivet, C., Boulonne, L., and Ratié, C. (2011). Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. Geoderma 161, 107–114. doi: 10.1016/j.geoderma.2010.12.002

[2] Bohanec, M., and Rajkovic, V. (1990). DEX: An Expert System Shell for Decision Support. Sistemica 1, 145–157.

[3] Bohanec, M. 2017. Multi-criteria DEX models: an overview and analyses. In The 14th International Symposium on Operational Research in Slovenia, eds L.Z. Stirn,M. K. Borštar, J. Žerovnik, and S. Drobne (Ljubljana: Slovenian Society Informatika - Section for Operational Research), 155–60.

[4] Bohanec, M. (2019) DEXi: A Program for Multi-Attribute Decision Making. Available online at: http://kt.ijs.si/MarkoBohanec/dexi.html.

[5] Cordell, D., Drangert, J.O. & White, S. 2009. The story of phosphorus: global food security and food for thought. Global Environment Change, 19, 292–305.

[6] Creamer, R., and Holden, N. (2010). Special Issue: Soil Quality. Soil Use and Management 26(3), 197-197. doi: 10.1111/j.1475-2743.2010.00299.x.

[7] Creamer, R.E., Brennan, F., Fenton, O., Healy, M.G., Lalor, S.T.J., Lanigan, G.J., et al. (2010). Implications of the proposed Soil Framework Directive on agricultural systems in Atlantic Europe - a review. Soil Use and Management 26(3), 198-211. doi: 10.1111/j.1475-2743.2010.00288.x.

[8] Erisman, J.W., Van Grinsven, H., Grizzetti, B., Bouraoui, F., Powlson, D., Sutton, M.A., Bleeker, A. & Reis, S. 2011. The European Nitrogen Problem in a Global Perspective. In: The European Nitrogen Assessment (eds M.A. Sutton, C.M. Howard, J.W. Erisman, G. Billen, A. Bleeker, P. Grennfelt, H. van Grinsven & B. Grizzetti), pp. 9–31. Cambridge University Press, Cambridge.

[9] Schröder, J.J. 2005. Revisiting the agronomic benefits of manure: a correct assessment and exploitation of its fertilizer value spares the environment. Bioresource Technology 92 (2), 253-261

[10] Schröder, J.J., Schulte, R.P.O., Creamer, R.E., Delgado, A., van Leeuwen, J., Lehtinen, T., et al. (2016). The elusive role of soil quality in nutrient cycling: a review. Soil Use and Management 32(4), 476-486. doi: 10.1111/sum.12288.

[11] Schröder, J.J., Kuzmanovski, V., Picaud, C., Saby, N. & Debeljak, M. (2018) Validation of the DEXi model of Nutrient Cycling -3nd exploration based on a selection of 534 site-years with wheat (T. aestivum and T. durum) from France-, Note for LANDMARK, Wageningen UR, Wageningen, the Netherlands, 5 pp.

[12] Schulte, R. P. O., Creamer, R. E., Donnellan, T., Farrelly, N., Fealy, R., O'Donoghue, C., et al. (2014). Functional land management: A framework for managing soil-based ecosystem services for the sustainable intensification of agriculture. Environ. Sci. Policy 38, 45–58. doi:10.1016/j.envsci.2013.10.002.

[13] Stolte, J.M., Tesfai, L., Øygarden, S., Kværnø, J., Keizer, F., Verheijen, P., et al. (2016). Soil threats in Europe. EU Joint Research Centre. EUR 27607 EN; doi:10.2788/488054 (print); doi:10.2788/828742 (online).

[14] Withers, P.J.A., Van Dijk, K.C., Neset, T.S., Nesme, T., Oenema, O., Rubaek, G.H., Schoumans, O.F., Smit, A.L. & Pellerin, S. 2015. Stewardship to tackle global phosphorus inefficiency: the case of Europe. Ambio, 44, 193–206.

APPLICATION OF WEIGHTED GOAL PROGRAMMING METHOD FOR HYBRIDS SELECTION OF ENDIVES

Jožef Vinčec
University of Maribor Faculty of agriculture and life science
Pivola 11, 2311 Hoče
E-mail: jozef.vincec@student.um.si

Karmen Pažek
University of Maribor Faculty of agriculture and life science
Pivola 11, 2311 Hoče
E-mail: Karmen.pazek@um.si

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

Jernej Prišenk
University of Maribor Faculty of agriculture and life science
Pivola 11, 2311 Hoče
E-mail: jernej.prisenk@um.si

Abstract: Choosing a hybrid for sowing is one of the key tasks in the agricultural production planning. By properly planning, we can gain a key advantage over competition. We approach to the problem with construction of a model in which we are looking for an appropriate hybrid for a specific period from all of the available hybrids. We used a weighted goal programming method with criteria: growing area, growing season and the financial result of hybrid for a certain week. Results where choice for a single hybrid is expressed in binary form. The results of the weighted goal programming were compared with the scenario in which were selected those hybrids with the minimum occupied area (MIN). The results show that with the use of the MIN scenario we achieve a smaller required area for 0.04 ha compared to weighted goal programming 4.55 ha. However, use of weighted goal programming achieves 2 000.30 € better financial result and have 16 days shorter growing season compared to the MIN scenario.

Keywords: weighted goal programming, hybrid selection, endivie

1 INTRODUCTION

Application of operational research in agriculture usually means finding the optimum crop rotation. El-Nazer and McCarl (1986) used LP model to find optimal crop rotations after having built a regression model to estimate yield. Dogliotti et al. (2003) for example developed ROTAT to systematically generate all possible rotations from given number of crops. Another important area of operational research in agriculture is diet planning. Anderson and Earle (1983) take step further and provided possibility of applying goal programming to diet planning instead of conventional linear programming. Prisenk et al. (2013) use weighted goal programming to determine optimal feed rations for sport horses. Other popular areas of research in agriculture mean minimizing the cost of machine or labour work, fertilization costs, and others.

The approach to planning production within a certain culture is less present. In agricultural production farmers are faced with a large number of different hybrids available at the market, which can be cultivated during same period. Hybrids in vegetable production have a certain requirement for temperature and duration of daylight. Based on that conditions they are sorted in growing periods when they can be cultivated in certain area.

Other important fact is that different hybrids have different length of growth and different amount of yield at the same area compared to other hybrids. Therefore, in theory we can achieve higher yield just by choosing appropriate hybrid. Selection of an appropriate hybrid for a particular week is the key task and goal for developing of our modelling tool. Farmers are usually faced with the choice of most suitable hybrid for a particular week in the year once a year for whole growing season. Choosing appropriate hybrid for a particular week is associated with the estimated demand (in kg) per culture for a particular week.

We based this research on the case of endives. The selection procedure for a suitable hybrid consists of 2 parts. In 1st part, we make a list of wider range of hybrids for a particular culture that exist on the market. Data on hybrids are obtained from catalogues of hybrid seeds suppliers and on the basis of these data we estimate economic viability of the production for certain hybrid. The second part is the selection of a specific hybrid with the help of a developed methodology based on weighted goal programming method and is calculated for each week of the year. Based on the agronomic requirements for each hybrid, then we compare different hybrids and choose the appropriate one using the developed model.

2 METHODOLOGY

We first approached to the problem by choosing appropriate hybrid for a particular week by analyzing the economic viability of a broader set of hybrids. We used a methodology of calculating total costs for estimation a production by particular hybrid. The results enter to the model for choosing the appropriate hybrid as a restriction. We used a weighted goal programming which is according to Chang (2007) most widely used multi-objective technique in management science because of its inherent flexibility in handling decision-making problems with several conflicting objectives and incomplete or imprecise information.

Our model allows direct compromise between all unwanted deviations of the variables by combining them into a weighted, normalized goal function. Assuming the linearity of the target function, then we can present a linear weighted target program as:

$$\text{Min a} = \sum_{q=1}^{Q}\left(\frac{u_q n_q}{k_q} + \frac{v_q n_q}{k_q}\right) \tag{1}$$

Subject to:

$$f_q\,(\underline{x}) + n_q - p_q = b_q \quad q = 1, \dots ,Q$$

$$\underline{x} \in F$$

$$n_q,\, p_q \geq 0 \quad q = 1, \dots , Q$$

Where is:

- nq is the negative deviational variable of the qth goal. It represents the level at which the target level is not sufficiently achieved.
- pq is the positive deviational variable of the qth goal. It represents the level by which the target level is over-achieved
- bq is numeric target level for each goal
- fq(x) the achieved value according to the basic criterion.
- kq is the normalisation constant associated with the qth goal.

Variable definitions are the same as in priority programming, except that the u_q and v_q weights are no longer indexed by priority levels.

Limitations in the assessment for each method are:

- Financial result (€ / area), which must be greater than the average of all FR hybrids / varieties in individual crops.
- Growing period (number of days), which must be less than the average of all growing periods of hybrids / varieties in a particular culture.
- Required area (ha), which must be less than the highest estimated area of hybrids / varieties in the individual crop.
- The sum of the decisions can only be 1.

Weighted goal programming is based on Archimedes' goal function, which minimizes the sum of weighted deviations from individual goals. Consensus is thus achieved by minimizing the weighted sum of disagreements (Gonzales-Pachon and Romero 1999, cited in Žgajnar 2011).

Using the model, we analysed 5 hybrids of endives with a growing period from 60 to 80 days and are intended for production for fresh consumption in the period from 25 to 42 weeks of the year.

The goal of the developed model is to select the appropriate hybrid in a given week according to the criteria of the financial result, the growing season and the required area for the production of the desired quantity of endives in kilograms for a certain week of the year. Table 1 show the necessary demands of endives for our case on the basis of which according to the recommended plant set-up, the estimated weight of the endive head is calculated the required area (ha) for each endive hybrid to meet the estimated needs.

Table 1: List of endive needs for each week of the year in the growing season

Week	25	26	27	28	29	30	31	32	33
Needs	6.466,71	6.466,71	6.466,71	6.466,71	6.466,71	6.466,71	6.466,71	6.466,71	6.466,71
Week	34	35	36	37	38	39	40	41	42
Needs	6.466,71	6.466,71	6.466,71	6.466,71	6.734,38	7.489,11	6.466,71	6.466,71	6.466,71

On the basis of the demand (Table 1) for endives in a given week, the specific endives hybrid, the recommended planting set, the predicted weight of the individual endives head are indicated in Table 2, the estimated area required (ha), with which of all available hybrids for each week in a year, meet the estimated needs from Table 1. When certain hybrid it’s not allowed to grow (by the supplier of certain hybrid) in certain week, we add large number (999) and prevent to be chosen.

Table 2: Estimated required areas (ha) to meet the demand (in kg) in each week per individual hybrid

Week	Eros F1	Amos F1	Géante maraîchère - Bossa	Anconi RZ	Mikado RZ
25	999.00	0.28	999.00	999.00	999.00
26	999.00	0.28	0.28	999.00	999.00
27	999.00	0.27	0.28	999.00	999.00
28	999.00	0.26	0.27	999.00	999.00
29	999.00	0.26	0.27	999.00	999.00
30	999.00	0.25	0.26	999.00	999.00
31	999.00	0.25	0.26	999.00	999.00
32	999.00	0.24	0.26	999.00	999.00
33	999.00	0.24	0.25	999.00	999.00
34	999.00	0.23	0.25	0.32	0.32
35	999.00	0.23	0.24	0.31	0.31
36	0.28	999.00	0.24	0.30	0.30
37	0.26	999.00	0.24	0.30	0.30
38	0.25	999.00	0.24	0.30	0.30
39	0.26	999.00	0.27	0.32	0.32
40	999.00	999.00	0.23	0.27	0.27
41	999.00	999.00	999.00	0.27	0.27
42	999.00	999.00	999.00	0.26	0.26

As a limitation to the modelling tool, the financial result for each hybrid is calculated by method of calculating total costs. Besides financial result, duration of growing for each hybrid are shown in Table 3.

Table 3: Estimated financial result for each hybrid and the growing season from planting to harvest

	Eros F1	Amos F1	Géante maraîchère - Bossa	Anconi RZ	Mikado RZ
FR	9 906.56	7 049.42	5 620.85	3 156.56	2 406.56
Grow duration	60	61	66	80	78

The model tool was developed in the Microsoft Excel (Figure 1) and uses the Solver add-in. The procedure for choosing the appropriate hybrid in each week was designed using the code written in the Microsoft Visual Basic for Applications, which is considered as a stand-alone problem, which we solve individually each week.

Figure 1: Section from developed model.

The model tool enables the selection of suitable hybrids of the endive, which provides such a hybrid at the required quantity.

3 RESULT AND DISCUSSION

The result of our developed model is the selection of the appropriate hybrid in each week. In the developed model, the selection of the appropriate hybrid is indicated in Table 4. Results are shown in binary mode, which means, that when 1 appears the hybrid is selected.

Table 4: Result of modelling tool

Week	Eros F1	Amos F1	Géante maraîchère - Bossa	Anconi RZ	Mikado RZ
25	0	1	0	0	0
26	0	1	0	0	0
27	0	1	0	0	0
28	0	1	0	0	0
29	0	1	0	0	0
30	0	1	0	0	0
31	0	1	0	0	0
32	0	1	0	0	0
33	0	1	0	0	0
34	0	1	0	0	0
35	0	1	0	0	0
36	0	0	1	0	0
37	1	0	0	0	0
38	1	0	0	0	0
39	1	0	0	0	0
40	0	0	1	0	0
41	0	0	0	0	1
42	0	0	0	0	1

Table 5: Comparism between results based on weighted goal programming and basic scenario (MIN)

Week	Chosen WGP	Area WGP	FR WGP	Grow WGP	Chosen MIN	Area MIN	FR MIN	Grow MIN
25	Amos F1	0.28	1 994.41	61	Amos F1	0.28	1 994.41	61
26	Amos F1	0.28	1 945.77	61	Amos F1	0.28	1 945.77	61
27	Amos F1	0.27	1 899.44	61	Amos F1	0.27	1 899.44	61
28	Amos F1	0.26	1 855.27	61	Amos F1	0.26	1 855.27	61
29	Amos F1	0.26	1 813.10	61	Amos F1	0.26	1 813.10	61
30	Amos F1	0.25	1 772.81	61	Amos F1	0.25	1 772.81	61
31	Amos F1	0.25	1 734.27	61	Amos F1	0.25	1 734.27	61
32	Amos F1	0.24	1 697.37	61	Amos F1	0.24	1 697.37	61
33	Amos F1	0.24	1 662.01	61	Amos F1	0.24	1 662.01	61
34	Amos F1	0.23	1 628.09	61	Amos F1	0.23	1 628.09	61
35	Amos F1	0.23	1 595,53	61	Amos F1	0.23	1 595.53	61
36	Géante maraîchère - Bossa	0.24	1 349.30	66	Géante maraîchère - Bossa	0.24	1 349.30	66
37	Eros F1	0.26	2 587.15	60	Géante maraîchère - Bossa	0.24	1 329.16	66
38	Eros F1	0.25	2 501.79	60	Géante maraîchère - Bossa	0.24	1 363.82	66
39	Eros F1	0.26	2 596.70	60	Eros F1	0.26	2 596.70	60
40	Géante maraîchère - Bossa	0.23	1 272.19	66	Géante maraîchère - Bossa	0.23	1 272.19	66
41	Mikado RZ	0.27	638.46	78	Anconi RZ	0.27	837.44	80
42	Mikado RZ	0.26	622.50	78	Anconi RZ	0.26	816.50	80
TOTAL		4.55	31 166.17	1139		4.51	29 163.17	1 155

The results of weighted goal programming (the left side of Table 5) are compared with the selection of that hybrid, which occupies the minimum required area (the right side of the table).

Purpose of comparing both scenarios is to show what is best to use: presented methodology in agricultural planning or to stay with old technique which means usage of selection hybrid that occupies minimum required area. The area to meet the needs listed in

Table 1 is greater by 0.04 ha when using weighted goal programming compared to the use of the smallest possible areas (MIN scenario). When comparing the financial result, using results from a weighted goal programming method, the financial result of 2 003.00 € is higher compared to the results from MIN scenario. An analysis of growing season by a period from planting to harvest shows that the use of weighted goal programming has a lower total amount of growing period by 16 days compared to the MIN scenario.

4 CONCLUSIONS

Production planning and the selection of appropriate hybrids are key steps in the agricultural production planning and potential search for a competitive advantage within the production of a particular culture. This is best confirmed by the fact that in the production of the same amount of crop with the choice of the appropriate hybrid, a smaller area can be occupied and less growth days spent. To make an appropriate decision and to reduce decision maker preference over particular decision on hybrid selection we have developed a model. Model was developed in Microsoft Excel with Solver add-in and uses weighted goal programming method to find the appropriate hybrid. The results show that the use of the developed model is justified, since in comparison with the reference scenario, results from developed model achieve a higher financial result of 2 003.00 € and a 16-day shorter growing season.

References

[1] Chang C-T. 2007. Efficient structures of achievement functions for goal programming models. Asia Pac J Oper Res 24(6):755–764

[2] Dogliotti, S., Rossing, W. A. H., Van Ittersum, M. K., 2003. ROTAT, a tool for systemati-cally generating crop rotations. European Journal of Agronomy 19, 239–250.

[3] El-Nazer, T., McCarl, B. A., 1986. The choice of crop rotation: A modeling approach and case study. American Journal of Agricultural Economics 68 (1), 127–136.

[4] Gonzales-Pachon J. and Romero C. 1999 Distance-based consensus methods: a goal programming aprocah. Omega: 27(3). 341−347 p.

[5] Prišenk J., Pažek K., Rozman Č., Turk J., Janžekovič M. and A. Borec. 2013. Application of weighted goal programming in the optimization of rations for sport horses. Journal of Animal and Feed Sciences, 22, 335–341 p.

[6] Žgajnar J. 2011. Večkriterijsko optimiranje na kmetijskih gospodarstvih v razmerah tveganja. Ponatis doktorske disertacije DAES. 194 p

THE SYSTEM DYNAMICS MODEL FOR DIVERSIFICATION OF AGRICULTURAL HOLDINGS INTO FARM TOURISM

Maja Žibert
University of Maribor, Faculty of agriculture and life sciences
Pivola 10, 2311 Hoče, Slovenia
E-mail: maja.zibert@student.um.si

Črtomir Rozman
University of Maribor, Faculty of agriculture and life sciences
Pivola 10, 2311 Hoče, Slovenia
E-mail:crt.rozman@um.si

Boris Prevolšek
University of Maribor, Faculty of tourism
Cesta prvih borcev 36,8250 Brežice, Slovenia
E-mail: boris.prevolsek@um.si

Andrej Škraba
University of Maribor, Faculty of organizational sciernces
Kidričeva cesta 55a,4000 Kranj, Slovenija
E-mail: andrej.skraba@fov.uni-mb.si

Abstract: The aim of this paper is to research, by means of systems thinking and the model of system dynamics, the main variables and their causal relationships in the system structure which presents the diversification of farming establishments to non-agricultural industries on the farm. In the theoretical part, the theory of systems thinking and system dynamics is presented. The second part of the article represents the system structure – CLD. The model presents information on the main feedback loops and their dynamics in addition.

Keywords: CLD – causal loop diagram, system dynamics, diversification of farming establishments, tourism, rural tourism

1 INTRODUCTION

Farm tourism is not a new phenomenon [4]. It is a form of countryside tourism which dates back a century in some destinations [6]. The developmental trends show that more supplementary activities are registered within farming establishments every year. Their common denominator is tourism. It is definitely the consequence of the increasing number of tourists, i.e. lodgings in the country. In 2018, Slovenia recorded 5.93 millions of tourists' arrivals and 15.96 millions of lodgings of the tourists. A part of them in tourist farms too. A rich history of the development of tourism in the countryside is recorded in Germany [10] and in Austria. State policies are positively oriented towards the development of tourist facilities in the countryside with subsidies and programs of development also in Italy [8] and France [2]. Along with all the advantages and disadvantages which are represented by the development of supplementary activities related to tourism, there is a clear tendency of the use of modern supports in decision-making by which the proper directives can be ensured before bigger investment and activities affecting the environment.

In the case of our article, we speak about the support in agricultural policy and development decision-making within the procedures of diversification of the basic agricultural activities in farming establishments. Such types of dynamics were already used by numerous researchers [13], [3], [1].

The document is divided into two parts. First, we become acquainted with the theory of systems thinking and system dynamics. Then, in the second part, we discuss the problem of spreading and introduction of a supplementary activity in the farming establishment – the farm tourism by the use of causal loop diagram as one of the ways of support in decision-making.

2 SYSTEMS THINKING, SYSTEM DYNAMICS AND CAUSAL LOOP DIAGRAM

Sterman[14] says that systems thinking is necessary for efficient decision-making. Such thinking approach has been used for a number of years (more than 60). In time, it has been developed and improved all the time. Richmond [12] speaks about systems thinking as about a multidimensional system where:

a) We can *think with models*, which mean the ability to build a model and transfer the acquired knowledge into a real circumstance.
b) We speak about *dynamic thinking* which enables anticipation of future behavior of systems with all the delays, fluctuations, and feedback loops.
c) We can understand a system as *interrelated thinking* where a single cause does not mean a single consequence. Consequences depend on a multitude of indirect influences.
d) *The system management* – we understand the dimension of systems thinking as the most pragmatic component.

Systems thinking and system dynamics observe the same types of problems. Contrary to the systems thinking, the system dynamics enables us – by means of computer simulations of the models – a depiction of the behavior of the real system when testing the effects of alternative decisions through time [5]. As the model is presented later, causal loop diagrams contain variables and causal relationships presented with arrows which are labeled with a mark or reinforcing or balancing. Reinforcing (R) means that the effect over value increases over the value it would normally have if the cause increases. Balancing (B), however, means that the consequence is reduced below the value it would normally have if the cause increases. When the system elements are interconnected and form a closed sequence of causes and consequences, we speak about the causal loop.

3 CAUSAL LOOP DIAGRAM OF SYSTEM STRUCTURE

As explained introductory, we seek additional sources of income by diversification to non-agricultural industries in farming establishments. This action, however, does not influence positively only a farming establishment but also offers numerous advantages for the broader region: the quality of life in the countryside, culture, tradition, and, last but not least, employment. Due to all the specifics of the agrarian structures and lowering the factor incomes per employee in the agriculture [15], the farmers have to think hard about all the factors, not only economic ones, when they think about the step of diversification, especially if an investment would require more significant financial input. A contribution by the use of system dynamics (SD) presents the method for the support in decision-making. Figure 1 represents the causal loop diagram of system structure – diversification of farming establishments in tourist farms with lodging with important consequences for the region and the farming establishment.

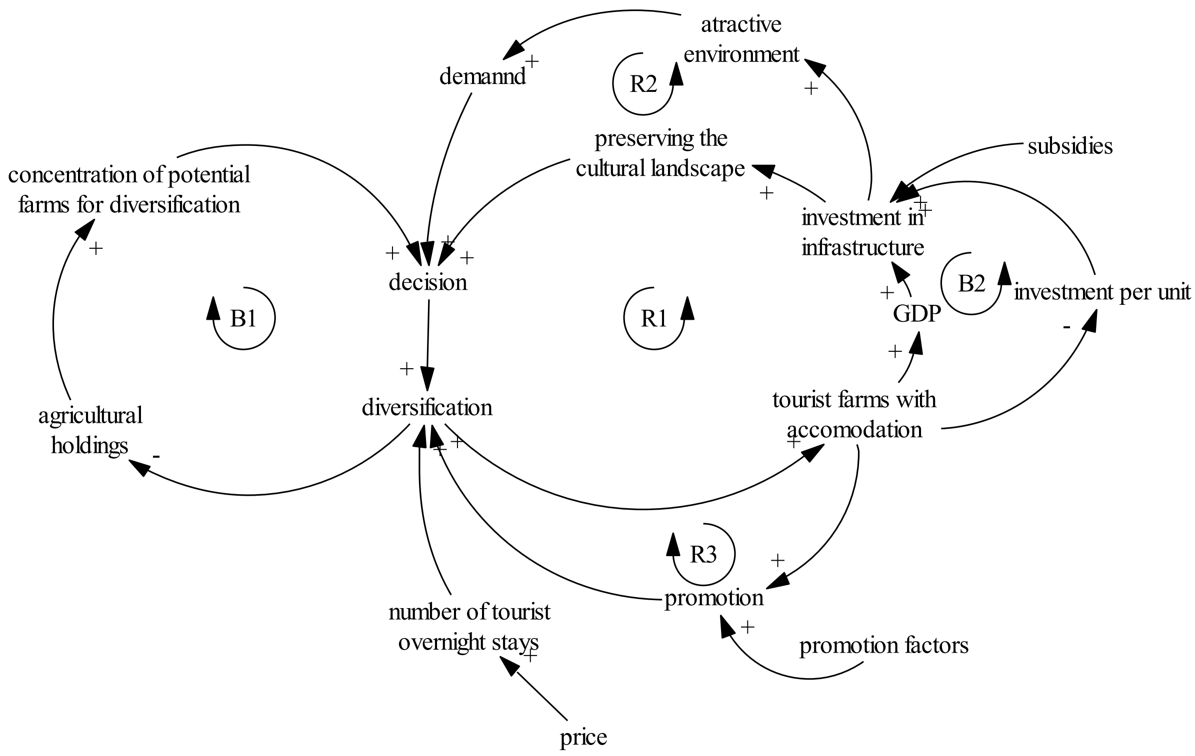


Figure 1: Causal loop diagram of system structure – diversification of farming establishments in tourist farms

In the system dynamics model, we can see several main feedback loops which represent reinforcing and balancing. The loops R1, R2, and R3 indicate the developmental activity. In 2018, the industry of tourism and traveling created 3.3% of gross domestic product (GDP) in Slovenia directly. With respect to broader influences, however, it created 11.9% of GDP [17].

Increasing or growth of GDP influences investments in infrastructure directly, which has positive consequences on the environment mostly, as this is the way it is preserved more easily. In addition, the destinations are more easily accessible. At the same time, it influences environmental attractiveness. The attractiveness of the environment increases the demand for lodging and/or visiting destinations. Increasing the demand influences positively the decision of farming establishment whether it will diversify its primary industry. As already mentioned, this diversification influences economic effects, employability, and the quality of life positively. By the development of supplementary activity – farm tourism, the opportunities emerge for the development of other supplementary activities related to the cultivation of primary agricultural crops, the sales of agricultural crops and products of farms, activities which are connected to traditional knowledge on farms, and social security services.

In a broader perspective, not all farming establishments are appropriate for a step of this type of diversification. Žibert et. al. [18] researched the attributes of farming establishments for diversification to non-agricultural industries.

An important variable of system structure is also the promotion factor. Not in the sense of promotion of the industry that tourism is the catalyst which would help in economic challenges of the countryside [9], [16], but in the sense of the promotion of tourist farms, destinations, the tradition of cultural habits, events, and environment whose part is the farming establishment itself. These are, therefore, the tools which are available to farms or broader groups of entities, and through which they communicate with their target publics about all the matters which influence the profitability and, primarily, the decision for the step of diversification [11].

In spite of everything, however, the share of GDP cannot entirely cover the investments in infrastructure which helps in the development of the tourist industry. Gartner [7] reports on numerous support rates in the development of the industry. In spite of that, however, the share of investments in infrastructure per unit shows one of the decisive equalization loops (B2) which influence the fact in the system whether the farm will diversify its industry or not. Figure 1 represents a quality model and discusses the important relationships of causal loops which influence the decision for diversification of farming establishments into supplement activities – farm tourism.

4 CONCLUSIONS

The model seeks answers to the strategic questions connected with the dynamics of transfer of farming establishments with the potential for the tourist activities to tourist farms. The model will be used for the depiction of the behavior of the real system when testing the effects of alternative decisions through time.

References

[1] Bastan, M., Ramazani Khorshid-Doust, R., Delshad Sisi, S., & Ahmadvand, A. (2018). Sustainable development of agriculture: a system dynamics model. *Kybernetes*, *47*(1), 142-162.

[2] Bel, F., Lacroix, A., Lyser, S., Rambonilaza, T., & Turpin, N. (2015). Domestic demand for tourism in rural areas: Insights from summer stays in three French regions. *Tourism Management*, *46*, 562-570.

[3] Blumberga, A., Bazbauers, G., Davidsen, P. I., Blumberga, D., Gravelsins, A., & Prodanuks, T. (2018). System dynamics model of a biotechonomy. Journal of Cleaner Production, 172, 4018-4032.

[4] Busby, G., & Rendle, S. (2000). The transition from tourism on farms to farm tourism. Tourism Management, 21(6), 635–642.

[5] Dangerfield, B. (2014). Systems thinking and system dynamics: A primer. *Discrete-event simulation and system dynamics for management decision making*, 26-51.

[6] Dernoi, L. (1983). Farm tourism in Europe. Tourism Management, 4(3), 155–166.

[7] Gartner, W. (2004). Rural tourism in the USA. International Journal of Tourism Research, 6(3), 151–164.

[8] Giaccio, V., Mastronardi, L., Marino, D., Giannelli, A., & Scardera, A. (2018). Do Rural Policies Impact on Tourism Development in Italy? A Case Study of Agritourism. *Sustainability*, *10*(8), 2938.

[9] Hoggart, K., Buller, H., & Black, R. (1995). Rural Europe; identity and change. London: Arnold.

[10] Oppermann, M. (1996). Rural tourism in southern Germany. Annals of Tourism Research, 23(1), 86–102.

[11] Podnar, K., & Golob, U. (2001). *The problem of advertorial and commercialization of Slovene press*.

[12] Richmond, B. (1993). System thinking: critical thinking skills for the 1990s and beyond, System Dynamics Review, vol. 9, nr. 2, summer 1993, 113 – 133.

[13] Rozman, Č., Pakež, K., Kljajić, M., Bavec, M., Turk, J., Bavec, F., … & Škraba, A. (2013). The dynamic simulation of organic farming development scenarios – A case study in Slovenia. Computer and Electronics in Agriculture, 96, 163 – 172, http://dx.doi.org/10.1016/j.compag.2013.05.005

[14] Sterman, D. J. (2000). Business Dynamics: Systems Thinking and Modeling for a complex World, Irwin McGraw-Hill, Boston, MA, USA, 2000.

[15] SURS. (2017). https://www.stat.si/statweb/News/Index/7109 [Accessed 7/6/2019].

[16] Williams, A., & Shaw, G. (Eds.). (1998). Tourism and economic development: European experiences (3rd ed.). Chichester: Wiley.

[17] WTTC. (2018). https://www.wttc.org/economic-impact/country-analysis/country-reports/ [Accessed 7/6/2019].

[18] Žibert, M., Rozman, Č., Prevolšek, B., Škraba, A. (2019). Attributes of the agricultural holding for diversification of farm activities. Paper presented at conference Ekologija za boljši jutri. Rakičan.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 2:

Formal and Behavioral Issues in MCDM

Overconfidence in electronic reverse auctions

Ayşegül Engin
ayseguel.engin@univie.ac.at
Department of Business Decisions and Analytics, University of Vienna, Oskar Morgenstern Platz 1, Vienna, 1090, Austria
Rudolf Vetschera
rudolf.vetschera@univie.ac.at
Department of Business Decisions and Analytics, University of Vienna, Oskar Morgenstern Platz 1, Vienna, 1090, Austria

Abstract: Overconfidence, which is mainly considered as a cause for the winner's curse in common value auctions, can have a significant impact on bidding behavior in auctions. This paper experimentally studies optimistic overconfidence in the electronic reverse auction context. We argue that overconfidence in this context may lead to misperceptions of the bidder's own cost affecting his probability of winning. We also study interactions of overconfidence with other psychometric properties of the bidder, and the format of information presentation. The results indicate that overconfidence leads to more aggressive bidding, but the effect can be mitigated in participants with higher risk literacy.

Keywords: Reverse auctions, Overconfidence, Information representation, Experiment

1 INTRODUCTION

Overconfidence is an important bias that influences decisions, whenever subjective probability judgements are involved. This paper focuses specifically on optimistic overconfidence, which is defined as overestimating the likelihood of a favorable outcome in a risky situation [12]. Particularly in competitive situations like auctions such distortions may lead to aggressive bidding behavior. Existence of aggressive bidding behavior in an auction does not only harm the decision maker in question, but also can destabilize the whole market.

Even though biases are important concerns in any type of auctions, we argue that systematic decision behavior distortions are especially important for electronic auctions, as the bidders can only interact with the market through an electronic medium. Electronic reverse auctions (ERAs) have gained considerable popularity as procurement tools [see for review 16], because of their advantages (i.e. reduced sourcing time, lower selling costs, easier access to new markets and customers, increased market transparency [e .g. 17]). Yet suppliers perceive many disadvantages of ERAs [5].

Overconfidence is a frequently discussed bias in auction context (i.e.winner's curse [2]). Several empirical studies have empirically confirmed that overestimation of the auctioned object's value result in winner's curse [15, 4]. One problem of empirical studies of actual auctions is that it is not possible to relate the outcome of the auction (i.e. the fact that the winner actually incurs a loss) to individual characteristics of the winning bidder. Such characteristics can be measured in laboratory studies. Many experimental studies on overconfidence in auctions both in the financial domain [e .g. 19] as well as in other domains [e.g. 8] do not include psychometric factors.

In reverse auctions winner's curse and overconfidence are less studied, even though from a theoretical perspective the potential impact of overconfidence is argued [13]. In our view ERAs we consider in the procurement setting differ in various aspects from the frameworks that are considered in the literature. Most of the literature on overconfidence is based on the common value model and argues that overconfident bidders will have a higher estimate of that ex post revealed value [14]. Nevertheless, a similar argument can also be made for the independent private values model [7].

Uncertainty of the true cost is only one aspect in the decision problem in ERASs, where overconfidence may affect the decision process. Another aspect is that the bidders have uncertainty about other bidders' behaviors and costs. Therefore overconfidence may lead the bidders to overestimate their winning chances in the auction.

In a reverse auction setting, overconfidence concerning these two uncertainties influences bidding behavior in the opposite direction. Being overconfident about one's own cost would lead to too low bids, on the other hand being overconfident about the competitors' costs can lead to too high bids. The first goal of the present study is to analyze the relative strength of these two effects. The second goal of the present study is to explore possible relationships between overconfidence and other psychometric properties of the bidders.

Finally, we explicitly consider the way in which (stochastic) information is presented to bidders as another factor that can influence bidding behavior, in particular in an electronic auction. Since different ways of presenting (otherwise identical) information have an impact on decision behavior [9] and thus possibly on the occurrence of biases, there is a risk that buyers can influence bidding behavior in a subtle way via the interface design of ERAs.

The remainder of the paper proceeds as follows: section two discusses the research model and the hypotheses, section three explains the study design and data collection in detail, section four contains analyses of the experimental data and results and section five concludes the paper.

2 RESEARCH MODEL AND HYPOTHESES

Figure 1 represents the framework underlying our research. Individuals with optimistic overconfidence are likely to overestimate the quality of their own judgments [18]. We therefore expect overconfident bidders to also overestimate their ability to make the optimal bid that will win the auction. This should be reflected in their subjective probability of winning the auction.

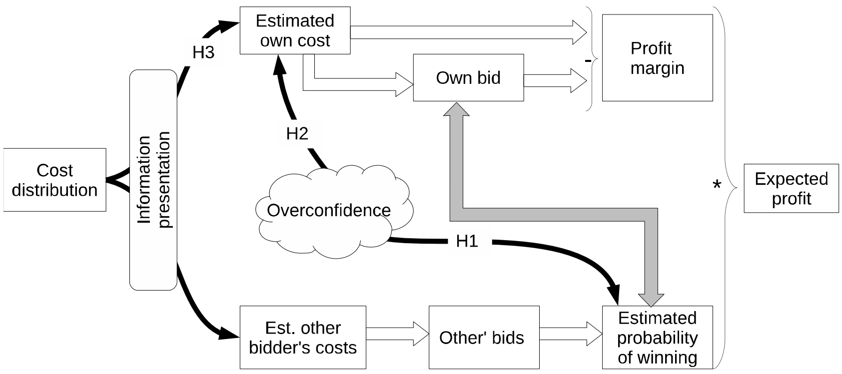


Figure 1: Research framework

We therefore formulate our first hypothesis as follows:

H1: The higher individuals score on the psychometric overconfidence measure, the higher they estimate their probability of winning the auction.

Overconfident and thus overly optimistic bidders will underestimate their true costs of completing a project. This misjudgment of their costs will influence their bidding behavior and will lead them to bid a lower value. We therefore formulate our second hypothesis as follows:

H2: The higher individuals score on the psychometric overconfidence measure, the lower the value they bid in relation to their true costs, and therefore the lower their true profit margin.

By focusing on the profit margin rather than the actual bid, we take into account that the true costs (and therefore the "correct" bid values) of bidders are different.

Our hypotheses *H1* and *H2* lead to opposite predictions concerning the actual value of the bid. Taking into account the relationship between probability of winning and bid value, overconfident bidders according to H1 should make higher bids. They overestimate their probability of winning the auction and thus, they would assign the same probability of winning to a higher bid than bidders who are less overconfident. They will thus be more willing to take the risk of a higher bid (asking for a higher price) because they still perceive the probability of winning the auction to be high. In contrast, according to *H2*, overconfident bidders underestimate their true costs, and bid a lower value. By referring to the bid value in *H2*, we implicitly assume that the more direct effect on bid values via a misperception of costs is stronger than the indirect effect via misjudging the probability of winning.

As previous experiments [9] have shown, there is not one best way of presenting information, but decision makers perform best when information is presented to them in a way that matches their cognitive style. We therefore formulate our third hypothesis as

H3: Individuals, who receive information in a format matching their cognitive style, submit bids which have higher profit margin.

3 STUDY DESIGN AND DATA COLLECTION

In order to investigate the hypotheses developed in the previous chapter, we designed and performed a computerized laboratory experiment using student subjects. In our study, we used a performance based incentive in the form of course credits.

Our research model shown in Figure 1 and the hypotheses derived from it involve two independent variables: Overconfidence in H1 and H2, and the match between information presentation and cognitive style in H3. Neither of these variables can directly be controlled as an experimental factor. Overconfidence is a psychometric characteristic of individuals, which we measured using the instrument of Blavatskyy [3].

Hypothesis H3 postulates that the match between information presentation and cognitive style has a positive effect on performance in the bidding task, which we measure by the profit margin. Matching between information presentation and cognitive style involves the information format, which is a controllable factor in the experiment, and cognitive style, which is a psychometric characteristic of individuals.

Our approach to determine whether information presentation matches the cognitive style of an individual follows [9]. Information on costs was presented either as a table or as a graph. Subjects were assigned randomly to one of these two presentation formats. In accordance with the results of [9], we consider that information presentation matches the cognitive style if subjects in the analytical half of the CSI scale receive information in tabular form, or subjects in the intuitive half of the scale receive information in graphical form. Otherwise, there is a mismatch.

The dependent variable in H1 is the subjective probability of winning the auction, which was directly elicited during the auction. The dependent variable in H2 and H3 is the profit margin of bids, which is the difference between the bid value and the true costs. Apart from the factors that we analyze in this study, the bid value might depend on several other characteristics of subjects. We therefore included several control variables in our study to control for the effect of these factors. Since determining the optimal bid value is a complex task, we included several controls that are related to the rational processing of numerical information: the Berlin numeracy and literacy test [6] for measuring participants' ability to decide on the basis of probabilistic information, the cognitive reflection test [11] to measure the subject's ability to reflect on decisions they make, and the rational−experiential inventory (REI) by Epstein et al. [10] to measure cognitive orientation of subjects. Furthermore, since bidding is a decision taken under risk, individual risk attitudes could influence bidding behavior. We control for risk attitudes by including a general measure of risk attitudes based on Abdellaoui et al. [1].

In total, 149 students from the faculty of business and economics of a central European university participated in the study. From the sample of 149, 21 participants were excluded due to their wrong answers in the control question. The control question examined, whether the participants are able to understand under which scenario they have lower costs. This left data from 128 subjects, 62 female (age: $\mu = 24.13, \sigma = 2.85$) and 66 male (age: $\mu = 24.3, \sigma = 3.99$) for analysis, of whom 67 received the graphical and 61 the textual information display.

4 RESULTS & DISCUSSION

Our data set contains observations (bid and estimated probability of winning) for five auctions from each of the eight subjects in each market. To account for possible effects of unobserved variables at the subject and market level, we performed our analyses using random effect models with subject and market as grouping variables.

To test hypothesis *H1*, we performed a regression on the self reported bid certainty value as dependent variable. We regress this variable on the optimistic overconfidence characteristic of the individual. To control for the objective probability of winning the auction, we include the quantile that the subject's cost come from. This is also the information that was available to individuals about costs. We further control for the subjects' ability to make decisions under risk (risk literacy), their level of intuitive behavior in decision making, and risk aversion. Table 1 summarizes the result for this regression.

	Dependent variable: Bid Certainty	
	Estimate	Std. Error
(Intercept)	$^{***}0.60$	(0.06)
2. Interquantile range	$^{*} -0.15$	(0.06)
3. Interquantile range	$^{***} -0.32$	(0.05)
Overconfidence	-0.02	(0.05)
Intuition score	-0.00	(0.00)
Risk literacy score	$^{**} -0.04$	(0.01)
Risk aversion	$^{*} -0.07$	(0.03)
2. Interquantile range:Overconfidence	0.08	(0.05)
3. Interquantile range:Overconfidence	$^{***}0.16$	(0.04)
BIC	-181.23	
Marginal R2	0.157	
Conditional R2	0.483	
Num. obs.	618	
Num. groups: participants (markets)	125 (18)	

***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

Table 1: Regression results for H1

To test hypotheses *H2* and *H3*, we performed a regression using the profit margin (i.e., the difference of bid value minus the bidder's true costs) as dependent variable. According to *H2*, the profit margin is influenced by overconfidence, so the psychometric measure of overconfidence is our first dependent variable. According to *H3*, bidders who receive information in a format matching their cognitive style should perform better than bidders receiving information in a mismatching format. Following Engin and Vetschera [9], matching is represented by the interaction term between information presentation (a binary variable set to one if information was received as a table), and the subject's CSI score. Similar to the regression in Table 1, we control for the bidder's true costs by including dummy variables representing the cost quantile. Bidders who have already high costs have to use a lower profit margin than bidders with low costs in order not to exceed the maximum bid value. Furthermore, we again control for risk literacy, risk aversion and intuitive behavior.

	Dependent variable: Profit Margin	
	Estimate	Std. Error
(Intercept)	$^{***}1526.50$	(343.30)
Bid certainty	$^{*}-2.60$	(1.31)
2. Interquantile range	$^{***}-598.60$	(63.82)
3. Interquantile range	$^{***}-1108.36$	(55.36)
Intuition score	$^{**}-12.28$	(4.41)
Risk literacy score	-161.68	(119.89)
Overconfidence	$^{***}-629.73$	(167.41)
Risk aversion	$^{*}250.03$	(101.80)
Treatment (Text)	-389.06	(371.00)
Total CSI Score	-7.15	(6.15)
Treatment (Text):Total CSI Score	11.53	(8.34)
Risk literacy score:Overconfidence	$^{*}238.68$	(105.64)
BIC	9856.46	
R-squared:	0.666	
Num. obs.	618	
Num. groups: participants (markets)	125 (18)	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Regression results for *H2* and *H3*

5 CONCLUSION

In this paper, we report on a study that investigates the effects of optimistic overconfidence of decision makers in an electronic auction framework when bidders themselves have only stochastic information about their true situation. This is a quite realistic setting in reverse (procurement) auctions, when potential suppliers or contractors have to bid for delivering inputs or executing a project which will take place in an uncertain future. In such a situation, bidders might easily fall into a winner's curse. Overconfidence leads to a shift in participants' perceived probability, leading to different decisions about the bid and thus their outcomes.

Our results first of all confirm that overconfidence indeed leads to more aggressive bidding behavior. Overconfident and thus more aggressive bidders then also are more likely to win the auction, but only by bidding a low price. This diminishes their profit margin, sometimes even below zero. Thus, considering the effects of overestimating the probabilities of winning an auction, and of interpreting one's own situation in an overly optimistic way (i.e., a lower cost situation), the latter effects seems to be the stronger one. This result is in accordance with the well known winner's curse effect, while overconfident overestimation of the probability of winning would lead to the opposite outcomes.

References

[1] Mohammed Abdellaoui, Han Bleichrodt, and Olivier L'Haridon. A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty*, 36(3):245–266, 2008.

[2] Max H Bazerman and William F Samuelson. I won the auction but don't want the prize. *Journal of Conflict Resolution*, 27(4):618–634, 1983.

[3] Pavlo R Blavatskyy. Betting on own knowledge: Experimental test of overconfidence. *Journal of Risk and Uncertainty*, 38(1):39–49, 2009.

[4] James A Brander and Edward J Egan. The winner's curse in acquisitions of privately-held firms. *The Quarterly Review of Economics and Finance*, 65:249–262, 2017.

[5] Marjolein CJ Caniëls and Erik M van Raaij. Do all suppliers dislike electronic reverse auctions? *Journal of Purchasing and Supply Management*, 15(1):12–23, 2009.

[6] Edward T Cokely, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero. Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, 7(1): 25–47, 2012.

[7] Olivier Compte. The winner's curse with independent private values. Technical report, ENPC Paris School of Economics, 2002.

[8] Ingolf Dittmann, Dorothea Kübler, Ernst Maug, and Lydia Mechtenberg. Why votes have value: Instrumental voting with overconfidence and overestimation of others' errors. *Games and Economic Behavior*, 84:17–38, 2014.

[9] Ayşegül Engin and Rudolf Vetschera. Information representation in decision making: The impact of cognitive style and depletion effects. *Decision Support Systems*, 103:94 – 103, 2017.

[10] Seymour Epstein, Rosemary Pacini, Veronika Denes-Raj, and Harriet Heier. Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, 71(2):390–405, 1996.

[11] Shane Frederick. Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42, 2005.

[12] Dale W Griffin and Carol A Varey. Towards a consensus on overconfidence. *Organizational Behavior & Human Decision Processes*, 65:227 – 231, 1996.

[13] Tarun Jain, Jishnu Hazra, and T.C.E. Cheng. Sourcing under overconfident buyer and suppliers. *International Journal of Production Economics*, 206:93–109, 2018.

[14] John H Kagel and Dan Levin. Bidding in common-value auctions: A survey of experimental research. In John H Kagel and Dan Levin, editors, *Common Value Auctions and the Winner's Curse*, pages 1–81. Princeton University Press, 2002.

[15] Luis E Pereiro. The misvaluation curse in mergers and acquisitions. *Journal of Corporate Accounting & Finance*, 27(2):11–15, 2016.

[16] Tobias Schoenherr and VM Rao Tummala. Electronic procurement: a structured literature review and directions for future research. *International Journal of Procurement Management*, 1(1-2):8–37, 2007.

[17] Alan Smart and Alan Harrison. Online reverse auctions and their role in buyer–supplier relationships. *Journal of Purchasing and Supply Management*, 9(5-6):257–268, 2003.

[18] Diemo Urbig, Julia Stauf, and Utz Weitzel. What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence. *Discussion Paper Series/Tjalling C. Koopmans Research Institute*, 9(20), 2009.

[19] Olufunke Rebecca Vincent, Ayodeji Samuel Makinde, and Adio Taofik Akinwale. A cognitive buying decision-making process in B2B e-commerce using Analytic-MLP. *Electronic Commerce Research and Applications*, 25:59–69, 2017.

A SCENARIO-BASED AHP METHOD FOR ONE-SHOT DECISIONS AND INDEPENDENT CRITERIA

Helena Gaspars-Wieloch
Poznan University of Economics and Business, Department of Operations Research
Al. Niepodległości 10, 61-875 Poznań, Poland
E-mail: Helena.gaspars@ue.poznan.pl

Abstract: The paper contains a brief description of the essence of AHP and a short analysis of existing AHP modifications for decision making under uncertainty, especially combined with scenario planning. The contribution also presents a novel scenario-based AHP approach which is only designed for one-shot decisions and independent criteria, i.e. targets influenced by totally different external factors (for each criterion a distinct set of scenarios is supposed to be defined). One of the advantages of the new approach is the possibility to generate a relatively smaller number of pairwise comparison matrices thanks to the reduction of the initial sets of scenarios.

Keywords: AHP, scenario planning, uncertainty, one-shot decisions, independent criteria, decision maker's preferences

1 INTRODUCTION

The Analytic Hierarchy Process (AHP) is a multi-criteria procedure which has been investigated and developed by many researchers and practitioners. AHP is one of the most popular approaches in decision analysis and one that is arguably more accessible for new users [5]. It is applied to diverse fields. The original version allows solving deterministic problems, but the necessity to operate in an uncertain environment has entailed the creation of numerous AHP modifications enabling one to take uncertainty factors into account.

This work also deals with AHP and multiple criteria decision making (MCDM) under uncertainty. Nevertheless, we focus on (1) aspects which are not thoroughly analysed in the literature and (2) issues which are handled here in a different way. We explore the case of independent criteria (for each criterion a distinct set of scenarios has to be defined) and one-shot decisions (the selected option is performed only once – hence, in such circumstances, the use of probabilities as known primary data is not appropriate). We suggest choosing the final decision on the basis of reduced sets of scenarios, which makes the procedure less complex.

The paper is organized as follows. Section 2 discusses the essence of AHP. Section 3 briefly describes diverse approaches handling uncertainty in AHP. Section 4 develops the idea of combining AHP with scenario planning. Section 5 presents a novel scenario-based AHP procedure designed for independent criteria and one-shot decisions. Section 6 contains a short illustration. Conclusions are gathered in the last section.

2 ANALYTIC HIERARCHY PROCESS – BRIEF DESCRIPTION

AHP is a well-known MCDM method. However it is worth underlining that MCDM involves two groups of problems: multiple attribute decision problems (MADP) and multiple objective decision problems (MODP). In MADP the number of decisions is precisely defined at the beginning of the decision making process and the levels of considered attributes are assigned to each option. Within the framework of MODP the cardinality of the set of possible decisions is not exactly known. The decision maker (DM) only knows the set of criteria and constraints that create the set of possible solutions [11]. AHP is designed for MADP.

The procedure consists in defining the problem hierarchy, i.e. decomposing the problem into significant criteria and decisions. Then the DM is supposed to evaluate their relative importance (i.e. priorities) within appropriate criteria and decisions pairwise comparison matrices. This step is done by transforming linguistic expressions into concrete values (from 1/9 to 9 where 1/9 denotes absolute inferiority, 1 signifies equal preference and 9 means absolute preference). After normalizing those evaluations and computing their averages (called overall priorities) for each criterion and option, a weighted value is calculated for each decision in the last step. This measure allows one to indicate the best solution. It is usually recommended to check the judgement consistency. If the matrices are not consistent enough (the consistency ratio exceeds 10%), the subjective judgements need to be revised.

One of the strengths of AHP is that it enables solving problems where criteria are difficult to quantify since in this method human judgements are sufficient. The second vital advantage is that AHP does not require declaring criteria weights – they are computed on the basis of a pairwise comparison performed by the DM. Nevertheless, drawbacks connected with AHP have been also found. One of them concerns the rank reversal phenomenon [4]. Moreover, AHP forces DMs to declare preferences for all pairs of criteria and decisions [1].

AHP was originally developed by Saaty [18], but it has been extensively studied and refined since then [4]. It is applied to such fields as business, government, education (operations research, management science), industry and healthcare. It supports for instance forecasting, management, planning and ranking [3], [15], [21].

3 ANALYTIC HIERARCHY PROCESS UNDER UNCERTAINTY

The original Analytic Hierarchy Process is a deterministic decision making tool. Nevertheless, both researchers and practitioners have tried for many years to take diverse types of uncertainty into consideration, which is entirely justified since it is usually difficult to precisely define ones preferences or predict future events.

Here are some examples. Beynon [2] suggests using DS/AHP which encompasses the Dempster-Shafer theory of evidence and gives the ability to assign probability measures to groups of decisions. Tacnet et al. [19] combine AHP with fuzzy sets, possibility and belief functions theories in order to handle imprecise and uncertain evaluations of quantitative and qualitative criteria. Uncertainty also can be taken into consideration on the basis of Monte Carlo AHP [1], [23]. Mimovic et al. [17] use Bayesian analysis to improve the accuracy of input data for AHP. Lin and Wang [14] formulate an uncertain variable method and show how to check the consistency of uncertainty comparison matrices. Ennaceur [7] describes in his doctoral thesis new uncertain AHP methods based on the belief function theory. Eskandari and Rabelo [8] suggest a stochastic approach to capture the uncertain behaviour of the global AHP weights. Yang et al. [22] adopt the normal Cloud model and the Delphi feedback method in order to handle the randomness and fuzziness of individual judgements. Other interesting uncertainty issues are discussed for instance in [4], [12], [16], [20], [21].

4 AHP COMBINED WITH SCENARIO PLANNING

There are of course numerous techniques enabling handling uncertainty in MCD analysis and optimisation such as fuzzy numbers, probabilities, probability-like quantities and explicit risk measures. However, according to Durbach and Stewart [6] uncertainties become increasingly so complex that the elicitation of those measures becomes operationally difficult for DMs to comprehend and virtually impossible to validate. In their opinion it is useful to construct scenarios describing possible ways in which the future might unfold. Thus, in this work we are mainly interested in methods integrating scenario planning (SP) into AHP.

Durbach discusses these issues in his recent contribution [5]. He emphasizes that the aggregation of criteria and scenarios may be performed in two fundamental ways: in the »meta-alternative« approach scenarios are combined with decisions and the joint »meta-alternatives« are evaluated over attributes. On the other hand, in the »meta-attribute« approach scenarios are combined with attributes and then the decisions are evaluated in terms of these »meta-attributes«. Both approaches use a standard implementation of the AHP, and thus are subject to the same concerns, for example regarding rank reversal and interpretability of weights. Both techniques are valuable, but the aforementioned paper does not explain the crucial application difference between them. Hence, it is worth underlining that the first approach is designed for independent criteria - the performance of particular targets can be analyzed totally seperately since the number (m) and type of scenarios can be different for each criterion: m_1, m_2, ..., m_k, ..., m_p, where p denotes the number of criteria. The second approach allows the criteria to be dependent. This time there is a strong relationship between scenarios assigned to particular criteria – the number and type of events ought to be the same for each criterion considered in the decision problem and evaluation a_{ij}^k only can be connected with evaluations a_{ij}^1, ..., a_{ij}^{k-1}, a_{ij}^{k+1}, ..., a_{ij}^p where a_{ij}^k describes the performance of criterion C_k by decision D_j provided that scenario S_i happens [10]. Furthermore, we would like to emphasize that the scenario-based AHP method suggested by Durbach [5] requires the evaluation of supplementary pairwise comparison matrices related to scenarios and representing for instance how likely a scenario is to occur. For p independent criteria, n decisions and m_1, m_2, ..., m_k, ..., m_p scenarios for particular criteria, at least $(p^2 + \sum_{k=1}^{p} (n \cdot m_k)^2)$ comparisons are needed. This leads to additional effort and makes the decision making process more complex and challenging. That is why, we state that it would be desirable to integrate SP in a less time-consuming way.

At the end of this section it is worth mentioning that some researchers recommend using probabilities in scenario-based AHP [13]. However Durbach [5] stresses that scenarios should not be treated as states of nature since the set of scenarios does not constitute a complete probability space – in a statistical sense scenario »likelihoods« are not probabilities. Moreover, states of nature are mutually exclusive and exhaustive, they are constructed from the same underlying dimensions, which is not the case of scenarios. Durbach [5] doen't use probabilitites in his scenario-based procedures - he applies relative importance of scenarios. In our work we also do not refer to probabilities as we focus on one-shot decisions [10].

5 SP/AHP FOR INDEPENDENT CRITERIA: SP/AHP(IC)

In this section we only investigate the problem of independent criteria. Such criteria occur when particular goals depend on totally different factors like weather, demography, diseases, fashion, prices, political decisions. We concentrate on one-shot decisions - the selected decision is supposed to be performed only once. This assumption signifies that for each criterion just one scenario has the chance to occur. In such circumstances the use of the probability calculus is unjustified [10]. Additionally, we do not intend to take into consideration all the scenarios till the end of the decision making process. Instead of it, we prefer reducing the initial sets of scenarios thanks to preferences (predictions) declared by the DM (optimism coefficients) and then choosing the final course of action on the basis of selected data. A similar approach, in the context of one-criterion problems, is applied in [9].
The proposed method – SP/AHP(IC) – consists of the following steps:

1) Define the set of decisions (D), the set of criteria (C) and the sets of scenarios separately for each criterion $S(k)$. Estimate pairwise comparison matrices for (1) criteria and (2) particular options in terms of scenarios, separately for each criterion.
2) Define for each criterion C_k the DM's coefficient of optimism β_k which belongs to the interval [0,1]. It is equal to 0 for extreme pessimists (expecting the occurrence of scenarios with the worst outcomes) and 1 for extreme optimists.

3) Normalize each value in matrices for (1) criteria and (2) particular options in terms of scenarios so that the sum of transformed comparisons in each column is equal to 1. Calculate the average of normalized values for each row of all the aforementioned matrices. These averages constitute weights (overall priorities): P_k (for criteria) and $M^{k,j}{}_i$ (for scenarios within a given criterion and decision). Normalize scenario weights $M^{k,j}{}_i$ (0 for the lowest weight, 1 for the highest weight), separately for each decision and criterion, and denote them by $M(n)^{k,j}{}_i$.
4) Reduce the initial sets $S(k)$ to sets $S(k)^r$ following the rules enumerated below:
   1. All the scenarios within a given criterion and with the normalized weight $M(n)^{k,j}{}_i$ equal to β_k create the set $S(k)^r$.
   2. If within a given criterion and decision there aren't any scenarios with $M(n)^{k,j}{}_i$ equal to β_k, set $S(k)^r$ can include scenario(s) with the closest normalized weight.
5) Estimate pairwise comparison matrices in terms of decisions only for scenarios from sets $S(k)^r$. Normalize their values applying the same way as in step 3. Compute decision weights, i.e. the averages of the normalized values for each row ($N^{k,i}{}_j$).
   1. If set $S(k)^r$ is a singleton, these averages do not need to be modified: $N^{k,i}{}_j = N^{k*}{}_j$ where $N^{k*}{}_j$ denotes final decision weights within a given criterion.
   2. If a given reduced set includes more than one scenario, compute the mean decision weights on the basis of the selected scenarios in the following way:

$$N_j^{k*} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{\left|S(k)_j^r\right|}\sum_{S(k)_j^r} N_j^{k,i}\right) \tag{1}$$

   where $S(k)^r{}_j$ denotes the subset of set $S(k)^r$. It only contains scenarios chosen for criterion C_k, but in terms of decision D_j. $|S(k)^r{}_j|$ is the cardinality of $S(k)^r{}_j$.
6) Choose the decision for which the following measure has the highest value.

$$N_j^{*} = \sum_{j=1}^{p} N_j^{k*} \cdot P_k \tag{2}$$

Note that we omit the consistency analysis in the procedure since we assume that all the matrices are consistent enough. Otherwise, an appropriate matrices transformation is required. We do not present all the equations in detail in the paper due to page limitations, but we hope that the example discussed in the next section will dispel possible doubts.

6 EXAMPLE

Let's assume that the decision problem includes 3 decisions (D_1, D_2, D_3) and 2 independent criteria (C_1, C_2). The DM takes into account 3 scenarios for the first criterion ($S^1{}_1$, $S^1{}_2$, $S^1{}_3$) and 2 scenarios for the second one ($S^2{}_1$, $S^2{}_2$). Pairwise comparison matrices for (1) criteria and (2) particular options in terms of scenarios are presented in Tables 1-3 (step 1). Let's analyse the case of a moderate pessimist who declares the following coefficient values: β_1=0.4 and β_2=0.3 (step 2). Step 3 has been already done (see the second part of Table 1 and values introduced next to the original preferences in Tables 2-3). Normalized averages $M(n)^{k,j}{}_i$ are gathered in additional columns. Now (step 4) initial sets $S(1)=\{S^1{}_1, S^1{}_2, S^1{}_3\}$ and $S(2)=\{S^2{}_1, S^2{}_2\}$ are reduced to $S(1)^r=\{S^1{}_2, S^1{}_3\}$ and $S(2)^r=\{S^2{}_1, S^2{}_2\}$. Scenario $S^1{}_2$ belongs to $S(1)^r$ since its normalized weight 0.308 for decision D_1 is the closest to 0.4. Scenario $S^1{}_3$ also belongs to $S(1)^r$ since its normalized weight 0.322 for decision D_2 is the closest to 0.4 etc. In step 5 the DM's estimations are used to calculate decision weights (Table 4), separately for each criterion and scenario. Our reduced sets are not unit sets. That is why, the use of equation (1) is recommended to obtain a single decision weight for each decisions within particular criteria: $N^{1*}{}_1$=1/2·(0.80+0.83)=0.82; $N^{1*}{}_2$=0.18; $N^{2*}{}_1$=0.61; $N^{2*}{}_2$=0.39 (values are rounded to two decimal places). After reducing the scenario sets, option D_1 gains better results than D_2 for both considered criteria. Hence, step 6 is useless in this case – the final solution is obvious: N^*_1=0.75·0.82+0.25·0.61=0.77; N^*_2=0.23. The DM should select D_1.

Table 1: Criteria comparison matrices

	Criteria comparison – original values		***Criteria comparison – normalized values***		
	C_1	C_2	C_1	C_2	A_k
C_1	1.00	3.00	0.75	0.75	0.75
C_2	0.33	1.00	0.25	0.25	0.25
Sum	1.33	4.00	1.00	1.00	

Table 2: Scenario comparison matrices for decision D_1 and D_2 within criterion C_1

	Scenario comparison for D_1					***Scenario comparison for D_2***				
	S_1	S_2	S_3	$M^{1,1}_i$	$M(n)^{1,1}_i$	S_1	S_2	S_3	$M^{1,2}_i$	$M(n)^{1,2}_i$
S_1	1.00/0.08	0.25/0.06	0.13/0.08	0.07	0.000	1.00/0.68	7.00/0.58	3.00/0.71	0.66	1.000
S_2	4.00/0.31	1.00/0.23	0.33/0.23	0.26	0.308	0.14/0,10	1.00/0.08	0.25/0.06	0.08	0.000
S_3	8.00/0.61	3.00/0.71	1.00/0.69	0.67	1.000	0.33/0.22	4.00/0.34	1.00/0.23	0.26	0.322
Sum	13.00/1.00	4.25/1.00	1.46/1.00			1.47/1.00	12.00/1.00	4.25/1.00		

Table 3: Scenario comparison matrices for decisions D_1 and D_2 within criterion C_2

	Scenario comparison for D_1				***Scenario comparison for D_2***			
	S_1	S_2	$M^{2,1}_i$	$M(n)^{2,1}_i$	S_1	S_2	$M^{2,2}_i$	$M(n)^{2,2}_i$
S_1	1.00/0.80	4.00/0.80	0.80	1.000	1.00/0.17	0.20/0.17	0.17	0.000
S_2	0.25/0.20	1.00/0.20	0.20	0.000	5.00/0.83	1.00/0.83	0.83	1.000
Sum	1.25/1.00	5.00/1.00			6.00/1.00	1.20/1.00		

Table 4: Decision comparison matrices for scenarios $S^1{}_2$, $S^1{}_3$ (within C_1) and $S^2{}_1$, $S^2{}_2$ (within C_2)

	Decision comp. for $S^1{}_2$			***Decision comp. for $S^1{}_3$***			***Decision comp. for $S^2{}_1$***			***Decision comp. for $S^2{}_2$***		
	D_1	D_2	$N^{1,2}_j$	D_1	D_2	$N^{1,3}_j$	D_1	D_2	$N^{2,1}_j$	D_1	D_2	$N^{2,2}_j$
D_1	1.00/0.80	4.00/0.80	0.80	1.00/0.83	5.00/0.83	0.83	1.00/0.89	8.00/0.89	0.89	1.00/0.33	0.50/0.33	0.33
D_2	0.25/0.20	1.00/0.20	0.20	0.20/0.17	1.00/0.17	0.17	0.13/0.11	1.00/0.11	0.11	2.00/0.67	1.00/0.67	0.67
Sum	1.25/1.00	5.00/1.00		1.20/1.00	6.00/1.00		1.13/1.00	9.00/1.00		3.00/1.00	1.50/1.00	

7 CONCLUSIONS

The advantages of SP/AHP(CI) are as follows: 1) it allows handling uncertainty by means of scenario planning – a relatively simple and well-known tool; 2) the scenarios sets reduction enables one to reduce the number of relative importance estimations to $(p^2 + \sum_{k=1}^{p}(m_k^2 \cdot n) + |S(k)^r| \cdot n^2)$; 3) the approach can be applied to different DMs (optimist, pessimist, moderate); 4) it does not require probability estimation; 5) it is useful for both quantitative and qualitative criteria. It may seem to be controversial due to the reduction of the number of scenarios, but such an approach partially considers the one-shot character of decisions and the fact that scenarios are not conscious opponents who alter their strategies depending on the outcomes, see [9]. SP/AHP(CI) is only designed for independent criteria. Therefore, it would be desirable to create in the future an analogous procedure for dependent criteria.

Acknowledgement
This work was supported by the National Science Center, Poland [grant number 2014/15/D/HS4/00771].

References

[1] Banuelas, R., Antony, J. 2007. Modified analytic hierarchy process to incorporate uncertainty and managerial aspects. International Journal of Production Research, 42(18): 3851–3872.

[2] Beynon, M. 2002. DS/AHP method: A mathematical analysis, including an understanding of uncertainty. European Journal of Operational Research, 140: 148-164.

[3] Biloslavo, R., Dolinsek, S. 2010. Scenario planning for climate strategies development by integrating group Delphi, AHP and dynamic fuzzy cognitive maps. Foresight, 12(2): 38-48.

[4] Brunelli, M. 2015. Introduction to the Analytic Hierarchy Process. SpringerBriefs in Operations Research. P. 83. Springer.

[5] Durbach, I.N. 2019. Scenario planning in the analytic hierarchy process. Wiley.

[6] Durbach, I.N., Stewart, T.J. 2012. Modeling uncertainty in multi-criteria decision analysis. European Journal of Operational Research, 223(1): 1-14.

[7] Ennaceur, A. 2015. New AHP methods for handling uncertainty within the belief function theory. Doctoral thesis. Université de Tunis, Institut Supérieur de Gestion.

[8] Eskandari, H., Rabelo, L. (2007). Handling uncertainty in the analytic hierarchy process: a stochastic approach. International Journal of Information Technology and Decision Making, 6(1): 177-189.

[9] Gaspars-Wieloch, H. 2015. On a decision rule supported by a forecasting stage based on the decision maker's coefficient of optimism. Central European Journal of Operations Research, 23: 579-594.

[10] Gaspars-Wieloch, H. 2017. A decision rule for uncertain multi-criteria pure decision making and independent criteria. Optimum. Economic Studies, 3(87): 77-92.

[11] Hwang, C.L., Yoon, K. 1981. Multiple attribute decision making, methods and applications. Lecture Notes in Economics and Mathematical Systems, 186. New York: Springer-Verlag.

[12] Lafleur, J.M. 2011. Probabilistic AHP and TOPSIS for multiattribute decision making under uncertainty. Aerospace Conference, 10.1109/AERO.2011.5747655

[13] Levary, R., Wan, K. 1999. An analytic hierarchy process based simulation model for entry mode decision regarding foreign dirrect investment. Omega, 27(6): 661–677.

[14] Lin, L., Wang, C. 2012. On consistency and ranking of alternatives in uncertain AHP. Natural Science, 4: 340-348.

[15] Longaray, A.A, Gois, J.D.R., Munhoz, P.R. (2015). Proposal for using AHP method to evaluate the quality of services provided by outsourced companies. Procedia Computer Science, 55: 715-724.

[16] Millet, I., Wedley, W.C. 2002. Modelling risk and uncertainty with the analytic hierarchy process. Journal of Multi-criteria Decision Analysis. Optimization, Learning and Decision Support, 11(2): 97-107.

[17] Mimovic, P., Stankovic, J., Jankovic Milic, V. 2015. Decision-making under uncertainty – the integrated approach of the AHP and Bayesian analysis. Economic Research-Ekonomska Istraživanja, 28(1): 868-878.

[18] Saaty, T.L. 1980. The analytic hierarchy process. New York: McGraw-Hill.

[19] Tacnet, J.M., Dezert, J., Batton-Hubert, M. 2011. AHP and uncertainty theories for decision making using the ER-MCDA methodology. Conference Paper. : https://www.researchgate.net/publication/257211032

[20] Toth, W., Vacik, H. 2018. A comprehensive uncertainty analysis of the analytic hierarchy process methodology applied in the context of environmental decision making. Journal of Multicriteria Decision Analysis, 25: 142-161.

[21] Wu, D., Yang, Z., Wang, N., Li, C., Yang, Y. 2018. An integrated multi-criteria decision making model and AHP weighting uncertainty analysis for sustainability assessment of coal fired power units. Sustainability, 10(6): 1700.

[22] Yang, X., Yan, L., Zeng, L. 2013. How to handle uncertainties in AHP: The Cloud Delphi hierarchical analysis. Information Sciences, 222: 384-404.

[23] Yaragi, N., Tabesh, P., Guan, P., Zhuang, J. 2015. Comparison of AHP and Monte Carlo AHP under different levels of uncertainty. IEEE Transactions on Engineering Management, 62(1).

CONSISTENCY OF ASSESSMENTS AND REVERSAL OF THE RANKING IN MULTI-CRITERIA DECISION MAKING

Sławomir Jarek
University of Economics in Katowice, Department of Operations Research
40-287 Katowice , ul. 1 Maja 50, Poland
slawomir.jarek@ue.katowice.pl

Abstract: In many discrete multicriteria methods, the problem arises of determining the ranking of decision variants and determining one or several of the best variants on its basis. While determining this ranking, it is possible to use methods based on Pairwise Comparisons (PC). Unfortunately, the practice shows the existence of the problem of disturbing the obtained ranking after adding or removing selected decision variants. Initially, this problem was associated only with AHP, however, further research has shown that this is a typical problem appearing in methods based on pairwise and pattern comparison and characteristic for commonly used multicriteria methods such as like PROMETHE, TOPSIS, ELECTRE. This problem is of great practical importance, because with small modifications to the initial conditions of the problem of decision making, it may happen that completely different final rankings will be obtained. This problem appears particularly clearly in the process of studying preferences of decisions based on AHP and ANP. The research showed that the reversal of the ranking could occur after adding or removing selected decision variants or as a result of imprecise description of the decision-maker's preferences (inconsistent PC Matrix). The work focuses on the impact of the coherence of assessments expressed in the PCM on the reversal of the ranking. The influence of the missing rating in the PC process on coherence of preferences expressed by the decision maker was also examined.

Keywords: MCDM, AHP, PROMETHE, TOPSIS, ELECTRE, PCM

APPLICATION OF PAGERANK CENTRALITY IN MULTI-CRITERIA DECISION MAKING

Nikola Kadoić
University of Zagreb, Faculty of organization and informatics, Pavlinska 2, HR-42000 Varaždin
E-mail: nkadoic@foi.hr

Nina Begičević Ređep
University of Zagreb, Faculty of organization and informatics, Pavlinska 2, HR-42000 Varaždin
E-mail: nbegicev@foi.hr

Blaženka Divjak
University of Zagreb, Faculty of organization and informatics, Pavlinska 2, HR-42000 Varaždin
E-mail: bdivjak@foi.hr

Abstract: Multicriterial analysis is a highly developed area since there is a large number of multi-criteria decision-making (MCDM) methods. The multi-criteria analysis enables a precise problem analysis and ensures the rationality of decision that is made. However, all problem analysis using different methods can give different results (decision), so it is important to recognize which MCDM method is appropriate for a particular situation. There are MCDM methods by using which we can model dependencies and influences between the criteria in decision-making problem. One of the most used MCDM methods that are used in terms of problem analysis is the analytic network process (ANP). Previous researches discussed some problems related to using the ANP in decision-making. As a solution to those problems using the PageRank centrality can be considered. In this paper, we are presenting several possibilities of applying the PageRank centrality for multi-criteria analysis. Presented possibilities are compared and discussed. As a result, using the weighted PageRank centrality is proposed as the optimal solution for multi-criteria analysis when dependencies (influences) between the criteria are examined.

Keywords: criteria, multi-criteria decision-making, MCDM, ANP, PageRank centrality, PageRank, influences, dependencies

1 INTRODUCTION

There are many methods that can be used in terms of multi-criteria analysis. Each of them models the problem differently, but with the main goal – to find an optimal solution to a problem that has been analysed. In this paper, we are analysing decision methods from the perspective of modelling the influences (dependencies) between the criteria. Those two concepts have the opposite meaning [1]: if the first criterion influences the second criterion, then the second criterion depends on the first criterion.

This paper is motivated with the research in the scope of the project "Development of a methodological framework for strategic decision-making in higher education – a case of open and distance learning (ODL) implementation." As a part of the research on the project, different MCDM methods were analysed from the position of applicability in the area of higher education. It is concluded that the area of higher education is characterized by the existence of influences between the criteria [2]. However, literature review analysis resulted with the conclusion that, in the analysis of MCDM problems in the area of higher education, methods which do not support modelling influences (dependencies) between the criteria (such as analytic hierarchy process (AHP)) are much more often used instead of methods which support this feature. In this project, special attention is given to modelling influences between the criteria and analysis of the method analytic network process (ANP). The ANP is the most often used method for modelling the influences (dependencies) between the criteria. However, it has many disadvantages, and this is the reason for such literature review results.

In the second section of this paper, we will shortly present the ANP method and discuss some of its characteristics. In the third section, we will present several types of PageRank centrality and its possibility for using in terms of multi-criteria analysis. Finally, we will discuss and compare the presented types of PageRank centrality and list the advantages of using PageRank comparing to the ANP.

2 THE ANALYTIC NETWORK PROCESS (ANP)

We presented the decision-making process using the ANP in our last SOR paper [3], and CJOR paper [4] which followed the SOR paper, and our further analysis will be demonstrated on the decision-making problem that is discussed in those papers. The problem is related to the evaluation of senior researchers (scientists). Senior researchers are active in both the research and teaching fields. In this analysis, we will not include alternatives. The decision-making problem is presented in Figure 1.

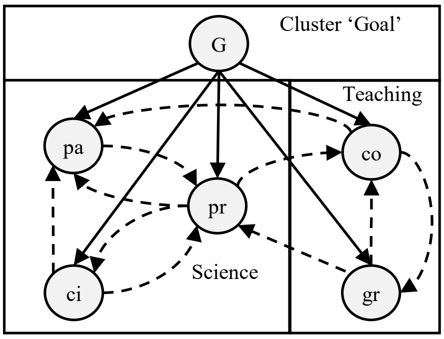


Figure 1: Network Structure of the problem evaluation of the scientists

The steps in ANP (also adapted from [5], [6]):

- Problem structuring phase: it is related to the creation of the network structure. It is presented in Figure 1. The model consists of five criteria (papers, pa; projects, pr; citations, ci; courseware, co; grades from students, gr) which are grouped into two clusters (Teaching and Science). The model also includes the decision-making goal (node G in cluster Goal). The arrows between the elements represent dependencies in the model.
- Pairwise comparisons procedure and creating the weighted supermatrix (Table 2):
  - Comparing criteria in order to reach unweighted supermatrix (Table 1),
  - Comparing clusters in order to reach clusters' weights which are needed to obtain the weighted supermatrix. In this example, we decided that all clusters are equally important,
  - Combining the unweighted supermatrix with clusters' weights. (Much more detailed procedure description is available in our previous papers [3], [4].)
- Creation the limit matrix (Table 3) by multiplying the weighted supermatrix with itself until it converges.

Table 1: Unweighted Supermatrix

	G	co	gr	pa	ci	pr
G	0	0	0	0	0	0
co	0.33	0	1	0	0	1
gr	0.67	1	0	0	0	0
pa	0.25	1	0	0	0.4	0.6
ci	0.25	0	0	0	0	0.4
pr	0.5	0	1	1	0.6	0

Table 2: Weighted Supermatrix

	G	co	gr	pa	ci	pr
G	0	0	0	0	0	0
co	0.08	0	0.5	0	0	0.5
gr	0.17	0.5	0	0	0	0
pa	0.186	0.5	0	0	0.4	0.3
ci	0.186	0	0	0	0	0.2
pr	0.383	0	0.5	1	0.6	0

Table 3: Limit matrix

	G	co	gr	pa	ci	pr
G	0.00	0.00	0.00	0.00	0.00	0.00
co	0.23	0.23	0.23	0.23	0.23	0.23
gr	0.11	0.11	0.11	0.11	0.11	0.11
pa	0.24	0.24	0.24	0.24	0.24	0.24
ci	0.07	0.07	0.07	0.07	0.07	0.07
pr	0.34	0.34	0.34	0.34	0.34	0.34

The final priorities can be found in any column of the limit matrix.

The ANP is much deeper analysed in paper [7]. In the paper, a list of weak points of the ANP is provided. Those characteristics result from the fact that ANP is much less used than it should be used. Most of them influence the complexity of the ANP implementation, misunderstanding of certain steps of the method, and long duration of the implementation process [1]. However, the most exciting three characteristics are [7]:

- The inseparability of the criteria and alternatives. In some decision-making problems, if there is no directed connection between any of two nodes, then it is possible that at least some nodes will weight 0.0, or even the whole limit matrix is zero-matrix. This is not the case in this decision-making problem.
- The influence of the goal node on the priorities. The interesting and slightly intriguing characteristic of the ANP is the fact that the priorities with respect to the goal (first column in unweighted and weighted supermatrix) do not influence the finale priorities in the limit matrix. So, if we change the numbers in the first column (respecting that the sum of the numbers equals 1), the finale priorities will remain the same. This means that the node goal is not necessary for the model and that only the dependencies between the elements determine the final priorities. Indeed, there is a large number of papers which do not include goal as an element of the network structure. However, the goal node is theoretically defined as a network element (cluster), and it is a necessary element in the AHP, which is a 'weaker' variant of the ANP.
- The stochasticity of the supermatrix in the ANP. The most interesting and the most intriguing characteristic of the ANP is related to the stochasticity of the supermatrix in the ANP. If we look at the connection between the pa and pr, in both, unweighted and weighted supermatrix, we do not know how strong this connection is. The element pr can influence the element pa weakly, strongly or very strongly. So, independently of the intensity of this influence, the final priorities will remain the same. In the DEMATEL method [8], we use scale 0-4 to describe the intensity (level) of the influence (dependency) between two elements. For all real values between 0 and 4, in this case, we will reach the same final priorities. Let us say that there are four elements in the model; each influence any other (not itself). All elements influence the first element with intensity 4, the second element with intensity 3, the third element with intensity 3, and finally the third element with intensity 1. We got the situation that the first element depends on mostly by others, and the last element depends at least by others. However, in ANP, all elements will have the same priority. The reason for that is 'forcing' the stochasticity of supermatrix. It relativizes the problem, and the solution (the decision) might not be optimal.

3 THE PAGERANK CENTRALITY

The PageRank centrality is a special type of eigenvalue centrality. The eigenvalue centrality for undirected and unweighted networks is calculated using Equation 1 [9].

$$\boldsymbol{C_E(i)} = \frac{1}{\lambda}\sum_{j\in M(i)} \boldsymbol{C_E(j)} = \frac{1}{\lambda}\sum_{j\in N} \boldsymbol{a_{ij} C_E(j)} \quad (1)$$

where $\boldsymbol{M(i)}$ is a set of neighbours of actor $\boldsymbol{i}$, $\boldsymbol{\lambda}$ is a constant (the maximum eigenvalue) and $\boldsymbol{a_{ij}}$ is an element of a matrix of neighbours $\boldsymbol{A}$. PageRank centrality is used for directed networks, and there are variants of this measure in terms of weighted and unweighted graphs.

The PageRank centrality can be calculated using the iterative procedure [10] or using Equation 2.

$$\lim_{k\to\infty} \boldsymbol{A^k Z_0} = \widetilde{\boldsymbol{A}} \tag{2}$$

where $\boldsymbol{A}$ is the matrix of neighbours, $\boldsymbol{Z_0}$ is a one-column matrix which contains elements $\frac{\mathbf{1}}{\boldsymbol{N}}$, and $\widetilde{\boldsymbol{A}}$ is a matrix of priorities.

In terms of decision-making with the ANP, matrix $\boldsymbol{A}$ can correlate with weighted supermatrix. Additionally, we can create weighted supermatrix avoiding the pairwise procedure on the node level as described in the paper [4]:

- the starting point is the identification of the intensities of influences between the elements in the network (Table 4),
- then, that matrix can be stochastically normalized using the normalization by sum (Table 6) or transition matrix (function).

Table 4: Matrix of influences intensities between the criteria

	co	gr	pa	ci	pr
co	0	3	0	0	2
gr	2	0	0	0	0
pa	2	0	0	2	3
ci	0	0	0	0	2
pr	0	3	4	3	0

Table 5: (Un)weighted supermatrix

	co	gr	pa	ci	pr
co	0	0.5	0	0	0.5
gr	0.5	0	0	0	0
pa	0.5	0	0	0.4	0.3
ci	0	0	0	0	0.2
pr	0	0.5	1	0.6	0

The problem that appears with the powering the supermatrix is already mentioned earlier when three the most interesting characteristics of the ANP were listed. The solution to that problem, the PageRank calculates the new matrix as in Equation 3 [11], [12].

$$\boldsymbol{G} = \boldsymbol{\alpha} \cdot \boldsymbol{A} + (\mathbf{1} - \boldsymbol{\alpha}) \cdot \boldsymbol{E} \tag{3}$$

In most cases, $\boldsymbol{\alpha} = \mathbf{0.85}$ [13]. (Note: If a certain column in $\boldsymbol{A}$ contains all 0, then this column has to be replaced with ca column whose values equal $\frac{\mathbf{1}}{\boldsymbol{N}}$.)

Adding the $\boldsymbol{E}$ in Equation 3 ensures that original matrix $\boldsymbol{A}$ converges to the non-zero matrix, and now it is no longer possible that we cannot calculate the global priorities. The role of $\boldsymbol{E}$ is making a matrix ($\boldsymbol{G}$) whose graph is strongly connected – there is a direct connection between any two nodes in $\boldsymbol{G}$. Additionally, the influence of $\boldsymbol{G}$ on the final priorities is negligible ($\mathbf{0.15}$). This is the first possible application of PageRank to eliminate at least one of the weak points of the ANP (the inseparability of the criteria and alternatives).

The PageRank centrality can be also interesting in terms of eliminating the issues that are the result of the stochasticity of the supermatrix in the ANP [14]. Then, the original PageRank centrality algorithm should be changed in a way that we sum powers of the non-stochastic supermatrix and then aggregate and normalize the results. The sums of the columns of the non-stochastic supermatrix should be less than 1 because - in only that case, it is possible to sum all the powers (using Equation 4).

$$\widetilde{\boldsymbol{A}} = \sum_{\boldsymbol{k}\to\infty} \boldsymbol{A^k} = \boldsymbol{A} \cdot (\boldsymbol{A} - \boldsymbol{I})^{-\mathbf{1}} \tag{4}$$

Consequently, the original PageRank for directed and weighted graphs (matrices) is transformed as follows:

1. The starting point is a matrix of influences between the criteria (Table 4)
2. In the second step, we are dividing each value in Table 4 with the maximum sum of columns, which is increased by 1. The maximum sum is in column pr and equals 7, which means that all values in Table 4 will be divided by 8. The result is presented in Table 6.
3. Respecting the Equation 3, we have to calculate the matrix $\boldsymbol{I} - \boldsymbol{A}$. The result is presented in Table 7.

4. Now we calculate the inverse of matrix $\boldsymbol{I}-\boldsymbol{A}$. (Table 8)
5. Multiplication of Tables 7 and 8 (Table 9)
6. Calculation of the sum of rows (ΣR) and columns (ΣC) of Table 9 and their difference, $\boldsymbol{d}$ (see Table 9). The difference should then be normalized. There are several ways to do it: using the absolute value of the smallest value (difference), or any other higher number. Increasing the normalization value will result in smaller differences between the priorities on end. When normalization value, $\boldsymbol{n}$, is chosen, it should be added to differences, $\boldsymbol{d}+\boldsymbol{n}$. Now, all values are positive, and it is possible to calculate the criteria weights (normalization by sum).
   In this example, we chose the normalization value as differences between the highest difference and the lowest difference.

Table 6: Step 2

	co	gr	pa	ci	pr
co	0	0.375	0	0	0.25
gr	0.25	0	0	0	0
pa	0.25	0	0	0.25	0.375
ci	0	0	0	0	0.25
pr	0	0.375	0.5	0.375	0

Table 7: Step 3

	co	gr	pa	ci	pr
co	1	-0.375	0	0	-0.25
gr	-0.25	1	0	0	0
pa	-0.25	0	1	-0.25	-0.375
ci	0	0	0	1	-0.25
pr	0	-0.375	-0.5	-0.375	1

Table 8: Step 4

	co	gr	pa	ci	pr
co	1.21	0.62	0.22	0.22	0.44
gr	0.30	1.15	0.05	0.05	0.11
pa	0.47	0.48	1.40	0.65	0.81
ci	0.10	0.19	0.20	1.20	0.40
pr	0.38	0.74	0.80	0.80	1.59

Table 9: Steps 5 and 6

	co	gr	pa	ci	pr	ΣR		ΣC	d	$d+n$	priorities
co	0.21	0.62	0.22	0.22	0.44	1.71		1.46	0.24	2.89	0.22
gr	0.30	0.15	0.05	0.05	0.11	0.68		2.18	-1.50	1.14	0.09
pa	0.47	0.48	0.40	0.65	0.81	2.82		1.68	1.14	3.78	0.29
ci	0.10	0.19	0.20	0.20	0.40	1.08		1.93	-0.85	1.80	0.14
pr	0.38	0.74	0.80	0.80	0.59	3.32		2.35	0.97	3.61	0.27
ΣC	1.46	2.18	1.68	1.93	2.35		highest		1.14	13.22	
							lowers		-1.50		
							norm. value. n		2.64		

If we compare the final priorities with the priorities in Table 3 (from the Limit matrix), we can identify some differences. Even though the ranks of the criteria remained the same, there are absolute differences between the criteria weights. Now, the problem with stochasticity of the supermatrix has been eliminated.

4 CONCLUSIONS

In this paper, we were dealing with the possibilities to use the PageRank centrality to diminish some of the weak points of the method ANP. Using the PageRank centrality in the process of calculating the limit matrix (from the weighted supermatrix), we can directly influence and eliminate the weak point of the ANP related to the converging to zero matrix - inseparability the criteria and alternatives. Indeed, there are real-world requests to calculate the criteria weights when the alternatives are still not known. In those cases, very often, when there is a small number of the connections in the model, some of the criteria, or even all would weight 0.0. If we found those criteria irrelevant; we will not put them into the model at all – so we cannot accept 0.0. as the final criteria weight of certain criteria. PageRank solves this situation.

The other benefit of the PageRank centrality is related to the dealing with stochasticity in supermatrix in ANP. When the matrix is stochastic, it is ensured that it will converge into the limit matrix from which we can directly take the criteria weights. However, we should not force the stochasticity of the supermatrix just because there is a 'great' mathematical property of stochastic matrix in terms of its powering. If we want to use the PageRank approach, which is not stochastic, it is important to have the original matrix of the intensities of the influences between the criteria – not pairwise comparisons priorities. This is not a problem since the pairwise comparisons are also resulting from those intensities between the elements. Additionally, in this approach, when we use original intensities of the influences and avoid making the pairwise comparisons, we lower some other ANP weak points.

The only open issue in terms of ANP characteristics is related to the influence of the goal on the criteria weights. In ANP, and presented approach, criteria weights are consequences of influences between the criteria, not consequences of their importance with respect to the goal, too. To solve those issues, we can use possible aggregate the obtained results with the AHP results by using ex. arithmetic mean.

References

[1] N. Kadoić, N. Begičević Ređep, and B. Divjak, "A new method for strategic decision-making in higher education," *Central European Journal of Operations Research*, no. Special Issue of Croatian Operational Research Society and Collaborators, Oct. 2017.

[2] N. Kadoić, N. Begičević Ređep, and B. Divjak, "E-learning decision making: methods and methodologies," in *Re-Imagining Learning Scenarios*, 2016, vol. CONFERENCE, no. June, p. 24.

[3] N. Kadoić, N. Begičević Ređep, and B. Divjak, "Decision Making with the Analytic Network Process," in *SOR 17 Proceedings*, 2017, pp. 180–186.

[4] N. Kadoić, B. Divjak, and N. Begičević Ređep, "Integrating the DEMATEL with the analytic network process for effective decision-making," *Central European Journal of Operations Research*, vol. 27, no. 3, pp. 653–678, Sep. 2019.

[5] T. L. Saaty and L. G. Vargas, *Decision Making with the Analytic Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks*. Springer; Softcover reprint of hardcover 1st ed. 2006 edition (December 28, 2009), 2006.

[6] T. L. Saaty and L. G. Vargas, "The Analytic Network Process," in *Iranian journal of operational research*, vol. 1, no. 1, 2013, pp. 1–28.

[7] N. Kadoić, "Characteristics of the Analytic Network Process, a Multi-Criteria Decision-Making Method," *Croatian Operational Research Review*, vol. 9, no. 2, pp. 235–244, 2018.

[8] N. Kadoić, N. Begičević Ređep, and B. Divjak, "Structuring e-Learning Multi-Criteria Decision Making Problems," in *Proceedings of 40th Jubilee International Convention, MIPRO 2017*, 2017, pp. 811–817.

[9] M. Mincer and E. Niewiadomska-Szynkiewicz, "Application of social network analysis to the investigation of interpersonal connections," *Journal of Telecommunications and Information Technology*, vol. 2012, no. 2, pp. 83–91, 2012.

[10] D. Munđar and D. Horvat, "Rangiranje ekipa i prognoziranje ishoda u rukometu korištenjem PageRank algoritma," *Poučak*, vol. 67, pp. 8–15, 2016.

[11] T. Csendes and E. Antal, "PageRank Based Network Algorithms for Weighted Graphs with Applications to," vol. 2, pp. 209–216, 2010.

[12] T. Kumari, A. Gupta, and A. Dixit, "Comparative Study of Page Rank and Weighted Page Rank Algorithm," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 2, pp. 2929–2937, 2014.

[13] W. Xing and A. Ghorbani, "Weighted PageRank algorithm," *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305–314, 2004.

[14] N. Kadoić, "Nova metoda za analizu složenih problema odlučivanja temeljena na analitičkom mrežnom procesu i analizi društvenih mreža," University of Zagreb, 2018.

BIPOLAR SORTING AND RANKING OF MULTISTAGE ALTERNATIVES

Tadeusz Trzaskalik
Department of Operations Research, University of Economics in Katowice,
ul. 1 Maja 50, 40-287 Katowice, Poland
tadeusz.trzaskalik@ue.katowice.pl

Abstract: Bipolar is one of the multiple criteria decision analysis methods, based on the concept of bipolar reference objectives, proposed by Ewa Konarzewska-Gubała. The essence of the analysis in the Bipolar method consists in a fact that alternatives are not compared directly to each other, but they are confronted to the two sets of reference objects: desirable and non-acceptable. In the paper, a new version of the method applicable to multistage decision processes is described. Multistage alternatives are sorted and ranked according to the stage comparisons to the elements of stage reference sets A numerical example illustrates the proposed approach.

Keywords: multistage decision processes, multiple criteria decision analysis, Bipolar method, sets of reference objectives.

1 INTRODUCTION

One of the multiple criteria decision analysis methods is Bipolar, proposed by Ewa Konarzewska-Gubała [2, 3, 4, 5, 6]. In the Bipolar method alternatives are not compared directly to each other, but by means of two sets of reference objects: desirable ("good") and non-acceptable ("bad"). These two disjoint sets form the bipolar reference system. It is assumed that the decision maker applying the Bipolar method in practice, on the basis of her/his experience, opinions gathered and studies undertaken, is able to create such a system. The Bipolar method has been used many times in practical applications. Theoretical aspects of the method were also analyzed in [8, 9,10] and in papers prepared by other authors.

The classical Bipolar approach consists of three phases. In the first phase of the procedure we compare alternatives to reference objects and create outranking indicators. In the second phase position of each alternative in relation to the bipolar reference system is established, applying success achievement degree and failure avoidance degree. In the third phase, considering jointly evaluation of success achievement degree and failure avoidance degree, three categories of alternatives: B1, B2 and B3 are defined. These categories are defined in such a way, that each alternative from the Category B1 is preferred to any alternative from Category B2 and each alternative from Category B2 is preferred to any alternative from Category B3. Linear order is given in each category.

In the present paper the extension of the Bipolar approach to multistage decision processes is considered. A description of the proposed procedure is given Steps in the first and the second stage of the proposed procedure are very similar to the classical Bipolar approach.

The early version of the Bipolar approach to multistage decision processes can be found in [7]. Bellman's optimality equations [1] were employed to find the best solutions. Since that time this direction of research has not been continued. The aim of the present paper is to refresh the idea and systematic description of the proposed approach in the multistage case, what has not been done yet.

The paper consists of 5 sections. After introduction, in section 2 assumptions and notation are presented. In Section 3 Bipolar procedure to multistage decision processes is proposed and systematically presented. In section 4 illustrative numerical example can be found. Some concluding remarks are given in Section 5.

2 ASSUMPTIONS AND NOTATION

We will consider a finite, discrete multistage decision process. Let us denote:

T – number of periods of the process,
t – number of the considered stages of the process (t = 1,…,T),
y_t – a feasible state of the process of the beginning of the stage t,
$\mathbf{Y}_t$ – the set of feasible states at the beginning of stage t,
x_t – a feasible decision for the state t,
$\mathbf{X}_t(y_t)$ – the seto of feasible decisions at the beginning of stage t for the state y_t
Ω_t – transition function for the stage t. We have: $y_{t+1}= x_t$,
$\mathbf{a}_t$ – feasible realization for the stage t ($\mathbf{a}_t= (y_t, y_{t+1})$) – stage alternative
$\mathbf{A}_t$ – the set of feasible realizations of the process for the period t,
$\mathbf{a}$ – feasible realization of the whole process ($\mathbf{a} = (\mathbf{a}_1,\ldots,\mathbf{a}_T) = (y_1,\ldots,y_{T+1})$) – multistage alternative .
$\mathbf{A}$ – the set of all feasible realizations of the process,
K – the number of the all considered criteria (k=1,…,K)
C_t^k – criterion number k in stage t.

It is assumed, that at the each stage of the process t there is given: a set of stage criteria functions $\mathbf{F}_t = \{f_t^1,\ldots,f_t^K\}$, where $f_t^k : \mathbf{A}_t \rightarrow \boldsymbol{K}^k$ for k=1, ..., K, and $\boldsymbol{K}_k$ is a cardinal, ordinal or binary scale.

Criteria are defined in such a way that higher values are preferred to lower values[1]. For each stage and each criterion the decision maker establishes weight w_t^k of relative importance (it is assumed, that $\sum_{k=1}^K w_t^k = 1$ and $w_t^k \geq 0$ for each k=1, ...,K), equivalence threshold q_t^k and veto threshold v_t^{k}[2]. The decision maker also establishes minimal criteria values concordance level s as the outranking threshold ($0.5 \leq s \leq 1$)[3].

The decision maker establishes for t=1,…T reference systems $\mathbf{R}_t = \mathbf{G}_t \cup \mathbf{B}_t$, which consist of the set of „good" objects $\mathbf{G}_t$ and the set of "bad" objects $\mathbf{B}_t$. We assume, that $\mathbf{G}_t \cap \mathbf{B}_t = \varnothing$. It is also assumed that for all t=1,…,T, k=1,...,K and $\mathbf{r}_t^k \in \mathbf{R}_t$ values $f_t^k(\mathbf{r}_t^k)$ are known.

3 DESCRIPTION OF THE MULTISTAGE PROCEDURE

3.1 Comparison of stage alternatives to stage reference objects

For the pair ($\mathbf{a}_t$, r_t), where $\mathbf{a}_t \in \mathbf{A}_t$, $\mathbf{r}_t \in \mathbf{R}_t$, the following values:

$$c_t^+(\mathbf{a}_t,\mathbf{r}_t) = \sum_{k=1}^n w_t^k \varphi_t^{k+}(\mathbf{a}_t,\mathbf{r}_t) \quad \text{where} \quad \varphi_t^{k+}(\mathbf{a}_t,\mathbf{r}_t) = \begin{cases} 1, & \text{if } f_t^k(\mathbf{a}_t) - f_t^k(\mathbf{r}_t) > q_t^k \\ 0, & \text{otherwise} \end{cases}$$

$$c_t^-(\mathbf{a}_t,\mathbf{r}_t) = \sum_{k=1}^n w_t^k \varphi_t^{k-}(\mathbf{a}_t,\mathbf{r}_t) \quad \text{where} \quad \varphi_t^{k-}(\mathbf{a}_t,\mathbf{r}_t) = \begin{cases} 1, & \text{if } f_t^k(\mathbf{r}_t) - f_t^k(\mathbf{a}_t) > q_t^k \\ 0, & \text{otherwise} \end{cases}$$

$$c_t^=(\mathbf{a}_t,\mathbf{r}_t) = \sum_{k=1}^n w_t^k \varphi_t^{k=}(\mathbf{a}_t,\mathbf{r}_t) \quad \text{where} \quad \varphi_t^{k=}(\mathbf{a}_t,\mathbf{r}_t) = \begin{cases} 1, & \text{if } \left|f_t^k(\mathbf{r}_t) - f_t^k(\mathbf{a}_t)\right| \leq q_t^k \\ 0, & \text{otherwise} \end{cases}$$

are calculated.

Case 1: $c_t^+(\mathbf{a}_t, \mathbf{r}_t) > c_t^-(\mathbf{a}_t, \mathbf{r}_t)$. Stage indicators are defined as follows:

$$d_t^+(\mathbf{a}_t, \mathbf{r}_t) = c_t^+(\mathbf{a}_t, \mathbf{r}_t) + c_t^=(\mathbf{a}_t, \mathbf{r}_t), \qquad d_t^-(\mathbf{a}_t, \mathbf{r}_t) = 0$$

---

[1] It is possible to transform remaining types of criteria to the form used here.
[2] In the version of the procedure presented in the paper we assume that all q_t^k and v_t^k are equal to 0.
[3] In the version of the procedure presented in the paper we assume, that s = 0,5.

Case 2: $c_t^+(\mathbf{a}_t, \mathbf{r}_t) < c_t^-(\mathbf{a}_t, \mathbf{r}_t)$. Stage indicators are defined as follows:

$$d_t^+(\mathbf{a}_t, \mathbf{r}_t) = 0, \qquad d_t^-(\mathbf{a}_t, \mathbf{r}_t) = c_t^-(\mathbf{a}_t, \mathbf{r}_t) + c_t^=(\mathbf{a}, \mathbf{r}_t)$$

Case 3: $c_t^+(\mathbf{a}_t, \mathbf{r}_t) = c_t^-(\mathbf{a}_t, \mathbf{r}_t)$. Stage indicators are defined as follows:

$$d_t^+(\mathbf{a}_t, \mathbf{r}_t) = c_t^+(\mathbf{a}_t, \mathbf{r}_t) + c_t^=(\mathbf{a}_t, \mathbf{r}_t) \qquad d_t^-(\mathbf{a}_t, \mathbf{r}_t) = c^-(\mathbf{a}_t, \mathbf{r}_t) + c^=(\mathbf{a}_t, \mathbf{r}_t)$$

By means of outranking indicators two stage relationships: preference L_t and indifference I_t are defined as follows:

$$\mathbf{a}_t \ \mathrm{L}_t\, \mathbf{r}_t \ \text{ iff } \ d_t^+(\mathbf{a}_t, \mathbf{r}_t) > \mathrm{s} \ \wedge d_t^-(\mathbf{a}_t, \mathbf{r}_t) = 0$$
$$\mathbf{r}_t \ \mathrm{L}_t\, \mathbf{a}_t \ \text{ iff } \ d_t^+(\mathbf{a}_t, \mathbf{r}_t) = 0 \ \wedge d^-(\mathbf{a}_t, \mathbf{r}_t) > \mathrm{s}$$
$$\mathbf{a}_t \ \mathrm{I}_t \ \mathbf{r}_t \text{ iff } d^+(\mathbf{a}_t, \mathbf{r}_t) > \mathrm{s} \ \wedge d^-(\mathbf{a}_t, \mathbf{r}_t) > \mathrm{s}$$

3.2 Position of an stage alternative in relation to the bipolar reference system

For a given $\mathbf{a}_t \in \mathcal{A}_t$ auxiliary sets of indices are defined as follows:

$$\mathcal{L}_t(\mathbf{a}_t, \mathbf{G}_t) = \{h\text{: } \mathbf{a}_t\, \mathrm{L}_t\, \mathbf{g_t}^{(h)},\ \mathbf{g}_t^{(h)} \in \mathbf{G_t}\}$$
$$\mathcal{I}_t(\mathbf{a}_t, \mathbf{G}_t) = \{h\text{: } \mathbf{a}_t\, \mathrm{I}_t\ \mathbf{g_t}^{(h)},\ \mathbf{g}_t^{(h)} \in \mathbf{G}_t\}$$
$$\mathcal{L}_t(\mathbf{G}_t, \mathbf{a}_t) = \{h\text{: } \mathbf{g_t}^{(h)}\, \mathrm{L}_t\, \mathbf{a}_t,\ \mathbf{g}_t^{(h)} \in \mathbf{G}_t\}$$

Defining the position of a stage alternative $\mathbf{a}_t$ in relation to the set $\mathbf{G}_t$ we consider:

Case S1. $\quad \mathcal{L}_t(\mathbf{a}_t, \mathbf{G}_t) \cup \mathcal{I}_t(\mathbf{a}_t, \mathbf{G}_t) \neq \varnothing$.

We calculate the value of stage success achievement degree as follows:

$$d_{\mathbf{G}}^+(\mathbf{a}_t) = \max\ \{d_t^+(\mathbf{a}_t, \mathbf{g}_t^{(h)})\text{: } h \in \mathcal{L}_t(\mathbf{a}_t, \mathbf{G}_t) \cup \mathcal{I}_t(\mathbf{a}_t, \mathbf{G}_t)\}$$

Case S2. $\quad \mathcal{L}_s(\mathbf{a}_t, \mathbf{G}_t) \cup \mathcal{I}_t(\mathbf{a}_t, \mathbf{G}_t) = \varnothing \wedge \mathcal{L}_s(\mathbf{G}_t, \mathbf{a}_t) \neq \varnothing$.

We calculate the value of stage success achievement degree as follows:

$$d_{\mathbf{G}}^-(\mathbf{a}_t) = \min\ \{d^-(\mathbf{a}_t, \mathbf{g_t}^{(h)})\text{: } h \in \mathcal{L}_t(\mathbf{G}_t, \mathbf{a}_t)\}$$

For a given $\mathbf{a}_t \in \mathbf{A}_t$ auxiliary sets of indices are defined as follows:

$$\mathcal{L}_t(\mathbf{B}_t, \mathbf{a}_t) = \{h\text{: } \mathbf{b_t}^{(h)}\, \mathrm{L}_t\, \mathbf{a}_t,\ \mathbf{b}_t^{(h)} \in \mathbf{B}_t\}$$
$$\mathcal{I}_t(\mathbf{B}_t, \mathbf{a}_t) = \{h\text{: } \mathbf{b_t}^{(h)}\, \mathrm{I}_t\, \mathbf{a}_t,\ \mathbf{b}_t^{(h)} \in \mathbf{B}_t\}$$
$$\mathcal{L}_t(\mathbf{a}_t, \mathbf{B}_t) = \{h\text{: } \mathbf{a}_t\, \mathrm{L}_t\, \mathbf{b}^{(h)},\ \mathbf{b}_t^{(h)} \in \mathbf{B}_t\}$$

Defining the position of an alternative $\mathbf{a}_t$ in relation to the set $\mathbf{B}_t$ we consider:

Case F1. $\quad \mathcal{L}_t(\mathbf{B}_t, \mathbf{a_t}) \cup \mathcal{I}_t(\mathbf{B}_t, \mathbf{a_t}) = \varnothing \wedge \mathcal{L}_t(\mathbf{a_t}, \mathbf{B}_t) \neq \varnothing$.

We calculate the value of stage failure avoidance degree as follows:

$$d_{\mathbf{B}}^+(\mathbf{a}_t) = \min\ \{d_t^+(\mathbf{a}_t, \mathbf{b}_t^{(h)})\text{: } h \in \mathcal{L}_s(\mathbf{a}_t, \mathbf{B}_t)\}$$

Case F2. $\quad \mathcal{L}_t(\mathbf{B}_t, \mathbf{a}_t) \cup \mathcal{I}_t(\mathbf{B}_t, \mathbf{a}_t) \neq \varnothing$.

We calculate the value of stage failure avoidance degree as follows:

$$d_{\mathbf{B}}^-(\mathbf{a}_t) = \max\ \{d_t^-(\mathbf{a}_t, \mathbf{b}_t^{(h)})\text{: } h \in \mathcal{L}_t(\mathbf{B}_t, \mathbf{a}_t) \cup \mathcal{I}_t(\mathbf{B}_t, \mathbf{a}_t)\}$$

3.3 Relationships in the set of multistage alternatives

According to the stage success achievement, for each multistage alternative **a** we define multistage success achievement degree:

$$\mathrm{d}_{\mathrm{S}}^+(\mathbf{a}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{d}_{\mathrm{G}}^+(\mathbf{a}_t) \qquad \mathrm{d}_{\mathrm{S}}^-(\mathbf{a}) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{d}_{\mathrm{G}}^-(\mathbf{a}_t)$$

According to the stage failure avoidance degree, for each multistage alternative **a** we define multistage failure avoidance degree:

$$\mathrm{d}_{\mathrm{F}}^{+}(\mathbf{a})=\frac{1}{T}\sum_{t=1}^{T}\mathrm{d}_{\mathbf{B}}^{+}(\mathbf{a}_t) \qquad \mathrm{d}_{\mathrm{F}}^{-}(\mathbf{a})=\frac{1}{T}\sum_{t=1}^{T}\mathrm{d}_{\mathbf{B}}^{-}(\mathbf{a}_t)$$

Taking into account values $d_S^+(\mathbf{a})$, $d_S^-(\mathbf{a})$, $d_F^+(\mathbf{a})$ and $d_F^-(\mathbf{a})$, multistage alternatives can be sorted to the nine categories:

Category M_1: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) = 0$, $d_F^+(\mathbf{a}) > 0$, $d_F^-(\mathbf{a}) = 0$.
Category M_2: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) = 0$, $d_F^+(\mathbf{a}) > 0$, $d_F^-(\mathbf{a}) > 0$.
Category M_3: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) = 0$, $d_F^+(\mathbf{a}) = 0$, $d_F^-(\mathbf{a}) > 0$.
Category M_4: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) > 0$. $d_F^-(\mathbf{a}) = 0$.
Category M_5: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) > 0$, $d_F^-(\mathbf{a}) > 0$.
Category M_6: $d_S^+(\mathbf{a}) > 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) = 0$, $d_F^-(\mathbf{a}) > 0$.
Category M_7: $d_S^+(\mathbf{a}) = 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) > 0$ $d_F^-(\mathbf{a}) = 0$.
Category M_8: $d_S^+(\mathbf{a}) = 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) > 0$ $d_F^-(\mathbf{a}) > 0$.
Category M_9: $d_S^+(\mathbf{a}) = 0$, $d_S^-(\mathbf{a}) > 0$, $d_F^+(\mathbf{a}) = 0$, $d_F^-(\mathbf{a}) > 0$.

A way of building above categories implies that if $k < l$, each multistage alternative from the Category M_k should be preferred to any multistage alternative from the Category M_l. Let:

$$d(\mathbf{a}^{(i)}) = d_S^+(\mathbf{a}^{(i)}) + d_S^-(\mathbf{a}^{(i)}) + d_F^+(\mathbf{a}^{(i)}) - d_F^-(\mathbf{a}^{(i)})$$

Inside categories alternatives are ordered as follows:

$\mathbf{a}^{(i)}$ is preferred to $\mathbf{a}^{(j)}$, iff $d(\mathbf{a}^{(i)}) > d(\mathbf{a}^{(j)})$

$\mathbf{a}^{(i)}$ is equivalent to $\mathbf{a}^{(j)}$, iff $d(\mathbf{a}^{(i)}) = d(\mathbf{a}^{(j)})$.

4 NUMERICAL ILLUSTRATION

We consider 3 stage decision process. The sets of feasible states and decisions are as follows:

$\mathbf{Y}_t = (0,1\}$ for $t = 1,\ldots,4$, $\mathbf{X}_t(0) = \{0, 1\}$, $\mathbf{X}_t(1) = \{0, 1\}$ for $t = 1,2.3$.

It means that we have four stage alternatives at each stage of the process :

$$\mathbf{A}_t = \{\mathbf{a}_t^{(0}, \mathbf{a}_t^{(1)}, \mathbf{a}_t^{(2)}, \mathbf{a}_t^{(3)}\}$$

$$\mathbf{a}_t^{(0)} = (0, 0), \quad \mathbf{a}_t^{(1)} = (0, 1) \quad \mathbf{a}_t^{(2)} = (1, 0) \quad \mathbf{a}_t^{(3)} = (1, 1).$$

Set **A** consists of 16 multistage alternatives: $\mathbf{a}^{(0)} = (0, 0, 0, 0)$, $\mathbf{a}^{(1)} = (0, 0, 0, 1)$, $\mathbf{a}^{(2)} = (0, 0, 1, 0)$, $\mathbf{a}^{(3)} = (0, 0, 1, 1)$, $\mathbf{a}^{(4)} = (0, 1, 0, 0)$, $\mathbf{a}^{(5)} = (0, 1, 0, 1)$, $\mathbf{a}^{(6)} = (0, 1, 1, 0)$, $\mathbf{a}^{(7)} =(0, 1, 1, 1)$, $\mathbf{a}^{(8)} = (1, 0, 0, 0)$, $\mathbf{a}^{(9)} = (1, 0, 0, 1)$, $\mathbf{a}^{(10)} = (1, 0, 1, 0)$, $\mathbf{a}^{(11)} = (1, 0, 1, 1)$, $\mathbf{a}^{(12)} = (1, 1, 0, 0)$, $\mathbf{a}^{(13)} = (1, 1, 0, 1)$, $\mathbf{a}^{(14)} = (1, 1, 1, 0)$, $\mathbf{a}^{(15)} = (1, 1, 1, 1)$.

At each stage we have two reference sets:

$$\mathbf{G}_t = \{\mathbf{g}_t^{(0)}, \mathbf{g}_t^{(1)}) \text{ and } \mathbf{B}_t = \{\mathbf{b}_t^{(0)}, \mathbf{b}_t^{(1)}).$$

The matrix of stage criteria weights is given in table 1.

Table 1: Values of stage weights of criteria

0.13	0,27	0,23	0,1	0,25	0	0	0
0,13	0	0,23	0,1	0	0,27	0	0,25
0	0	0,23	0	0,25	0,27	0,11	0,14

The results of comparisons between stage alternatives $\mathbf{a}_t \in \mathbf{A}_t$ and elements of reference sets are given in table 2.

Table 2: Results of comparisons between stage alternatives and elements of reference sets

t	$\mathbf{A_t}$	$\mathbf{G_t}$	C_t^1	C_t^2	C_t^3	C_t^4	C_t^5	C_t^6	C_t^7	C_t^8	$\mathbf{B_t}$	C_t^1	C_t^2	C_t^3	C_t^4	C_t^5	C_t^6	C_t^7	C_t^8
1	$\mathbf{a}_1^{(0)}$	$\mathbf{g}_1^{(0)}$	–	=	–	+	=				$\mathbf{b}_1^{(0)}$	+	+	=	–	–			
		$\mathbf{g}_1^{(1)}$	–	–	=	–	=				$\mathbf{b}_1^{(1)}$	=	+	=	–	+			
	$\mathbf{a}_1^{(1)}$	$\mathbf{g}_1^{(0)}$	=	+	=	–	–				$\mathbf{b}_1^{(0)}$	+	=	–	=	+			
		$\mathbf{g}_1^{(1)}$	+	–	=	–	+				$\mathbf{b}_1^{(1)}$	=	=	+	+	=			
	$\mathbf{a}_1^{(2)}$	$\mathbf{g}_1^{(0)}$	–	–	–	=	=				$\mathbf{b}_1^{(0)}$	–	=	–	=	–			
		$\mathbf{g}_1^{(1)}$	–	+	=	–	+				$\mathbf{b}_1^{(1)}$	+	=	+	–	–			
	$\mathbf{a}_1^{(3)}$	$\mathbf{g}_1^{(0)}$	=	–	=	+	=				$\mathbf{b}_1^{(0)}$	–	=	=	–	+			
		$\mathbf{g}_1^{(1)}$	–	+	–	–	+				$\mathbf{b}_1^{(1)}$	+	+	–	+	–			
2	$\mathbf{a}_2^{(0)}$	$\mathbf{g}_2^{(0)}$	–		–	+		=	–		$\mathbf{b}_2^{(0)}$	+		+	=		–	+	
		$\mathbf{g}_2^{(1)}$	–		–	–		–	=		$\mathbf{b}_2^{(1)}$	–		+	=		+	–	
	$\mathbf{a}_2^{(1)}$	$\mathbf{g}_2^{(0)}$	–		=	=		–	=		$\mathbf{b}_2^{(0)}$	–		=	=		=	=	
		$\mathbf{g}_2^{(1)}$	+		=	–		+	=		$\mathbf{b}_2^{(1)}$	–		+	=		=	=	
	$\mathbf{a}_2^{(2)}$	$\mathbf{g}_2^{(0)}$	=		–	+		=	=		$\mathbf{b}_2^{(0)}$	=		–	+		–	+	
		$\mathbf{g}_2^{(1)}$	–		+	–		+	–		$\mathbf{b}_2^{(1)}$	=		–	+		=	+	
	$\mathbf{a}_2^{(3)}$	$\mathbf{g}_2^{(0)}$	+		–	=		+	=		$\mathbf{b}_2^{(0)}$	+		+	=		+	–	
		$\mathbf{g}_2^{(1)}$	–		–	=		+	–		$\mathbf{b}_2^{(1)}$	+		=	–		–	+	
3	$\mathbf{a}_3^{(0)}$	$\mathbf{g}_3^{(0)}$			–		–	+	=	–	$\mathbf{b}_3^{(0)}$			+		–	–	+	–
		$\mathbf{g}_3^{(1)}$			–		+	=	+	+	$\mathbf{b}_3^{(1)}$			+		+	–	=	+
	$\mathbf{a}_3^{(1)}$	$\mathbf{g}_3^{(0)}$			=		=	–	–	–	$\mathbf{b}_3^{(0)}$			–		–	+	–	=
		$\mathbf{g}_3^{(1)}$			–		+	–	+	+	$\mathbf{b}_3^{(1)}$			–		+	–	+	+
	$\mathbf{a}_3^{(2)}$	$\mathbf{g}_3^{(0)}$			–		+	+	+	+	$\mathbf{b}_3^{(0)}$			+		–	–	–	–
		$\mathbf{g}_3^{(1)}$			–		–	+	+	+	$\mathbf{b}_3^{(1)}$			–		+	–	+	–
	$\mathbf{a}_3^{(3)}$	$\mathbf{g}_3^{(0)}$			+		–	+	=	=	$\mathbf{b}_3^{(0)}$			+		+	=	-	–
		$\mathbf{g}_3^{(1)}$			–		+	+	–	–	$\mathbf{b}_3^{(1)}$			=		=	+	+	=

We want to sort alternatives to classes $M_1 – M_9$ and rank them.

Applying of the procedure, we calculate values $d_{\mathbf{G}}^{+}(\mathbf{a}_t)$, $d_{\mathbf{G}}^{+}(\mathbf{a}_t)$, $d_{\mathbf{B}}^{+}(\mathbf{a}_t)$ and $d_{\mathbf{B}}^{-}(\mathbf{a}_t)$. They are given in table 3.

Table 3: Values of $d_{\mathbf{G}}^{+}(\mathbf{a}_t)$, $d_{\mathbf{G}}^{+}(\mathbf{a}_t)$, $d_{\mathbf{B}}^{+}(\mathbf{a}_t)$ and $d_{\mathbf{B}}^{-}(\mathbf{a}_t)$

Stage no	Stage realization	$d_{\mathbf{G}}^{+}(\mathbf{a}_t)$	$d_{\mathbf{G}}^{-}(\mathbf{a}_t)$	$d_{\mathbf{B}}^{+}(\mathbf{a}_t)$	$d_{\mathbf{B}}^{-}(\mathbf{a}_t)$
1	$\mathbf{a}_1^{(0)}$	0	0,88	0,63	0
	$\mathbf{a}_1^{(1)}$	0,69	0	0,77	0
	$\mathbf{a}_1^{(2)}$	0,75	0	0	0,88
	$\mathbf{a}_1^{(3)}$	0,52	0	0,52	0
2	$\mathbf{a}_2^{(0)}$	0	0,75	0,5	0
	$\mathbf{a}_2^{(1)}$	0,88	0	0,87	0
	$\mathbf{a}_2^{(2)}$	0,77	0	0,5	0
	$\mathbf{a}_2^{(3)}$	0,52	0	0,61	0
3	$\mathbf{a}_3^{(0)}$	0,77	0	0	0,66
	$\mathbf{a}_3^{(1)}$	0,5	0	0	0,73
	$\mathbf{a}_3^{(2)}$	0,77	0	0	0,77
	$\mathbf{a}_3^{(3)}$	0,64	0	0,75	0

For each multistage alternative $\mathbf{a}^{(0)}, \ldots, \mathbf{a}^{(15)}$ multistage values $d_S^{+}(\mathbf{a})$, $d_S^{-}(\mathbf{a})$, $d_F^{+}(\mathbf{a})$, $d_F^{-}(\mathbf{a})$, number of class the alternative **a** is sorted to and the position of the alternative **a** in the bipolar ranking are calculated. The results are given in table 4. Two multistage alternatives: $\mathbf{a}^{(7)}$ and

$\mathbf{a}^{(15)}$ are sorted to the class M_1, three of them: $\mathbf{a}^{(4)}$, $\mathbf{a}^{(5)}$, $\mathbf{a}^{(6)}$ – to the class M_2, one: $\mathbf{a}^{(3)}$ – to the class M_4 and the rest – to the class M_5. Multistage alternatives are ranked as follows: $\mathbf{a}^{(7)}$, $\mathbf{a}^{(15)}$, $\mathbf{a}^{(11)}$, $\mathbf{a}^{(4)}$, $\mathbf{a}^{(6)}$ $\mathbf{a}^{(5)}$, $\mathbf{a}^{(12)}$, $\mathbf{a}^{(14)}$, $\mathbf{a}^{(13)}$, $\mathbf{a}^{(10)}$, $\mathbf{a}^{(3)}$, $\mathbf{a}^{(2)}$, $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$, $\mathbf{a}^{(8)}$, $\mathbf{a}^{(9)}$.

Table 4. Bipolar sorting and bipolar ranking

A	$d_S^+(\mathbf{a})$	$d_S^-(\mathbf{a})$	$d_F^+(\mathbf{a})$	$d_F^-(\mathbf{a})$	$d(\mathbf{a})$	M	R	**A**	$d_S^+(\mathbf{a})$	$d_S^-(\mathbf{a})$	$d_F^+(\mathbf{a})$	$d_F^-(\mathbf{a})$	$d(\mathbf{a})$	M	R
$\mathbf{a}^{(0)}$	0,257	0,543	0,377	0,22	0,957	5	13	$\mathbf{a}^{(8)}$	0,507	0,25	0,167	0,51	0,41	5	15
$\mathbf{a}^{(1)}$	0,167	0,543	0,377	0,243	0,843	5	14	$\mathbf{a}^{(9)}$	0,417	0,25	0,167	0,537	9,297	5	16
$\mathbf{a}^{(2)}$	0,55	0,293	0,5	0,257	1,087	5	12	$\mathbf{a}^{(10)}$	0,8	0	0,29	0,55	0,54	2	10
$\mathbf{a}^{(3)}$	0,507	0,293	0,75	0	1,55	4	11	$\mathbf{a}^{(11)}$	0,757	0	0,54	0,293	1,003	2	3
$\mathbf{a}^{(4)}$	0,743	0	0,423	0,22	0,947	2	4	$\mathbf{a}^{(12)}$	0,687	0	0,34	0,22	0,807	2	7
$\mathbf{a}^{(5)}$	0,653	0	0,423	0,243	0,833	2	6	$\mathbf{a}^{(13)}$	0,597	0	0,34	0,243	0,693	2	9
$\mathbf{a}^{(6)}$	0,66	0	0,46	0,257	0,8633	2	5	$\mathbf{a}^{(14)}$	0,603	0	0,377	0,257	0,723	2	8
$\mathbf{a}^{(7)}$	0,617	0	0,71	0	1,327	1	1	$\mathbf{a}^{(15)}$	0,56	0	0,627	0	1,187	1	2

5 CONCLUDING REMARKS

In the version of the procedure presented in the current paper, for simplicity we assumed that all the values of equivalent and veto thresholds are equal to zero and concordance level is equal to 0.5. The next step which should be done is to cancel this assumptions.

In the general case it may happen, that some of stage alternatives will not be comparable with the stage reference sets. In such situations some multistage alternatives will be also non-comparable.

Future direction of the development of the method is preparation of the general case description. Another direction is to prepare software which will enable numerical simulations. As a good future field of applications seems to be regional sustainable development problems.

References

[1] Bellman R. (1957). Dynamic programming. Princeton.

[2] Konarzewska-Gubała E. (1987). Multicriteria Decision Analysis with Bipolar Reference System: Theoretical Model and Computer Implementation. *Archiwum Automatyk i Telemechaniki* vol. 32 no 4, 289-300.

[3] Konarzewska-Gubała E. (1989). BIPOLAR: Multiple Criteria Decision Aid Using Bipolar Reference System. LAMSADE, Cahier et Documents no 56, Paris.

[4] Konarzewska-Gubała E. (1991): *Multiple Criteria Decision Aid: System Bipolar.* Scientific Works of the University of Economics in Wrocław, no 551 (in Polish).

[5] Konarzewska-Gubała E. (1996). Supporting an effective performance appraisal system. *Argumenta Oeconomica* vol.1, 123-125.

[6] Konarzewska-Gubała E. (2002). Multiple Criteria Company Benchmarking Using the BIPOLAR Method. In Trzaskalik T., Michnik J.(eds.) Multiple Objective and Goal Programming. Recent Developments (pp 338-350). Physica-Verlag. Springer-Verlag Company, Heidelberg, New York.

[7] Trzaskalik T.(1987). Model of multistage multicriteria decision processes applying reference sets. Scientific Works of the University of Economics in Wrocław no 413, 73-93 (in Polish).

[8] Trzaskalik T., Sitarz S. (2012). How to Deal with Overgood and Underbad Alternatives in Bipolar Method. In: Watada J., Watanabe T.,, Phillips-Wren G., Howlett R., Jain L.C. (eds.). Proceedings of the 4th International Conference on Intelligent Decision Technologies (IDT´2012) , Intelligent Decision Technologies Smart Innovation, Systems and Technologies vol. 16, 345-354.

[9] Trzaskalik T., Sitarz S., Dominiak C. (2013). Unified procedure for Bipolar method. In: Zadnik L. Żerovnik J., Povh J., Drobne S., Lisec A. (eds.). The 12th International Sympodium on operational research, Slovenian Society Informatika – Section for Operational Research, 213-218.

[10] Trzaskalik T, Sitarz S., Dominiak C. (2019). Bipolar method and its modifications. Cent Eur J Oper Res 27: 625 https://doi.org/10.1007/s10100-019-00615-2

INVESTIGATING THE SELF-SERVING BIAS IN SOFTWARE SUPPORTED MULTIPLE CRITERIA DECISION MAKING PROCESS

Tomasz Wachowicz
University of Economics in Katowice, Department of Operations Research,
ul. 1 Maja 50, 40-287 Katowice, Poland, ORCID: 0000-0001-9485-6667
E-mail: tomasz.wachowicz@uekat.pl

Ewa Roszkowska
University of Białystok, Faculty of Economy and Management, ul. Warszawska 63,
15-062 Białystok, Poland, ORCID: 0000-0003-2249-7217
E-mail: e.roszkowska@uwb.edu.pl

Abstract: In this study we analyze if and to which extent the self-serving bias occurs when evaluating the results of the decision support tools that were offered to the decision makers in solving the multiple criteria decision making problem. Using experimental data we examine how do they explain the fact that the final decision support tool they chose as most accurate produced the ranking different from the one the decision maker declared beforehand, prior to decision making process, as reflecting their preferences best. We found that self-serving bias is not common in highly involving decision support system and may depend on the actual task and its consequentiality.

Keywords: decision support, multiple criteria decision making, results comparison, self-serving bias.

1 INTRODUCTION

The existence of heuristics and biases in social, negotiation and decision making processes has been widely discussed in many studies. Such phenomena as framing effect (Tversky, Kahneman 1986), anchoring effect (Kahneman et al. 1982) or availability heuristic (Tversky, Kahneman 1973) have already been thoroughly analyzed, which confirmed that the cognitive biases may negatively impact the efficiency of the these processes. Therefore a major goal of any decision support is to help decision makers (DMs) to switch from the intuitive to the analytical thinking, and hence to define their preferences, goals and aspirations in more accurate way and produce the more efficient and satisfying final decisions.

Unfortunately, despite being offered the decision support mechanisms, DMs may still make some heuristic-based mistakes. Some experiments in representative negotiation showed, that agents built the scoring systems that represented the actual preferences of their principals inaccurately (Roszkowska, Wachowicz 2015). This, in many situations, affected the agent's understanding of the negotiation process and outcomes (Wachowicz et al. 2019). Furthermore, DMs are still prone to make some errors that are not related to the problem content directly, but result from inattentional blindness (Kersten et al. 2017; Simons, Chabris 1999) or the usage of the round numbers (Kersten et al. 2018).

A question arises - vital from the viewpoint of developing new and reliable decision support approaches, why the heuristic-based errors still occur in formally supported decision making processes? One of many explanations may be that the DMs do not engage enough in decision support process offered by the software support tools since they are still convinced to be able to solve the problems themselves using their knowledge and skills. Thus, it could be interesting to confront the results of their individual and intuitive decision analysis with those obtained by means of formal decision support and asked them for the reasons of potential discrepancies. Their answers could shed a new light on how do they proceed during the decision process supported by some formal algorithms and how do they evaluate such a support. Especially, the occurrence of the self-serving bias (SSB) (Heider 1958) can indicate the DMs' negative attitude towards the formal decision support.

In this paper we present preliminary results of experimental study, in which SSB is examined. An online multiple criteria decision making (MCDM) experiment is designed, in which the participants define their preferences individually using holistic approach. Then they follow a formal decision support protocol, which implements TOPSIS, AHP and direct rating methods that differ in the amount and quality of the preferential information that needs to be provided by DM. They finally are asked to explain the results in a series of open-question questionnaires that allow to confirm the occurrence of SSB. In the next sections we describe the SSB and the experiment itself and discuss the initial results.

2 SELF-SERVING BIAS AND THE USE OF COMPUTER TECHNOLOGY

Self-serving bias, first articulated by Heider (Heider 1958), is a kind of explanatory pattern according to which various external factors are raised by an individual as the reasons for personal failures in any actions taken individually or within group tasks, but the internal characteristics and attributions are considered as the sources of its success (Campbell, Sedikides 1999). The typical way to investigate the existence of SSB in a particular social or decision making context is to organize an experiment, in which the participants perform a task (that requires some effort that allows to measure their intelligence, skills, etc.) and then give them the feedback on their success or failure in this task. Finally, the participant is asked to comment on the outcomes their obtained. Depending on the style of their comments, the existence of SSB is confirmed or not.

Since the computers started to play an important role in the DMs' everyday life, the existence of SSB in using computer technology has also been studied recently. Moon (Moon 2003) has analyzed how the consumers blame the software systems for they failures in purchasing decisions and found that usually SSB occurs, yet it may depend on the history of user-computer interactions. If the users are more involved in the information exchange with the computer (self-disclosure rises), the pattern of attribution changes significantly and the users start to consider the computers as having their contribution in the positive results and themselves as being more guilty for failures. Serenko (Serenko 2007) analyzed, how SSB occurs for the use of software agents with different level of autonomy. He showed that the SSB effect is mitigated here, and the majority of users credit the agents for positive outcomes (more, when the autonomy increases). However, they blame the agents for negative outcomes in similar proportions as they credited them for obtaining a good support.

These studies show that the SSB may be mitigated depending on the quality and scope of interaction with software support tool as well as on the users' subjective perception of how much they were involved and how much they may influence this process. When we now focus on decision support systems (DSS), which implement the formal methods and algorithms for supporting complex and multiple criteria decision making processes, analyzing the SSB effects may be an indicator of the real usefulness (not the user's self-declared one) of such a systems or show the potential necessity of its redesigning or further improvements. Note, that by involving the users in the process of preference elicitation the DSS requires a high level of self-disclosure regarding their needs, goals and aspirations. Simultaneously, it has a particularly low autonomy, which amounts to the selection of preference aggregation mechanism only. According to Moon and Serenko, this should reduce the potential occurrence of SSB. Yet, if SSB still exists in case of negative decision support results, it may indicate that the users do not realize that their failures may be caused by some other cognitive limitation they were unable to overcome while using this DSS. Contrary, the lack of SSB may result from an increase in user's awareness of the potential mistakes they made or limited diligence in preference elicitation process, which is a positive effect of using DSS for future decision making.

3 DESIGNING THE MCDM EXPERIMENT

In this study we analyze the existence of SSB using the results of online software supported MCDM experiment. The experiment was conducted by means of Electronic Survey Platform (ESP) (Roszkowska, Wachowicz 2016), which allows to design a classic survey based on the close- and open-question questionnaires combined with the decision support modules coded in PHP. In our experiment, we predefined a MCDM problem of a flat to rent by the student for the forthcoming academic year. There were five alternatives considered, each evaluated by means of five criteria: two of them defined by means of crisp numerical values (monthly rent and living area), one by interval value (commuting time) and last two – qualitatively using verbal description (flat layout and furnishings).

The experimental protocol required the participants to read the case and start with holistic declaration of preferences for the alternatives in a form of ordinal ranking. Then they were offered three alternative approaches to analyze the problem deeper at the disaggregated level. The process of the decision analysis consisted of three steps. In step one the user determined the importance of the criteria by means of two tools: (1) using pairwise comparisons and verbal scale (AHP-like); and (2) using 7-point linguistic scale (numerical equivalents were not disclosed to users). In step two they were offered the three preference elicitation methods to declare their preferences for options within each criterion: TOPSIS, AHP and direct rating (DR) (SMART-like approach). In TOPSIS 7-point Likert scale with stars representation was used to declare the preferences for qualitative criteria. The AHP interface used sliders with verbal descriptions to define the preferences for pairs. In DR the users had to assign points to options within each issue using the range [0; 100] with obligatory assignments of 0 and 100 to worst and best options respectively. Finally, in step three the rankings and corresponding ratings for all three methods were displayed and the users had to choose the one that reflect their preferences best.

The users were also asked to answer a series of questions regarding the use and usefulness of the decision analysis they went through. Among many questions, two were focused on confronting the results obtained by means of decision analysis with the prior holistic declaration of preferences. In Q1 the users had to explain the reasons for potential differences in rankings obtained by means of these the decision support approaches used in step 3 of the decision analysis phase. In Q2 they were asked to explain the potential differences in the ranking defined holistically and the one resulting with the method they had chosen in the last step (if the rankings matched, they explained the reason for ideal match). These were the open questions so no predefined answers existed and users had to write their own explanations not being bounded by any suggestions from the questionnaire.

The participants were 190 students of four Polish universities. To increase the consequentiality of participation in the experiment they were offered extra credits for their academic course performance. The preference elicitation process in ESP requires the students to involve quite deeply (some in-class observations made during the usage of ESP confirmed that they could spend even 90 minutes on completing all phases of the experiment). Therefore, we can expect that they will consider these results as the potential success or failure in task completion, and the SSB may be investigated while analyzing their answers.

4 PRELIMINARY RESULTS AND FUTURE WORK

We analyzed the answers of ESP users, which allowed to define some general categories of answers for questions Q1 and Q2 separately. Since the answers could be detailed and complex, it was possible to classify one answer to two different categories simultaneously.

When analyzing the answers for Q1 we found surprisingly, there were only two users who obtained the same ranking of alternatives for all three MCDA algorithms (i.e., successful in their task), though they were unable to clearly describe the reasons for this ideal concordance of their results. Hence, the detection of SSB was impossible for them. All the remaining 188 DMs obtained at least two different rankings from their decision analysis conducted by means of TOPSIS, AHP and DR (a kind of task failure) and the structure of their answers for Q1 is shown in Table 1.

Table 1: The categories of answers for Q1 explaining the discordances in rankings.

Answer category	Number of answers	% of users
Cat. 1. My fault	27	14%
Cat. 2. Methods' error	14	7%
Cat. 3. Methods work differently	58	31%
Cat. 4. No differences as such	27	14%
Cat. 5. I do not know why they are different	67	36%

Note, that there were only 27 answers (14%), in which the differences in rankings were explained as resulting from own errors or mistakes (which is the opposite reaction to SSB). These were the people, who pointed out their mistakes directly (14 users), as well as those who claim their error was strengthen by the specificity of the preference elicitation, characteristic to each of these methods (5), and those who claimed that their mistakes were cause by the difficulty of the decision support algorithms (5). The Self-serving bias, shown in clear and unambiguous declarations that the differences were caused by the faults of methods, was revealed only in 14 answers only (7%). However, there were another 58 answers (31%), in which the blame for the differences in rankings' results was easily laid on the differences in the philosophy of the preference aggregation used by these methods, i.e. the different scales they use and the operations they perform on the quantitative representation of users preferences, the preference aggregation mechanisms, etc. Let us note that this really may be a true reason for the differences the users obtained, since such situations are often reported in many experiments (Mela et al. 2012; Zanakis et al. 1998). Yet, this may also be an easy and very convenient excuse to avoid responsibility for being not diligent enough or make the mistakes in providing the concordant preference information in each preference elicitation mechanism. Therefore, these group of users can be considered as partially revealing the SSB, yet the true reasons require further and deeper investigation.

An interesting group of answers is category 4. It consists of 27 people who claimed there were actually no differences in the results the methods had produced (14%). Some of them reckoned, they were only focused on fining best alternative, so the rest of the ranking was not important (6), some others wrote, that despite different rankings the lists of alternatives identify similar categories of their quality, i.e. that they sorted the alternatives accurately (6); 15 others claimed that the rankings are nearly or almost the same, and claiming that they are different would be simply an exaggeration. Note, that they may be the people who might be biased by inattentional blindness (they did not notice they had had to build an accurate ranking of offers) or demonstrate the repression or post-decision dissonance (Vroom 1966), another biases that we were not expecting to find in our experiment.

Please also note that there were as much as 67 people (36%) that were unable to explain the differences using convincing arguments: 43 of them simply wrote they do not know the reasons, while remaining 24 acknowledged the differences but described them only without presenting any rationale for them.

The results of analysis of the answers for Q2 differ significantly. Here, there were 22 DMs who determined analytically the same ranking to the one they had defined holistically at the beginning of the experiment (step 1). Nine of them claimed this was due to the accuracy of the decision support algorithm (no SSB), while 12 that the only reason was their accuracy in providing the preference information in decision support process (SSB occurs). Only one DM reckoned, the concordance of rankings results from both his and method's accuracy. For remaining 168 DMs the holistic and analytical rankings were different (task failure). The categories of their explanations are shown in Table 2.

Table 2: The categories of answers for Q2 explaining the discordances in DM's holistic and analytically built ranking.

Answer category	Number of answers	% of users
Cat. 1. My fault	33	20%
Cat. 2. Methods' error	16	10%
Cat. 3. Methods work differently	8	5%
Cat. 4. No differences as such	29	17%
Cat. 5. I do not know why they are different	8	5%
Cat. 6. I thoroughly reconsidered my preferences	74	44%

The major difference in fractions between answers for Q1 and Q2 can be observed within category 3. Previously, when all three rankings were compared as much as 31% of answers that blamed the methods themselves for failure. Here there is no much than 5% DMs raising this issue as a major reason for failure. Surprisingly, there is as much as 44% DMs who do not consider the discrepancy as a failure (new category 6), but acknowledge the positive influence of the decision support method in better recognition of their preferences. They claim that by using the method they could realize and understand their true needs better, and their new analytical rankings are now the ones that represent their preferences more accurately. Please also note, that the SSB (full or partial, i.e. represented by categories 2 and 3) is significantly less frequently occurring here (15% in Q2 vs. 38% in Q1).

Our preliminary results show, that the occurrence of SSB may be dependent on the actual task and its subjectively perceived consequentiality by the DMs. When all three rankings were compared and their differences had to be explained, the ESP users were more prone to blame external factors for the differences. It could be due to the fact that there were not only them (i.e. DMs themselves) but also three other methods that could be made responsible for the discordances. Additionally, making three perfectly concordant rankings by means of three different methods was not the initial goal for them to achieve. Yet, when they had to compare their initial holistic ranking with the one they determined themselves and chosen as the best, the personal responsibility for the discrepancy could be perceived as higher (there was only one method that could be blamed, but fed with the preferential information provided by the DMs themselves). The users, having possibility of revealing SSB and blaming the algorithm for the discordances, have noticed the added value of using such methods in preference elicitation. They have truly acknowledged the fact that while using the analytical support they could rethink and conceptualize their preferences better, and thus, make more appropriate decisions. This finding is very important (and positive) from the viewpoint of designing new decision support mechanisms and systems.

Our future work will be focused on the detailed analysis of the results. We will try to find, if the SSB may depend on the decision making profile determined by means of REI test (Epstein et al. 1996). Some problems we faced while analyzing and categorizing the answers will also require consideration of using the closed-questions in future experiments in the post-decision questionnaires, that would help us identify the biases more accurately. Finally, we found that some other biases and heuristics different form SSB may occur while explaining the performance of DMs and their evaluation of the use and usefulness of the decision support mechanisms. The new questionnaires should also consist of the questions allowing for their unambiguous recognition and analysis.

Acknowledgements. This research was supported with the grants from Polish National Science Centre (2016/21/B/HS4/01583).

References

[1] Campbell, W.K., Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of general Psychology, 3*(1), 23.

[2] Epstein, S., Pacini, R., Denes-Raj, V., Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of personality and social psychology, 71*(2), 390-405.

[3] Heider, F. (1958). The psychology of interpersonal relations. *The Journal of Marketing, 56*, 322.

[4] Kahneman, D., Slovic, P., Tversky, A. (Eds.). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

[5] Kersten, G., Roszkowska, E., Wachowicz, T. (2017). The Heuristics and Biases in Using the Negotiation Support Systems. In M. Schoop & D. M. Kilgour (Eds.), *Group Decision and Negotiation. A Socio-Technical Perspective* (pp. 215-228): Springer.

[6] Kersten, G., Roszkowska, E., Wachowicz, T. (2018). Representative Decision-Making and the Propensity to Use Round and Sharp Numbers in Preference Specification. In P. D. G. K. Ye Chen, Rudolf Vetschera, Prof. Dr. Haiyan Xu (Ed.), *Group Decision and Negotiation in an Uncertain World. Lecture Notes in Business Information Processing* (Vol. 315, pp. 43-55): Springer.

[7] Mela, K., Tiainen, T., Heinisuo, M. (2012). Comparative study of multiple criteria decision making methods for building design. *Advanced Engineering Informatics, 26*(4), 716-726.

[8] Moon, Y. (2003). Don't blame the computer: When self-disclosure moderates the self-serving bias. *Journal of Consumer Psychology, 13*(1-2), 125-137.

[9] Roszkowska, E., Wachowicz, T. (2015). Inaccuracy in defining preferences by the electronic negotiation system users. *Lecture Notes in Business Insformation Processing, 218*, 131-143.

[10] Roszkowska, E., Wachowicz, T. (2016, June 20-24, 2016). *Analyzing the Applicability of Selected MCDA Methods for Determining the Reliable Scoring Systems.* Paper presented at the The 16th International Conference On Group Decision And Negotiation. , Bellingham.

[11] Serenko, A. (2007). Are interface agents scapegoats? Attributions of responsibility in human–agent interaction. *Interacting with computers, 19*(2), 293-303.

[12] Simons, D.J., Chabris, C.F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059-1074.

[13] Tversky, A., Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology, 5*(2), 207-232.

[14] Tversky, A., Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business, 59*(4), S251-S278.

[15] Vroom, V.H. (1966). Organizational choice: A study of pre-and postdecision processes. *Organizational behavior and human performance, 1*(2), 212-225.

[16] Wachowicz, T., Kersten, G.E., Roszkowska, E. (2019). How Do I Tell You What I Want? Agent's Interpretation of Principal's Preferences and Its Impact on Understanding the Negotiation Process and Outcomes. *Operational Research: International Journal*, (in print).

[17] Zanakis, S.H., Solomon, A., Wishart, N., Dublish, S. (1998). Multi-attribute decision making: a simulation comparison of select methods. *European journal of operational research, 107*(3), 507-529.

DECISION MAKING PROFILE AND THE CHOICES OF PREFERENCE ELICITATION MODE – A CASE OF USING GDMS INVENTORY

Tomasz Wachowicz
University of Economics in Katowice, Department of Operations Research,
ul. 1 Maja 50, 40-287 Katowice, Poland, ORCID: 0000-0001-9485-6667
E-mail: tomasz.wachowicz@uekat.pl

Ewa Roszkowska
University of Bialystok, Faculty of Economy and Management, ul. Warszawska 63,
15-062 Białystok, Poland, ORCID: 0000-0003-2249-7217
E-mail: e.roszkowska@uwb.edu.pl

Marzena Filipowicz-Chomko
Bialystok University of Technology, ul. Wiejska 45A,
15-351 Białystok, Poland, ORCID: 0000-0003-3041-4063
E-mail: m.filipowicz@pb.edu.pl

Abstract: In this paper we analyze how the decision making profile may affect the decision makers' (DMs) choices regarding the most preferable modes of both declaring the preferences (in the preference elicitation process) and representing the preference elicitation results by the software support systems. We use the General Decision Making Style Inventory (GDMS) that allows to describe the profile using a mix of five styles: rational, intuitive, dependent, avoidant and spontaneous. Using the dataset of online multiple criteria decision making experiment we identify the clusters of respondents with similar decision making profiles and analyze the differences in expected preference elicitation mode. Our results partially confirm earlier findings of behavioral studies in decision making that more intuitive DMs prefer the rankings and non-numerical ways of defining preferences, while the more rational ones operate more willingly with numbers and ratings. However, there is another group of GDMS-specific DMs, highly avoidant and non-spontaneous, who also differ from others with respect of preferring pictorial definition of preferences.

Keywords: multiple criteria decision making, decision making styles, preference elicitation.

1 INTRODUCTION

The early research works in a field of economics and psychology suggested that some behavioral aspects of decision making processes may affect the decision maker's (DM) behavior and the results they obtain (Gilovich et al. 2002; Kahneman, Tversky 1979; Simon 1955). This made the behavioral operations research to focus on the psychological elements of decision making such as the decision makers' (DMs) abilities for using fast and slow thinking styles (Stanovich 1999) and their potential inclinations for occurring of some heuristics and biases. From the viewpoint of operations research these findings seem to be of crucial importance, since they underline the necessity of including of these behavioral elements in the process of designing of the decision support algorithms and tools. As shown in some experimental studies, offering the DMs a support tool without any prior analysis of its cognitive requirements and the DMs abilities of using it, may result in many mistakes and a final misuse of this tool (Kersten et al. 2017, 2018; Wachowicz et al. 2015). Consequently, the reliability of a decision support and the usefulness of the support tool itself may be questioned.

It seems crucial then to identify the DMs' information processing (or decision making) style and try to offer them the support that fits this style and their cognitive capabilities best. Such a style, naturally, may be identified using various psychometric tools. One of them suggests to consider the decision making style as a unidimensional bipolar structure with experientially and rationality as poles (Allinson, Hayes 1996). Some others consider the potential modes as orthogonal, suggesting measuring the style as a mix of various modes (Epstein 1998; Scott, Bruce 1995). The discussion, which of these approaches is better describing the nature of decision making style is unfinished (Hayes et al. 2003; Hodgkinson, Sadler- Smith 2003), yet some application research still indicate the problems of using these inventories, especially in constructing the factors describing the modes according to the assumptions of the authors (Bavol'ár, Orosová 2015; Engin, Vetschera 2017).

In one of our earlier study we tried to identify, how the decision making style may affect the DMs use and preferences toward the decision support tools measuring the former with the Rational-Experiential Inventory (Roszkowska, Wachowicz 2019). It confirmed the necessity of using orthogonal approach, yet still some nuances in differences of styles and their impact were unable to be significantly captured by means of this tool. Therefore, in this paper we try to describe the decision making profile in more detailed way using the General Decision Making Style (GDMS) inventory (Scott, Bruce 1995). The goal of this paper is to examine the relationship between the decision making profile described by GDMS and the most preferred way of declaring the preferences and organizing the preference elicitation results by DMs.

The remaining of this paper is organized in three more sections. In section 2 we provide a brief description of GDMS inventory. In section 3 the multiple criteria decision making experiment is described. Finally, in section 4 we present the experimental results followed by their discussion.

2 INDENTIFYING THE DECISION MAKING STYLES BY GDMS INVENTORY

Scott and Bruce (1995) defined the decision-making style as "the learned habitual response pattern exhibited by an individual when confronted with a decision situation. It is not a personality trait, but a habit-based propensity to react in a certain way in a specific decision context". They defined five decision-making styles: rational, intuitive, dependent, avoidant and spontaneous designed the General Decision-Making Style Inventory to measure them as a five-mode profile. The rational style has been "characterized by the search for logical evaluation of alternatives", intuitive – "by attention to detail and a tendency to rely on feeling", dependent - "by the search for and reliance on the advice of others", avoidant - by "the tendency to avoid decisions", and spontaneous has been" characterized by a sense of immediacy". The GDMS inventory contains 25 questions, all measured on a 5-point Likert scale. Each style can be identified by the answers for five questions, which was confirmed by the authors using exploratory factor analysis.

Several studies have focused on the relation between personality and decision-making and discussed the predictive validity of individual differences in decision making. Galotti and Tinkelenberg (2009) found that the avoidant style correlated positively with the number of sources used to collect information, the number of criteria used in the decision process and negatively with the number of options considered. Sager and Gastil (1999) reported that in small group the members' preference for consensual decisions correlated positively with their score on the rational and dependent decision-making styles. Curşeu and Schruijer (2012) reported that the rational style has strong positive impact on making the logically correct choices in decision making as well as strong negative impact on indecisiveness, whereas avoidant and dependent styles are positively associated with indecisiveness. Parker et al. (2007) reported the dependence of the decision-making styles and decision outcomes, Significantly better outcomes are reported for decision makers who are less maximizing. The depending and intuitive decision styles are very week positive correlated with outcomes, while avoidant and spontaneous are weekly negative correlated with outcomes. Finally, Delaney (et al. 2015) used cluster analysis to identify three decision-making profiles: affective/experiential, independent/self-controlled, and an interpersonally-oriented dependent ones and reported how these profiles differed by age and gender.

Acknowledging the impact of the decision making profile on various elements of the decision making process and results we are interested in finding if the GDMS profiles differentiate the decision makers with respect to their preferences regarding the preference elicitation mode that can be offered to them in the software supported multiple-criteria decision making.

3 THE MCDM SOFTAWRE SUPPORTED EXPERIMENT

To reach the research goal the multiple criteria decision making experiment was organized in the online survey system (OSS) (Roszkowska, Wachowicz 2019). The predefined problem required of DMs to build the ranking of flats to rent, and was chosen purposely in a view of potential participants, who were the students of five Polish universities. There were five alternatives to rank with respect to five criteria (both quantitative and qualitative in nature).

The experiment consisted of four phases. In phase 1 the participants read the problem and were asked to define their preferences holistically without any decision support offered by OSS. Then, in phase 2, they started the preference elicitation process with analyzing the issue weights. They were ask to define them both verbally and using pair-wise comparisons, or – if unsatisfied with the weights obtained these ways – define them individually in direct rating approach. In phase 3 the alternatives were compared in series of single-criterion analyses. Three approaches were used in OSS that used the algorithms of TOPSIS (Hwang, Yoon 1981), AHP (Saaty 1980) and SMARTS-like direct rating (Edwards, Barron 1994). For each of these algorithms different interfaces were designed that required operating with graphics (for TOPSIS), verbal evaluations (for AHP) or directly with numbers (for SMART). In the last phase the results of decision analyses were displayed to DMs and they were asked to choose the approach that allowed to define their preferences in most accurate way[1]. The respondents were also asked to answer series of pre-decision making profiling questionnaires (including GDMS inventory), and post-decision making evaluation questionnaires, in which they descried their opinion regarding the support offered.

Among many questions three of them were of special focus in this study:

- Q1. What is your most preferred way of declaring your preferences while using the decision support tool (numerical, verbal, pictorial or other)?
- Q2. What way of representing the final results of decision analysis by the decision support tool is a sufficient for you (ranking, rating, other)?
- Q3. What is the most preferred way of representing the final attractiveness of alternatives (numerical, non-numerical (verbal or pictorial), mixed, other)?

After eliminating incomplete records we obtained the dataset consisting of 266 records that we used in further analyses.

4 PRELIMINARY RESULTS AND FUTURE WORK

Using the confirmatory factor analysis we were unable to build the GDSM profile model for the assumed five factors (styles) out of 25 questions at the statistically acceptable level. Both chi-square measure (χ^2<0.001) and the goodness-of-fit indexes (GFI=0.82 and AGFI=0.78) recommended the model to be rejected. Taking into account the fact, that the GDMS inventory was translated to Polish and used in non-original version for Polish native students, we decided

[1] For details see (Roszkowska, Wachowicz 2019).

to verify if the fundamental assumptions of the instrument regarding the factors hold here, using the exploratory factor analysis (EFA). EFA identified high cross-loadings for the 24th GDMS question, associated to spontaneous style in the original Scott and Bruce's model; and here with the highest loading to the factor representing the intuitive style (0.560). Therefore we decided to remove it from the model, as well as all the other questions that did not received significant loadings higher than 0.65 with the differences in loadings higher than 0.2 (Howard 2016). Using this approach five questions appeared to be eliminated: one for the rational style, and two for each intuitive and spontaneous ones (questions: 5, 10, 11, 16 and 24 from the original GDMS inventory).

For the reduced GDMS inventory, EFA produced the model of statistically satisfactory fit (KMO=0.819, variance explained at 66.1%, Bartlett's test confirmed with p<0.001). The factor values obtained this way were used in cluster analysis to generate the groups of similar decision making profiles. We aimed at generating the clusters that differ in profiles significantly, yet in a view of rather small sample the number of clusters could not have been too big. Using different clusters approach the most promising results were obtained for average-linking clustering with Pearson correlation as the similarity measure. The analysis of the linkage distances across clustering steps revealed the highest distance of 0.8, which corresponded to moving from four- to three-cluster structure. Thus, we decided to use four clusters of profiles, with the mode values as shown in Fig. 1.

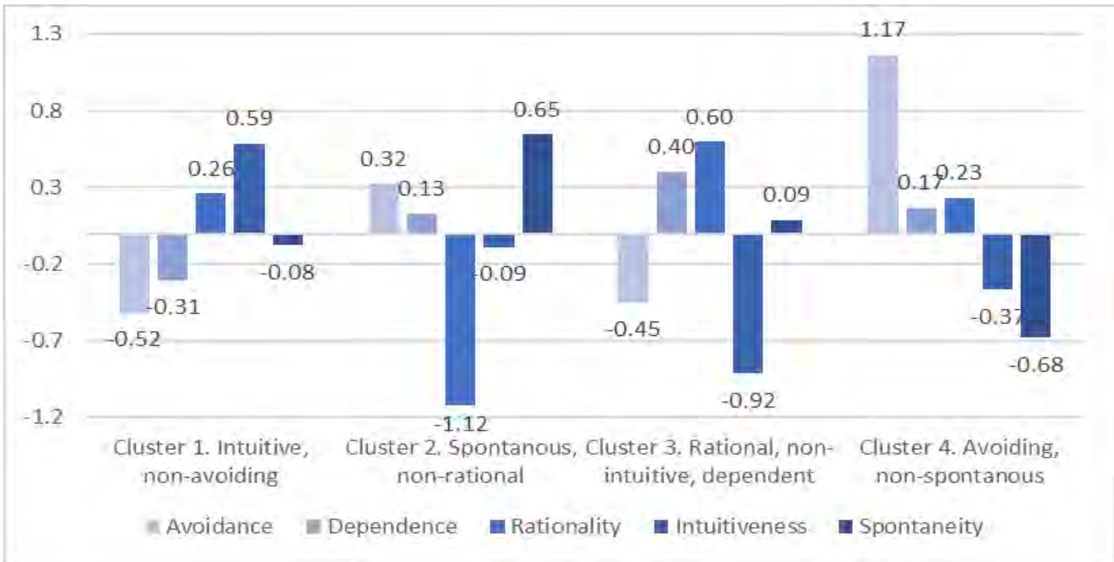


Figure 1: Clusters of modified GDMS-based decision-making profiles.

We named the clusters taking into account the significance of differences for most extreme modes across clusters. Cluster 1 was similar to cluster 3 with respect to avoidance (p=0.604) and spontaneity (p=0.433), and to cluster 4 in rationality (p=0.861). It was simultaneously the only one that differed from others significantly with respect to dependence (p<0.013). Clusters 2 and 4 were similarly intuitive (p=0.109). Such clustering makes our profiles to be incomparable to ones that could be determined using other profiling mechanisms. For instance, there are no equivalents of versatile and avoiding profiles that can be computed based on REI test (see Roszkowska, Wachowicz 2019), one maximizing and the other minimizing the rationality and intuitiveness modes simultaneously. Conversely, we identify here a cluster 4, in which a medium level of both those modes are observed with a highly differentiating styles as avoiding and spontaneous.

According to our research goal we were interested in examining an impact of such profiles on the most preferred mode of preference analysis, which was examined in three post-decision making questions (see section 3). The differences for clusters are shown in Tables 1-3.

Table 1: Most preferred way of declaring preferences in preference elicitation phase

Declaration of preferences	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Numerical	69 (62.2%)	36 (60.0%)	**31 (68.9%)**	29 (58.0%)
Pictorial	33 (29.7%)	12 (20.0%)	11 (24.4%)	**19 (38.0%)**
Verbal	8 (7.2%)	**12 (20.0%)**	2 (4.4%)	2 (4.0%)
In other way	1 (0.9%)	0 (0.0%)	1 (2.2%)	0 (0.0%)

The chi-square test confirms the dependence between the preference declarations and clusters (p=0.049). As shown in Table 1, the spontaneous and non-rational DMs (cluster 2), significantly more frequently from others choose verbal declarations (p<0.013). They also choose pictorial declarations significantly less frequently (p=0.012) than avoiding and non-spontaneous ones (cluster 4). The latter choose the pictures more frequently (p=0.077) than rational and dependent (cluster 3), and less frequently (yet with p=0.136) opt for numerical declarations.

Table 2: Sufficient way of representing final results

Final results represented as	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Rating	50 (45.0%)	30 (50.0%)	**26 (57.8%)**	26 (52.0%)
Ranking	**59 (53.2%)**	29 (48.3%)	18 (40.0%)	24 (48.0%)
Other information needed	2 (1.8%)	1 (1.7%)	1 (2.2%)	0 |(0.0%)

Table 3: Most preferred way of representing the attractiveness/evaluation of alternatives

Attractiveness of offers represented	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Numerically	52 (46.8%)	29 (48.3%)	23 (51.1%)	21 (42.0%)
Non-numerically (verbally/pictures)	22 (19.8%)	17 (28.3%)	**3 (6.7%)**	12 (24.0%)
In mixed way	36 (32.4%)	14 (23.3%)	**19 (42.2%)**	17 (34.0%)
Other	1 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Unfortunately, chi-square tests did not confirm the general dependence of sufficient or most preferred representation of results of decision support process (Q2 and Q3) and the decision-making profile (p=0.775 and p=0.267 respectively). However, the individual one-tail fraction tests confirmed significance of some differences.

Regarding the efficient way of representing the results (Q2) we were able to find the slightly significant differences in fractions for clusters 1 and 3. As in earlier works, more intuitive DMs more frequently opted for ranking than those more rational and dependent (p=0.068) and less frequently for rating. Regarding the details of representing the final attractiveness of alternatives in the results of decision support process, the rational and dependent DMs (cluster 3) preferred significantly less frequently non-numerical representations of results from all other DMs (p<0.022). They far more often (p=0.019) choose mixed (numerical and non-numerical) way of representation than spontaneous and non-rational DMs (cluster 2). The latter ones differ also somewhat significantly (p=0.107) in choosing the mixed way than the avoidant and non-spontaneous DMs (cluster 4).

As shown in this study, there are some differences in choices of preference elicitation mode depending on the decision making style recognized by GDSM. Some of them are in line to the classic presumptions or earlier findings, i.e. like those that rational DMs prefer numbers (here ratings), while the more intuitive ones do not (here, choose rankings). However, we found that some of differences could not be explained using some other profiling mechanisms, such as REI. Our cluster 4 identifies the avoiding and non-spontaneous DMs, who significantly differ from others in higher willingness to operate with pictorial representation of preferences. We hope that by increasing the sample size we will be able to specify other cluster and find further interesting differences among them.

Acknowledgements. This research was supported with the grants from Polish National Science Centre (2016/21/B/HS4/01583).

References

[1] Allinson, C.W., Hayes, J. (1996). The cognitive style index: A measure of intuition‐ analysis for organizational research. *Journal of Management studies, 33*(1), 119-135.

[2] Bavol'ár, J., Orosová, O.g. (2015). Decision-making styles and their associations with decision-making competencies and mental health. *Judgment and Decision making, 10*(1), 115-122.

[3] Curşeu, P.L., Schruijer, S.G. (2012). Decision styles and rationality: An analysis of the predictive validity of the General Decision-Making Style Inventory. *Educational and Psychological Measurement, 72*(6), 1053-1062.

[4] Edwards, W., Barron, F.H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes, 60*(3), 306-325.

[5] Engin, A., Vetschera, R. (2017). Information representation in decision making: The impact of cognitive style and depletion effects. *Decision Support Systems, 103*, 94-103.

[6] Epstein, S. (1998). Cognitive-experiential self-theory. *Advanced personality*, 211-238.

[7] Galotti, K.M., Tinkelenberg, C.E. (2009). Real-life decision making: Parents choosing a first-grade placement. *The American journal of psychology*, 455-468.

[8] Gilovich, T., Griffin, D., Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*: Cambridge University Press.

[9] Hayes, J., Allinson, C.W., Hudson, R.S., Keasey, K. (2003). Further reflections on the nature of intuition‐ analysis and the construct validity of the Cognitive Style Index. *Journal of Occupational and Organizational Psychology, 76*(2), 269-278.

[10] Hodgkinson, G.P., Sadler‐ Smith, E. (2003). Complex or unitary? A critique and empirical re‐ assessment of the Allinson‐ Hayes Cognitive Style Index. *Journal of Occupational and Organizational Psychology, 76*(2), 243-268.

[11] Howard, M.C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction, 32*(1), 51-62.

[12] Hwang, C.-L., Yoon, K. (1981). *Multiple Attribute Decision Making: Methods and Applications*. New York: Springer-Verlag.

[13] Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 263-291.

[14] Kersten, G., Roszkowska, E., Wachowicz, T. (2017). The Heuristics and Biases in Using the Negotiation Support Systems. In M. Schoop & D. M. Kilgour (Eds.), *Group Decision and Negotiation. A Socio-Technical Perspective* (pp. 215-228): Springer.

[15] Kersten, G., Roszkowska, E., Wachowicz, T. (2018). Representative Decision-Making and the Propensity to Use Round and Sharp Numbers in Preference Specification. In P. D. G. K. Ye Chen, Rudolf Vetschera, Prof. Dr. Haiyan Xu (Ed.), *Group Decision and Negotiation in an Uncertain World. Lecture Notes in Business Information Processing* (Vol. 315, pp. 43-55): Springer.

[16] Parker, A.M., De Bruin, W.B., Fischhoff, B. (2007). Maximizers versus satisficers: Decision-making styles, competence, and outcomes. *Judgment and Decision making, 2*(6), 342.

[17] Roszkowska, E., Wachowicz, T. (2019). *Cognitive Style and the Expectations Towards the Preference Representation in Decision Support Systems.* Paper presented at the International Conference on Group Decision and Negotiation.

[18] Saaty, T.L. (1980). *The Analytic Hierarchy Process*. New York: McGraw Hill.

[19] Sager, K.L., Gastil, J. (1999). Reaching consensus on consensus: A study of the relationships between individual decision‐ making styles and use of the consensus decision rule. *Communication Quarterly, 47*(1), 67-79.

[20] Scott, S.G., Bruce, R.A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement, 55*(5), 818-831.

[21] Simon, H.A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99-118.

[22] Stanovich, K.E. (1999). *Who is rational?: Studies of individual differences in reasoning*: Psychology Press.

[23] Wachowicz, T., Kersten, G.E., Roszkowska, E. (2015). *The impact of preference visualization and negotiators' profiles on scoring system accuracy*. Paper presented at the 27th European Conference on Operational Research EURO 2015, University of Strathclyde, Glasgow.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 3:

Graph Theory and Algorithms

SPECTRAL CLUSTERING OF SURVIVAL CURVES

Kolos Csaba Ágoston
Department of Operations Research and Actuarial Sciences,
Corvinus University of Budapest, H-1093, Fővám tér 8., Budapest, Hungary,
Institute of Economics, Research Centre for Economic and Regional Studies,
Hungarian Academy of Sciences, H-1097, Tóth Kálmń u. 4., Budapest, Hungary,
E-mail: kolos.agoston@uni-corvinus.hu

Snježana Majstorović
Department of Mathematics, Josip Juraj Strossmayer University of Osijek
Trg Ljudevita Gaja 6, 31000 Osijek, Croatia,
E-mail: smajstor@mathos.hr

Ágnes Vaskövi
Department of Finance, Corvinus University of Budapest,
H-1093, Fővám tér 8., Budapest, Hungary,
E-mail: agnes.vaskovi@uni-corvinus.hu

Abstract. Mortality analysis is a well-studied area of statistics, but clustering methods on mortality data are not available. Furthermore, curve clustering, especially on survival curves, is a relatively new field in data analysis. In this paper we use spectral methods for survival curve clustering that might have relatively high importance for insurance companies. We discuss the key points in implementation and present a numerical study on mortality data of EU countries.

Keywords: curve clustering, spectral clustering, Makeham mortality law, survival function, Laplace matrix

1 INTRODUCTION

Insurance, as well as other financial products needs homogeneous risk communities. If insurance companies in EU operate in multiple countries and do not want to develop different products for each country, it is a straightforward question which countries can be grouped and covered by a common insurance product (annuity for instant). Therefore, it is desirable to divide heterogeneous risk communities into smaller parts that can be considered homogeneous. These homogeneous groups could be compared with numerous statistical methods to test the dissimilarity in certain aspects.

In our study we focus on mortality analysis, where several methods are available for comparing and/or testing the equivalence of two populations. However, it seems that there is a lack of results concerning the application of clustering methods on mortality data.

Mortality of a community can be described in multiple ways. For countries the age specific probability of death is the most common way of representing mortality profile. In case of institutional data Kaplan-Meier survival function estimation is used [3], which is a nonparametric method. Parametric methods can also be applied. Very often some functional forms are assumed for the hazard or intensity function. The classical form is Makeham mortality law which appears also in smoothing crude death probabilities.

In cluster analysis it is a key point how to define the distance or dissimilarity measure between instances. Arató et al. [1] describe dissimilarity measures between mortality tables but these measures were applied in other context. In this paper we adjusted Majstorović et al. [5] methodology to survival curves, so we chose a parametric method for estimation to get a smooth survival function, thus a similarity measure described in [7] can be used.

The next step is to choose a clustering algorithm appropriate for the chosen similarity measure. We decided to use spectral methods which give possibility to detect clusters by analyzing the eigenvectors and eigenvalues of a special matrix associated to a similarity graph constructed from the given set of curves by using an appropriate similarity measure [4]. The main advantages of spectral clustering methods are simplicity of implementation and their ability to cope with arbitrary input data.

In this paper we used spectral methods to cluster survival curves estimated for 23 out of 28 European Union countries and we compared them with some other clustering methods such as k-means and k-median. We partitioned the curves into $2 \leq k \leq 6$ clusters and found the most appropriate one by using some standard internal clustering validation measures. We compared k-partitions of EU countries obtained from clustering survival and hazard curves.

2 SURVIVAL CURVE CLUSTERING PROBLEM

2.1 Survival and hazard function

Let T be a nonnegative continuous real-valued random variable which denotes a person's time of death and let F denote a cumulative distribution function of T on the interval $[0, +\infty)$. The survival function $S(t)$ gives the probability that the investigated person did not die until age t, that is

$$S(t) = P(\{T > t\}) = 1 - F(t).$$

The hazard function (or intensity) is defined as

$$\mu(t) = -\frac{dS}{dt}\frac{1}{S(t)}. \tag{1}$$

The most common way for estimating the survival function is the Kaplan-Meier method ([3]), which is a nonparametric method and the estimated survival function is stepwise function. There are numerous parametric methods, too [8]). We choose the Makeham mortality law [6] which assumes that the force of mortality $\mu(x)$ at attained age x during calendar year t has the following form

$$\mu(x; \mathbf{q}) = \alpha e^{\beta x} + \lambda, \tag{2}$$

where $\mathbf{q} = (\alpha, \beta, \lambda)$ is a vector of constants, i.e. parameters that capture the essential properties of the progression of mortality. The survival function for Makeham mortality law is

$$S(x; \mathbf{q}) = e^{-\lambda x - \frac{\alpha}{\beta}(e^{\beta x} - 1)}, \tag{3}$$

and it is a solution of the ordinary differential equation (1) with the initial condition $S(0;\mathbf{q}) = 1$.

2.2 Survival curve clustering: basic idea

Survival curve clustering is a problem of partitioning a set of survival curves into k subsets called clusters, $1 \leq k \leq m$, in such a way that curves in the same cluster are in some sense more similar to each other (they are closest to each other) than to those in other clusters.

Let $\mathcal{S} = \{S(t\,;\,\mathbf{q}_i) \,:\, i = 1,\ldots,m\}$ be the set of survival curves defined with (3). If we wish to apply some standard clustering algorithm on $\mathcal{S}$ such as k-means, then we need to choose k arbitrary distinct initial curve-centers $s(t;c_1),\ldots,s(t;c_k)$ and assign each curve from $\mathcal{S}$ to the curve-center to which it is at least dissimilar. After the k-partition is constructed, for each cluster we need to update its curve-center by solving global optimization problem on the space of parameters and repeat the assignment and updating process until new centers are no different from the ones obtained in the previous step. Since the global optimization problem in the update phase is very difficult to solve, we can avoid it by using spectral clustering methods which allow us to make a new representation of $\mathcal{S}$ so that any simple clustering algorithm such as k-means can be applied.

3 SPECTRAL APPROACH FOR SURVIVAL CLUSTERING PROBLEM

To apply spectral methods in clustering survival curve set $\mathcal{S}$, first we need to choose some similarity function on $\mathcal{S}$ and then construct similarity graph G with vertex set $\mathcal{S}$. Vertices u and v of G are connected by an edge in if and only if some condition on the value of similarity between the corresponding data elements is satisfied.

Since we wish to compare survival velocities, we choose cosine similarity measure introduced by Sangalli et al. [7]. The cosine similarity function $c : \mathcal{S}\times\mathcal{S} \to \mathbb{R}$ is defined as

$$c(s(t;\mathbf{q}_i), s(t;\mathbf{q}_j)) = \int_a^b \frac{s'(t;\mathbf{q}_i)\cdot s'(t;\mathbf{q}_j)}{\sqrt{\int_a^b (s'(t;\mathbf{q}_i))^2\,dt}\cdot\sqrt{\int_a^b (s'(t;\mathbf{q}_j))^2\,dt}}\,dt, \quad 0 \leq a < b < \infty.$$

This measure is applicable on $\mathcal{S}$ since $ds(t\,;\,\mathbf{q}_i)/dt$ is square-integrable function on an arbitrary segment $[a,b]$, $a,b \in \mathbb{R}$, $a < b$. Now, the edge that connects vertices $s(t;\mathbf{q}_i)$ and $s(t;\mathbf{q}_j)$ in G is weighted with $w_{ij} = c(s(t;\mathbf{q}_i), s(t;\mathbf{q}_j))$.

After the similarity graph is constructed, the curve clustering problem becomes graph partitioning problem that is well known in the literature under the name min k-cut problem [4]. There are many options for finding its solution and we choose spectral approach that uses random walk normalized Laplacian matrix $\mathbf{L}_{rw}$ associated to a similarity graph G. Matrix $\mathbf{L}_{rw}$ is defined as

$$\mathbf{L}_{rw} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W},$$

where $\mathbf{W}$ is weighted adjacency matrix of G and $\mathbf{D}^{-1}\mathbf{W}$ is the transition matrix of a random walk on G. Spectral clustering approach uses $\mathbf{L}_{rw}$ to make a new point-based representation of each curve in $\mathcal{S}$ which belongs to a k-dimensional space, where k is the number of clusters [4].

Spectral survival curve algorithm (`SSCA`)

INPUT: `Set of survival curves` $\mathcal{S}$, `number` k `of clusters.`

1. `Construct the similarity graph` G `associated to` $\mathcal{S}$;
2. `Compute the matrix` L_{rw};
3. `Compute` k `eigenvectors` $u_1, \ldots, u_k$ `of` L_{rw} `which correspond to the` k `smallest eigenvalues and place them as columns in matrix` $U \in \mathbb{R}^{m \times k}$.
4. `Cluster the points` $(x_i)_{i=1,\ldots,m}$ `in` $\mathbb{R}^k$, `which are vectors corresponding to the rows of` U, `with` k`-means algorithm, into clusters` $C_1, \ldots, C_k$.

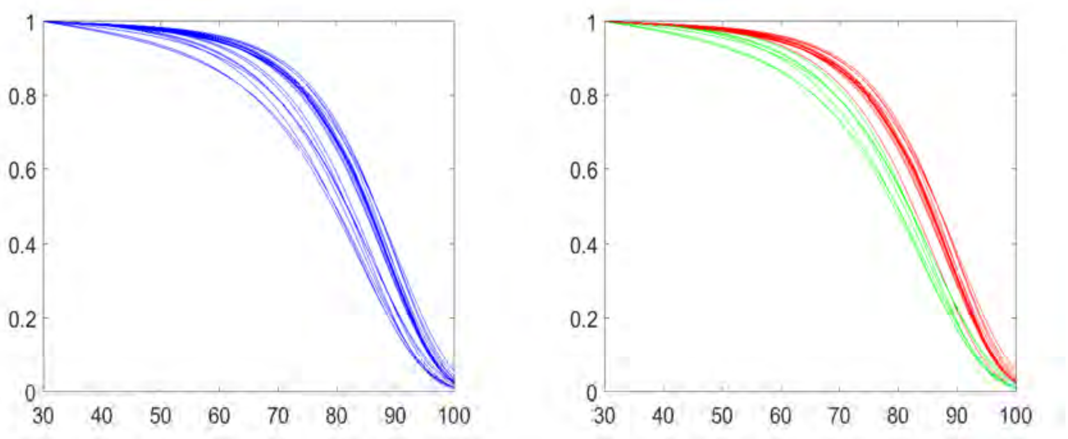
OUTPUT: `Clusters` $C_1, \ldots, C_k$.

---

4 NUMERICAL RESULTS

In this section we apply the method described in Section 3. We collected data for 23 out of 28 EU countries from the Human Mortality Database ([2]). We chose the unisex tables for year 2014. We fitted Makeham mortality law for ages 30-100 where we used `R` statistical software (version 3.5.1) and package `MortalityLaws` ((version 1.7.0). We obtained parameters based on function `MortalityLaws`. From the available options we chose loss function `binomialL` since it is more sensitive to the population size.

The set $\mathcal{S}$ of survival curves is illustrated in Figure 1(a). Plot indicates that the clustering structure seems obvious for the case of $k = 2$ clusters, while for $k \geq 3$ the k-partition is not predictable. This is justified by using `SSCA` algorithm from Section 3 which was implemented in `Matlab` programming platform (version R2016a). We considered k-

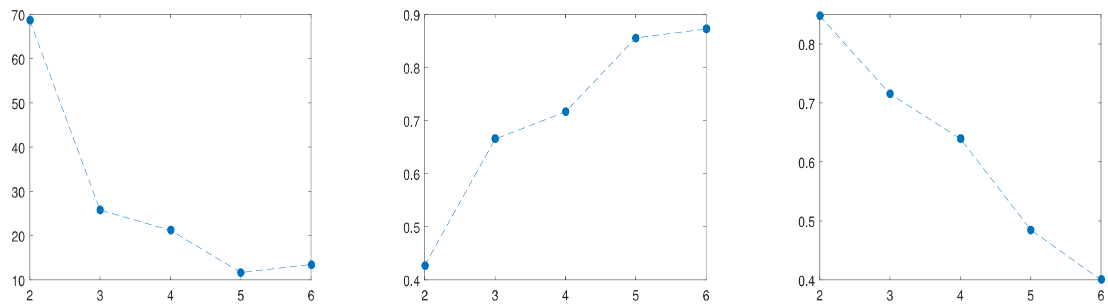


(a) (b)

Figure 1: **a** Set of survival curves for *EU* countries, **b** 2-clustering of survival curves (one color per cluster)

partitions of $\mathcal{S}$, with $2 \leq k \leq 6$. In each case k-means was ran up to 100000 times, while

the maximum number of iterations was set to 1000. Results for some choices of k are presented in in Table 1.
For $k = 2$ SSCA algorithm gives stable results, while for higher number of clusters the results are inconsistent and unstable. Similar results are obtained by using k-median algorithm. To measure the goodness of a k-clustering structure, $2 \leq k \leq 6$, we used Calinski-Harabasz, Davies-Bouldin and Silhouette (see Figure 2). Because of a tight



(a) (b) (c)

Figure 2: **a** Calinski-Harabasz (CH), **b** Davies-Bouldin (DB) and **c** Silhouette (S) values for $2 \leq k \leq 6$ clusters. High values of CH and S indices indicates that the set $\mathcal{S}$ of curves has been appropriately clustered. For DB index, lower value indicates better partition.

k	k-partition
2	{{AUT,BEL,DEU,DNK,ESP,FIN,SVN,NLD,SWE,IRL,ITA,LUX,FRA,PRT,GBR}, {CZE,EST,SVK,LTU,LVA,HRV,HUN,POL}}
3	{{AUT,DEU,ESP,FIN,SVN,NLD,SWE,IRL,ITA,FRA,PRT}, {BEL,DNK,LUX,GBR}, {CZE,EST,SVK,LTU,LVA,HRV,HUN,POL}}
5	{{BEL,DEU,DNK,ESP,FIN},{CZE,EST,SVN,NLD,IRL,FRA,} {AUT},{SVK,LTU,LVA,HRV,HUN,POL},{SWE,ITA,LUX,PRT,GBR}}

Table 1: K-partitions of survival curves

connection between survival and hazard function, we also performed spectral clustering on hazard functions into $2 \leq k \leq 6$ clusters and compared the obtained k-partitions with the ones obtained for survivals curves. For comparison, we used the adjusted Rand index [5], see Table 2.

	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
ARI	0.6646	0.7233	0.7233	0.7470	0.0907

Table 2: Values of ARI for k-partitions of survival and hazard curves. High values of ARI indicates the higher similarity between partitions.

5 CONCLUSION

In this paper we developed a cluster analytical method for mortality data. We fitted Makeham mortality law for the data and calculated the survival function based on this law.

We defined a similarity measure between the survival curves and used spectral clustering method for grouping these survival curves.

We applied survival curve clustering method for 23 EU countries for varying number of clusters, where we also investigated the optimal number of clusters. All the used clustering criteria resulted in 2 clusters, which seem to match the empirical assumptions: Western European countries belong to the first cluster and Central and Eastern European countries form the second cluster (Austria and Slovakia are the representative members of the two groups). This grouping structure proves to be stable, however forming more clusters the results are less stable and the homogeneity of clusters is not increased.
Our future work is based on using or developing some new approches to survival curve clustering problem, as well as using some other methods for modelling the survival function.

Acknowledgements. This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00017, titled "Sustainable, intelligent and inclusive regional and city models".

References

[1] Arató, M., Bozsó, D., Elek, P. and Zempléni, A. (2009). Forecasting and simulating mortality tables. Mathematical and Computer Modelling, Vol. 49, pp. 805–813.

[2] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [22/03/2019]).

[3] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, Vol. 53, No 282, 457–481.

[4] Luxburg, U (2007). A tutorial on spectral clustering. Statistics and Computing, Vol. 17, No 4, 395–416.

[5] Majstorović, S., Sabo, K., Jung, J. and Klarić, M. (2018). Spectral methods for growth curve clustering. Central European Journal of Operations Research, vol. 26 (3). pp. 715–737.

[6] Makeham, W. M. (1867). On the law of mortality. Journal of the Institute of Actuaries (1866),pp 325–358.

[7] Sangalli, L. M., Secchi P, Vantini S. and Vitelli V. (2010). K-mean alignment for curve clustering. Computational Statistics and Data Analysis. vol. 54. pp. 1219–1233.

[8] Tabeau, E. (2001). A Review of Demographic Forecasting Models for Mortality. In: Tabeau E., van den Berg Jeths A., Heathcote C. (eds) Forecasting Mortality in Developed Countries. European Studies of Population, vol 9. Springer, Dordrecht.

ROBUST CLUSTERING IN SOCIAL NETWORKS

Immanuel Bomze, **Michael Kahr** and **Markus Leitner**
University of Vienna, Department of Statistics and Operations Research (ISOR),
Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria
immanuel.bomze@univie.ac.at, m.kahr@univie.ac.at, markus.leitner@univie.ac.at

Abstract: During the last decades the importance of considering data uncertainty in optimization problems has become increasingly apparent, since small fluctuations of input data may lead to comparably bad decisions in many practical problems when uncertainty is ignored. If the probability distribution of the uncertain data is not known (or cannot be sufficiently estimated), a common technique is to estimate bounds on the uncertain data (i.e., define uncertainty sets) and to identify optimal solutions that are robust against data fluctuations within these bounds. This approach leads to the robust optimization paradigm that allows to consider uncertain objectives and constraints [1].

Optimization problems where only the objective is uncertain arise, for instance, prominently in the analysis of social networks. This stems from the fact that the strength of social ties (i.e., the amount of influence individuals exert on each other) or the willingness of individuals to adopt and share information can, for example, only be roughly estimated based on observations. A fundamental problem arising in social network analysis regards the identification of communities (e.g., work groups, interest groups), which can be modelled as a Dominant Set Clustering Problem [5,6,7] which in turn leads to a Standard Quadratic Optimization Problems (StQP); see [2]. Here the link strengths enter the objective while the constraints are familiar probability constraints, so that they can be considered certain.

Hence we investigate data uncertainty in the objective function of StQPs, considering different uncertainty sets, and derive implications for the complexity of robust variants of the corresponding deterministic counterparts. We can show that considering data uncertainty in a StQP results in another StQP of the same complexity if ellipsoidal, spherical or boxed uncertainty sets are assumed [4]. Moreover we discuss implications when considering polyhedral uncertainty sets, and derive rigorous bounds for this case, based upon copositive optimization [3].

Keywords: graph clustering, community detection, dominant set, robust optimization, quadratic optimization

References

[1] Ben-Tal A, El Ghaoui L, Nemirovski AS (2009) Robust optimization. Princeton Series in Applied Mathematics (Princeton NJ: Princeton University Press).

[2] Bomze IM (1998) On standard quadratic optimization problems. Journal of Global Optimization, 13(4):369–387.

[3] Bomze IM (2012) Copositive optimization – Recent developments and applications. European Journal of Operational Research, 216(3):509–520.

[4] Bomze IM, Kahr M, Leitner M. (2018) Trust your data or not - StQP remains StQP: Community Detection via Robust Standard Quadratic Optimization. Submitted, available at http://www.optimization-online.org/DB_HTML/2018/04/6586.html

[5] Pavan M, Pelillo M (2007) Dominant sets and pairwise clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(1):167–172.

[6] Rota Bulò S, Pelillo M (2017) Dominant-set clustering: A review. European Journal of Operational Research 262(1):1–13.

[7] Rota Bulò S, Pelillo M, Bomze IM (2011) Graph-based quadratic optimization: A fast evolutionary approach. Computer Vision and Image Understanding 115(7):984–995.

HARDNESS OF MINIMUM ACTIVATION PATH*

Sergio Cabello
University of Ljubljana and IMFM, Slovenia

Éric Colin de Verdière
Université Paris-Est, LIGM, CNRS, ENPC, ESIEE Paris, UPEM, Marne-la-Vallée, France

Abstract: In the MINIMUM ACTIVATION PATH problem, we are given a graph G with edge weights $w(\cdot)$ and two vertices s, t of G. We want to assign a non-negative power $p\colon V \to \mathbb{R}_{\geq 0}$ to the vertices of G, so that the activated edges $\{uv \in E(G) \mid p(u) + p(v) \geq w(uv)\}$ contain some s-t-path, and minimize the sum of assigned powers. We show that the problem is weakly NP-hard.

Keywords: activation path, activation network, NP-hardness, barrier problem

1 Introduction

In this work we show that the following optimization problem is weakly NP-hard.

> MINIMUM ACTIVATION PATH.
> Input: a graph $G = (V, E)$ with positive edge weights $w\colon E \to \mathbb{R}_{>0}$ and two vertices s and t of G.
> Task: find an assignment of ***powers*** $p\colon V \to \mathbb{R}_{\geq 0}$ to the vertices such that its ***cost*** $\sum_{v \in V} p(v)$ is minimum and the ***active edges*** $E(p) = \{uv \in E \mid p(u) + p(v) \geq w(uv)\}$ contain an s-t-path.

The MINIMUM ACTIVATION PATH problem is considered by Panigrahi [10], where an algorithm with running time $O(\text{poly}(n, |D|))$ is given. Here n is the number of vertices of G and D is a set of possible values for the power assignments (the *domain*). Thus, in that scenario one considers power assignments $p\colon V \to D$.

Our result is consistent with the result of Panigrahi. In our reduction, we use large *integer* weights: they have a polynomial bit length, but they are exponentially large. Taking $D = \mathbb{Z} \cap \max_{uv} w(uv)$ in the algorithm of Panigrahi, one only gets a pseudopolynomial time algorithm for such instances. This is consistent with a weakly NP-hardness proof.

There is a rich literature on so-called Activation Network problems. The task is to assign powers $p(v)$ to the vertices of an edge weighted graph $G = (V, E)$ so that the active edges $E(p) = \{uv \in E \mid p(u) + p(v) \geq w(uv)\}$ satisfy certain property, such as for example spanning the whole graph. See the survey by Nutov [9] for an overview of the area. It seems that the computational complexity of the MINIMUM ACTIVATION PATH problem has not been set before.

Another motivation to consider the computational complexity of the problem is its relation to an optimization problem considered by Cabello et al. [3]. Let x and y be points in the plane, and let $\mathbb{S}$ be a family of shapes in the plane. An x-y *curve* is a curve in $\mathbb{R}^2$ with endpoints x and y. We say that $\mathbb{S}$ ***separates*** x and y if each x-y curve intersects some shape from $\mathbb{S}$. Let $D(c, r)$ denote the *open* disk centered at c with radius r. Consider the following problem, which is a variant of the problem considered in [3].

*Research partially supported by the Slovenian Research Agency (P1-0297, J1-8130, J1-8155). Part of the research was done while Sergio was invited professor at Université Paris-Est.

MINIMUM BARRIER SHRINKAGE.
Input: a family $\{D(c_i, r_i) \mid i = 1, \dots, n\}$ of n open disks and two points $x, y \in \mathbb{R}^2$.
Task: find values $\delta_1, \dots, \delta_n \geq 0$ such that $\sum_i \delta_i$ is minimum and the family of open disks $\{D(c_i, r_i - \delta_i) \mid i = 1, \dots n\}$ does not separate x and y in $\mathbb{R}^2$.

The MINIMUM BARRIER SHRINKAGE and the MINIMUM ACTIVATION PATH seem related because of the following classical property: in a planar graph $G = (V, E)$, a set of edges $F \subseteq E$ is a minimum s-t cut if and only if the dual edges $\{e^* \mid e \in F\}$ form a shortest cycle separating the face s^* from the face t^* in the dual graph G^*. In fact, it seems that our proof of NP-hardness for the MINIMUM ACTIVATION PATH problem can be adapted to show the NP-hardness of the MINIMUM BARRIER SHRINKAGE.

The MINIMUM BARRIER SHRINKAGE is in turn related to the MINIMUM BARRIER RESILIENCE problem introduced by Kumar et al. [8]. In this problem the task is to find an x-y path that touches as few disks of $\{D(c_i, r_i) \mid i = 1, \dots, n\}$ as possible, without counting multiplicities. The MINIMUM BARRIER RESILIENCE problem models the number of total failures of the sensors without compromising the possibility to detect movement along x-y paths. In contrast, the MINIMUM BARRIER SHRINKAGE problem models the fact that sensors are less reliable when we move away from their placement. Despite the claim in the preliminary version [7] of [8], we do not know whether the MINIMUM BARRIER RESILIENCE problem can be solved exactly in polynomial time. See [1, 6, 11] for proofs of NP-hardness for non-disk shapes, and [2, 4, 6] for approximation algorithms.

2 Greedy solution in a path

Consider a graph G and a path $\pi = v_0, \dots, v_n$ in G. We define greedily a power assignment p^*_π on the vertices of G to activate π, in a way that power is pushed forward along π as much as possible. Formally, the ***greedy power assignment along*** π is

$$p^*_\pi(v) = \begin{cases} 0 & \text{if } v \text{ does not belong to } \pi \text{ or } v = v_0, \\ 0 & \text{if } v = v_i,\, i > 0 \text{ and } p^*_\pi(v_{i-1}) \geq w(v_{i-1}v_i), \\ w(v_{i-1}v_i) - p^*_\pi(v_{i-1}) & \text{if } v = v_i,\, i > 0 \text{ and } p^*_\pi(v_{i-1}) < w(v_{i-1}v_i). \end{cases} \quad (1)$$

For a power assignment p, let $\operatorname{cost}(p)$ denote the total cost of p, namely, the sum of the powers at the vertices. For path π, let $\operatorname{opt}(\pi)$ be the cost of the minimum cost power assignment that activates π. The following lemma tells that the greedy power assignment along π has minimum cost to activate π.

Lemma 1. *For each path* π, $\operatorname{cost}(p^*_\pi) = \operatorname{opt}(\pi)$.

Proof. It is clear that p^*_π activates all the edges of π. Let p be another power assignment activating all edges of π. We have to show that $\operatorname{cost}(p^*_\pi) \leq \operatorname{cost}(p)$.

We can assume that $p(v) = 0$ at all vertices v outside π. Otherwise, we change p to have this property. This reassignment of power would decrease the cost and would keep activating the path π.

The strategy is to gradually transform p into p^*_π while keeping all edges of π activated and without increasing the value of $\operatorname{cost}(p)$. The property $\operatorname{cost}(p^*_\pi) \leq \operatorname{cost}(p)$ is trivially correct if $p = p^*_\pi$. So assume $p \neq p^*_\pi$ and let i be the smallest integer such that $p(v_i) \neq p^*_\pi(v_i)$. Because all edges are activated, and by construction of p^*_π, we must have $p(v_i) > p^*_\pi(v_i)$. Let $\Delta = p(v_i) - p^*_\pi(v_i) > 0$. There are two cases:

- Assume $i \leq n - 1$. Update p by decreasing $p(v_i)$ by Δ and increasing $p(v_{i+1})$ by Δ. Since each edge of π is activated by p^*_π and by p before this transformation, each edge of π is still activated by the new p. Moreover, $\operatorname{cost}(p)$ is unchanged.

- Assume $i = n$. Update p by decreasing $p(v_n)$ by Δ. Again, each edge of π is still activated. The cost has decreased by Δ.

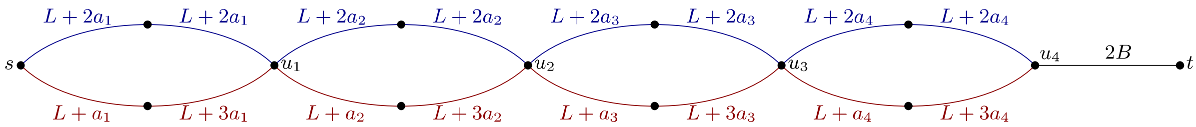


Figure 1: The graph G when $n = 4$.

This transformation does not increase the value of $\mathrm{cost}(p)$. Moreover, the new power assignment coincides with p^*_π on vertices $v_0, \ldots, v_i$. Thus, after a finite number of steps, $p = p^*_\pi$. This proves the lemma.

For the path $\pi = v_0, \ldots, v_n$, let $\varphi(\pi) = p^*_\pi(v_n)$. That is, the power assignment given by the greedy power assignment along π to the final vertex. Since $p^*_\pi(v_n)$ depends on $p^*_\pi(v_{n-1})$, we have the following.

Lemma 2. *Let π be the path $v_0, \ldots, v_n$ and let π' be the path $v_0, \ldots, v_n, u$. (Thus, π' extends π by an additional edge $v_n u$.) Then $\varphi(\pi') = \max\{0, w(v_n u) - \varphi(\pi)\}$ and $\mathrm{opt}(\pi') = \mathrm{opt}(\pi) + \varphi(\pi')$.*

Proof. From the definition of the greedy power assignment along π and π', the power assignments p^*_π and $p^*_{\pi'}$ differ only at vertex u, where we have $p^*_\pi(u) = 0$ and

$$\begin{aligned} p^*_{\pi'}(u) &= \max\{0, w(v_n u) - p^*_{\pi'}(v_n)\} = \max\{0, w(v_n u) - p^*_\pi(v_n)\} \\ &= \max\{0, w(v_n u) - \varphi(\pi))\}. \end{aligned}$$

From this we get the claim for $\varphi(\pi')$. Because of Lemma 1 for π and π' we also get

$$\begin{aligned} \mathrm{opt}(\pi') &= \mathrm{cost}(p^*_{\pi'}) = \mathrm{cost}(p^*_\pi) + p^*_{\pi'}(u) - p^*_\pi(u) \\ &= \mathrm{opt}(\pi) + \varphi(\pi') - 0. \end{aligned}$$

3 The reduction

Now we provide the reduction. The reduction is inspired by the reduction used to show that the restricted shortest path problem is NP-hard; this seems to be folklore and attributed to Megiddo by Garey and Johnson [5, Problem ND30]. We use the notation $[n] = \{1, \ldots, n\}$ and reduce from the following problem.

> SUBSET SUM
> Input: a sequence $a_1, \ldots, a_n$ of positive integers and a positive integer B.
> Question: is there a set of indices $I \subseteq [n]$ such that $\sum_{i \in I} a_i = B$?

The problem SUBSET SUM is one of the standard weakly NP-hard problems that can be solved in pseudopolynomial time via dynamic programming [5, Section 4.2]. In particular, when the numbers a_i are bounded by a polynomial in n, the problem can be solved in polynomial time.

Set L to be an integer strictly larger than $2\sum_{i \in [n]} a_i$. Then, for each $I \subseteq [n]$ we have $2\sum_{i \in I} a_i < L$.

We construct a graph $G = G(a_1, \ldots, a_n, B)$ as follows (see Figure 1). G will include vertices s, t, $u_1, \ldots, u_n$. Let us use the notation $u_0 = s$. For each $i \in [n]$, we put between u_{i-1} and u_i two 2-edge paths, one path with weights $L + 2a_i$ and $L + 2a_i$, and the other path with weights $L + a_i$ and $L + 3a_i$, as we go from u_{i-1} to u_i. Finally, we put the edge $u_n t$ with weight $2B$. This finishes the construction of G.

Lemma 3. *There exists a path π from s to u_n in G with $\mathrm{opt}(\pi) = c$ and $\varphi(\pi) = r$ if and only if there exists $I \subseteq [n]$ such that*

$$r = 2\sum_{i \in I} a_i \quad \textit{and} \quad c = nL + 2\sum_{i \in [n]} a_i + \sum_{i \in I} a_i.$$

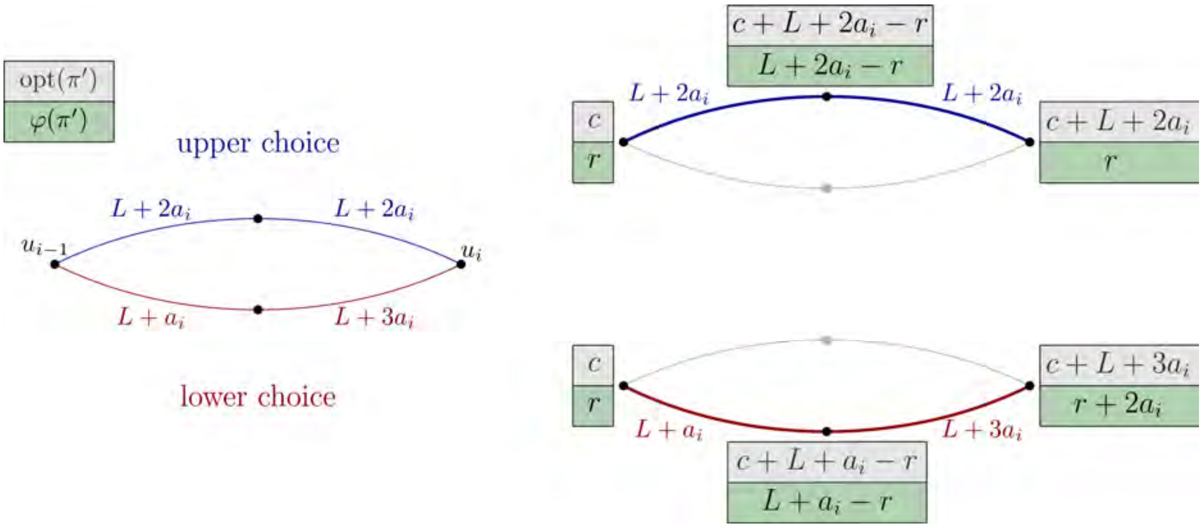


Figure 2: Left: Upper and lower choice at i. Right: the change in $\mathrm{opt}(\pi')$ and $\varphi(\pi')$ depending on whether the path is extended by the upper or the lower choice.

Proof. Consider the two 2-paths connecting u_{i-1} to u_i. We refer to the path with weights $L + 2a_i$ the *upper choice at* i, and the path with weights $L + a_i$ and $L + 3a_i$ the *lower choice at* i. See Figure 2.

Assume that we have a path π' that goes from $s = u_0$ to u_{i-1} with $\varphi(\pi') \leq L$. Let π'_u be the concatenation of π' with the upper choice, and let π'_ℓ be the concatenation of π' with the lower choice. Because of Lemma 2, we obtain that $\mathrm{opt}(\pi'_u) = \mathrm{opt}(\pi') + L + 2a_i$ and $\varphi(\pi'_u) = \varphi(\pi')$, while $\mathrm{opt}(\pi'_\ell) = \mathrm{opt}(\pi') + L + 3a_i$ and $\varphi(\pi_\ell) = \varphi(\pi') + 2a_i$. See Figure 2. Here, the assumption $\varphi(\pi') \leq L$ has been important to ensure that in using Lemma 2 the maximum defining $\varphi(\cdot)$ is not at 0. It easily follows by induction on i that, for each path π' from $s = u_0$ to u_i, we indeed have $\varphi(\pi') \leq \sum_{j=1}^{i} 2a_i$, and thus the hypothesis is fulfilled for each $i \in [n]$.

The intuition here is that the lower choice has a larger cost, but keeps more power at the extreme of the prefix path for later use.

Consider now a path π from $s = u_0$ to u_n. Let I be the set of indices $i \in [n]$ where the path takes the lower choice at i. From the previous discussion and a simple induction we have

$$\mathrm{opt}(\pi) \;=\; \sum_{i\in[n]\setminus I} (L + 2a_i) + \sum_{i\in I} (L + 3a_i) \;=\; nL + \sum_{i\in[n]} 2a_i + \sum_{i\in I} a_i$$

and

$$\varphi(\pi) \;=\; 2\sum_{i\in I} a_i \;\leq\; L.$$

Since all the paths from s to u_n must follow the upper or lower choice at each $i \in [n]$, the result follows.

Lemma 4. *For any real numbers A and B we have*

$$A + \max\{2B - 2A, 0\} \leq B \quad\Longrightarrow\quad A = B.$$

Proof. If $A \leq B$, then $B - A \geq 0$ and the assumption implies $A + (2B - 2A) \leq B$, which implies $B \leq A$, and thus $A = B$. If $A > B$, then $B - A < 0$ and the assumption implies $A + 0 \leq B$, which implies $A \leq B$. Thus this cannot happen.

Theorem 5. *The problem* MINIMUM ACTIVATION PATH *is NP-hard.*

Proof. We show that the instance for SUBSET SUM has a positive answer if and only if in the graph $G = G(a_1, \ldots, a_n, B)$ there is a power assignment with cost at most $C := nL + 2\sum_{i\in[n]} a_i + B$ that activates some path from s to t.

Assume that the exists a solution for the instance to the SUBSET SUM problem. This means that we have some $I \subseteq [n]$ such that $\sum_{i\in I} a_i = B$. Because of Lemma 3, there exists a path π from $s = u_0$ to u_n with optimal activation cost $\mathrm{opt}(\pi) = nL + 2\sum_{i\in[n]} a_i + \sum_{i\in I} a_i = C$ and $\varphi(\pi) = 2B$. Because of Lemma 1, this means that the power assignment p^*_π has cost $\mathrm{cost}(p^*_\pi) = C$, activates all edges of π, and assigns power $p^*_\pi(u_n) = \varphi(\pi) = 2B$ to vertex u_n. Such power assignment p^*_π also activates the edge $u_n t$ because it has weight $2B = p^*_\pi(u_n)$. (In particular, the vertex t gets power 0.)

Assume now that there is a power assignment $p' \geq 0$ with cost at most C that activates a path π' from s to t . Let π be the restriction of π' from s to u_n. Because of Lemma 2 and using that the power assignment p' activates π', we have

$$\mathrm{opt}(\pi) + \max\{2B - \varphi(\pi), 0\} \;=\; \mathrm{opt}(\pi') \;\leq\; \mathrm{cost}(p') \;\leq\; C. \qquad (2)$$

Because of Lemma 3, there exists some $I \subseteq [n]$ such that

$$\mathrm{opt}(\pi) \;=\; nL + 2\sum_{i\in[n]} a_i + \sum_{i\in I} a_i \quad \text{and} \quad \varphi(\pi) \;=\; 2\sum_{i\in I} a_i.$$

Substituting in (2), for such $I \subseteq [n]$ we have

$$nL + 2\sum_{i\in[n]} a_i + \sum_{i\in I} a_i + \max\left\{2B - 2\sum_{i\in I} a_i, 0\right\} \;\leq\; C \;=\; nL + 2\sum_{i\in[n]} a_i + B.$$

This means that

$$\sum_{i\in I} a_i + \max\left\{2B - 2\sum_{i\in I} a_i, 0\right\} \;\leq\; B.$$

Because of Lemma 4 we conclude that $\sum_{i\in I} a_i = B$, and the given instance to SUBSET SUM problem has a solution.

References

[1] Helmut Alt, Sergio Cabello, Panos Giannopoulos, and Christian Knauer. Minimum cell connection in line segment arrangements. *Int. J. Comput. Geometry Appl.*, 27(3):159–176, 2017. URL: https://doi.org/10.1142/S0218195917500017.

[2] Sergey Bereg and David G. Kirkpatrick. Approximating barrier resilience in wireless sensor networks. In Shlomi Dolev, editor, *Algorithms for Sensor Systems, ALGOSENSORS 2009*, volume 5804 of *Lecture Notes in Computer Science*, pages 29–40. Springer, 2009. URL: https://doi.org/10.1007/978-3-642-05434-1_5.

[3] Sergio Cabello, Kshitij Jain, Anna Lubiw, and Debajyoti Mondal. Minimum shared-power edge cut. *CoRR*, abs/1806.04742, 2018. URL: http://arxiv.org/abs/1806.04742, arXiv:1806.04742.

[4] David Yu Cheng Chan and David G. Kirkpatrick. Approximating barrier resilience for arrangements of non-identical disk sensors. In Amotz Bar-Noy and Magnús M. Halldórsson, editors, *Algorithms for Sensor Systems, ALGOSENSORS 2012*, volume 7718 of *Lecture Notes in Computer Science*, pages 42–53. Springer, 2013. URL: https://doi.org/10.1007/978-3-642-36092-3_6.

[5] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1979.

[6] Matias Korman, Maarten Löffler, Rodrigo I. Silveira, and Darren Strash. On the complexity of barrier resilience for fat regions and bounded ply. *Computational Geometry*, 72:34–51, 2018. doi:10.1016/j.comgeo.2018.02.006.

[7] Santosh Kumar, Ten H. Lai, and Anish Arora. Barrier coverage with wireless sensors. In *Proceedings of the 11th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 284–298. ACM, 2005. URL: http://doi.acm.org/10.1145/1080829.1080859, doi:10.1145/1080829.1080859.

[8] Santosh Kumar, Ten H. Lai, and Anish Arora. Barrier coverage with wireless sensors. *Wireless Networks*, 13(6):817–834, 2007. doi:10.1007/s11276-006-9856-0.

[9] Zeev Nutov. Activation network design problems. In Teofilo F. Gonzalez, editor, *Handbook of Approximation Algorithms and Metaheuristics, Second Edition*, chapter 15. CRC Press, 2018.

[10] Debmalya Panigrahi. Survivable network design problems in wireless networks. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1014–1027, 2011.

[11] Kuan-Chieh Robert Tseng and David Kirkpatrick. On barrier resilience of sensor networks. In Thomas Erlebach, Sotiris Nikoletseas, and Pekka Orponen, editors, *Algorithms for Sensor Systems, ALGOSENSORS 2011*, volume 7111 of *Lecture Notes in Computer Science*, pages 130–144. Springer, 2012.

ON THE COMPLEXITY OF A FILTERING PROBLEM FOR CONSTRAINT PROGRAMMING: DECOMPOSITION BY THE STRUCTURE OF PERFECT MATCHINGS

Radoslaw Cymer
Universität Augsburg, Germany
E-mail: r.cymer@web.de

Miklós Krész[1]
InnoRenew CoE
Livade 6, 6310 Izola, Slovenia
E-mail: miklos.kresz@innorenew.eu

Abstract: A complexity analysis based on the structure of perfect matchings is given for the most efficient basic filtering algorithms in constraint programming with respect to the role of edges in matchings.

Keywords: constraint programming,matching theory, decomposition theory

1 INTRODUCTION

In matching theory it is a basic problem to determine all the edges in a given graph which can be extended to a maximum matching. Such edges are called maximally matchable or allowed edges. Apart from the graph theory community (see e.g [11]), researchers in constraint programming have also investigated this problem (cf. [3,4, 12]). The motivation for studying the question from constraint programming point of view is originating from certain constraint propagation methods ([12]), where the applied filtering algorithmic scheme is based on the above question. In this paper we will study the efficient algorithms for perfect matchings only with respect to the above problem, which is related to the symmetric alldiff constraint introduced in [12]. However, as it was shown in [4], the scheme of constraint propagation based on perfect matchings can be extended to a more general framework.

As a main result of [4] a decomposition algorithm was worked out for identifying the allowed edges. In this paper we will give a detailed running time analysis for the decomposition algorithm presented in [4]. It turns out that the complexity bound given in that paper is not precise.

The organization of this paper is as follows. In Section 2 we will present the necessary formal background on matching theory. We collect here some basic material needed later on and include contents of almost all the required results. In Section 3 we analyze the iterative version of the algorithm to compute the category of edges into mandatory (covered by all perfect matchings), allowed and forbidden (i.e. not allowed). The obtained results are mainly based on the Structure Theorem of Gallai & Edmonds. In Section 4 an algorithm using divide-and-conquer paradigm is analyzed. Finally, in Section 5 we will give a short conclusion. Because of space constraints proofs are omitted.

2 MATCHING THEORY AND STRUCTURAL DECOMPOSITION

In this paper we will consider undirected general graphs and our main focus will be on graphs with perfect matchings. Our terminology will be standard, the set of vertices and set of edges will be denoted by *V(G)* and *E(G)*, respectively. . A good reference for any undefined terms is [10].

---

[1] Also at University of Primorska, Slovenia and University of Szeged, Hungary.

We will call an edge of a graph G *allowed* if it occurs in some maximum matching (respectively, perfect matching, if exists) and any edge which is not allowed will be called *forbidden*. An edge which belongs to every maximum (respectively, perfect) matching will be called *mandatory*.

A graph G with a perfect matching is said to be *elementary* if its allowed edges form a connected spanning subgraph of G. A *matching covered* graph is an elementray graph without forbidden edges.

Let M be a maximum matching of G. An edge $e \in E(G)$ is said to be *M-positive* if $e \in M$, otherwise e is called M-negative. An *M-alternating path* in G is a path stepping on M-positive and M-negative edges in an alternating fashion.

The deficiency of G denoted by $\delta(G)$ is defined as the number of vertices left unmatched by a maximum matching. A graph G is said to be *factor-critical* if $\delta(G-x)=0$ for every $x \in V(G)$. Maximum matchings of factor-critical graphs are called *near-perfect matchings*.

The bipartite graph $G = (V_1 \cup V_2, E)$ has positive surplus (viewed from V_1) if $|\Gamma(X)| > |X|$ for all $\emptyset \neq X \subseteq V_1$, where $\Gamma(X)$ denotes the set of neighbours of X. Bipartite graphs with positive surplus are connected.

For a general graph $G = (V, E)$ we define subsets $A(G)$, $C(G)$ and $D(G)$ of $V(G)$ as follows:

$D(G)$ = {the set of vertices in G not covered by at least one maximum matching of G},

$A(G) = \Gamma(D) \setminus D(G)$,

$C(G) = V(G) \setminus (A(G) \cup D(G))$.

The following theorem (see e.g. [10]) is fundamental in the structure theory of matchings.

Theorem 1 (Gallai-Edmonds Structure Theorem). *If G is a graph and $A(G)$, $C(G)$ and $D(G)$ are defined as above, then the following statements hold:*

1. *The components of the subgraph induced by $D(G)$ are factor-critical,*
2. *The subgraph induced by $C(G)$ has a perfect matching,*
3. *The bipartite graph obtained from G by deleting the vertices of $C(G)$ and the edges spanned by $A(G)$, and by contracting each component of $D(G)$ to a single vertex has positive surplus (when viewed from $A(G)$),*
4. *Every maximum matching of G contains a perfect matching of each component of $C(G)$, a near-perfect matching of each component of $D(G)$, and a complete matching from $A(G)$ into distinct components of $D(G)$.*

The decomposition has the following properties:

- Edges spanned by $A(G)$ are forbidden
- Edges connecting $A(G)$ to $C(G)$ are forbidden
- Every edge incident with a vertex of $D(G)$ is allowed
- There is no edge between $C(G)$ and $D(G)$
- Vertices of $A(G)$ and $C(G)$ are vital
- Vertices of $D(G)$ are allowed
- Each connected component of $C(G)$ has even cardinality
- Each connected component of $D(G)$ has odd cardinality

Note that every component of $G[C]$ has a perfect matching, the bipartite subgraph $G[A, base(D)]$ obtained from G by deleting edges spanned by $A(G)$ and by contracting each component of $D(G)$ to a single vertex has a complete matching from $A(G)$ to $D(G)$, and every connected component of $G[D]$ has a near-perfect matching.

A set X of vertices in G is *extreme* if $\delta(G-X) = \delta(G) + |X|$. In [4] a general pruning routine was introduced to aid in the investigation of extreme sets of graphs which have perfect matchings. It was also shown that finding an extreme set can be accomplished in linear time. The algorithm of [4] is an iterative graph decomposition method by which we can mark the forbidden edges with taking advantage of the structure provided by the decomposition. In each step maximal extreme sets of the subgraphs obtained by the previous steps of the algorithm are determined and some edges are marked as forbidden according to the structure defined by the extreme sets. Using the structure with respect to the extreme sets new subgraphs are defined and the iteration is continued.

The decomposition method is based on the following theorem [4].

Theorem 2. *Let $G = (V, E)$ be any graph with a perfect matching M, $x \in V$, and let (D, A, C) be the Gallai-Edmonds canonical decomposition of $G-x$. Then the following statements hold:*

1. *The set $X = A \cup \{x\}$ is extreme in G,*
2. *Edges spanned by X or joining X with C are forbidden,*
3. *The bipartite graph G_0 obtained from $G - C$ by contracting each connected component of D to a single vertex and by deleting each edge spanned by X has a perfect matching,*
4. *Edges belonging to all, some (but not to all) or none of perfect matchings in the bipartite graph G_0 are, respectively, edges belonging to all, some (but not to all) or none of perfect matchings in G.*
5. *The graph G_i obtained from $G - C$ by contracting the set $V(G) - D_i$ to a single vertex has a perfect matching,*
6. *The mandatory, allowed, or forbidden edges of G are precisely those edges which are, respectively, mandatory, allowed, or forbidden in one of the graphs G_i, $i = 0, \ldots, t$, where $t = |X|$.*

In the next two sections we will present and analyse the "pure iterative" and the "divide and conquer" strategy of the decomposition method. In Section 3 we will consider the naïve approach for identifying forbidden edges with the help of maximal extreme sets. In this case we will make use of the properties of maximal extreme sets as direct consequences of its definition. We will show that this algorithm provides an improved worst-case running time, then the previously developed methods [11,12] which could ensure $O(|V|\cdot|E|)$ only. In Section 4 the algorithm of [4] will be presented and by detailed running time analysis we will show that the divide and conquer strategy is more efficient than the pure iterative approach.

3 PURE ITERATIVE ALGORITHM

The following is the pure iterative version of the decomposition algorithm:

Algorithm 1 The iterative approach to determine the partition of edges

Require: General graph $G = (V, E)$ with an initial perfect matching M
Ensure: Partition of edges
while there are UNSCANNED vertices in G **do**
Select one UNSCANNED vertex x
Relabel x as SCANNED
Compute the maximal extreme set X in $G - x$ (see algorithm in [2, Section 2.1]) Let (A, C, D) be the Gallai-Edmonds Decomposition after the last step of the previous routine
Mark all edges spanned by X as forbidden
Form the gluing bipartite graph G_0 with bipartition $(X, base(D_i))$ Let $M_0 \leftarrow M \cap E(G_0)$
Determine the partition of edges in G_0 with respect to M_0 (s. Algorithm 2 in [3])
Mark all vertices of X as SCANNED
end while

The crucial point in the analysis of the above algorithm is the number of the required iterations. It is clear that the number of iterations are related to the number of maximal extreme sets.

However, according to [2, Theorem 2.1] the problem of finding the maximum Tutte set, and thus the maximum extreme set is intractable. This means that in general there may exist exponential number of maximal extreme sets. To demonstrate this fact with a very simple example consider the graph consisting of K_2 and *n* triangles *0* attached to one of its endpoints. It is easy to check that such a graph has 2^n maximal extreme sets. In order to overcome this drawback we introduce the following new concepts.

Definition 1 (Elementary extreme set). *An extreme set, such that each element of it belongs to the same elementary component.*

In general, the number of maximal elementary extreme sets in a non-elementary graph is lower than the number of maximal extreme sets in its elementary components. The similar holds true when we remove the forbidden edges: the number of extreme sets in matching covered graphs maybe greater than the number of extreme sets in elementary components. The following result about maximal elementary extreme sets was proved by Bartha & Krész [1].

Theorem 3. *Let G be a graph with a perfect matching. The maximal elementary extreme sets of G form a partition on $V(G)$.*

Therefore, extending the concept of [10] from elementary graphs, maximal elementary extreme sets will be called *canonical classes*. The set of all canonical classes will be denoted by *P(G)*. The following result from [13] clarifies the number of maximal extreme sets in a non-elementary graph.

Theorem 4. *Any maximal extreme set of a non-elementary graph is the union of some maximal elementary extreme sets***.**

The following concept is equivalent to the one of "strong proof" by Király [5].

Definition 2 (Extreme closure) *For any vertex u, the extreme closure of u, denoted by Ext(u), is the intersection of all maximal extreme sets containing u.*

As a straightforward consequence of Theorem 4, *Ext(u)* can be also characterized with canonical classes.

Proposition 1. *Ext(u) is a union of some canonical classes.*

Corollary 1. *$Ext(u)$ can be found in linear time for each $u \in V(G)$.*

Now let *Ext(G)* denote the set of distinct extreme closures of *G*.

Theorem 5. *There is a one-to-one correspondence between Ext(G) and P(G); consequently $|Ext(G)| = |P(G)|$ holds.*

During the algorithm we identify the forbidden edges spanned by *A(G)* and between *C(G)* and *A(G)*. Next we build the gluing bipartite graph and perform alternating depth-first search starting from the color class *X* and a free edge. It is necessary since considering only canonical classes is not satisfactory to find the partition of edges (it can happen that the forbidden edge lies between two different maximal elementary extreme sets).

Theorem 6. *Let* $p=|Ext(G)|$ and $m=|E(G)|$. Then *Algorithm 1 uses maximum p iteration steps, consequently the upper bound for the pure iterative algorithm is* $O(p \cdot m)$.

4 DIVIDE AND CONQUER STRATEGY

Recall that the algorithm developed in [4] is based directly on Theorem 2. It uses a divide-and-conquer paradigm which is a natural consequence of the result.

The procedure first constructs a perfect matching in a given general graph, then decomposes the graph, according to the Gallai-Edmonds Structure Theorem (Theorem 1) and successively

identifies allowed edges and eliminates forbidden edges reducing the remainder graph in a suitable way. The method is summarized as *Algorithm 2*.
The goal is now to determine the running time of this algorithm. For this first we review some concepts from [1].
Let C be an elementary component and M be a perfect matching. Then a *C-ear* is an M-alternating path α connecting two vertices of C such that no vertex of α, other than its endpoints, lies in C. It is easy to see that a C-ear starts and ends with an M-negative edge. Furthermore, it can be shown (cf. [1]) that the existence of a C-ear is independent from the choice of the matching M.
We say that elementary component C' is *two-way accessible* from component C, in notation $C\rho C'$, if C' is covered by a C-ear. It was shown in [1] that the reflexive and transitive closure ρ^* of ρ is a partial order on the set of elementary components. For a similar approach see also [6].
Now let us introduce some new concepts.

Definition 3. ***Let*** *$C_0, C_1, \dots, C_k$ be distinct elementary components of G such that $C_0\rho C_1\rho C_2\rho \dots \rho C_k$. Then we say that $(C_0, C_1, \dots, C_k)$ forms a ρ-chain. Moreover, let $P(C_0, C_1, \dots, C_k)$ denote the set of canonical classes of the elementary components forming the chain. The canonical length of $(C_0, C_1, \dots, C_k)$ is given by $|P(C_0, C_1, \dots, C_k)|-1$. Finally,* ***t****he canonical diameter of G is the maximum canonical length concerning all ρ-chains in G.*

Algorithm 2 The divide-and-conquer approach to determine the partition of edges

Require: General graph $G = (V, E)$ with an initial perfect matching M
Ensure: Partition of edges
if $|V| = 2$ **then** {base case}
  Mark vertices in V as SCANNED
  if $|E(G)| = 1$ **then**
    Mark edge in E as mandatory
  else
    Mark edge in E as allowed
  end if
  return
end if
Select one UNSCANNED vertex x
Relabel x as SCANNED
Compute the Gallai-Edmonds Decomposition (A, B, C, D) of $G-x$ (s. Algorithm 1 in [4])
Let $X \leftarrow A \cup \{x\}$ {extreme set}
Mark all edges spanned by X as forbidden
Mark all edges between X and C as forbidden
if $|C(G-x)| > 0$ **then**
  Find connected components $C_1, C_2, \dots, C_k$ of $G[C]$
  for every connected component C_i **do**
    Let $M_i \leftarrow M \cap E(C_i)$
    Recursive call of this procedure with $G = C_i$ and $M = M_i$
end for end if
Form the gluing bipartite graph G_0 with bipartition $(X, B \cup base(D_i))$
Let $M_0 \leftarrow M \cap E(G_0)$
Determine the partition of edges in G_0 (s. Algorithm 2 in [3])
Remove forbidden edges from G
Mark all vertices of $X \cup B$ as SCANNED
Mark all edges in G_0 as TRAVERSED
Mark vertices incident with all TRAVERSED edges as SCANNED
Let t be the number of connected components of $G[B \cup D]$, i.e. $t \leftarrow |X|$

if $|D(G-x)|>0$ **then**
  Form the pieces $G_1, G_2, \ldots, G_t$ of G at extreme set X
  for every piece G_i with at least one UNSCANNED vertex **do**
    Let $M_i \leftarrow M \cap E(G_i)$
    Recursive call of this procedure with $G = G_i$ and $M = M_i$
  end for
end if

Now using the above concepts, we are ready to give the complexity analysis of Algorithm 2.

Theorem 7. *Let m=|E(G)| and let λ denote the canonical diameter of graph G with perfect matchings. Then Algorithm 2 uses maximum λ iteration steps, consequently the upper bound for the divide and conquer strategy is O(λ·m).*

As a final result we show that the improvement of the complexity with the divide and conquer strategy can be expressed formally by the parameters used for the analysis of Algorithms 1 and 2.

Theorem 8. *Let p_{min} denote the cardinality solution of the minimum set cover for Ext(G), i.e. the minimum number of sets of Ext(G) the union of which covers V(G). Then $\lambda \leq p_{min}$, where λ denotes the canonical diameter of G.*

5 CONCLUSION

In this paper we have provided a detailed analysis of the state-of-the-art filtering algorithms for constraint propagation with respect to the role of edges in perfect matchings. We could characterized the worst-case complexity of both the pure iterative method and the divide-and-conquer strategy with graph parameters defined by the matching structure. With the help of this concept, we have shown formally that the divide-and-conquer strategy is indeed more efficient than the pure iterative algorithm.

Acknowledgement
This research was partially supported by the National Research, Development and Innovation Office-NKFIH (Hungary) Fund No. SNN-117879. The second author also acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program, the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European regional Development Fund), and the support of the ARRS grant N1-0093.

References

[1] Miklós Bartha, and Miklós Krész. *Structuring the elementary components of graphs having a perfect internal matching.* Theoretical Computer Science, 299(1-3):179-210, 2003.

[2] D. Bauer, H.J. Broersma, N. Kahl, A. Morgana, E. Schmeichel, T. Surowiec. *Tutte sets in graphs II: The complexity of finding maximum Tutte sets.* Discrete Applied Mathematics, 155(10):1336-1343, 2007.

[3] Radosław Cymer. *Dulmage-Mendelsohn Canonical Decomposition as a generic pruning technique.* Constraints, 17(3):234-272, 2012.

[4] Radosław Cymer. *Gallai-Edmonds Decomposition as a Pruning Technique.* Central European Journal of Operations Research, 23(1):149-185, 2015.

[5] Zoltán Király. *The Calculus of Barriers.* unpublished

[6] Nanao Kita. *A partially ordered structure and a generalization of the canonical partition for general graphs with perfect matchings.* Volume 7676 of the Lecture Notes in Computer Science, pages 85-94, 2012. Springer.

[7] Anton Kotzig. *On the Theory of Finite Graphs with a Linear Factor I.* Matematicko-Fyzikálny Časopis Slovenskej Akadémie Vied, 9(2):73-91, 1959. In Slovak.

[8] Anton Kotzig. *On the Theory of Finite Graphs with a Linear Factor II.* Matematicko-Fyzikálny Časopis Slovenskej Akadémie Vied, 9(2):136-159, 1959. In Slovak.

[9] Anton Kotzig. *On the Theory of Finite Graphs with a Linear Factor III.* Matematicko-Fyzikálny Časopis Slovenskej Akadémie Vied, 10(4):205-215, 1960. In Slovak.

[10] László Lovász and Michael D. Plummer. *Matching Theory.* Annals of Discrete Mathematics (29), North-Holland, Amsterdam, 1986.

[11] M. de Carvalho, J. Cheriyan, *An O(VE) algorithm for ear decompositions of matching-covered graph*s, ACM Transactions on Algorithms 1(2):324-337, 2005.

[12] J.C. Regin: *The Symmetric Alldiff Constraint*. IJCAI 1999: 420-425.

[13] Miklós Bartha, and Miklós Krész. *Splitters and barriers in Open graphs having a perfect internal matching*. Acta Cybernetica, 18:697-718, 2008.

FIVE HEURISTICS FOR THE k-MATCHING PROBLEM

Boštjan Gabrovšek
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
FMF, University of Ljubljana, Jadranska ulica 19, 1000 Ljubljana, Slovenia
bostjan.gabrovsek@fs.uni-lj.si
Tina Novak
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
tina.novak@fs.uni-lj.si
Janez Povh
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
janez.povh@fs.uni-lj.si
Darja Rupnik Poklukar
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
darja.rupnik-poklukar@fs.uni-lj.si
Janez Žerovnik
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
janez.zerovnik@fs.uni-lj.si

Abstract: We introduce five algorithms which give an approximate solution to the k-matching problem. The algorithms rely on iterating the Hungarian method over subgraphs of graphs obtained by contracting sub-matchings from previously known non-optimal solutions. Three of the algorithms are deterministic and two of them are parametric and incorporate randomization. We test the algorithms using two datasets and compare their performance timewise and qualitywise.

Keywords: k-matching problem, k-assignment problem, heuristic algorithm, Hungarian method

1 MOTIVATION

Let us consider multi-associations between the vertices in the sets $V_1, V_2, \ldots, V_k$. For example, in bioinformatics, V_1 can correspond to the set of known (relevant) diseases, V_2 to the set of known drugs and V_3 to the set of known genes. A Multi-association is a triple $(p_1, p_2, p_3) \in V_1 \times V_2 \times V_3$ which means that disease p_1, drug p_2 and gene p_3 are related. Such a triple may imply that the gene p_3 is activated in disease p_1 and is usually silenced by drug p_2, hence drug p_2 may be considered to cure disease p_1. This is related to the very vibrant area of drug re-purposing and precision medicine, see e.g. [8, 7, 11].

We can represent the data as a complete 3-partite graph where the vertex set is $V_1 \cup V_2 \cup V_3$ and the edges have weights equal to the strength of the association between the ending vertices. If we want to find a decomposition of this graph into disjoint triangles with maximum (minimum) total weight, we obtain the 3-assignment problem (3-AP) [1, 2, 3, 5]. The 3-AP can serve as a model also in production planning when we assign workers, machines and tasks in a way that each worker gets exactly one task at one machine and the total cost is minimum.

Next we describe the notation used, the algorithms and results. Note that this is a survey paper based on results in [6], where all details are provided.

2 PRELIMINARIES & NOTATION

Let $G = (V, E, w)$ be a complete weighted k-partite graph where $V = V_1 \cup V_2 \cup \cdots \cup V_k$ is the vertex set, V_i are vertices of the i-th partition with cardinality n, $E = \bigcup_{1 \le i < j \le k} \{uv \mid u \in V_i, v \in V_j\}$ is the edge set and $w : E \to \mathbb{R}$ is the weight function.

A *k-clique* is a subgraph of G, induced by a subset $Q \subset V$ with cardinality k, which is isomorphic to the complete graph K_k. We denote it by $G[Q]$. The *weight of* $G[Q]$, $w(Q)$, is the sum of the edge weights of $G[Q]$: $w(Q) = \sum_{e \in E(G[Q])} w(e)$.

A *disjoint k-clique system* of a complete k-partite graph is a set of n subsets

$$\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_n\}$$

inducing pairwise disjoint k-cliques, i.e., $G[Q_i]$ is a k-clique for every i and $Q_i \cap Q_j = \emptyset$, for all $i \neq j$. Clearly, the weight of $\mathcal{Q}$ is $w(\mathcal{Q}) := \sum_{i=1}^{n} w(Q_i)$.

Every disjoint k-clique system $\mathcal{Q}$ defines a set of non-overlapping k-associations, which we call k-matching or k-assignment and will denote it by the same letter $\mathcal{Q}$.

We can thus formulate the k-matching problem as finding

$$\min\{\sum_{\mathcal{Q}} w(\mathcal{Q}) \mid \mathcal{Q} \text{ is a } k\text{-matching}\}.$$

Given a 2-matching $M_{i,j}$ between partitions V_i and V_j, let $G/M_{i,j}$ be the graph obtained by contracting $M_{i,j}$. In detail, let $U' = \{\{u, v\} \mid uv \in M_{i,j}\}$ and $V'_\ell = \{\{v\} \mid v \in V_\ell\}$. The graph $G/M_{i,j} = (V', E', w')$ has the vertex and edge set as following

$$V' = U' \cup \bigcup_{\substack{1 \le l \le k \\ \ell \neq i, \ell \neq j}} V'_\ell, \qquad E' = F' \cup \bigcup_{\substack{1 \le i' < j' \le k \\ \{i', j'\} \cap \{i, j\} = \emptyset}} E_{i',j'}$$

where $F' = \{\{u\}\{v, v'\} \mid u \in V \setminus (V_i \cup V_j), vv' \in M_{i,j}\}$ and $E_{i',j'} = \{\{u\}\{v\} \mid uv \in E \setminus M_{i,j}\}$.

The new weight function is defined by summing the weights on the edges adjacent to the identified vertices:

if $u, v \in V \setminus (V_i \cup V_j)$, we have

$$w'(\{u\}\{v\}) = w(uv)$$

and if $u \in V \setminus (V_i \cup V_j) \wedge vv' \in M_{i,j}$, we have

$$w'(\{u\}\{v, v'\}) = w(uv) + w(uv').$$

Let $M_{i,j}$ be an arbitrary matching between the i-th and j-th partition and let $\mathcal{M}'$ be a $(k-1)$-matching for the $(k-1)$-partite graph $G/M_{i,j}$. We denote by $\mathcal{M}' * M_{i,j}$ the unique k-matching for G reconstructed from $\mathcal{M}'$ and $M_{i,j}$, i.e.,

$$\mathcal{M}' * M_{i,j} = M_{i,j} \cup \{uv \mid \{u\}\{v\} \in \mathcal{M}'\} \cup \left(\bigcup_{\substack{\{u\}\{vv'\} \in \mathcal{M}' \\ vv' \in M_{i,j}}} \{uv, uv'\} \right).$$

Let A be a set and f a function $f : A \longrightarrow \mathbb{R}$. Denote by

$$\arg\min_{x \in A} f(x) = \{x \mid \forall y \in A : f(y) \ge f(x)\},$$

the set of minimal elements in A under the function f.

3 ALGORITHMS

In this section we describe the algorithm A1 from [9] and our algorithms B1–F1. The algorithm names are adopted from [9] and [6]. Recall that in the case $n = 2$ an optimal 2-matching can be found using the Hungarian algorithm [10]. The result of this algorithm on a bipartite graph G will be denoted by $\textsc{Hungarian}(G)$.

3.0.1 Algorithm A1 [9]

The following algorithm (A1) for finding an approximate solution for the minimal k-matching problem of k-partite graph G has been proposed in [9].

Algorithm A1 [9]

function A1(G)
    if $k = 1$ **then**
        return $\emptyset$
    else
        $M_{1,2} = \textsc{Hungarian}(G[V_1 \cup V_2])$
        return A1$(G/M_{1,2}) * M_{1,2}$
    end if
end function

In short, the algorithm finds an optimal 2-matching for $G[V_1 \cup V_2]$, takes the quotient by this matching and recursively repeats this process. The final result is a complete k-matching $\mathcal{M}$ reconstructed from the previously computed $(k-1)$-matchings.

3.0.2 Algorithms B1, C1, and D1

First we propose three new algorithms based on A1, namely algorithms B1, C1, and D1.

Algorithm B1 (improvement of A1)

function B1(G)
    if $k = 1$ **then**
        return $\emptyset$
    else
        $\{a, b\} \in \arg\min_{1 \le i < j < k} w\Big(B1(G/\textsc{Hungarian}(G[V_i \cup V_j]))\Big)$
        return B1$(G/M_{a,b}) * M_{a,b}$
    end if
end function

Algorithm B1 searches through all possible pairs (V_i, V_j) of partitions, recursively runs on the quotient graph G/M, where M is the optimal 2-matching on $V_i \cup V_j$, and among these partitions chooses the one with the best matching of G/M. If there are more than one minimal partition, the algorithm chooses a random partition of minimal weight.

The third algorithm, denoted as Algorithm C1, can be considered as a steepest descent algorithm on the set of all k-matchings, where the next solution is chosen to be a minimal solution of $G/M_{i,j}$ by B1 for any i and j.

Algorithm C1 (steepest descent based on B1)

function C1(G)
  $\mathcal{M}$ =B1(G)
  repeat
    $M_{a,b} = \arg\min_{M_{i,j} \in \mathcal{M}} w(B1(G/M_{i,j}))$
    $\mathcal{M} = B1(G/M_{a,b}) * M_{a,b}$
  until the solution $\mathcal{M}$ has not improved
  return $\mathcal{M}$
end function

Denote by D1 the faster greedy algorithm, which takes the minimal 2-matching $M_{i,j}$ in the k-partite graph G, and continues considering the $(k-1)$-partite graph $G/M_{i,j}$ until only one partition is left.

3.0.3 Algorithms E1 and F1

Besides breaking ties randomly in case of more than one minimal choice, all the steps in algorithms B1, C1, and D1 are precisely defined, i.e. the algorithms are deterministic.

The algorithms E1 and F1, are alternative randomized versions of B1 and C1. If we, in the main loop of C1, instead of searching for the minimal 2-matching $M_{i,j}$, by which we contract the graph, loop through all 2-matchings in random order and accept the first matching $M_{i,j}$ that yields a better solution, and stop when there is no improvement possible, we obtain a new algorithm, which we call E1.

We denote by E1(n) the algorithm that takes the best solution of E1 out of n trials (restarts).

The algorithm E1 stops when there is no better solution, even if there are solutions of the same quality (weight) in the neighbourhood of the current solution. These neighborhood solutions might later give improvement, therefore we introduce another variant, called F1(n), which in such case chooses randomly one solution of equal weight and continues, but stops after at most n steps, since it does not necessarily terminate (e.g. it can iterate between two equally good solutions).

4 RESULTS & CONCLUSION

In the first experiment, with results presented in Table 1, we compare the algorithms used in [9] (namely, the Random and A1 algorithm) with algorithms B1−F1 on the first set of random k-partite graphs consisting of two sets of 1000 random complete k-partite graphs with $k = 3,\ 4$ and $n = 30,\ 100$, respectively. For $k = 3$ the weights are chosen randomly from the set $\{0, \dots, 9\}$ and for $k = 4$ the weights are chosen from $\{1, \dots, 100\}$.

The second test is performed on the dataset of hard problems from [4], which includes 15 instances of complete 3-partite graphs (9 graphs with 33 vertices and 6 graphs with 66 vertices),

k	n	Random	A1	B1	C1	D1	E1	E1(10)	F(100)
3	30	405	60.7	56.2	50.8	60.6	50.9	50.3	49.8
4	100	41806	6175	5966	5859	7752	5862	5847	5861

Table 1: Comparison of Random, Greedy and A2 algorithms from [9] with algorithms B2 − F2. Each row contains average values of solutions obtained by these algorithms, computed over 1000 random instances of complete k-partite graphs with n vertices.

where the optimal value is known (see [12]). We ran algorithms A1 − F1 sequentially over all test several times and present the final average results in Table 2.

Algorithm	A1	B1	C1	D1	E1	E1(10)	F1(100)
Error	1.2%	0.9%	0.1%	1.1%	0.1%	0.07%	0.09%
Time (rel.)	1.0	2.7	10.4	2.1	8.1	81	147

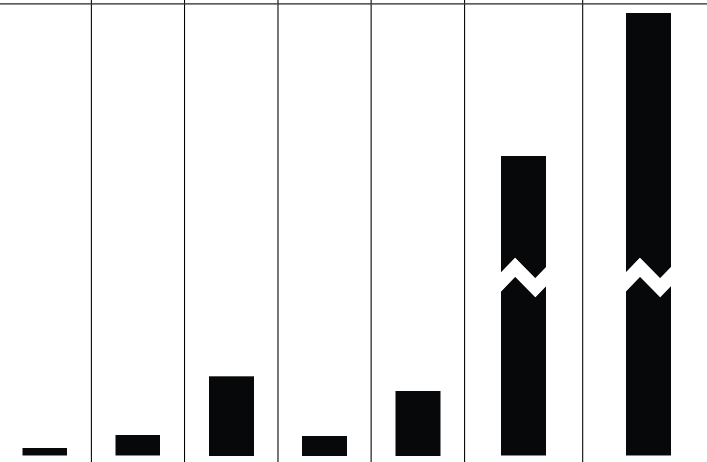


Table 2: We compare algorithms A1–F1 based on the average of 100 trials performed on 15 graphs. The second row represents how much the average solution differed from the optimal. The third row represents average run time.

Since E1 has relatively good speed and on average misses the optimal solution by merely 0.1%, we procliam it as our method of choice. For a deterministic algorithm we suggest C1.

5 Acknowledgements

This work was supported in part by ARRS, the Research agency of Slovenia, grants P1-0285 and ARRS-1000-15-0510.

References

[1] E. Balas and M. J. Saltzman. An algorithm for the three-index assignment problem. *Operations Research*, 39(1):150 – 161, 1991.

[2] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems, revised reprint*, volume 106. Siam, 2012.

[3] Rainer E Burkard, Rüdiger Rudolf, and Gerhard J Woeginger. Three-dimensional axial assignment problems with decomposable cost coefficients. *Discrete Applied Mathematics*, 65(1-3):123–139, 1996.

[4] Y. Crama and F. C. R. Spieksma. Approximation algorithms for three-dimensional assignment problems with triangle inequalities. *European Journal of Operational Research*, 60:273 – 279, 1992.

[5] Alan M Frieze. Complexity of a 3-dimensional assignment problem. *European Journal of Operational Research*, 13(2):161–164, 1983.

[6] B. Gabrovšek, T. Novak, J. Povh, D. Rupnik Poklukar, and J. Žerovnik. Multiple hungarian method for k-assignment problem. *Submitted to IEEE Access*, 2019.

[7] Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. Fuse: multiple network alignment via data fusion. *Bioinformatics*, 32(8):1195–1203, 2015.

[8] Vladimir Gligorijević, Noël Malod-Dognin, and Nataša Pržulj. Integrative methods for analyzing big data in precision medicine. *Proteomics*, 16(5):741–758, 2016.

[9] George He, Jiping Liu, and Cheng Zhao. Approximation algorithms for some graph partitioning problems. In *Graph Algorithms And Applications 2*, pages 21–31. World Scientific, 2004.

[10] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83 – 97, 1955.

[11] Noël Malod-Dognin, Julia Petschnigg, Sam FL Windels, Janez Povh, Harry Hemmingway, Robin Ketteler, and Nataša Pržulj. Towards a data-integrated cell. *Nature communications*, 10(1):805, 2019.

[12] Frits C. R. Spieksma. Instances of the 3-dimensional assignment problem. Available at `https://www.win.tue.nl/~fspieksma/instancesEJOR.htm`, last visited on 15.02.2019.

ON THE INDEPENDENT RAINBOW DOMINATION NUMBERS OF GENERALIZED PETERSEN GRAPHS $P(n,2)$ AND $P(n,3)$

(EXTENDED ABSTRACT)

Boštjan Gabrovšek
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
FMF, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia
bostjan.gabrovsek@fs.uni-lj.si

Aljoša Peperko
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
aljosa.peperko@fs.uni-lj.si

Janez Žerovnik
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
janez.zerovnik@fs.uni-lj.si

Abstract: We report on new results on 2-independent rainbow domination numbers of generalized Petersen graphs $P(n,k)$ for certain values of $n,k \in \mathbb{N}$. By applying a well established technique of tropical algebra (path algebra) we obtain exact 2-independent rainbow domination numbers of generalized Petersen graphs $P(n,2)$ and $P(n,3)$ thus confirming a conjecture of Shao et al.

Keywords: independent rainbow domination, independent rainbow domination number, generalized Petersen graphs, tropical algebra, path algebra

1 Introduction

As a combinatorial optimization issue, ordinary domination consists of determining the minimum number of places in which to keep a resource such that every place either is adjacent to the place in which a resource exists or has a resource. In practical applications, some additional constraints or desires must be taken into account. For example [12], if we are given a large computer network which consists of some clients and servers with t distinct resources $s_1, s_2, \cdots, s_t$, we need to seek the minimum number of servers each one possessing a non-empty subset of these resources in order that any client can be connected directly to a subset of servers that together have each resource s_i ($1 \leq i \leq t$). On the assumption that all resources have an identical cost, the goal is to seek the minimum value of the number of copies of such t resources. This application naturally can be modeled by the concept of t-rainbow domination. In addition, if a constraint prevents any pair of servers from occupying adjacent locations, then we have the independent t-rainbow domination problem.

In this report we announce new results on exact values of independent 2−rainbow domination number of generalized Petersen graphs $P(n,2)$ and $P(n,3)$. This is a preliminary paper where we report on some ideas of the proofs. The details of these proofs will appear in a full paper [6] that will be published elsewhere.

2 Independent rainbow domination

For a graph F, $S \subseteq V(F)$ and $w \in V(F)$, let $N_S(w)$ denote the open neighborhood of w in S, i.e. $\{u \mid uw \in E(F), u \in S\}$, and let $N_S[w]$ denote the closed neighborhood of w, i.e. $N_S[w] = \{w\} \cup N_S(w)$. If $S = V(F)$ and no confusion can occur, $N_S(w)$ and $N_S[w]$ will be denoted shortly by $N(w)$ and $N[w]$, respectively. If $S' \subseteq V(F)$, then the definition $N(S') = \cup_{x \in S'} N(x)$ is applied. The degree of a vertex w is the total number of edges incident to w.

Inspiblack by several facility location problems, Brešar, Henning and Rall [2, 3] initiated the study of the k-rainbow domination problem, and the problem is proved to be NP-complete even if the input graph is a chordal graph or a bipartite graph (see Chang [4]). This problem has attracted considerable attention many other types of domination are widely applied to real-world scenarios, see for example [9, 7].

An independent set S of a graph F is a subset of $V(F)$ for which vertices are pairwise non-adjacent. The independence number of F, denoted as $\alpha(F)$, is the maximum size of an independent set in F. Given a positive integer t and a graph F, the goal is to assign a subset of the color set $\{1, 2, \cdots, t\}$ to every vertex of F such that every vertex with the empty set assigned has all t colors in its neighborhood. Such an assignment is called the t-rainbow dominating function ($tRDF$) of the graph F. A $tRDF$ is independent ($ItRDF$) if vertices assigned nonempty sets are pairwise non-adjacent. The weight of a $tRDF$ g of a graph F is the value $w(g) = \sum_{v \in V(F)} |g(v)|$. If H is a vertex induced subgraph of $V(F)$, the weight restricted to H is $w_H(g) = \sum_{v \in H} |g(v)|$. The *independent t-rainbow domination number* $i_{rt}(F)$ is the minimum weight of an $ItRDF$s in F. The *upper t-rainbow domination number* of F, denoted by $I_{rt}(F)$, is the maximum weight of a minimal t-rainbow dominating function. (In other words, there is a dominating function f of weight $I_{rt}(F)$ on F such that no proper restriction of f is dominating.) We use $i(F)$ to denote the independent domination number, i.e. the size of a smallest independent dominating set, of a graph F.

3 Generalized Petersen graphs, polygraphs and rotographs

For positive integers $n \geq 3$ and $k \in \{1, 2, \cdots, n-1\}$, the generalized Petersen graph $P(n, k)$ is defined to be a graph with the vertex set $\{h_i^1, h_i^2 \mid i \in \{0, 1, \cdots, n-1\}\}$ and the edge set $\{h_i^1 h_i^2, h_i^1 h_{i+k}^1, h_i^2 h_{i+1}^2 \mid i \in \{0, 1, \cdots, n-1\}\}$, in which the subscripts are computed modulo n (see [17, 5]).

An important observation for the application of our techniques is that $P(n, k)$ are in fact rotagraphs. Let $G_1, \ldots, G_n$ be arbitrary mutually disjoint graphs and $X_1, \ldots, X_n$ a sequence of sets of edges such that an edge of X_i joins a vertex of $V(G_i)$ with a vertex of $V(G_{i+1})$ ($X_i \subseteq V(G_i) \times V(G_{i+1})$ for $i = 1, \ldots, n$). For convenience we set $G_{n+1} = G_1$. A *polygraph* $\Omega_n = \Omega_n(G_1, \ldots G_n; X_1, \ldots X_n)$ over monographs $G_1, \ldots, G_n$ has the vertex set $V(\Omega_n) = V(G_1) \cup \ldots \cup V(G_n)$, and the edge set $E(\Omega_n) = E(G_1) \cup X_1 \cup \ldots \cup E(G_n) \cup X_n$.

If all graphs G_i are isomorphic to a fixed graph G (i.e. there exists an isomorphism $\varphi_i : V(G_i) \longrightarrow V(G)$ for $i = 1, \ldots, n+1$, and $\varphi_{n+1} = \varphi_1$) and all sets X_i are equal to a fixed set $X \subseteq V(G) \times V(G)$ ($(u, v) \in X \iff \left(\varphi_i^{-1}(u), \varphi_{i+1}^{-1}(v)\right) \in X_i$ for all i), we call such a graph *rotagraph*, $\omega_n(G; X)$. A rotagraph without edges between the first and the last copy of G is a *fasciagraph*, $\psi_n(G; X)$.

More precisely, in a fasciagraph, $X_n = \emptyset$ and $X_1 = X, \ldots, X_{n-1} = X$.

If G is a monograph of $P(n,k)$, we call f a *partial ItRDF* if f is an $ItRDF$ for G, except that vertices of degree less than 3 and are assigned the empty set, do not need to have all colors in their neighbourhood.

By combining the fact that $P(n,k)$ are rotagraphs with a well established tropical (path) algebra technique, we calculate 2-independent rainbow domination numbers of generalized Petersen graphs $P(n,2)$ and $P(n,3)$. This proves a conjecture proposed in [17].

4 Tropical (path) algebra technique and main results

Tropical algebra (together with its isomorphic versions) is an algebraic setting with applications in a variety of classically non-linear problems appearing for instance in manufacturing and transportation scheduling, information technology, discrete event-dynamic systems, combinatorial optimization, mathematical physics, DNA analysis and graph theory (see e.g. [1, 11, 13, 14, 16, 10, 15], and the references cited there).

Tropical algebra (min-plus algebra) is an algebra (in fact, semialgebra) over the ordeblack, idempotent semiring (in fact, semifield) $\mathbb{R} \cup \{\infty\}$, equipped with the operations of addition $a \oplus b = \min(a,b)$ and multiplication $a \odot b = a + b$, with the unit elements ∞ (for addition) and 0 (for multiplication). As in standard arithmetic, the operations of addition and multiplication are associative and commutative, and multiplication is distributive over addition. Matrix and polynomial operations are defined similarly to their standard counterparts, with the min-plus operations replacing the standard operations. In particular, for two $n \times n$ matrices A, B with entries from $\mathbb{R} \cup \{\infty\}$ we define the product AB by

$$(AB)_{ij} = \min_{k=1,\ldots,n} \{a_{ik} b_{kj}\}$$

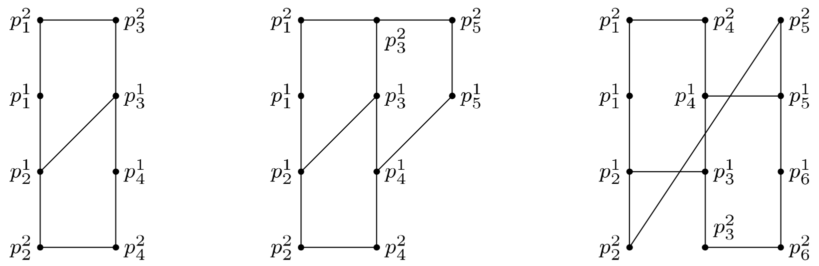
for all $i, j = 1, \ldots, n$. The mth tropical power is denoted by A^m. In fact, in our application we will consider matrices over a semisubring $\mathbb{N} \cup \{0\} \cup \{\infty\}$ equipped with the above operations (sometimes called path algebra, see e.g. [10, 15])

Min-plus algebra is isomorphic to max-plus algebra, which is the semifield $\mathbb{R} \cup \{\infty\}$, where addition is replaced by maximum and multiplication by addition, and also to max-times algebra $\mathbb{R}_+$, where addition is replaced by maximum and multiplication is the same as in standard arithmetic. For more on tropical algebra we refer to the monograph of Butkovič [1]. Tropical algebra is a part of a broader branch of mathematics, "idempotent mathematics", which was developed mainly by Maslov and his collaborators (see e.g. [13], [11]).

Next we describe how we apply a path algebra technique to obtain exact values $i_{r2}(P(n,2))$ and $i_{r2}(P(n,3))$ for $n \geq 7$.

The graphs $P(4n,2)$ are rotagraphs consisting of monographs G_A from Figure 1a with $X = \{(p_3^2, p_1^2), (p_4^1, p_1^1), (p_4^2, p_2^2)\}$. Graphs $P(4n+1,2)$ are polygraphs of the form $\Omega_n(H_1, G_A, \ldots, G_A\,;\, X_1, X, \ldots X)$, where H_1 is depicted in Figure 1b and $X_1 = \{(p_4^2, p_1^2), (p_5^1, p_1^1), (p_5^2, p_2^2)\}$. Similarly we define graphs H_2 and H_3 such that $P(4n+2,2)$ and $P(4n+3,2)$ are polygraphs with the first monograph H_2 and H_3, respectively.

By the same method one can see that $P(6n,3)$ is the rotagraph consisting of monographs G_B from Figure 1c. As before we can find starting monographs H_i' for the polygraphs $P(6n+i,3)$, $i = 1, \ldots, 5$.



(a) Monograph G_A (b) Monograph H_1 (c) Monograph G_B

Figure 1: Monographs of $P(n, 2)$ and $P(n, 3)$

Let $S = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}^8$ be the sequence of all possible colorings of G_A (with ordeblack vertices). By choosing two colorings S_i and S_j from S, we can color the graph $\Omega(G_A, G_A; X_1, \emptyset)$ by coloring the first G_A component by S_i and the second G_A component by S_j. We define $w(S_i, S_j)$ to be $w(S_i)+w(S_j)$ if the combined coloring defines a partial $I2RDF$ on the monographs and ∞ otherwise.

We now associate to G_A a $4^8 \times 4^8$ matrix $M' = (w(S_i, S_j))_{i,j}$. To simplify the computations, one observes that most rows (and columns) in M' have all entries equal to ∞, i.e. colorings that do not define an $ItRDF$. Removing these rows and columns, we are left with a 281×281 submatrix M.

Taking graphs H_1, H_2, and H_3, we analogously define 281×281 matrices A_1, A_2, and A_3, respectively.

Let M^l be the l-th tropical power of M. It follows by construction that $i_{r2}(P(4l, 2)) = \min_{i=1,\ldots,281} M^l_{ii}$.

In the cases $s = 1, 2, 3$ it holds that $i_{r2}(P(4(l-1)+s, 2)) = \min_{i=1,\ldots 281} A_s M^l_{ii}$ for $s = 1, 2, 3$.

We repeat an analogue procedure for $P(n, 3)$. Computations show that in this case, the $4^{12} \times 4^{12}$ matrix N' associated to G_B blackuces to a 5066×5066 matrix N. To the monographs H'_s, we respectively associate matrices B_s, $s = 1, \ldots, 5$. Again we have by construction $i_{r2}(P(6l, 3)) = \min_{i=1,\ldots 5066} N^l_{ii}$ and $i_{r2}(P(6(l-1)+s, 3)) = \min_{i=1,\ldots 5066} B_s N^l_{ii}$ for $s = 1, \ldots, 5$.

It is well known that the tropical powers of matrices have a cyclic behaviour (see e.g. [1, 10, 15]) We prove (calculate) the following result:

$M^{l+5} = M^l + [16]^{281}_{i,j=1}$ for all $l \geq 25$ and $N^{l+4} = N^l + [21]^{5066}_{i,j=1}$ for all large enough l.

The arguments briefly sketched above imply the next two theorems, which confirm that the upper bounds proved in [17] are actually equalities (as conjectublack there).

Theorem 4.1 *Let $n \geq 7$.*

$$i_{r2}(P(n, 2)) = \begin{cases} \left\lceil \frac{4n}{5} \right\rceil, & n \equiv 0 \bmod 10 \\ \left\lceil \frac{4n}{5} \right\rceil + 1, & n \equiv 9 \bmod 10 \\ \left\lceil \frac{4n}{5} \right\rceil + 2, & n \equiv 2, 3, 4, 5, 7, 8 \bmod 10 \\ \left\lceil \frac{4n}{5} \right\rceil + 3, & n \equiv 1, 6 \bmod 10 \end{cases}$$

Theorem 4.2 *Let $n \geq 7$.*

$$i_{r2}(P(n,3)) = \begin{cases} \left\lceil \frac{7n}{8} \right\rceil, & n \equiv 0,2,4,14 \bmod 16 \\ \left\lceil \frac{7n}{8} \right\rceil + 1, & n \equiv 5,7,8,10,12,13,15 \bmod 16 \\ \left\lceil \frac{7n}{8} \right\rceil + 2, & n \equiv 1,3,6,9,11 \bmod 16 \end{cases}$$

Since it is known (see e.g. [17]) that $P(2k+1,k+1) \cong P(2k+1,2)$ and $P(3k+2,k+1) \cong P(3k+2,3)$ for all $k \geq 2$ we obtain also the following two consequences.

Corollary 4.3 *Let $k \geq 2$.*

$$i_{r2}(P(2k+1,k+1)) = \begin{cases} \left\lceil \frac{8k+4}{5} \right\rceil + 1, & k \equiv 4 \bmod 5 \\ \left\lceil \frac{8k+4}{5} \right\rceil + 2, & k \equiv 1,2,3 \bmod 5 \\ \left\lceil \frac{8k+4}{5} \right\rceil + 3, & k \equiv 0 \bmod 5 \end{cases}$$

Corollary 4.4 *Let $k \geq 2$.*

$$i_{r2}(P(3k+2,k+1)) = \begin{cases} \left\lceil \frac{21k+14}{8} \right\rceil, & 3k \equiv 0,2,12,14 \bmod 16 \\ \left\lceil \frac{21k+14}{8} \right\rceil + 1, & 3k \equiv 1,3,4,6,8,9,11 \bmod 16 \\ \left\lceil \frac{21k+14}{8} \right\rceil + 2, & 3k \equiv 5,7,10,13,15 \bmod 16 \end{cases}$$

5 Further work

There are several possible directions of further research. For example, the calculation of t-independent rainbow domination numbers of generalized Petersen graphs $P(n,k)$ for other t and k. The generalized Petersen graphs are of course just a popular examples of graph families with natural polygraph structure. The methods employed in this research can be applied to any such family.

6 Acknowledgements

This work was supported in part by the Slovenian Research Agency (grants P1-0222 and J1-8133).

References

[1] Butkovič, P. (2010). Max-linear systems: theory and algorithms, Springer-Verlag, London.

[2] Brešar, B., Henning, M. A., and Rall, D. F. (2005). Paired-domination of Cartesian products of graphs and rainbow domination. Electronic Notes in Discrete Mathematics. 22, 233–237.

[3] Brešar, B. and Šumenjak, T. K. (2007). On the 2-rainbow domination in graphs. Discrete Applied Mathematics 155 (1), 2394–2400.

[4] Chang, G. J., Wu, and J., Zhu, X. (2010). Rainbow domination on trees. Discrete Applied Mathematics. 158, 8–12.

[5] Ebrahimi, B. J., Jahanbakht, N., and Mahmoodianc, E. S. (2009). Vertex domination of generalized Petersen graphs. Discrete Mathematics. 309, 4355–4361.

[6] Gabrovšek, B., Peperko, A., and Žerovnik, J. (2019). manuscript in preparation.

[7] Garey, M. R. and Johnson, D. S. (1979).Computers and Intractability: a Guide to the Theory of NP-Completeness, W. H. Freeman and Co., San Francisco, Calif.

[8] Goddard, W. and Henning, M. A. (2013). Independent domination in graphs: A survey and recent results. Discrete Mathematics. 313(7), 839–854.

[9] Haynes, T. W., Hedetniemi, S. T., and Slater, P. J. (1998). Fundamentals of Domination in Graphs, Marcel Dekker, New York.

[10] Klavžar, S. and Žerovnik, J. (1996). Algebraic approach to fasciagraphs and rotagraphs. Discrete Applied Mathematics. 68, 93–100.

[11] Kolokoltsov, V.N. and Maslov, V.P. (1997). Idempotent analysis and its applications, Kluwer Acad. Publ.

[12] Kraner Šumenjak, T., Rall, D.F., Tepeh, A. (2013). Rainbow domination in the lexicographic product of graphs. Discrete Applied Mathematics, Volume 161, Issues 13–14, 2133–2141.

[13] Litvinov G.L. (2007). The Maslov dequantization, idempotent and tropical mathematics: A brief introduction, *J. Math. Sci.*(N. Y.) 140, no.3 , 426–444.

[14] Müller, V. and Peperko, A. (2015). On the spectrum in max-algebra, Linear Algebra Appl. 485, 250–266.

[15] Repolusk, P. and Žerovnik, J. (2018). Formulas for various domination numbers of products of paths and cycles. Ars Comb. 137, 177-202.

[16] Rosenmann, A., Lehner, F., and Peperko, A. (2019). Polynomial convolutions in max-plus algebra, Lin. Alg. Appl. 578, 370–401.

[17] Shao, Z., Li, Z., Peperko, A., Wan, J., and Janez Žerovnik, J. (2019). Independent Rainbow Domination of Graphs, Bulletin Malaysian Math. Sci. Soc. 42 (2), 417–435.

An Exact Penalty Method over Discrete Sets

Nicolò Gusmeroli and Angelika Wiegele
Alpen-Adria-Universität, Klagenfurt

Binary quadratic problems (BQP) are very general and have several applications in different fields. Many problems can be modeled as BQPs, e.g. the shortest quadratic path problem, the max k-cluster problem and the quadratic assignment problem. We consider BQPs with (integer valued) equality constraints, hence we solve

$$f^* := \min \left\{ c^\top x + x^\top F x \colon Ax = b,\ x \in \{0,1\}^n \right\}.$$

In order to solve this problem we apply an *exact penalty method over discrete sets*, proposed by Lasserre [1] in 2016. This method transforms an equality constrained BQP in n variables into a *max-cut* instance of dimension $n+1$ by using a Lagrangian approach. This method also provides a *threshold* parameter from which it is possible to determine whether the original problem is feasible. Since we solve the *max-cut* instance by using a branch-and-bound procedure this parameter is very important, because it allows us to have earlier pruning.

In the algorithm we first compute a lower bound and an upper bound on the problem, we then define the penalty and the threshold parameters that depend on these bounds. Then we penalize the equality constraints and we reformulate the resulting problem. Hence we increase the dimension by one in order to get the equivalent *max-cut* instance.

In the original formulation the bounds are calculated by solving the standard SDP relaxation of the problems, where all the equality constraints are relaxed. Hence it follows that the penalty parameter can be potentially large. For some instances, computing the maximum cut is more difficult than solving the original problem, thus we focus on the algorithm in order to make the value of the penalty parameter as small as possible.

It is possible to decrease the penalty parameter in two ways: either by giving a different formulation of the parameters or by improving the bounds used in its definition. We explain the main features of these two parameters and derive their optimal formulations. Then we show which are the extremal bounds in the definition of the penalty parameter and we present some possible relaxations, thus we improve both the parameters.

The *max-cut* instances are solved using *BiqMac*, a solver developed by Rinaldi, Rendl and Wiegele in 2010, see [2]. We conclude by comparing our algorithm to some of the current best solvers available, e.g. *CPLEX* and *BiqCrunch*, see [3], showing the efficiency of our method.

References

[1] Lasserre, J. B. (2016) *A MAX-CUT formulation of 0/1 programs.* Operations Research Letters, 44(2)

[2] Rendl F., Rinaldi G. and and Wiegele A. (2010) *Solving Max-Cut to optimality by intersecting semidefinite and polyhedral relaxations.* Math. Program. 121

[3] Krislock N., Malick J. and Roupin F. (2017) *BiqCrunch: a semidefinite branch-and-bound method for solving binary quadratic problems.* ACM Transactions on Mathematical Software (TOMS)

THE GENERAL POSITION PROBLEM ON GRAPHS

Sandi Klavžar
University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia
University of Maribor, Faculty of Natural Sciences and Mathematics, Maribor, Slovenia
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
E-mail: sandi.klavzar@fmf.uni-lj.si

Abstract: A general position problem in graph theory is to find a largest set of vertices that are in a general position. More precisely, if $G=(V(G), E(G))$ is a graph, then a subset S of its vertices is a *general position set* if for any triple of pairwise different vertices u, v, and w from S we have $d(u,v) \neq d(u,w) + d(w,u)$, where d is the standard shortest path distance function. S is called a gp-set of G if S has the largest cardinality among the general position sets of G. The *general position number* (*gp-number* for short) $gp(G)$ of G is the cardinality of a gp-set of G. These concepts were recently introduced in (Paul Manuel, Sandi Klavžar, A general position problem in graph theory, Bulletin of the Australian Mathematical Society 98 (2018) 177-187).
In this talk we will quickly present different motivations for the concepts and briefly survey state of the art on the general position problem on graphs. Despite a short period of time since the general position problem was introduced, quite a fascinating number of results have been obtained, including bounds on the gp-number, complexity issues, exact results for several families, a characterization of general position sets, the the gp-number of different graph operations, and more.

Keywords: graph distance, general position problem, cliques in graphs, computational complexity

k-FAIR DOMINATION PROBLEM IN CACTUS GRAPHS

Tina Novak
University of Ljubljana, Faculty of Mechanical Engineering, Aškerčeva 6
SI-1000 Ljubljana, Slovenia
tina.novak@fs.uni-lj.si

Janez Žerovnik
University of Ljubljana, Faculty of Mechanical Engineering, Aškerčeva 6
SI-1000 Ljubljana, Slovenia
janez.zerovnik@fs.uni-lj.si

Abstract. For $k \geq 1$, a set of vertices $D \subset V(G)$ of graph G is a k-fair dominating set if every vertex $v \notin D$ is dominated by exactly k vertices in D. To solve the weighted minimum k-fair dominating problem means to find the minimum of all weights $w(D)$ over all k-fair dominating sets D. For $k = 1$ the set D is called perfect dominating set and the problem is the weighted minimum perfect dominating problem. In the paper, the linear time algorithm for solving the weighted minimum k-fair dominating problem in cactus graphs is given.

Keywords: linear time algorithm; cactus graph; DFS structure; weighted k-fair domination; weighted perfect domination.

1 Introduction

In graph theory, there are many NP-hard problems on general graphs that can be solved efficiently on trees [6]. One of the well known examples is the graph domination problem [5]. Often, a problem that can be solved in linear time on trees can also be solved in linear time on cactus graphs. Recent example include the weighted minimum independent k-domination problem (also studied in [13]), which can be solved in linear time for cactus graphs (see [4, 10]). It is also known that the weighted domination number of a cactus graph and the Hosoya polynomial of a cactus graph can be calculated in linear time (see [11] and [8] respectively). Upper bounds for fair and k-fair domination of cacti are studied in [7]. The linear algorithms for all above mentioned problems in cactus graphs share some common ideas. The general properties of quantities of graphs that provide the linearity of algorithms (using the opposite DFS order and transferred from trees to cacti) for computing these quantities in cacti are studied in [9]. The motivation for study the weighted minimum k-fair dominating problem came by the paper [3] and its ArXiv version [2] in which an open problem asks for a polynomial time algorithm for computing the perfect domination for trees.

In this short paper we recall the main ideas from [8, 10, 11] that give rise to a linear time algorithm for k-fair domination. We claim that the algorithm can be modified to solve the fair domination problem on cacti within the polynomial time complexity. Ideas will be discussed in the conference talk, whereas further details and proofs will be provided in the full version.

2 Preliminaries

Let $G = G(V, E, w)$ be an undirected connected weighted nonempty graph with the set of vertices $V = V(G)$, the set of edges $E = E(G)$, and the positive real function $w : V \to \mathbb{R}_+$ (weight). For a subset $S \subseteq V$ the weight of S is $w(S) = \sum_{v \in V(S)} w(v)$. Let $N(v)$ be the set of vertices adjacent to the vertex v and $N[v] = \{v\} \cup N(v)$; sets $N(v)$ and $N[v]$ are the open and the closed neighborhood of a vertex v, respectively. For a subset $S \subseteq V$ denote by $N(S)$ the set of vertices adjacent to any vertex in S and similarly, as for the $N[v]$, let $N[S] = S \cup N(S)$. A subset $D \subseteq V$ is a dominating set if $N[D] = V$. Let k be an integer and $k \geq 1$. A subset D is *a k-fair dominating set* if every $v \in V \setminus D$ is adjacent to exactly k vertices in D, i.e. $|N(v) \cap D| = k$. The weighted k-fair domination number $\mathrm{wfd}_k(G)$ is defined as the minimum weight $w(D)$ among all k-fair dominating sets D of the graph G, i.e.

$$\mathrm{wfd}_k(G) = \min \big\{ w(D) \,\big|\, D \text{ is a } k\text{-fair dominating set of } G \big\}. \tag{1}$$

A weighted k-fair dominating set D of the weight $\mathrm{wfd}_k(G)$ is called a $\mathrm{wfd}_k(G)$-set. Suppose that G is a rooted graph, $G = (G, r) = G_r$, with the root r. Define two additional parameters wfd_k^+ and wfd_k^- as

$$\begin{aligned} \mathrm{wfd}_k^+(G_r) &= \min \big\{ w(D) \,\big|\, D \text{ is a } k\text{-fair dominating set of } G_r,\ r \in D \big\} & (2)\\ \mathrm{wfd}_k^-(G_r) &= \min \big\{ w(D) \,\big|\, D \text{ is a } k\text{-fair dominating set of } G_r,\ r \notin D \big\}. & (3) \end{aligned}$$

If there is no such a k-fair dominating set D with $r \notin D$, we set $\mathrm{wfd}_k^-(G_r) = \infty$. It is obvious that $\mathrm{wfd}_k(G_r) = \min \big\{ \mathrm{wfd}_k^+(G_r), \mathrm{wfd}_k^-(G_r) \big\}$. Let us define k additional parameters which will be later needed in the recursive formulas

$$\begin{aligned} \mathrm{wfd}_k^0(G_r) &= \mathrm{wfd}_k(G_r - r\,; N(r) \cap D = \emptyset) \\ \mathrm{wfd}_k^\ell(G_r) &= \min_{|S|=\ell} \big\{ \mathrm{wfd}_k(G_r - r\,; N(r) \cap D = S) \big\} \text{ for } \ell = 1, 2, \ldots, k-1. \end{aligned}$$

Remark 1 *Additional parameters* $\mathrm{wfd}_k^1(G_r), \mathrm{wfd}_k^2(G_r), \ldots, \mathrm{wfd}_k^{k-1}(G_r)$ *are well defined if* $|N(r)| \geq k-1$. *If* $|N(r)| < j \leq k-1$, *we set* $\mathrm{wfd}_k^j(G_r) = \infty$

Remark 2 *Let* $G_r = (\cup_{j=1}^m G_{v_j}) \cup \{rv_1, rv_2, \ldots, rv_m\}$ *where* $N(r) = \{v_1, v_2, \ldots, v_m\}$ *and* $G_{v_i} \cap G_{v_j} = \emptyset$ *for all pairs* $i, j \in \{1, 2, \ldots, m\}$, $i \neq j$. *If* $k \leq m$, *then*

$$\begin{aligned} \mathrm{wfd}_k^0(G_r) &= \sum_{j=1}^m \mathrm{wfd}_k^-(G_{v_j}) \\ \mathrm{wfd}_k^\ell(G_r) &= \min_{\substack{S \subseteq N(r) \\ |S|=\ell}} \Big\{ \sum_{v \in S} \mathrm{wfd}_k^+(G_v) + \sum_{v \in N(r)-S} \mathrm{wfd}_k^-(G_v) \Big\} \text{ for } \ell = 1, 2, \ldots, \mathrm{k}-1. \end{aligned}$$

Remark 3 *According to the definitions above we can use notation* $\mathrm{wfd}_k^k(G_r)$ *instead of* $\mathrm{wfd}_k^-(G_r)$.

3 Auxiliary results

Following the idea of [11], we calculate all above defined parameters (wfd_k^+, $\mathrm{wfd}_k^- = \mathrm{wfd}_k^k$, wfd_k^ℓ for $\ell = 0, 1, \ldots, k-1$) of some special examples of weighted graphs from the given values of parameters of the given subgraphs.

Lemma 4 *Let $\widetilde{G}_u$ and G_v be disjoint rooted graphs with roots u and v respectively. Let G_u be a disjoint union of $\widetilde{G}_u$ and G_v joined by the edge uv. Then*

$$\begin{aligned}
\mathrm{wfd}_k^0(G_u) &= \mathrm{wfd}_k^0(\widetilde{G}_u) + \mathrm{wfd}_k^-(G_v) \\
\mathrm{wfd}_k^\ell(G_u) &= \min\left\{\mathrm{wfd}_k^\ell(\widetilde{G}_u) + \mathrm{wfd}_k^k(G_v), \mathrm{wfd}_k^{\ell-1}(\widetilde{G}_u) + \mathrm{wfd}_k^+(G_v)\right\}; \text{ for } \ell = 1, 2, \ldots, k \\
\mathrm{wfd}_k^+(G_u) &= \mathrm{wfd}_k^+(\widetilde{G}_u) + \min\left\{\mathrm{wfd}_k^+(G_v), \mathrm{wfd}_k^{k-1}(G_v)\right\}.
\end{aligned}$$

Lemma 5 *Let G_u be a union of two rooted graphs $\widetilde{H}_u$ and H_u where $\widetilde{H}_u - u$ and $H_u - u$ are disjoint. Then*

$$\begin{aligned}
\mathrm{wfd}_k^0(G_u) &= \mathrm{wfd}_k^0(\widetilde{H}_u) + \mathrm{wfd}_k^0(H_u) \\
\mathrm{wfd}_k^\ell(G_u) &= \min_{0\le j\le \ell}\left\{\mathrm{wfd}_k^{\ell-j}(\widetilde{H}_u) + \mathrm{wfd}_k^j(H_u)\right\} \qquad \text{for } \ell = 1, 2, \ldots, k \\
\mathrm{wfd}_k^+(G_u) &= \mathrm{wfd}_k^+(\widetilde{H}_u) + \mathrm{wfd}_k^+(H_u) - w(u).
\end{aligned}$$

Definition 6 (Path-like graph) *A rooted graph $G_{v_n} = (G, v_n)$ is a path-like graph if G_{v_n} is a disjoint union of n rooted graphs $(G_1, v_1), (G_2, v_2), \ldots, (G_n, v_n)$, joined by the edges $v_1v_2, v_2v_3, \ldots, v_{n-1}v_n$.*

Definition 7 (Cycle-like graph) *A rooted graph $G_{v_n} = (G, v_n)$ is a cycle-like graph if G_{v_n} is a disjoint union of $n-1$ rooted graphs $(G_1, v_1), (G_2, v_2), \ldots, (G_{n-1}, v_{n-1})$, joined by the edges $v_1v_2, v_2v_3, \ldots, v_{n-1}v_n$ and v_nv_1 .*

Remark 8 *If H_u in Lemma 5 is a cycle-like graph, the sets whose elements we want to minimize contain only three elements, i.e.*

$$\mathrm{wfd}_k^\ell(G_u) = \min_{0\le j\le 2}\left\{\mathrm{wfd}_k^{\ell-j}(\widetilde{H}_u) + \mathrm{wfd}_k^j(H_u)\right\} \qquad \text{for } \ell = 2, 3, \ldots, k$$

Let (G, v_n) be a cycle-like graph as in definition above. Denote by (G', v_n) a path-like graph which is the disjoint union of the rooted graphs $(G_{v'_n} = v'_n, v'_n), (G_1, v_1), (G_2, v_2), \ldots, (G_{n-1}, v_{n-1})$, and $(G_n = v_n, v_n)$, joined by the edges $v'_nv_1, v_1v_2, v_2v_3, \ldots, v_{n-1}v_n$. For the graph (G', v_n) we say that it is obtained from the cycle-like graph (G, v_n) by doubling the vertex v_n (the duplicate is marked by v'_n) and setting $w(v'_n) = w(v_n)$. For the path-like graph G'_{v_n}, obtained from a cycle-like graph G_{v_n}, we define for every $i \in \{0, 1, 2, \ldots, n\}$ a rooted graph $H'_i = (H'_i, v_i)$ as follows

$$\begin{aligned}
H'_0 &:= G_0 := v'_n \\
H'_i &:= \left(\bigcup_{j=0}^{i} G_j\right) \cup (v_1, v'_n) \cup \left(\bigcup_{j=1}^{i-1}(v_j, v_{j+1})\right).
\end{aligned} \tag{4}$$

For convenience we denote a singleton graph (a graph with one vertex v and no edges) by v.

Lemma 9 *Let G_{v_n} be a cycle-like graph and G'_{v_n} the corresponding path-like graph obtained from the cycle-like graph G_{v_n}. Let H'_i, $i = 0, 1, \ldots, n$, be the rooted subgraphs of G'_{v_n} as defined in (4). Then*

$$\begin{aligned}
\mathrm{wfd}_k^0(G_{v_n}) &= \mathrm{wfd}_1^-(H'_{n-1} - v'_n, v_1 \notin D)\\
\mathrm{wfd}_k^1(G_{v_n}) &= \min\left\{\mathrm{wfd}_k^-(H'_{n-1} - v'_n, v_1 \in D), \mathrm{wfd}_k^+(H'_{n-1} - v'_n, v_1 \notin D)\right\}\\
\mathrm{wfd}_k^2(G_{v_n}) &= \mathrm{wfd}_k^+(H'_{n-1} - v'_n, v_1 \in D)\\
\mathrm{wfd}_k^-(G_{v_n}) &= \begin{cases} \mathrm{wfd}_k^+(H'_{n-1} - v'_n, v_1 \in D) & ;\ k = 2\\ \infty & ;\ k > 2 \end{cases}\\
\mathrm{wfd}_k^+(G_{v_n}) &= \mathrm{wfd}_k^+(G'_{v_n}, v'_n \in D) - w(v_n)
\end{aligned}$$

and $\mathrm{wfd}_k^3(G_{v_n}) = \mathrm{wfd}_k^4(G_{v_n}) = \ldots = \mathrm{wfd}_k^{k-1}(G_{v_n}) = \infty$.

4 Algorithm

4.1 Structure of cacti and DFS algorithm

Each cactus graph can be arranged in the skeleton structure as it is explained in [1]. Each vertex of a cactus graph is either a C-vertex, i.e. a vertex on a cycle of degree 2 or a G-vertex, i.e. a vertex not included in any cycle or a H-vertex or a hinge, i.e. a vertex on a cycle of degree ≥ 3. The depth first search (DFS) algorithm is a well known method for exploring graphs. It can be used for recognizing cactus graphs providing the data structure (see [11, 12]). Consider, K_r is a rooted cactus graph. After running the DFS algorithm, the vertices of K_r are DFS ordered. The order is given by the order in which DFS visits the vertices. The depth first number, $\mathrm{DFN}(v)$, is the position of v in the DFS order. We set $\mathrm{DFN}(r) = 0$. Following [14], it is useful to store the information recorded during the DFS run in four arrays, called the DFS (cactus) data structure:

- $\mathrm{FATHER}(v)$ is the unique predecessor (father) of the vertex v in the rooted tree, constructed with the DFS.
- $\mathrm{ROOT}(v)$ is the root vertex of the cycle containing v i.e. the first vertex of the cycle (containing v) in the DFS order. If v does not lie on a cycle, then $\mathrm{ROOT}(v) = v$. We set $\mathrm{ROOT}(r) = r$. (In any DFS order, if $\mathrm{DFN}(w) < \mathrm{DFN}(v)$ and w is the root of the cycle containing v and v is the root of another cycle (it is a hinge), then $\mathrm{ROOT}(v) = w$.)
- For vertices on a cycle (i.e. $\mathrm{ROOT}(v) \neq v$), orientation of the cycle is given by $\mathrm{ORIEN}(v) = z$, where z is the son of $\mathrm{ROOT}(v)$ that is visited on the cycle first. If $\mathrm{ROOT}(v) = v$, then $\mathrm{ORIEN}(v) = v$.
- $\mathrm{IND}(v) := |\{u \mid \mathrm{FATHER}(u) = v\}|$ is the number of sons of v in the DFS tree.

The pseudocode of the DFS algorithm is written in detail for example in [11, 12]. We refer to [11] for elaboration of some properties of the DFS ordered vertices of a cactus and the relationship between the definitions of C, G, H-vertices in a rooted cactus K_r and arrays in the DFS table.

4.2 Main result

***Step 1*: Cactus recognition**

Using the DFS algorithm on the rooted cactus G (any vertex chosen for a root) the data structure of cactus graph can be derived, including arrays $\mathrm{DFN}(v)$, $\mathrm{FATHER}(v)$, $\mathrm{ROOT}(v)$, $\mathrm{ORIEN}(v)$ and $\mathrm{IND}(v)$.

***Step 2*: Initialization**

For every vertex v we set $\mathrm{wfd}_k^0(v) = 0$, $\mathrm{wfd}_k^\ell(v) = \infty$ for $\ell = 1, 2, \ldots, k$, and $\mathrm{wfd}_k^+(v) = w(v)$.

***Step 3*: Computation**

Start with v, the last vertex in the DFS order. Let $u = \mathrm{FATHER}(v)$.
While $v \neq u$ (i.e. $v = u$ is not the root of G) do (3a) or (3b):

(3a) If the edge $e = uv$ is not an edge of a cycle of G (i.e. $\mathrm{ROOT}(v) = v$):

- If $\mathrm{DFN}(u) \neq \mathrm{DFN}(v) - 1$ (i.e. $\mathrm{DFN}(u) < \mathrm{DFN}(v) - 1$), there exists rooted subgraph $\tilde{G}_u$. The algorithm calls itself recursively for the subgraph $\tilde{G}_u$, the rooted subcactus with root u and vertices in DFS table with $\mathrm{DFN}(u), \ldots, \mathrm{DFN}(v) - 1$. Temporally, we set $\mathrm{wfd}_k^0(u) = \mathrm{wfd}_k^0(\tilde{G}_u)$, $\mathrm{wfd}_k^\ell(u) = \mathrm{wfd}_k^\ell(\tilde{G}_u)$ for $\ell = 1, 2, \ldots, k$, and $\mathrm{wfd}_k^+(u) = \mathrm{wfd}_k^+(\tilde{G}_u)$.
- After the recursion or when u and v are the sequential vertices in the DFS order, the values of parameters $\mathrm{wfd}_k^0(u)$, $\mathrm{wfd}_k^\ell(u)$ for $\ell = 1, 2, \ldots, k$, and $\mathrm{wfd}_k^+(u)$ must be corrected, according to Lemma 4, as:
  $\mathrm{wfd}_k^+(u) = \mathrm{wfd}_k^+(u) + \min\{\mathrm{wfd}_k^+(v), \mathrm{wfd}_k^{k-1}(v)\}$
  $\mathrm{wfd}_k^\ell(u) = \min\{\mathrm{wfd}_k^\ell(u) + \mathrm{wfd}_k^k(v), \mathrm{wfd}_k^{\ell-1}(u) + \mathrm{wfd}_k^+(v)\}$; $\ell = k, k-1, \ldots, 2, 1$
  $\mathrm{wfd}_k^0(u) = \mathrm{wfd}_k^0(u) + \mathrm{wfd}_k^k(v)$.
- $v = u$ and $u = \mathrm{FATHER}(v)$.

(3b) If the edge $e = uv$ lies on a cycle C (i.e. $r = \mathrm{ROOT}(v) \neq v$):

- We have to read and remember all cycle's vertices. Denote them by $a_1, \ldots, a_n$ where $a_1 = v$, $a_{n-1} = \mathrm{ORIEN}(v)$ and $a_n = r = \mathrm{ROOT}(v)$.
- If $\mathrm{DFN}(a_j) < \mathrm{DFN}(a_{j-1}) - 1$, denote by K_{a_j} the rooted subcacti on vertices with DFN: $\mathrm{DFN}(a_j) \leq \mathrm{DFN} < \mathrm{DFN}(a_{j-1})$ for $j = 2, 3, \ldots, n-1$. Recursively calculate (and repair) values $\mathrm{wfd}_k^+(a_j) = \mathrm{wfd}_k^+(K_{a_j})$, $\mathrm{wfd}_k^\ell(a_j) = \mathrm{wfd}_k^\ell(K_{a_j})$ for $j = k, k-1, \ldots, 2, 1$, and $\mathrm{wfd}_k^0(a_j) = \mathrm{wfd}_k^0(K_{a_j})$.
- According to the definition of the path-like graph obtained from a cycle-like graph above and the mentioned tags of subgraphs (4), we calculate all parameters for graphs $(H'_{n-1} - v'_n, v'_1 \notin D)$, $(H'_{n-1} - v'_n, v_1 \in D)$ and $(G'_{v_n}, v'_n \in D)$. Then calculate values of all parameters of the cycle-like graph G_{a_n} as is given in Lemma 9.
- Repair the values of the parameters $\mathrm{wfd}_k^+(r)$, $\mathrm{wfd}_k^\ell(r)$ for $\ell = k, k-1, \ldots, 2, 1$, and $\mathrm{wfd}_k^0(r)$ by adding the values of the parameters of the cycle-like graph above by Lemma 5:
  $\mathrm{wfd}_k^+(r) = \mathrm{wfd}_k^+(G_{a_n}) + \mathrm{wfd}_k^+(r) - w(r)$

$\mathrm{wfd}_k^\ell(r) = \min_{0\le j\le 2}\{\mathrm{wfd}_k^j(G_{a_n}) + \mathrm{wfd}_k^{\ell-j}(r)\};\ \ell = k, k-1, \ldots, 2, 1$

$\mathrm{wfd}_k^0(r) = \mathrm{wfd}_k^0(G_r) + \mathrm{wfd}_k^0(r)$.

- u is the vertex with $\mathrm{DFN}(u) = \mathrm{DFN}(\mathrm{ORIEN}(v)) - 1$.
- $v = u$ and $u = \mathrm{FATHER}(v)$.

Acknowledgments

The work was supported in part by ARRS, Research Agency of Slovenia (Grant number P2-0248).

References

[1] Burkard, R. E. and Krarup, J. (1998). A Linear Algorithm for the Pos/Neg-Weighted 1-Median Problem on a Cactus. *Computing.* **60.** 193-215.

[2] Caro, Y., Hansberg, A. and Henning, M. (2011). Fair Domination in Graphs. *ArXiv:* 1109.1150v1.

[3] Caro, Y., Hansberg, A. and Henning, M. (2012). Fair domination in graphs. *Discrete Mathematics.* **312.** 2905-2914.

[4] Chaluvaraji, B., Chellali, M. and Vidya, K. A. (2010). Perfect k-domination in graphs. *Australasian journal of combinatorics.* **48.** 175-184.

[5] Cockayne, E., Goodman, S. and Hedetniemi, S. (1975). A linear algorithm for the domination number of a tree. *Information Processing Letters.* **4.** 41-44.

[6] Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: a Guide to the Theory of NP-completeness.* W. H. Freeman, San Francisco.

[7] Hajian, M. and Jafari Rad, N. (2019). Fair domination number in cactus graphs. *Discussiones Mathematicae Graph Theory.* **39.** 489-503.

[8] Novak, T., Rupnik Poklukar, D. and Žerovnik, J. (2018). The Hosoya polynomial of double weighted graphs. *Ars Mathematica Contemporanea.* **15.** 441-466.

[9] Novak T. and Žerovnik, J. (in progress). Linear time algorithms in cactus graphs.

[10] Novak, T. and Žerovnik, J. (to appear). Weighted Independent k-Domination of Cactus Graphs. *Ars Combinatoria.*

[11] Novak, T. and Žerovnik, J. (2016). Weighted domination number of cactus graph. *International journal of applied mathematics.* **29**. 401-423.

[12] Thulasiraman, K. and Swamy, M. N. S. (1992). *Graphs: theory and algorithms.* John Wiley & Sons, Inc.

[13] Yen, W. C-K. (2011). Algorithmic Results of Independent k-Domination on Weighted Graphs. *Chiang Mai Journal of Science.* **38.** 58-70.

[14] Zmazek, B. and Žerovnik, J. (2003). Computing the weighted Wiener and Szeged number on weighted cactus graphs in linear time. *Croatica Chemica Acta.* **76.** 137-143.

NETWORKS WITH EXTREMAL CLOSENESS

Darja Rupnik Poklukar
University of Ljubljana, Faculty of mechanical engineering, Aškerčeva 6, Ljubljana
e-mail: darja.rupnik@fs.uni-lj.si
Janez Žerovnik
University of Ljubljana, Faculty of mechanical engineering, Aškerčeva 6, Ljubljana
e-mail: janez.zerovnik@fs.uni-lj.si

Abstract: Closeness is a measure of centrality, an important feature of communication and social networks. Extremal networks among all graphs and among several subclasses of graphs including trees and cacti are given. The new concept of generalized closeness and its properties are explored.

Keywords: Closeness, graph operations, extremal graphs.

1 Introduction

In graph theory and network analysis, the concept of centrality is one of the most important indicators to understand the structure and dynamics of networks. Of course, "importance" has a wide number of meanings, leading to many different definitions of centrality.

As mentioned in [6, 19], there are many graph theoretical parameters depending upon the distance such as vertex and edge betweenness, average vertex and edge betweenness, normalized average vertex and edge betweenness, closeness, vertex residual closeness. The aim of closeness and residual closeness is to measure the vulnerability even when the actions disconnect the graph. It was explained in [1, 3] that residual closeness is considered to be more sensitive vulnerability measure than some other known measures. Closeness of some graph classes has been studied recently in [3, 5]. Several interesting results on closeness of graph transformations, regarding vertex residual closeness, normalized vertex residual closeness and closeness centrality for some classes of graphs have been obtained in [2, 18, 19]. This paper points out the measure of closeness vertex centrality and its generalizations.

In [11], closeness (or closeness centrality) of a connected graph was defined as a measure of centrality of a node in a network as

$$\tilde{\mathcal{C}}(v_i) = \sum_{j \neq i} \frac{1}{d(v_i, v_j)}, \tag{1}$$

where $d(v_i, v_j)$ is the distance between vertices v_i and v_j. Thus, the more central a node is, the closer it is to all other nodes. For example, in an information network, closeness is a useful measure that estimates how fast the flow of information would be from a given node to other nodes. In the social network analysis, closeness can be used for finding the individuals who are best placed to influence the entire network most quickly.

Dangalchev in [3], proposes a rather different definition, which is used effectively for disconnected graphs and allows to create convenient formulae for graph operations. In [3], point closeness of vertex v_i is defined by

$$\mathcal{C}_G(v_i) = \mathcal{C}(v_i) = \sum_{j \neq i} 2^{-d(v_i, v_j)}, \tag{2}$$

with $d(v_i, v_i) = 0$. The graph closeness is then defined as

$$\mathcal{C}(G) = \sum_{i} \mathcal{C}(v_i). \tag{3}$$

Dangalchev emphasizes that, in contrast to (1), the graph closeness based on (2) can be used also for disconnected graphs, because if $d(v_i, v_j) = \infty$ then $2^{-d(v_i,v_j)} = 0$.

For example, it is known (see [3]) that for the path on n vertices $\mathcal{C}(P_n) = 2n - 4 + 2^{2-n}$, while closed-form expressions for the star and the complete graph on n vertices are

$$\mathcal{C}(S_n) = \frac{(n-1)(n+2)}{4}, \qquad \mathcal{C}(K_n) = \frac{n(n-1)}{2}.$$

The expression for the cycle on n vertices can be written

$$\mathcal{C}(C_n) = 2n\left(1 - \left(\frac{1}{2}\right)^{\lfloor \frac{n-1}{2} \rfloor} \cdot \left(\frac{3}{4}\right)^{r}\right),$$

where $\lfloor x \rfloor$ denotes the floor function of x and $r \equiv n-1 \pmod 2$.

In general, distance based graph invariants (including centrality measures mentioned above) have natural applications in information and communication networks. However, it is less known, but rather interesting that some distance based invariants are also very popular topological descriptors that are extensively studied in chemical graph theory. In particular, the Wiener number [21], which is just the sum of all distances in a molecular graph, was proved to surprisingly well correlate the structure of molecules to their physicochemical properties and biological activity. Since then, various topological indices are widely used for quantitative relationship studies in mathematical chemistry [8, 22]. The polynomial (see, for example [7])

$$H(G,x) = \sum x^{d(v_i,v_j)}$$

that is associated with a connected graph G was first studied by Hosoya [10]. In the literature this polynomial is called the Hosoya polynomial, the Wiener polynomial [17], and sometimes the Hosoya-Wiener polynomial because its derivative at $x = 1$ equals the Wiener number. We wish to only note in passing that the theory of the Wiener number and Hosoya-Wiener polynomials naturally generalizes to weighted graphs [25], double weighted graphs [12], and may also be generalized to the reliability Wiener number and polynomial [13, 14, 15].

When considering the closeness, it is easy to see that

$$\mathcal{C}(G) = 2H(G, 1/2).$$

Observing two arbitrary graphs, often it is not easy to say which of them has lower (or, higher) closeness. Of course, the closeness of each graph can be computed in polynomial time, as only the distances in the graph are needed. Nevertheless, it makes sense to characterize the graphs with extremal closeness among a given class of graphs.

Here we first recall our recent results [16] on extremal graphs among some subclasses of cacti, and then discuss possible generalizations of these results.

In order to find the extremal graphs with respect to closeness, we first studied certain operations on graphs, following previous work where extremal graphs with respect to some other graph invariants were obtained.

The rest of the paper is organized as follows. In the next section, operation C is defined and it is proven that operation C always increases the closeness. In Section 3, extremal graphs within several classes of graphs are determined. In Section 4 some ideas of generalized closeness and its properties are given.

2 Closeness and operation C

Let G_0, G_1 and G_2 be simple connected graphs, with $|V(G_0)|, |V(G_1)|, |V(G_2)| \geq 2$. Let us define a graph H_1 as a union of graphs G_0, G_1 and G_2 such that $V(G_0) \cap V(G_1) = \{u\}$ and

$V(G_0) \cap V(G_2) = \{v\}$ and $V(G_1) \cap V(G_2) = \emptyset$. Operation C (see [20]) is defined as follows: Graph H_2 is obtained from the graph H_1, such that graph G_2 is transferred from vertex v to vertex u. Similarly, graph H_3 is obtained from the graph H_1, such that graph G_1 is transferred from vertex u to vertex v. The graphs H_1, H_2 and H_3 are shown on Figure 1.

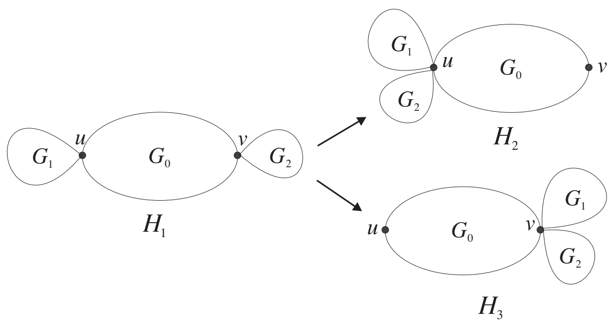


Figure 1: Operation C

Theorem 2.1 *[16] Let G_0, G_1 and G_2 be simple connected graphs, with $|V(G_0)|$, $|V(G_1)|$, $|V(G_2)| \geq 2$ and let H_1, H_2 and H_3 be graphs depicted in Figure 1. Then either*

$$\mathcal{C}(H_2) > \mathcal{C}(H_1) \quad or \quad \mathcal{C}(H_3) > \mathcal{C}(H_1).$$

3 Graphs with extremal closeness

Denote with $\mathcal{G}_n$ the set of all simple graphs with n vertices and let $G \in \mathcal{G}_n$. It was shown in [16]

Observation 3.1 *Let G be an arbitrary graph of the set $\mathcal{G}_n$. Then $\mathcal{C}(N_n) \leq \mathcal{C}(G) \leq \mathcal{C}(K_n)$, with equality on the left if and only if $G \cong N_n$ and equality on the right if and only if $G \cong K_n$.*

Here N_n denotes the null graph and K_n the complete graph on n vertices.

One of the simplest classes of connected graphs are trees. In [16] it was shown that

Theorem 3.2 *[16] Let T be an n-vertex tree with $n \geq 3$, then we have*

$$\mathcal{C}(P_n) \leq \mathcal{C}(T) \leq \mathcal{C}(S_n).$$

Here, P_n and S_n denotes a path and a star on n vertices, respectively.

An interesting class of connected graphs are cacti that first appeared in the scientific literature about 65 years ago as Husimi trees [9]. There are many applications of cacti in chemistry and in the theory of communication networks, see e.g. [23, 24]. Recall that cactus is a graph in which every edge is a part of at most one cycle. Consider a generalization of star SC_n that consists of a number of triangles that all share one (central) vertex. Clearly the number of vertices will be $n = 2t + 1$ when t triangles are used. For n even, add a pendant edge to the central vertex. We have proved in [16]

Theorem 3.3 *Let G be a n-vertex cactus with $n \geq 3$. Then*

$$\mathcal{C}(P_n) \leq \mathcal{C}(G) \leq \mathcal{C}(SC_n).$$

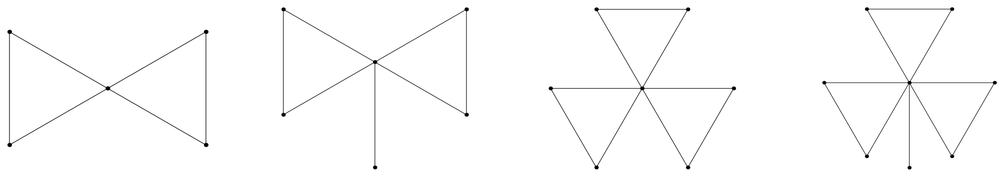


Figure 2: Cacti with maximal closeness with fixed number of vertices $n = 5, 6, 7, 8$.

Note that using operation C we can transform any cactus into a cactus in which all the cycles meet in one vertex. Operation C can also be used to move each pendant edge to the central vertex.

Comparing the cacti with a single vertex that meets all the cycles and all edges not on a cycle, it is straightforward to observe that the maximal closeness is obtained when all the cycles are triangles.

Let us define generalized star $SC_n(k)$ to be a graph with k triangles sharing a vertex called central vertex, and $n-2k-1$ edges that also meet central vertex. Some examples are shown in Figure 3. Clearly $SC_n = SC_n(\lfloor \frac{n-1}{2} \rfloor)$ and $S_n = SC_n(0)$.

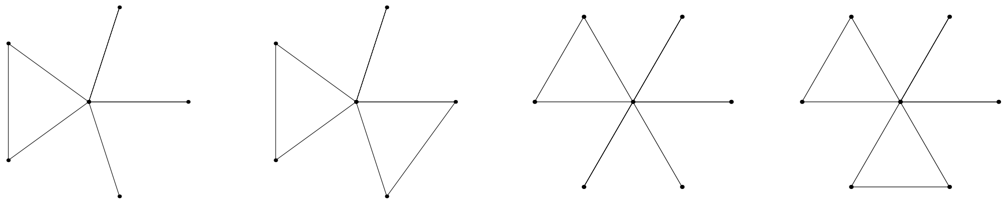


Figure 3: Examples of generalized star $SC_6(1), SC_6(2), SC_7(1), SC_7(2)$.

Theorem 3.4 *[16] Let G be a cactus on n vertices with k cycles. Then*

$$\mathcal{C}(G) \leq \mathcal{C}(SC_n(k)).$$

The minimal cacti however seem to be more difficult to find. It seems that the question may be a challenging open problem.

Let $\mathcal{G}_{n,k}$ denote a set of simple connected graphs with n vertices and k cut-edges. Let K_{n-k}^k be a complete graph on $n-k$ vertices with k additional pendant edges, all adjacent to one of the vertices of K_{n-k}.

Theorem 3.5 *Let $G \in \mathcal{G}_{n,k}$. Then $\mathcal{C}(G) \leq \mathcal{C}(K_{n-k}^k)$, with equality if and only if $G \cong K_{n-k}^k$.*

We conjecture that the minimal cacti among the class of cacti with given number of cut edges is the family constructed as follows. Start with a cycle on $n-k$ vertices. Find k vertices on the cycle so that the sum of all distances is maximal, and attach the edges to these vertices. A proof of our conjecture would solve this open problem.

4 Generalized closeness

It is worth to mention that the results can be naturally generalized. The base 1/2 in the definition of closeness (2) can be replaced with any constant $0 < \alpha < 1$ giving rise to a definition a generalized closeness as

$$\mathcal{C}_\alpha(G) = \sum_i \sum_{j \neq i} \alpha^{d(v_i, v_j)}, \qquad 0 < \alpha < 1.$$

Dangalchev in [4] argues that choosing the proper base for the closeness depends on the properties of the network we want to investigate. For example, for a graph with n vertices we can choose $\alpha = 10^{-k}$ ($n < 10^k$) to separate vertices with different distances. From that kind of generalized closeness we can directly determine the number of vertices on every distance and also determine the radius and the diameter of a graph.

Comparing other measures of vertex centrality (betweenness, degree centrality) with the generalized vertex closeness

$$\mathcal{C}_\alpha(v_i) = \sum_{j \neq i} \alpha^{d(v_i, v_j)}$$

for variable $0 < \alpha < 1$, or with the average value $\int_0^1 C_\alpha(v_i)\, d\alpha$ seems to give some very interesting results.

Therefore, considering closeness for an arbitrary base α can give some very interesting issues for the future work. Note that all results in this paper would easily generalize to any $0 < \alpha < 1$ and hence we claim that the results mentioned here are valid for arbitrary $\mathcal{C}_\alpha$. Also, we clearly have $\mathcal{C}_\alpha(G) = 2H(G, \frac{1}{\alpha})$. Therefore one can naturally extend the questions regarding closeness to analogous questions about the Hosoya-Wiener polynomials. Moreover, based on the definition of closeness, it may be worth noting that the definition of Hosoya-Wiener polynomial for $x \in (0, 1]$ is sound for arbitrary graphs (not only connected graphs).

References

[1] A. Aytaç and Z. Odabaş. Residual closeness of wheels and related networks. *Int. J. Found. Comput. S.*, 22(05):1229–1240, 2011.

[2] V. Aytaç and T. Turaci. Closeness centrality in some splitting networks. *Comput. S. J. of Moldova*, 3(78):251–269, 2018.

[3] C. Dangalchev. Residual closeness in networks. *Physica A*, 365:556–564, 2006.

[4] C. Dangalchev. Residual closeness and generalized closeness. *Int. J. Found. Comput. S.*, 22(8):1939–1948, 2011.

[5] C. Dangalchev. Residual closeness of generalized thorn graphs. *Fundam. Inform.*, 126:1–15, 2018.

[6] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[7] I. Gutman. Some properties of the Wiener polynomials. *Graph Theory Notes New York*, 25:13–18, 1993.

[8] I. Gutman. Degree-based topological indices. *Croat. Chem. Acta*, 86(4):351–361, 2013.

[9] F. Harary and G. Uhlenbeck. On the number of Husimi trees, I. *Proceedings of the National Academy of Sciences of the USA*, 39(4):315–322, 1953.

[10] H. Hosoya. On some counting polynomials in chemistry. *Discrete Appl. Math.*, 19:239–257, 1988.

[11] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):1–4, 2001.

[12] T. Novak, D. Rupnik Poklukar, and J. Žerovnik. The Hosoya polynomial of double weighted graphs. *Ars Math. Contemp.*, 15(2):441–466, 2018.

[13] D. Rupnik Poklukar and J. Žerovnik. On the reliability Wiener number. *Iranian J. Math. Chem.*, 5(2):107–118, 2014.

[14] D. Rupnik Poklukar and J. Žerovnik. The reliability Wiener number of cartesian product graphs. *Iranian J. Math. Chem.*, 6(2):129–135, 2015.

[15] D. Rupnik Poklukar and J. Žerovnik. Reliability Hosoya - Wiener polynomial of double weighted trees. *Fundam. Inform.*, 147:447–456, 2016.

[16] D. Rupnik Poklukar and J. Žerovnik. Networks with extremal closeness. *Fundam. Inform.*, 167:219–234, 2019.

[17] B. E. Sagan, Y. N. Yeh, and P. Zhang. The Wiener polynomial of a graph. *Int. J. Quantum Chem.*, 60:959–969, 1996.

[18] T. Turaci and V. Aytaç. Residual closeness of splitting networks. *Ars Combinatoria*, 130:17–27, 2017.

[19] T. Turaci and M. Ökten. Vulnerability of mycielski graphs via residual closeness. *Ars Combinatoria*, 118:419–427, 2015.

[20] P. Šparl and J. Žerovnik. Graphs with given number of cut-edges and minimal value of Wiener number. *J. Chem. Mod.*, 3(1-2):131–137, 2011.

[21] H. Wiener. Correlation of heats of isomerization and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.*, 69(11):2636–2638, 1947.

[22] Wikipedia contributors. Wiener index — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Wiener_index&oldid=868963324`, 2017. [Online; accessed 22-June-2018].

[23] B. Zmazek and J. Žerovnik. Computing the weighted Wiener and Szeged number on weighted cactus graphs in linear time. *Croat. Chem. Acta*, 76(2):137–143, 2003.

[24] B. Zmazek and J. Žerovnik. The obnoxious center problem on weighted cactus graphs. *Discrete Appl. Math.*, 136:377–386, 2004.

[25] B. Zmazek and J. Žerovnik. On generalization of the Hosoya - Wiener polynomial. *MATCH Commun. Math. Comput. Chem.*, 55:359–362, 2006.

GRAPH THEORY APPLICATIONS IN COMPUTER NETWORK SECURITY: A LITERATURE REWIEV

Gregor Rus, Alenka Brezavšček
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj
E-mail: gregor.rus4@um.si, alenka.brezavscek@um.si

Abstract: The paper comprise a comprehensive review of existing literature on graph theory applications in computer network security. Published paper are analysed, discussed and arranged into four groups according to security issue they are dealing with: network vulnerability analyses, detection of anomalies in the network traffic, protection against malware, and applications in cryptography.

Keywords: graph theory, information security, computer network, application, literature review

1 INTRODUCTION

Graph theoretical ideas are highly utilized by computer science applications [32]. Graph based techniques are found to be especially useful in modelling and routing in the computer networks (e.g., [34]). Representing a computer problem as a graph can provide a different point of view and make a problem much simpler [26]. As computer networks continue to grow in size and complexity, the security aspects have become a critical issue in ensuring business continuity in an organization. In recent years statements in professional literature proved, that graph theory approaches represent a useful tool also for modelling different aspects of computer network security.

The aim of this paper is to review the key applications of graph theory in computer network security. Published papers will be analysed, discussed, and arranged according to security issues they are dealing with. This will provide the reader a comprehensive overview on this attractive and important research area, as well as reveal some opportunities for further research.

2 NETWORK VULNERABILITY ANALYSES

As computer networks continue to grow, evaluating their vulnerability to attacks becomes increasingly important. A large network builds upon multiple platforms and diverse software packages and supports several modes of connectivity. Therefore, when evaluating the security of a network, it is not enough to consider the presence or absence of isolated vulnerabilities. In a complex computer network, a security analyst must take into account the effects of interactions of local vulnerabilities and find global vulnerabilities introduced by interconnections [18]. Consequently, it is desirable to quantify the likelihood of potential multi-step attacks that combine multiple vulnerabilities [35]. Many authors (e.g., [9,11,13,18,20,23,27]) found that this becomes feasible by using a model of causal relationships between vulnerabilities, called *attack graph*. Such a graph is a succinct representation of all paths through a computer network that end in a state where an intruder (i.e., attacker) has successfully achieved his goal.

In general, an attack graph is a sort of *scenario graph*. Such graphs represent all possible scenarios (or paths) that lead to a particular state. Consequently, the attack graphs focus on scenarios where the security incident occurs due to intended and malicious attack from the intruder [30].

Formal definition of attack graph is presented in [18]. Given a set of atomic propositions AP, an attack graph is a tuple $G=(S,T,S_0,S_s,L)$, where S is a set of states, $T \subseteq S \times S$ transition relation between states, $S_o \subseteq S$ is a set of initial states, $S_s \subseteq S$ is a set of success states (i.e., intruder goal), and L: $S \rightarrow 2^{AP}$ a labelling of states with a set of propositions true in that state. A successful attack represents a path $s_1, ..., s_n$ where s_i is connected with s_{i+1}, and s_1 is an initial state while s_n a success state (the attack has been realized).

An example of simple computer network and corresponding attack graph is shown in Fig. 1 [30].

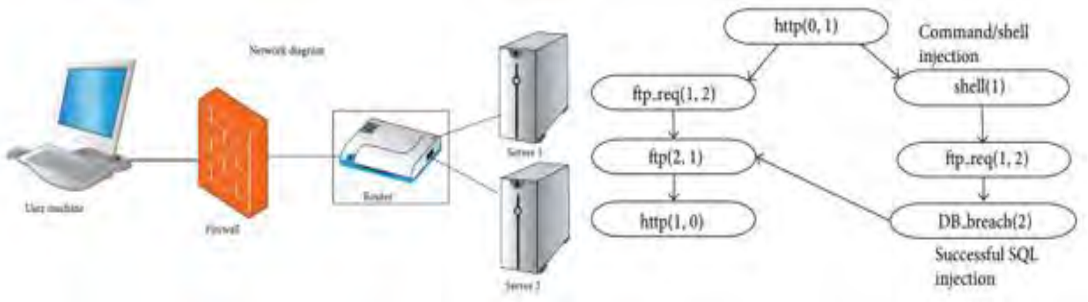


Figure 1: Simple computer network and corresponding attack graph [30].

Authors discuss the potential applications of attack graphs to different areas of network security. As mentioned in [18], attack graphs can serve as a tool for detection an attack on the network, either as a tool to protect network from attack or as a tool for forensic analysis of the network process. Besides, the attack graphs enable assessment or quantifying security risks of enterprise networks (e.g., [9,36]).

Depending on analyses required, there are some variations of attack graphs that can be applied to specific analyses of computer network security. For example, Jha et al. [18] used *probabilistic attack graphs* enabling consideration that some attacks are more likely to occur then other. Edges in such a graph are weighted by the probability of a particular attack occurrence. To be more cost efficient, only more probable paths can be secured (depending on cost/safety priorities).

Furthermore, authors of [15,23] applied the *attack trees* which similarly as the attack graphs present the attacker point of view. The main difference is that the structure of graph is a tree. The root of the tree presents the main goal of the attacker, the paths to leafs represent different ways of attack, while the vertices in between correspond to sub goals the attacker has to achieve to come to the desired outcome. Similarly, as the attack trees, also the *attack-defence trees* are used, which have defence solutions added to possible state transitions [23]. A comprehensive overview of attack graphs applications to computer network security analyses is provided in [20].

In general, the attack graphs can be constructed manually. However, even in very small networks the number of states and their transitions can be very large. Therefore, there is a need for software to generate the attack graphs/trees automatically based on a given network topology. In [31] one of the earliest such implementation is presented. In [9] the algorithm for attack graph generation software is presented, while in [39] different generators (both commercial and open source) are described and reviewed. Szpyrka et al. [33] for example, applied the tool MulVAL (Multi-host, Multi-stage Vulnerability Analysis Language) which provides a reasoning engine for automatically identifying vulnerabilities in an enterprise network.

3 DETECTION OF ANOMALIES IN THE NETWORK TRAFFIC

In recent years the research area of network flow behaviour analysis has attracted many authors. Anomalous traffic detection has become an indispensable component of any network security infrastructure. Traditional methods for detecting traffic anomalies are typically based on machine learning, data mining or the statistical analysis of network models. However, these techniques often generate a huge number of false alarms, and as a result, further work is necessary in order to improve detection accuracy and performance [21].

In this paper we limited our scope to the literature on graph-based network flow behaviour methods. Hu et al. [16] divided them into two categories: 'network-wide based approaches' and 'host-level based approaches'. On the other hand, Noble and Cook [24] classify the graph-based anomaly detection techniques into two types. The first, called 'anomalous substructure detection', searches for specific, unusual substructures within a graph, while the second, denoted as 'anomalous sub-graph detection', partitions the graph into distinct sets of vertices (subgraphs), which are tested against each other in the search of unusual patterns [16]. Akoglu [2] presented several real-world applications of graph-based anomaly detection in diverse domains, including computer traffic, financial, auction, and social networks.

Many authors (e.g., [11,16,17,21]) agree that the concept of *traffic dispersion graphs* (TDGs) can be successfully used to monitor, analyse, and visualize the network traffic. The formal definition of TDGs can be found in [17]. A TDG is a graphical representation of the various interactions ("who talks to whom") of a group of network nodes. For example, in a computer network, a node represents an entity with a distinct IP address and the graph captures the exchange of packages between different network nodes. Informal way of understanding the concept of TDG is to look at them as a sort of network's "social network". As seen in [17], TDGs are useful by using them as application classification, as intrusion deception, and most importantly, they can be implemented at high speeds with low memory requirements.

TDGs capture many of well-known graph theory concepts. Node degrees help us understand relationships between the nodes and can potentially pin point specific nodes with high degrees. By examining TDG it is possible to establish a role of nodes by the difference between in and out degree (for example, it is easy to spot Webserver in TDG, since they have non zero in-degree). One interesting property of TDGs is also the number and size of connected components. Different types of networks have different properties and the number (and size) of connected components vary significantly.

The anomaly detection process using TDGs is described in detail in [21]. The process consists of 4 steps: 1) Sampling network traffic and generating network flows; 2) Creating TDG from network flows in time sampling intervals; 3) Calculating adjacency matrices of the TDG and calculating graph metrics of the TDG; and 4) Comparing values of graph metrics of the TDG with their threshold values. If a value of one of graph metrics exceeds its threshold level then the TDG is an abnormal, otherwise it is a normal one. This process is illustrated in Fig. 3.

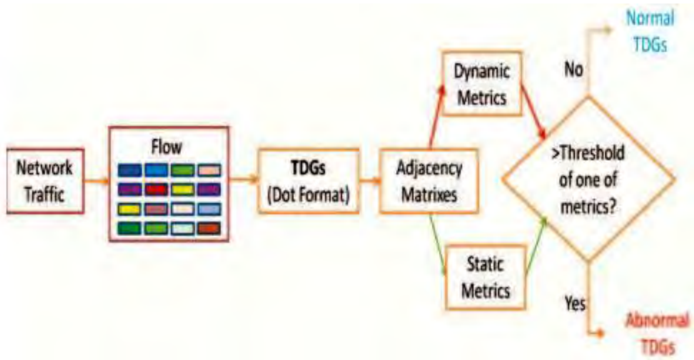


Figure 3: Anomaly detection process using TDGs [21].

Besides TDGs applications, also some related concepts were found to be useful for anomaly detection in the network traffic. For example, Jin et al. [19] introduced the concept of *traffic activity graphs* (TAGs), which are similar to TDGs, but have the structure of bipartite graphs (where bipartite sets correspond to in-hosts and out-hosts respectively). Furthermore, the authors [19] also presented the orthogonal nonnegative matrix tri-factorization method (tNMF), which is used to get from the set of all nodes, the "core" of nodes, where the most traffic is happening. Moreover, some authors (e.g., [6,7]) applied the concept of *traffic* *causality graphs* (TCGs) where the flow in the network is additionally classified by its nature (communication, propaganda, etc.) This type of graph is frequently used in companies and campus network in order to analyse both, the amount and the type of interaction between network members. Finally, a recent work of Hu et al. [16] represented the use of *network flow connectivity graph* (NFCG), which is working on smaller networks inside a big network. The aim is to catch the topology of the network, states and compactness of nodes and statistically present probability of interactions and degree distribution. In "normal" environment the statistical features are random patterns, but some attacks on the network have effect on those features. The comprehensive overview of related work on traffic anomaly detection using graph theory is provided in [16,21].

4 PROTECTION AGAINST MALWARE

One of the important areas of computer network security is related to the protection of a system from harmful, malicious software (i.e., malware). The traditional malware detection schemes rely on a signature-based approach to classify programs as being either malicious or benign. Signature-based approaches are popular due to their low false positive rate and low computational complexity on the end host, both of which are appealing for daily usage. Unfortunately, these schemes have been shown to be easily defeated by simple code obfuscation techniques. With the ease of creating a new malware through these techniques and polymorphic malware becoming more prevalent, non-signature based methods are becoming more attractive [5]. Non-signature malware detection schemes have to rely on malware behaviour and fall into two categories: 'sequence based schemes' and 'graph based schemes' [28]. In the paper, we limited our attention to the latest one.

Most graph based schemes store a large number of *behaviour graphs* of known malware, and for each given program, search the behaviour graphs within the database to find similar graphs. When we can found a malware behaviour graph which is similar to the behaviour graph of the given program, then this program is classified as malware. Otherwise, it is classified as benign software. Such graph based schemes are more robust to malware obfuscation, but they are inefficient in terms of both processing speed and storage overheads. Searching a graph database for graphs similar to a given graph is computationally expensive [28].

In the literature we found many graph theoretic approaches to ensure accurate, efficient and robust non-signature malware detection at end host. For example, Shafiq and Liu [28] applied so called *GZero approach*, Dam and Touili [12] used abstract *API graphs*, while Zhao et al. [40] developed their own *graph-based data mining* method to detect unknown malware.

However, it turns out that already mentioned TDGs (see Section 3) are useful also within the scope of protection against malware. Due to their ability to detect abnormal network traffic, they can be also used to detect abnormal traffic cased by malicious purpose, and can be therefore used as a powerful tool to detect the presence of unwanted applications [11,14,17].

Furthermore, some authors deal with the problem of finding the source of malware in a network [29]. For this purpose, the SIR model (Susceptible-Infected-Recovered) can be used. It is interesting to note that this model was originally developed for detecting the source of human viruses [8] and was later successfully applied to computer networks too. In the SIR model, the networks are classified into three types: susceptible (capable of being infected), infected nodes (that can spread the malware) and recovered (which cannot be infected again). The goal of such software is to determine the source of infection, which is often named the centre of the network. The network is represented by a finite undirected graph, and known statistical methods such as maximum likelihood estimation (MLE) are used [29].

Another approach in graph based malware analysis is proposed in [10]. Authors develop a framework to collect insights and intelligence out of dynamic malware analysis. Malware samples tend to exhibit a cooperative strategy with remote malicious domains and IPs to perpetrate malicious activities (e.g., stealing credentials, spam propagation, advanced DDoS attacks, etc.). They designed and integrate an approach to generate cyber-threat intelligence (CTI) for the purpose of identifying the infrastructure used by malware to threaten the cyber-space. Authors considered different measures of centrality and described a method with pattern recognition. With statistical analysis and decomposition of graph in smaller subgraphs, they present "badness scoring" which ranks IP by how dangerous they are (the badness score takes into account also the degree of IP).

5 APPLICATIONS IN CRYPTOGRAPHY

With the increase use of Internet and other new telecommunication technologies, cryptography has become a key area to research and improve in order to transfer data securely between two or more entities, especially when the data transferred are classified as sensitive and confidential [3]. Even many traditional encryption algorithms exist, there is a need for new, non-standard encryption algorithms enabling decrease of the risk of non-authorized disclosure of data. As there is a need for more secure cryptographic schemes, the application of graph theory in cryptography is going to increase. In this section we highlight recent results of graph theory application in this field.

Al Etaiwi [3] proposed a new encryption algorithm to encrypt and decrypt data securely with the benefits of graph theory properties. The structure of this symmetric encryption algorithm is presented in Fig. 4. The algorithm uses the concepts of *cycle graph*, *complete graph* and *minimum spanning tree* to generate a complex cipher text using a shared key.

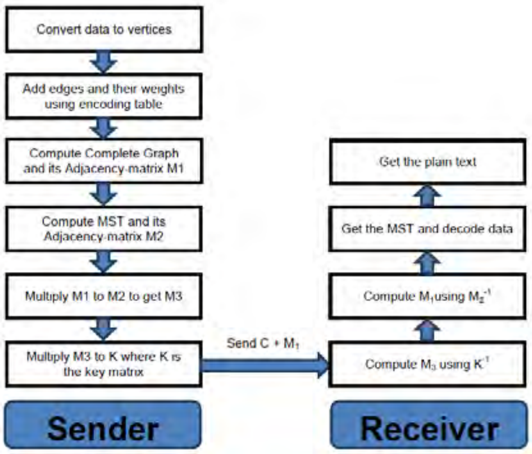


Figure 4: An example of symmetric encryption algorithm based on graph theory concepts [3].

Lu et al. [22] present their solution with *weighted graphs* with weights on set of both vertices and edges, but only with weights of edges being publicly known. Yamuna et al. [38] proposed an encryption mechanism where the nodes are organized in *Hamiltonian path*. Their method aims to use the adjacency matrix as an additional parameter to encrypt and forward the data, and use matrix properties for decryption. Priyadarsini [25] reviewed and analysed some of the cryptographic algorithms based on general graph theory concepts, extremal graph theory and expander graphs.

Yamuna et al. [37] proposed selective encryption mechanism for wireless ad hoc networks based on specific key and spanning tree concept. Since only selected packets are encrypted using selective encryption scheme, it reduces the communication overhead. The nodes are being organized in a minimum spanning tree fashion using Prim's algorithm. This mechanism provides protection of privacy in communication as it avoids the formation of self-loops and parallel edges and key is exchanged only among the authenticated neighbours only.

Agarwal and Uniyal [1] developed a scheme for secure communication using *prime weighted graph*, while in recently published paper an encryption-decryption algorithm using *Euler graphs* and *Hamiltonian circuits* has been proposed and discussed [4].

6 CONCLUSION

Graph theory has become a very essential component in many applications in the computing field including networking and security. Unfortunately, it is also amongst the most complex topics to understand and apply. This paper gives a brief overview of the subject and the applications of graph theory in computer network security, and provides pointers to key research and recent survey papers in the area. According to the security issue they are dealing with, the published papers were classified into four groups: network vulnerability analyses, detection of anomalies in the network traffic, protection against malware, and applications in cryptography. Findings of the paper can serve as a guide for the applied mathematicians and other interested professionals who would like to know more about network security, cryptography and cyber security based of graph theory, and reveals some ideas for further research.

References

[1] Agarwal, S., Uniyal, A. S. (2015). Prime weighted graph in cryptographic system for secure communication, International Journal of Pure and Applied Mathematics, 105 (3): 325-338.

[2] Akoglu, L. et al. (2014). Graph-based Anomaly Detection and Description: A Survey, Cornell University, arXiv: 1404.4679.

[3] Al Etaiwi, W. M. (2014). Encryption Algorithm Using Graph Theory. Journal of Scientific Research and Reports, 3 (19): 2519-2527.

[4] Amudha, P. et al. (2018). An Application of Graph Theory in Cryptography. International Journal of Pure and Applied Mathematics, 119 (13): 375-383.

[5] Anderson, B. at al. (2011). Graph-based malware detection using dynamic analysis. J Comput Virol, 7:247–258, doi: 10.1007/s11416-011-0152-x.

[6] Asai, H. et al. (2011). Traffic causality graphs: profiling network applications through temporal and spatial causality of flows, Proceedings of the 23rd International Teletraffic Congress, September 06-09, 2011, San Francisco, California.

[7] Asai, H. et al. (2014). Network application profiling with traffic causality graphs. Int. J. Network Mgmt, 24: 289-303. doi:10.1002/nem.1865.

[8] Bailey, N. T. J. (1975). The Mathematical Theory of Infectious Diseases and its Applications. Griffin, London.

[9] Barik, M. et al. (2016). Attack Graph Generation and Analysis Techniques. Defence Science Journal, 66(6): 559-567.

[10] Boukhouta A. et al. (2015). Graph-Theoretic Characterization of Cyber-threat Infrastructures, Digital Investigation: The International Journal of Digital Forensics and Incident Response, 14 (1): S3-S15.

[11] Cheema, A. S. et al. (2013). Network Security Using Graph Theory, International Journal of Innovations in Engineering and Technology (IJIET), 2 (3): 131-138.

[12] Dam, K. H. T. and Touili, T. (2017). Malware Detection based on Graph Classification. In Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP 2017), 455-463 doi: 10.5220/0006209504550463.

[13] Doynikova, E. V. and Kotenko, V. (2018). Improvement of attack graphs for cybersecurity monitoring: handling of inaccuracies, processing of cycles, mapping of incidents and automatic countermeasure selection”, Tr. SPIIRAN, 57: 211–240.

[14] Faloutsos, M. (2013). Detecting Malware with Graph-based Methods: Traffic Classification, Botnets, and Facebook Scams. WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web: 495-496. May 13–17, 2013, Rio de Janeiro, Brazil.

[15] Haque, M. S. and Atkison, T. (2017). An Evolutionary Approach of Attack Graph to Attack Tree Conversion. International Journal of Computer Network and Information Security. 9: 1-16. 10.5815/ijcnis.2017.11.01.

[16] Hu, H. et al. (2018). Graph analysis of network flow connectivity behaviors. Turkish Journal of Electrical Engineering & Computer Sciences,

http://online.journals.tubitak.gov.tr/openInPressDocument.htm?fileID=1099831&no=195715&fileType=Report%20Document.
[17] Iliofotou, M. et al. (2007). Network monitoring using traffic dispersion graphs (TDGs. IMC '07 Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 315-320.
[18] Jha, S. et al. (2002). Two Formal Analys s of Attack Graphs. CSFW '02 Proceedings of the 15th IEEE workshop on Computer Security Foundations, June 24 - 26, 2002.
[19] Jin, Y. et al. (2009), Unveiling core network-wide communication patterns through application traffic activity graph decomposition, Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems, June 15-19, 2009, Seattle, WA, USA .
[20] Kotenko, I. and Stepashkin, M. (2006). Attack Graph Based Evaluation of Network Security. CMS 2006, LNCS 4237, Eds. Leitold, H. in Markatos, E., 216 – 227.
[21] Le, D. Q. et al. (2011). Traffic dispersion graph based anomaly detection. V SoICT '11 Proceedings of the Second Symposium on Information and Communication Technology, Hanoi, Vietnam - October 13 - 14, 2011, ACM New York, 36-41.
[22] Lu, S. et al. (2008). Visual Cryptography on Graphs, V: Hu X., Wang J. (Eds) Computing and Combinatorics. COCOON 2008. Lecture Notes in Computer Science, 5092. Springer, Berlin, Heidelberg.
[23] Nagaraju, V. et al. (2017). A survey of fault and attack tree modeling and analysis for cyber risk management, IEEE International Symposium on Technologies for Homeland Security (HST), Waltham, MA, 1-6.
[24] Noble, C. C. and Cook, D. J. (2003). Graph-based anomaly detection. Proceedings of the 9th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, D.C. - August 24 - 27, 2003, ACM Press, 631–636.
[25] Priyadarsini, P. L. K. (2015). A Survey on some Applications of Graph Theory in Cryptography. Journal of Discrete Mathematical Sciences and Cryptography, 18 (3): 209-217.
[26] Rangaswamy, K. D. and Gurusamy, M. (2018). Application of Graph Theory Concepts in Computer Networks and its Suitability for the Resource Provisioning Issues in Cloud Computing-A Review. Journal of Computer Science, 14(2): 163-172.
[27] Sahakyan, H. and Alipour, D. (2015). On Attack Model of Network Security. International Journal Information Content and Processing, 2 (1): 26-42.
[28] Shafiq, Z. and Liu, A. (2017). A graph theoretic approach to fast and accurate malware detection. 2017 IFIP Networking Conference (IFIP Networking) and Workshops, Stockholm, 2017, 1-9. doi: 10.23919/IFIPNetworking.2017.8264865.
[29] Shah, D. and Zaman, T. (2010). Detecting sources of computer viruses in networks: theory and experiment. Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems, 203-214.
[30] Shandilya, V. et al. (2014). Use of Attack Graphs in Security Systems. Journal of Computer Networks and Communications, Vol. 2014, Article ID 818957, 13 pages.
[31] Sheyner, O. et al. (2002). Automated generation and analysis of attack graphs. Proceedings 2002 IEEE Symposium on Security and Privacy, Berkeley, CA, USA, 2002, 273-284.
[32] Shirinivas, S. G. et al. (2010). Applications of Graph Theory in Computer Science an Overview. International Journal of Engineering Science and Technology. 2 (9): 4610-4621.
[33] Szpyrka M., et al. (2013). Telecommunications Networks Risk Assessment with Bayesian Networks. In: Saeed K., Chaki R., Cortesi A., Wierzchoń S. (Eds) Computer Information Systems and Industrial Management. CISIM 2013. Lecture Notes in Computer Science, vol 8104. Springer, Berlin, Heidelberg.
[34] Vallivaara, V. (2015). Designing Information Secure Networks with Graph Theory. Short Paper on CyCON 2015.
[35] Wang, L. et al. (2008). An Attack Graph-Based Probabilistic Security Metric. IFIP Annual Conference on Data and Applications Security and Privacy, DBSec 2008: Data and Applications Security XXII. 283-296.
[36] Xie, L. et al. (2013). Network Security Risk Assessment Based on Attack Graph. Journal of Computers. 8. 10.4304/jcp.8.9.2339-2347.
[37] Yadhu , R. (2013). Graph Theory Application in Selective Encryption Mechanism for Wireless Ad Hoc Network, International Journal of Emerging Trends and Technology in Computer Science (IJETTCS), 2(2): 363-365.
[38] Yamuna, M. et al. (2012). Encryption Using Graph Theory and Linear Algebra. International Journal of Computer Application, 5 (2): 102-107, https://rspublication.com/ijca/OCT12/12.pdf
[39] Yi S. et al. (2013). Overview on attack graph generation and visualization technology, International Conference on Anti-Counterfeiting, Security and Identification (ASID), Shanghai, 2013, 1-6. doi: 10.1109/ICASID.2013.6825274.
[40] Zhao, Z. et al. (2013). An unknown malware detection scheme based on the features of graph. Security Comm. Networks, 6:239–246. doi: 10.1002/sec.52.

FRAUD DETECTION IN TRANSACTIONS USING SOCIAL NETWORK ANALYSIS

Anja Žnidaršič
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, 4000 Kranj, Slovenia
E-mail: anja.znidarsic@fov.uni-mb.si

Manja Krajnčič
University of Maribor, Faculty of Natural Sciences and Mathematics
Koroška cesta 160, 2000 Maribor, Slovenia
E-mail: manja.krajncic@student.um.si

Drago Bokal
University of Maribor, Faculty of natural Sciences and Mathematics
Koroška cesta 160, 2000 Maribor, Slovenia
E-mail: drago.bokal@um.si

Abstract: We apply social network analysis to the transaction network of subsidized student meals to identify anomalous patterns that could be interpreted as instances of fraud. Our approach is unsupervised, and consists of component identification, their statistical analysis, and identification of suspiciously frequent and tightly consecutive subsidy which may indicate either subscription or superposition fraud.

Keywords: fraud detection, social network, social network analysis, subsidized student meals.

1 INTRODUCTION

Fraud can be operationalized in different ways, hence the field where it appears has a great influence to the definition of fraud. Fraud could be defined as "a crime where the objective is to gain money by an illegal form", and almost any enterprise that involves money or service (e.g. insurance, telecommunications, financial, credit cards etc.) can be attacked by fraudulent acts [9].

1.1 Fraud detection methods

Bolton and Hand [5] emphasized that fraud detection comes into play once fraud prevention has failed. And, since this is a continuous process, a newly developed successful fraud detection methods are crucial to improve fraud prevention [12].

Different anomaly detection techniques could be used to detect fraudulent behaviour. The main objective of anomaly detection (also known as outlier detection, deviation detection, novelty detection, and exception mining) is to identify anomalous or unusual data, but on the other hand interesting and rare patterns, from a given dataset [2].

Fraud detecting methods can generally be classified into supervised and unsupervised methods [5]. Supervised methods are those in which we create fraud detecting model with the help of historical data where the dataset contains also both a variable specifying normal and fraudulent behavior. The main disadvantage of that approach is that datasets are usually unbalanced with majority of entities being classified as non-fraudulent, and therefore, standard classification techniques may have limited performance and have to be adapted [9]. On the other hand, unsupervised methods look for unusual subsets in data that are most different from normal ones. The main advantage of unsupervised methods is that they do not rely on accurate identification of fraudulent behaviour in dataset which could be often in short supply or even non-existent and therefore they are able to detect new types of fraud that have not been discovered before [5].

The most commonly known methods for anomaly detection are profile generators, rule generators, Bayesian classification, neural networks, support vector machine, etc. They are developed for both supervised and unsupervised methods and usually they are combined together.

1.2 Graph based fraud detection

Pironet and coworkers [9] pointed out that "[fraud is] perpetuated by people or organizations that are not separated from the rest of the world". In order to successfully and effectively detect the fraudulent behaviour, it is important not to look only on individual entities at the time, but to examine connections among those entities. In social network analysis (SNA), anomaly detection focus is given to the manner how interactions between individuals influence other interactions between and with other individuals in comparison to the 'traditional' analysis where individuals' attributes are assumed to follow some population level distribution which do not take into account the relationships among individuals [11]. Therefore, SNA methods [8, 15] could be a key approach to detect fraudulent behaviour or anomalies in connected environments. First, a basic definition of a social network and related concepts is provided, with known application of SNA techniques to detect fraudulent behaviour.

Connections among entities are mathematically described by graphs, more precisely by a set (or sets) of vertices (entities, also named as units, nodes or actors) and relations (ties) among pairs of vertices. If additional information on relations and/or vertices are added, the structure is called a social network (SN) [15]. If vertices in the network are classified into two sets, then we have two-mode network, and usually units from the first set are connected to units in the other, while units inside the individual set are unconnected. Prior to the analysis we have to consider that: "Exactly how a social network is chosen to be represented will depend of course on the type of anomalies to be detected" [11].

Savage with coworkers [11] proposed a two step approach for detection of anomalies in online social networks, which can be applied to any anomaly detection based on network data: (i) the selection and calculation of network features, and (ii) the classification of observations from this feature space. Magomedov with coworkers [7] design an algorithm for feature calculation, outlier detection and identifying specific sub-graph patterns. They found out, based on regression analysis and decision tree algorithm, that top five most important features are to detect fraudulent entities are: amount of transaction, degree of the sender vertex, total amount of transactions (incoming and outgoing) corresponding to the sender, number of vertices in the sender's egonet, and outdegree of the sender vertex.

In the literature, network analysis is used to reveal fraudulent behaviour in numerous situations: telecommunication frauds [11, 16], insurance fraud [12], credit cards [1, 4, 6, 17], joint medical fraud [13], non-paying of VAT [9], fraud in online auction system [3], social security fraud [14], fake profiles on SN [10, 11], etc. Methods of SNA and their results could be used also in data preparation phase to standard classification methods such as decision trees and SVM. Pironet with coworkers [9] analyzed fraudulent organizations and they found out that information obtained with SNA methods from the patterns common to fraudulent organizations contributed significantly to the creation of a better classifier and therefore revealed fraud.

2 DATA

A subsidized student meal (hereinafter referred to as SSM) is a partially funded meal by the Republic of Slovenia, which is intended for students during their study. The purpose of the SSM is to provide each student with at least one warm meal per day. A price of each meal consists of a subsidized part amounting to EUR 2.63 (2018) which is settled by the state, as well as an additional fee which is determined by the individual provider of SSM. The information system for the SSM was developed by Margento R & D in 2010. The system enables to cash in student vouchers by mobile phones or contactless chip cards.

Our data consists of 80891 anonymous students and 575 anonymous providers with 4825401 transactions from January till December 2015. For each transaction, additional data such as date and time of transaction, and amount of surcharge are available. Data are arranged as a 2-mode network (or bipartite graph), where vertices represent providers (P) of student meals (restaurants and other dining facilities) and students (S) and edges represent transactions (T) among provider and students. Clearly, there could be several transactions (in different time points) between a given consumer and provider, hence in general, our network has multiple edges.

The transactions are logged by a central information system, hosted by the company Margento. The management of the information system is the responsibility of the Student association of Slovenia, who is responsible for managing contractual relations with the service providers and taking care of status information of service consumers (students). They are also responsible to provide a supervision of the service by managing a cohort of inspectors, who observe the service providers and consumers and may issue a warning following observed misbehavior or even recommend further processes, such as detailed investigation. Our fraud detection results aim at guiding these inspectors towards specific providers and specific anomalous behaviors.

Network of transactions, described above, was preprocessed to be suitable for answering our research question with network analysis methods. First, from the 2-mode network (or bipartite transaction graph), we identify all students with a transaction within less than 30 minutes by the same provider. Due to complexity of original 2-mode network, we decided to produce a special network of proximate incidence graph transactions, where a pair consumer-provider is presented by a vertex and an edge between two vertices is present if transactions of two students with the same provider were executed within 30 minutes. This could mean that a pair (or group) of students was having lunch together, while large number of such transactions could be suspicious.

The obtained network of proximate transactions consists of 2683 vertices with 4329 edges. The network is valued meaning that weights on edges represent the number of transactions two students have had with the same provider within 30 minutes. 2771 edges has value one, while 1558 edges have values greater than 1.

3 RESULTS

From the proximate transaction network, we omit all the edges with weight less than 50 and remove isolated vertices. Reduced network (Figure 1) has only 23 vertices with 10 separate components consisting of at least two vertices. Vertex label consists of "consumer ID", "000", and "provider ID". To emphasize, weight on an edge represent number of transactions two students have had with the same provider within 30 minutes. Size of a vertex is proportional to number of visits student had for the corresponding provider (represented by that vertex). For example, component in the right bottom corner in Figure 1 represent two students, "149355" and "149356", whose transactions with provider "2302" were executed within 30 minutes together with 21 and 19 other students, respectively. Line value between these two vertices is 53 meaning that those two students visited provider "2302" within 30 minutes 53 times. The component next to that (second from the right in the bottom) represents students "141585" and "141594" who visited provider "1188" for 96 times within 30 minutes and they have not visited that provider with any other student.

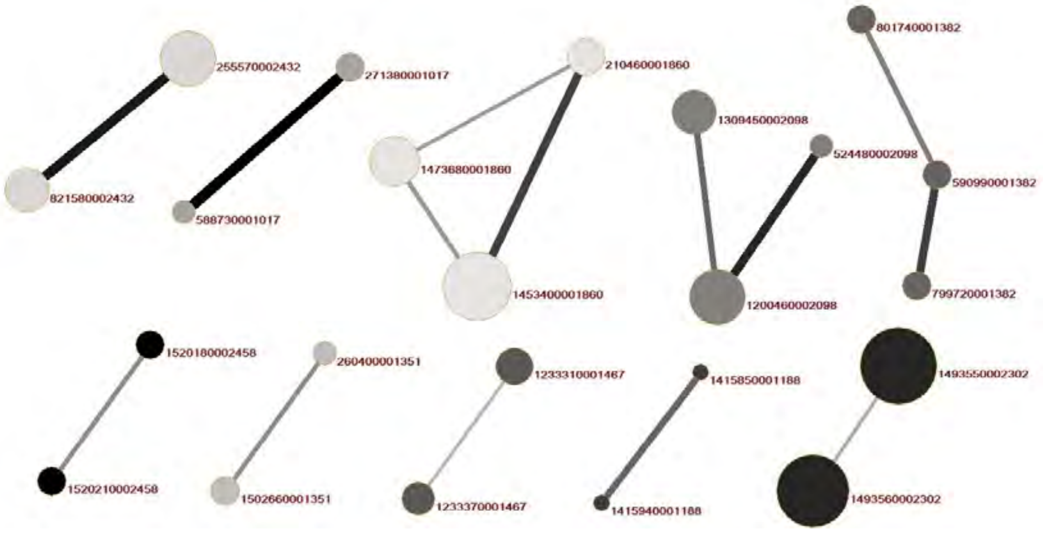


Figure 1: Network with 10 components where vertices represent pairs students-provider and weight on edges number of transactions within 30 minutes.

In the next step, we examined all the revealed suspicious providers closely. Due to space limitation, here we provide detailed analysis for three of them. Most of them seem to represent providers with a group of students, who came to them because of similar schedules, but three of them stand out. Provider "1188", whose transactions were described above, was within the time frame of 30 minutes visited only by two students ("141585" and "141594").

Figure 2 presents a network of consumers of provider "1017" where edges weights represents number of transactions for two consumers which occur less than 30 minutes apart. The network consists of 10 vertices where a group of two students ("27138" and "58873") stands out. They visited that provider 151 times within 30 minutes. Successive transactions for these two consumers have on average occurred within 9 seconds, while the average of all transactions with this provider was 22 seconds.

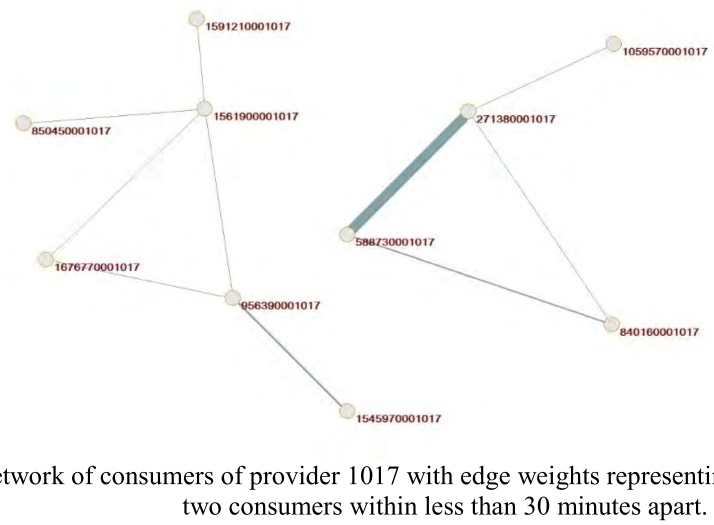


Figure 2: Network of consumers of provider 1017 with edge weights representing number of transactions for two consumers within less than 30 minutes apart.

Network for provider “2432” consist of 26 vertices (Figure 3) and only the previously revealed consumers “25557” and “82158” stand out. The observed edge weight is 136 meaning that in 136 out of 365 days the transactions of those two consumers were provided within 30 minutes. It could be observed from Figure 3 that provider “2432” has a group of four consumers. Detailed analysis of time stamps revealed that in the first part of the year all four were having lunch together, while from October on only two remain, meaning that the two consumers did not have a student status anymore. Transactions of those four students occurred on average within 9 seconds and the average of all transactions was 26 seconds.

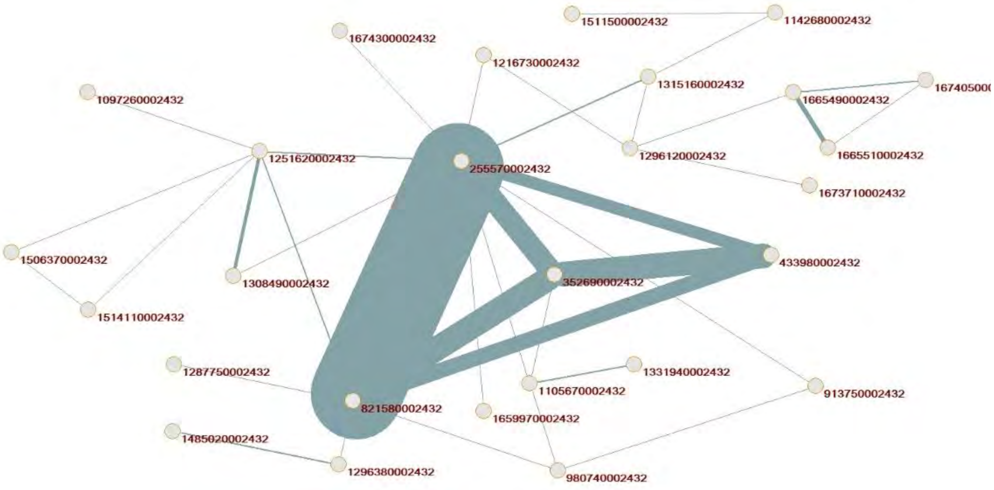


Figure 3: Network of consumers of provider 1017 with edge weights representing number of transactions for two consumers within less than 30 minutes apart.

4 DISCUSSION AND CONCLUSIONS

Presented results of our social network analysis identified three instances of providers, who may have committed either subscription fraud (applying for subsidies identifying as students without actually providing for the service, where the student ceased to use the service) or superimposed fraud (similarly, but while the student was still occasionally using the service). The suspiciousness is based on exceptional frequency of pairwise transactions and on exceptionally short interval between transactions. The techniques used are a specific instance of a more general approach that adapts detector-constructor fraud detection method to social network analysis and uses suspiciousness indicators as behavior profiling basis to detect anomalous patterns in social networks, which is part of our further research. The results of our unsupervised analysis could be applied to direct the investigation by inspectors who would provide legal evidence about identified suspiciousness.

Acknowledgement

We acknowledge the support of the team in Margento R&D.

References

[1] Abdallah, A., Maarof, M. A., Zainal, A. 2016. Fraud detection system: A survey. Journal of network and computer applications, 68, 90-113, doi=10.1016/j.jnca.2016.04.007.

[2] Ahmed, M. Mahmood, A. N., Islam, M. R. 2016. A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278-288, doi.org/10.1016/j.future.2015.01.001.

[3] Bangcharoensap, P., Kobayashi, H., Shimizu, N., Yamauchi, S., & Murata, T. (2015). Two Step graph-based semi-supervised Learning for Online Auction Fraud Detection. Lecture Notes in Computer Science, 165–179. doi:10.1007/978-3-319-23461-8_11

[4] Bolton, R. J., Hand, D. J. 2001. Unsupervised profiling methods for fraud detection. In Proceedings of Credit Scoring Credit Control, 235–255

[5] Bolton, R. J., Hand, D. J. 2002. Statistical Fraud Detection: A Review. Statistical Science 17(3), 235-249.

[6] Lebichot, B., Braun, F., Caelen, O., & Saerens, M. 2016. A graph-based, semi-supervised, credit card fraud detection system. Complex Networks & Their Applications V, 721–733. doi:10.1007/978-3-319-50901-3_57

[7] Magomedov, S., Pavelyev, S., Ivanova, I., Dobrotvorsky, A., Khrestina, M., Yusubaliev, T. 2018. Anomaly Detection with Machine Learning and Graph Databases in Fraud Management. *International Journal of Advanced Computer Science and Applications*, 9(11). http://dx.doi.org/10.14569/IJACSA.2018.091104

[8] de Nooy, Wouter, Andrej Mrvar, and Vladimir Batagelj. 2018. Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software (Structural Analysis in the Social Sciences). Cambridge University Press.

[9] Pironet, M., Antunes, C., Moura, P., & Gomes, J. 2009. Classification for Fraud Detection with Social Network Analysis. Technical University of Lisbona. http://pironet.pt/miguel/TeseMestrado.pdf [Accessed date: 21/05/2019]

[10] Ramalingam, D., Chinnaiah, V. 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review, Computers and Electrical Engineering, 65:165-177. https://doi.org/10.1016/j.compeleceng.2017.05.020.

[11] Savage, D., Zhang, X., Yu, X., Chou, P., Wang, Q. 2014. Anomaly detection in online social networks. Social networks, 39, 62-70.

[12] Stripling, E., Baesens, B., Chizi, B., & vanden Broucke, S. (2018). Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud. Decision Support Systems, 111, 13–26. doi:10.1016/j.dss.2018.04.001

[13] Sun, C., Yan, C., Li, Z., Zheng, Q., Lu, X. Cui, L. 2018. Abnormal Group based Joint Medical Fraud Detection. IEEE Access, 1–1. doi:10.1109/access.2018.2887119

[14] Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-Based Fraud Detection for Social Security Fraud. Management Science, 63(9), 3090–3110. doi:10.1287/mnsc.2016.2489

[15] Wasserman, S., Faust, K. 1998. Social Network Analysis: Methods and Applications. Cambridge, New York, Melbourne: University Press.

[16] Yan, H., Jiang, Y., & Liu, G. (2018). Telecomm Fraud Detection via Attributed Bipartite Network. 2018. 15th International Conference on Service Systems and Service Management (ICSSSM). doi:10.1109/icsssm.2018.8464982

[17] Zanin, M., Romance, M., Moral, S., Criado, R. 2018. Credit Card Fraud Detection through Parenclitic Network Analysis. Complexity, 2018, 1–9. doi:10.1155/2018/5764370

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 4:

High-Performance Computing and Big Data

RUNNING DEEP LEARNING EXPERIMENTS OVER THE PRACE 5IP INFRASTRUCTURE

Agnès Ansari, CNRS/IDRIS France, agnes.ansari@idris.fr
Alberto Garcia Fernandez, CNRS/IDRIS France, alberto.garciafernandez@idris.fr
Bertrand Rigaud, CNRS/CC-IN2P3 France, bertrand.rigaud@cc.in2p3.fr
Marco Rorro, CINECA Italy, m.rorro@cineca.it
Andreas Vroutsis, EPCC UK, A.Vroutsis@epcc.ed.ac.uk

Abstract: The PRACE Data Analytics service relies on a set of coherent components: frameworks, libraries, tools and additional features to support, facilitate and promote the data analytics activities in PRACE. We present the results we obtained for a set of deep learning benchmarks and real use cases we ran on the different PRACE architectures while using these components that confirm their efficiency and present the additional features we put in place.

Keywords: Data analytics, deep learning, tensorflow, keras, horovod

1 INTRODUCTION

The purpose of the Data Analytics Service was to investigate the deployment of new Data Analytics technologies on HPC systems, and develop prototype solutions by means of pilot scientific use cases to access the functionalities. To achieve these objectives, we carried out two main actions in order to develop the prototypes and think about a global data analytics service for the purpose of a future deployment and production service:

- Identify a set of Deep Learning and Machine Learning services, frameworks, tools and libraries to be made available on the PRACE Data Analytics infrastructure, composed of systems running different architectures
- Offer any additional tools or services that facilitate the use of the Deep Learning and Machine Learning services on the PRACE infrastructure and allow users to quickly gain expertise in these fields

In chapter 2, we present the deep learning frameworks and libraries we identified to be the most appropriate. Chapter 3 describes the experiments we ran. In chapter 4, we present the results. Chapter 5 describes the additional features we have identified following the experiences we got from running the prototypes.

2 DEEP LEARNING FRAMEWORKS AND LIBRARIES

We initially browse through a large promising set of frameworks and libraries whose popularity has changed in a different way since the beginning of PRACE-5IP. We finally consider a reduced set of them, composed of the most popular ones. Most of these components were already available at each partner site, but in order to offer a data analytics service, it was important to test the feasibility, efficiency, reliability, ease of use and scalability of these tools by analysing their behaviours from small to large systems.

The deep learning experiments have been conducted with TensorFlow and Caffe. Apart these frameworks, some additional libraries and components have been considered such as Keras, Horovod. Note that Keras is now fully integrated to TF2.0. TF2.0 also includes a new distributed training feature.

We based our evaluation on two main objectives:

- the capability to run standard benchmarks as a preliminary validation phase
- the identification of real use cases that we got through the launch of a call for prototypes, in order to complete the preliminary phase

3 DESCRIPTION OF THE DEEP LEARNING BENCHMARKS AND USE CASES

In the first phase, we used these frameworks and libraries to run standard benchmarks ranging from basic, to small, and then larger, varying the dataset size and the model size.

Table 1 and Table 2 below show the main characteristics of the different benchmarks we ran and the HPC resources we used:

Table 1: Benchmarks description

	Model	***Dataset***	***Framework***
Basic benchmark	AlexNet [3], GoogleLeNet [8], Overfeat [5], VGG11 [6]	No (image created in memory)	Tensorflow - Caffe PowerAI
CIFAR-10 benchmark	Simple CNN, 12 layers	CIFAR10 163MiB-10 classes	Tensorflow - PowerAI
ImageNet benchmark	VGG-19 [7], ResNet50 [2]	ImageNet-138GB (training) 1000 classes	Tensorflow- Caffe PowerAI

Table 2: HPC resources

Site	***Cluster Name***	***Architecture/CPU***	***Accelerators***
CNRS/IDRIS	Quessant	OpenPower 8 – 2 Power 8/node	4 GPUs/node – Nvidia Tesla P100
CINECA	Davide	OpenPower 8 – 2 Power 8/node	4 GPUs/node – Nvidia Tesla P100
CNRS/CC-IN2P3	K80 cluster	Intel – 2 Xeon E5-2640v3/node	4 GPUs/node – Nvidia K80
EPCC	Urika-GX	Intel – 2 Xeon E5-2695	-
EPCC	Cirrus	Intel – 2 Xeon E5-2695	-
CINECA	Marconi	Intel – Xeon Phi 7250	-
CINECA	Galileo	Intel – 2 Xeon E5-2630/node	2 GPUs/node – Nvidia K80

In order to extrapolate the execution of standard benchmarks to real use cases, we also ran an Astrophysics use case. It comes from a deep learning challenge. It uses 20 000 images (2 GB) made of two galaxies overlapping. These galaxies have been extracted from real images from the Hubble Space Telescope and combined manually to create the blends. The goal of the challenge is to train a model that can automatically detect the contiguous region where the light of the two galaxies overlap. The model predicts a probability for each pixel to belong to such region, and the probabilistic image is then thresholded to obtain an actual prediction. It uses the UNet model [4].

4 DEEP LEARNING BENCHMARKS AND USE CASES RESULTS

The results of all prototype services we mentioned in the previous sections, are described in details in the white paper [1]. We present hereafter a subset of these results.

Except for the Basic benchmarks below, the plots show the Tensorflow results. We did not plot the Caffe results as they are very poor compared to the Tensorflow ones. Indeed and although it depends on the benchmarks, the Caffe performance can appear to be between 4 and 5 times worse for 1 and 2 GPUs and between 10 and 12 times worse for 4 GPUs.

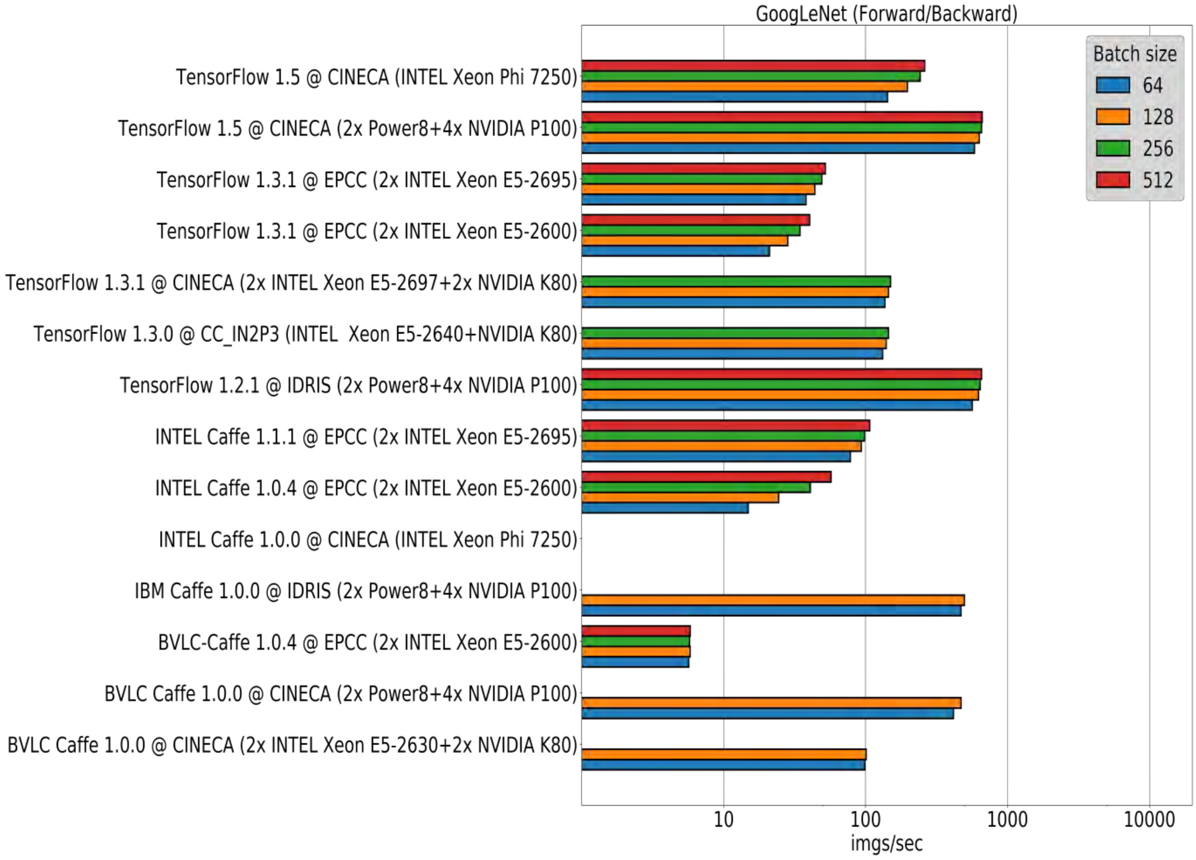


Figure 1: Basic benchmarks – Tensorflow and Caffe with the trained model GoogLeNet (X-axis: number of images per second [log scale], Y-axis: framework and architecture) and for a varying set of batch size with one GPU only

The results reported in Figure 1 above show that the performance levels depend mostly on the hardware configuration: the GPU NVIDIA P100 gets the best result with TensorFlow, followed by the performance obtained with Caffe with either the BVLC and IBM release. The INTEL Xeon Phi 7250 with TensorFlow follows, doing better than the GPU NVIDIA K80.

Figure 2 and Figure 3 below show the performance of Tensorflow (X-axis: batch size, Y-axis: execution time (seconds)) and for different architectures for the CIFAR-10 benchmark and the Astrophysics use case. As for the basic benchmarks, the 2 figures show that the performances which depend mostly on the hardware i.e. the GPU NVIDIA P100 (IDRIS and CINECA) outperform the GPU NVIDIA K80 (CC-IN2P3), as expected.

The results show also the huge impact of the batch sizes with small batch sizes producing poor execution time performance compared to large batch sizes. Unlike the P100, the K80 shows a significant performance gain between one and two GPUs whereas the gain is less clear with four GPUs, that show a slight performance, but far from linear with the number of GPUs. The 2 prototypes at IDRIS with or without containers don't show a significant difference in term of performance.

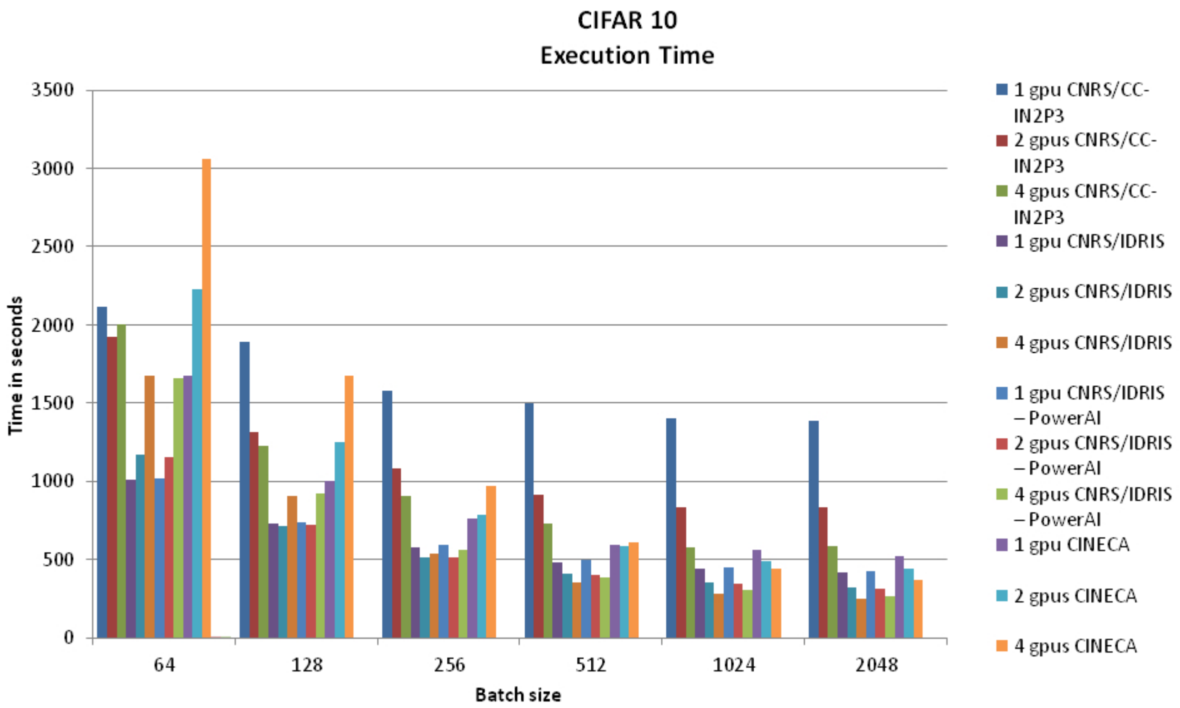


Figure 2: CIFAR-10 benchmark

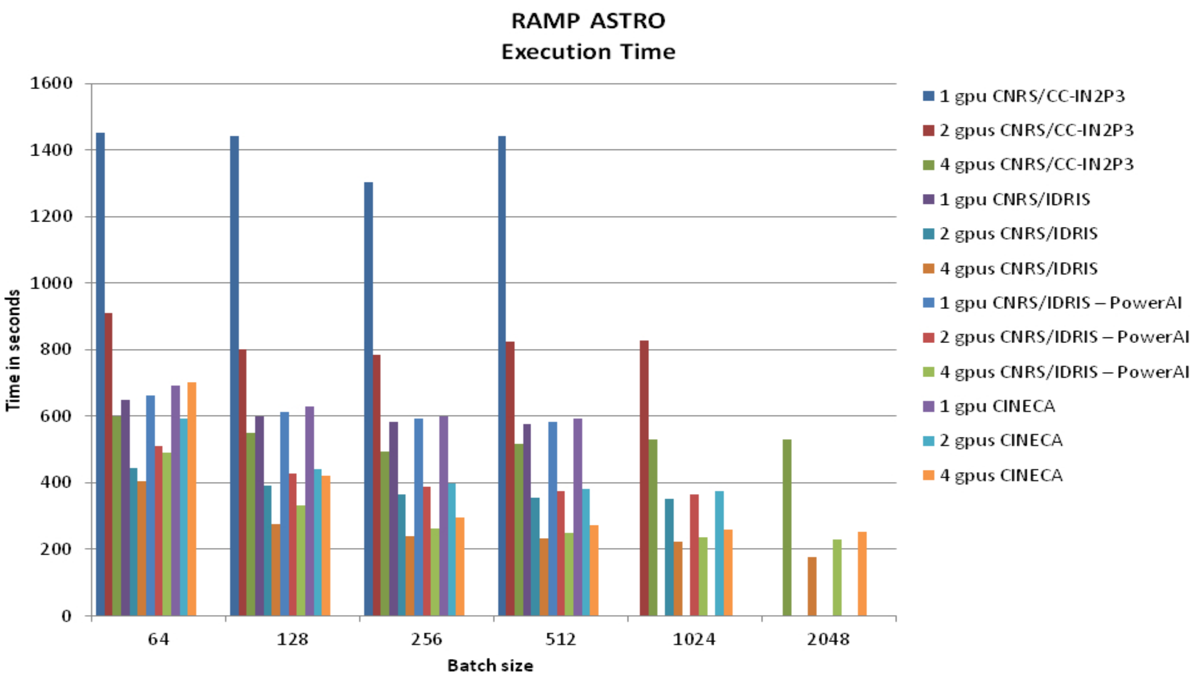


Figure 3: The Astrophysics use case

Figure 4 shows the performance of Tensorflow (X-axis: batch size and architectures, Y- axis: model bandwidth (images/second)) for the ImageNet benchmark.

The bandwidth increases as the number of GPUs and the batch sizes increases. The NVIDIA K80 runs Out Of Memory, the model becoming too large to manage for a batch size of 64 and 128. The use of container at IDRIS (PowerAI docker image) introduces an overhead in performance which tends to increase with the number of GPUs and larger batch sizes (around 5% overhead on the total execution time).

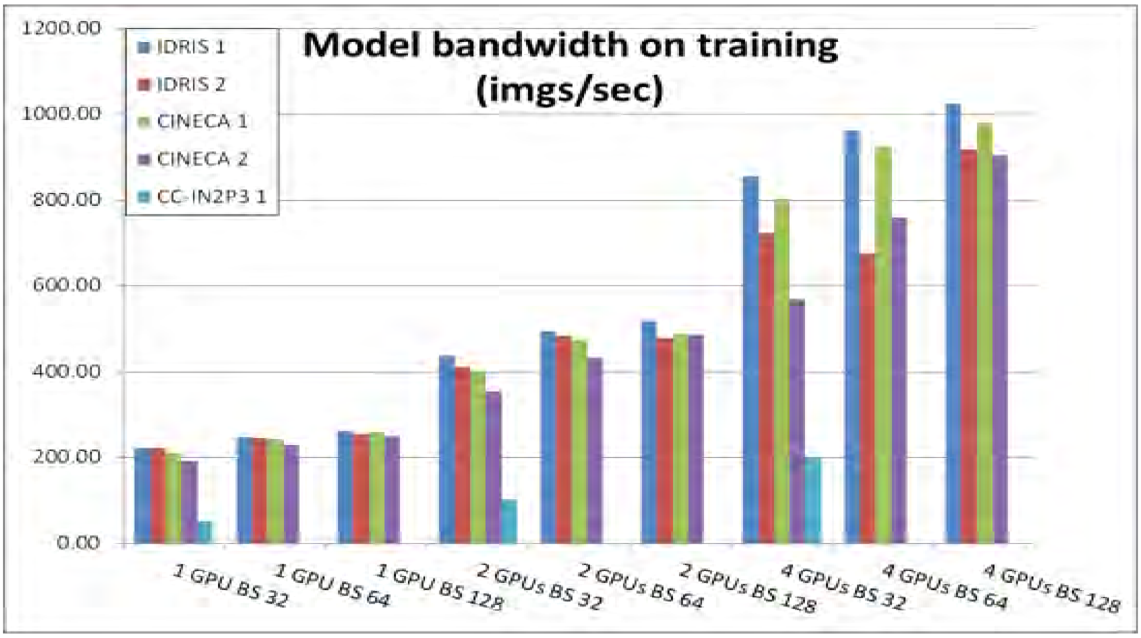


Figure 4: ImageNet - Intra-node model bandwidth

5 ADDITIONAL FEATURES

We have developed two additional features that are described in the following sections, to help users to perform their Data Analytics tasks.

5.1 The PRACE GitLab Data Analytics project

Given the large amount of work completed within the group, one of our concerns was to make this work available to users, so they can easily gain expertise, and, in turn, share their own experiences and results. For that, we defined the Data Analytics GitLab project within the PRACE GitLab to build some kind of a PRACE AI knowledge database. We started to build this data-base by gathering the material we used in the Data Analytics working group for the evaluation of the DL/ML tools over the different HPC architectures including standard benchmarks and small use cases. It is available at https://repository.prace- ri.eu/git/ Data-Analytics. It contains three projects: Benchmarks, Use-cases and Datasets. Each of them is described using metadata.

5.2 The dataset download service

During the development of our prototype services, we identified the need to offer users an easy access to datasets that are commonly used in the AI domain. Indeed, whereas small data files can be downloaded quickly from the internet, dataset downloads of large files can become tedious, lasting for several hours for a complete download. A possible solution is to share these datasets in a dedicated storage space in the PRACE infrastructure in order to benefit from the fast, reliable and secure PRACE VPN network and enhance the user productivity.

The dataset download prototype has been set up at CINECA with one TB of storage disk. It is implemented through iRODS (Integrated Rule-Oriented Data System) which offers a high transfer protocol and is currently used by the B2SAFE EUDAT service for the federation of data nodes.

6 CONCLUSION

We have described the PRACE Data Analytics service based on prototype services whose effectiveness has been studied by running benchmarks and use cases close. This study has shown that the performances depend mostly on the hardware. The GPU NVIDIA P100 outperforms the GPU NVIDIA K80, as expected. These prototype services include the most popular deep learning frameworks TensorFlow and Caffe with additional librairies: Keras, Horovod. TensorFlow remains the major player with a fast community growth, providing performance numbers far ahead from those of Caffe. Keras is now fully integrated to Tensorflow 2.0. TF2.0 provides also a new distributed training feature. Two additional services have been identified that can greatly enhance the user productivity: a dataset download service that can make standard datasets easily available to users and the Data Analytics GitLab project available within the PRACE GitLab which can help users to gain expertise and share their experience.

Acknowledgements
This work was financially supported by the PRACE project funded in part by the EU's Horizon 2020 Research and Innovation programme (2014-2020) under grant agreement 730913.

References

[1] Agnès Ansari - CNRS/IDRIS, Alberto Garcia Fernandez - CNRS/IDRIS, Bertrand Rigaux - CNRS/CC-IN2P3, Marco Rorro - CINECA, Andreas Vroutsis - EPCC. 2019. The PRACE Data Analytics service.

[2] Kaiming He, Xiangyu Zhang, Shaoqing RenJian. SunMicrosoft Research. Deep Residual Learning for Image Recognition.

[3] Krizhevsky, Sutskever and Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks - Proceedings of the 25th International Conference on Neural Information Processing Systems.

[4] Olaf Ronneberger, Philipp Fischer, Thomas Brox. Computer Science Department and BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany. U-Net: Convolutional Networks for Biomedical Image Segmentation.

[5] Sermanet, Eigen, Zhang, Mathieu, Fergus and LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. Computing Research Repository.

[6] Simonyan, Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computing Research Repository.

[7] Karen Simonyan, Andrew Zisserman. 2015. Very Deep Convolutional networks for large scale image recognition.ICLR 2015.

[8] Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke and Rabinovich. Going Deeper with Convolutions. Computing research repository.

OPPORTUNITIES OF CLOUD HIGH PERFORMANCE COMPUTING FOR SMES – A META-ANALYSIS

Blaž Gašperlin
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, Kranj, Slovenia
blaz.gasperlin@student.um.si

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

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

Abstract: High Performance Computing (HPC) is used to solve complex scientific, engineering and societal problems. Due to its high cost, it is mainly reserved for the large companies and institutes. However, with HPC services offered in the cloud it can become an interesting technology for small and medium sized enterprises (SMEs). The problem lies in the lack of awareness of benefits and availability of resources. We have conducted a meta-analysis of the results of CloudFlow, Fortissimo and SesameNet projects. Results suggest that SMEs can benefit using cloud HPC services, but there are also challenges that need to be addressed.

Keywords: high-performance computing, high-performance cloud computing, manufacturing SMEs

1 INTRODUCTION

In recent years new technologies and services have emerged, such as Blockchain, Artificial Intelligence (AI), Internet of Things (IoT) etc. and, one of them, cloud High Performance Computing (HPC). HPC services available through cloud is believed to be one of the key-enabling technologies for boosting competitive advantage for small and medium sized enterprises (SMEs). Until recently, HPC was mainly used for solving complex engineering, scientific and societal problems by large companies and institutes [1]. European Commission assesses that the manufacturing SMEs would benefit best from the cloud HPC services. They represent about 2 million SMEs (of the whole 24,5 SMEs) [2]. The trends in Europe show, that the investment in high performance computing has only grown in last years [3]. The investments in this field has been relatively steady, but in 2017, there was greater leap, which points to importance of using HPC. Many small and medium sized companies carries out simulations on ordinary workstations, when developing their product, which takes great amount of time and can lead to high costs [4]. Thus, it can be rented by SMEs from cloud HPC provider, which gives them powerful infrastructure only when they need it, which means lower costs, because there is no need for implementation of own infrastructure and company pays only for the service that it is used in that moment.

The problem is that most SMEs still lag behind with digital transformation [5], cloud HPC services and business models are not yet fully explored [6], and there is lack of efficient public support in this area. Reasons lie in the lack of technological knowledge and general awareness of benefits, that can bring faster product development, more cost efficient production, and greater competitiveness on the market [7]. In addition, the question arises, which sectors are mature enough to use these services and where we need to put more effort to embrace the use of HPC in the cloud. Several EU initiatives and projects addressed awareness raising, exploring the potential of HPC services and underlying business models in the cloud by conducting experiments (reader can find detailed information in [8, 9]). One of the problems was identification of SMEs with potential to use cloud HPC services. For this purpose the multi-criteria model was developed [9] and implemented within the SesameNet project [10]. Ziegler et al. [7] describe the results of an initiative to bring the whole ecosystem (companies, ISV, HPC infrastructure providers, experts, etc.) and develop sustainable business models that includes modelling services, software adaptation (parallelization), implementation and maintenance, i.e. "Simulation Workflow as a Service", that would be accessible through a »one-stop-shops« . These kinds of holistic services can become of interest to other industrial segments, also for the SMEs. In 2019 an analysis of the 60 organizations that were assessed using multi-criteria assessment tools showed that most of the organizations still have difficulties in assessing their computational needs, but most of them perceive that they are ready to take on the cloud HPC services [11].

Within all of these projects and initiatives there had been over 95 experiments conducted and more than 60 organizations assessed [8, 11]. The aim of this paper is to conduct a meta-analysis of the experiments and assessment results, which are publicly available on the websites of the projects [6, 12, 13]. On the basis of analysis we present the benefits of using cloud HPC for SMEs, but also the challenges and recommendations for future actions.

2 METHODOLOGY

We have conducted a meta-analysis of publicly available results of the cloud HPC experiments, conducted in the projects CloudFlow (20 experiments) [6], Fortissimo (79 experiments) [12] and SesameNet (36 use cases) [13]. Experiments have been selected in several waves of Open Calls of the projects CloudFlow and Fortissimo between 2013 and 2016 [6, 12], conducted and evaluated in cooperation with a specific HPC centre, whereas in SesameNet project, successful cases of cloud HPC implementation were selected to raise awareness and foster implementation of these services in EU. We classified each experiment into corresponding sector (industry). We analysed in total 6 different sectors: aeronautics, automotive industry, building industry, field of energy and environment, health and manufacturing industry (Table 1). From Table 1 we can see, that most experiments refers to automotive and manufacturing sectors. The smallest part of experiments refers to aeronautic sector, health and electro industry. The actual number of experiments in the case of Fortissimo project slightly differs from total number of experiments (equal to 62), due to repetition (overlapping) of some experiments on different sectors. We further narrowed the selection of projects to be analysed based on the reported numerical indicators: simulation time, time-to-market, costs/profits (in Table 1 showed as analysed experiments). We limited our analysis to the three indicators that are easily understood across all sectors, but there are other indicators that should be considered in the future (i.e. electricity consumption, natural resources consumption, increase of sold licences [6]).

Table 1: Number and date of experiments with corresponding projects (CloudFlow, Fortissimo, SesameNet)

Project	No. of experiments	Sector	No. of experiments per sector	No. of analysed experiments	Date of experiments
CloudFlow	20	Aeronautics			2014 - 2015
		Automotive	3	2	
		Building	1		
		E&E*	7	3	
		Health	1	1	
		Manufacturing	7	2	
		Electro	1		

Fortissimo	79*	Aeronautics	4	4	2013 - 2016
		Automotive	15	6	
		Building	7	3	
		E&E*	9	3	
		Health	5	1	
		Manufacturing	19	3	
		Electro	3		
SesameNet	36	Aeronautics			2015 - 2017
		Automotive			
		Building	1	1	
		E&E*	15	3	
		Health	6	2	
		Manufacturing	21	3	
		Creative and new Industries	4		

*Energy and environment
* Actual number of experiments is slightly different, due to repetition (overlapping) of some of the experiments
Building = Construction

3 RESULTS

In this section we present results of meta-analysis of more than 95 experiments conducted in the EU projects. Furthermore, we analysed three cases in depth in order to showcase the benefits of using cloud HPC services in SMEs.

Summarized benefits of using cloud HPC services by industry sector in reduction of planning/development costs (increased yearly revenue) and time/speed of product development (simulation time) is shown in Figure 1. On X axis, the industry sectors are presented, on Y axis values of two variables present: a) average savings for planning/development of product per year in 10 000 EUR, and b) average time savings of product/service development in percentages per development.

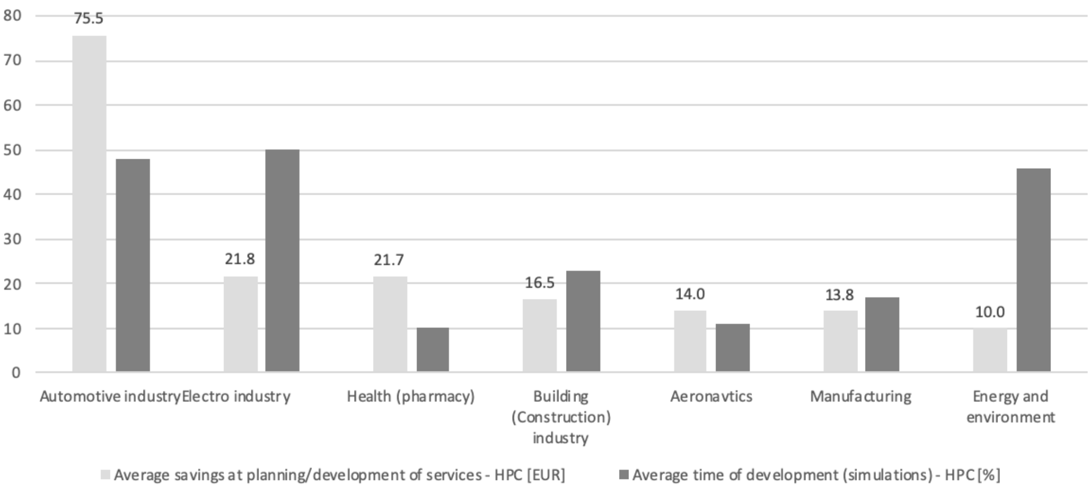


Figure 1: Comparison of average savings per year and average speeds of product/service development by industry sectors (adapted from [14])

Overall assessment of HPC contribution to reduce time of product development is presented in Figure 1. Detailed analysis for a case of aeronautics sector is shown in Figure 2. Savings in the development time of the product (measured in days) is showcased for two experiments (company Pipistrel [12] and MT-Propeller [12]). Pipistrel designs and manufactures light and ultra-light aircrafts. MT-Propeller produces propellers for smaller aircrafts. From Figure 2 we can observe, that the time for product development (measured in days, presented on X axis) has reduced, when using cloud HPC. Compared to the classical way of product development, the cloud HPC simulation provided a reduction by 27.5 days for the Pipistrel, while in the MT-Propeller the reduction of time was 6 days.

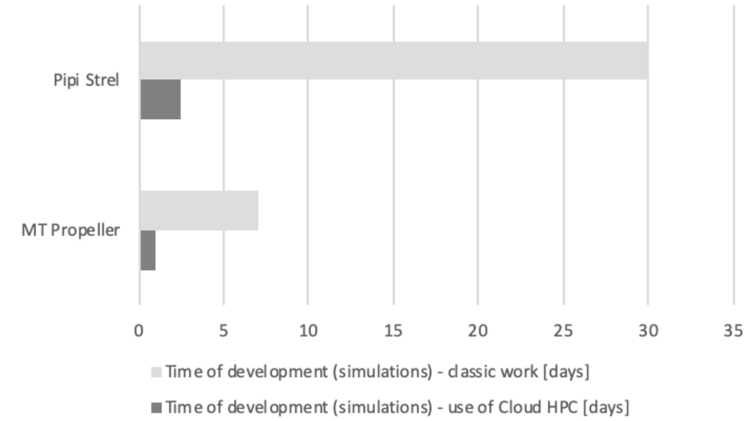


Figure 1: Time of product development (simulation) in days - classical work vs. cloud HPC (in days) (adapted from [14])

In the second case, we present the time-to-market for the experiments in the automotive sector, for companies Btech [6], which manufactures car lights and Elaphe Propulsion Technologies Ltd. [12], which develops in-wheel propulsion electric motors for cars (Figure 3). We have presented time-to-market before the cloud HPC use (set to 100%) and time-to-market after the use of cloud HPC services (Btech 25% and Elaphe Propulsion 20%, percentages in relation to the time needed before the cloud HPC product development). From Figure 3 it can be seen, that complete product can be put on the market by at least 75% to 80% faster, when using cloud HPC, compared to the classical way of manufacturing. Time-to-market comparison vs. "average speeds of product/service development by industry sectors from Fig. 1 show that indirect relation between both values exsist.

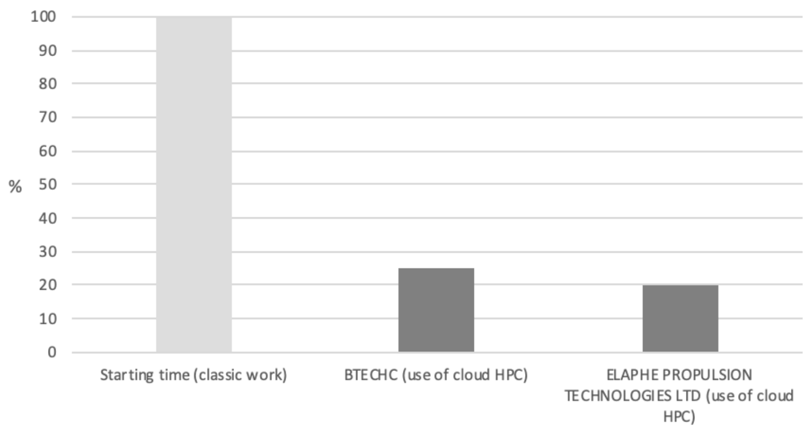


Figure 2: Time-to-market comparison - classical work vs. cloud HPC (adapted from [14])

The third case shows financial benefits, when using cloud HPC, where we present financial benefits of four different companies: Matrici [12], which manufactures and designs complex metal panels for the automotive and aerospace industry, Elaphe Propulsion Technologies Ltd., which area of work (described above) [12], Nolan group [12], which designs and develops motor helmets, and Borit [6], which manufactures bipolar fuel cell panels for electric vehicles.

From Figure 4 we can observe, that development costs are reduced considerably: Matrici company reported a yearly savings in planning and/or development in 1 500 000 EUR, Elaphe Propulsion Technologies Ltd. reported the savings of 135 000 EUR, Nolan group 52 000 EUR and Borit company savings of 10 000 EUR. In Figure 4 we present a detailed insight of yearly savings for planning and development for companies within a single industry sector (four companies in automotive sector), whereas in Figure 1 the presented average values refer to the average savings per year for all industry sectors.

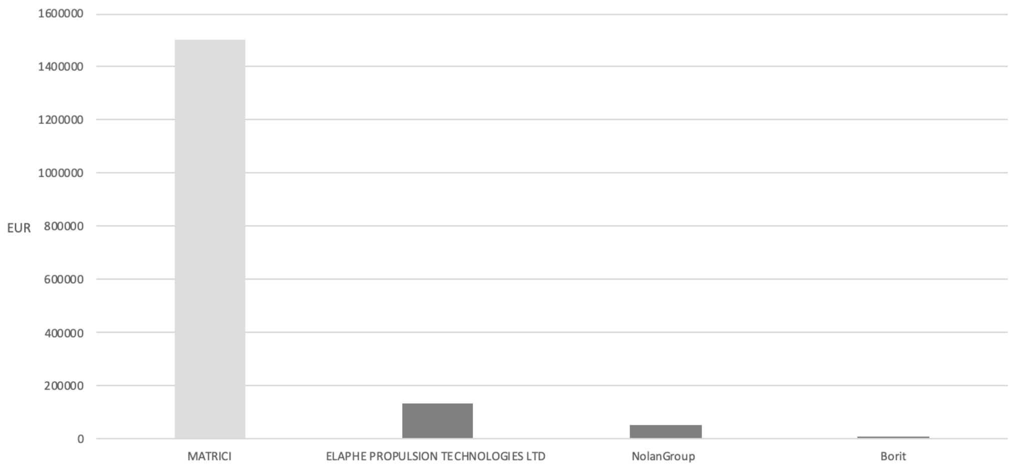


Figure 3: Yearly savings of money at planning and development (in EUR) – use of cloud HPC (adapted from [14])

4 CONCLUSION

Meta-analysis of more than 95 experiments conducted in the projects CloudFlow, Fortissimo, and SesameNet use cases between the years 2013 and 2017 have been addressed. Limitation of this study is in the variety of the reported numbers, which in the first place narrowed down units for the analysis, and secondly made comparison difficult. Results of the analysed experiments and used cases explicated that the use of cloud HPC services reduce development costs, product development time and time-to-market. The results of previous research confirm that awareness of cloud HPC is growing [11]. Furthermore, the use of HPC can be found in other, non-manufacturing sectors, i.e. creative industries and tourism and is on the rise with the data-driven economy [11, 15]. We could see, that it is evolving on disperse areas of use and therefore is becoming emerging technology, which can be used on daily basis for solving complex problems much faster and to speed up development times and lowering the costs.

In this paper, we showcased some examples of benefits that use of cloud HPC brings, which is important for awareness raising activities. However, for the cloud HPC services (and other key enabling technologies of the digital transformation) to be used broadly, we need to strengthen the support environment – not only the infrastructure – but also systematic and holistic support on the European, national, regional and local level. One such approach could be the voucher system of consultation services, through the Digital Innovation Hubs [16], updated formal and informal educational programmes, and the rest of business support environment stakeholders involvement.

Acknowledgement
Mirjana Kljajić Borštnar was supported by the Slovenian Research Agency, ARRS, through research programme P5-0018.

References

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

[2] European Comission. (2018). Annual Report on European SMEs 2017/2018. (K. Hope, Ed.). Luxembourg: European Commission. https://doi.org/10.2873/248745 [Accessed 24/06/2019]

[3] Joseph, E. C., Dekate, C., & Conway, S. (2014). EESI-2 Special Study To Measure And Model How Investments In HPC Can Create Financial ROI And Scientific Innovation In Europe. http://www.eesi-project.eu/wp-content/uploads/2015/05/EESI2_D7.4_Final-report-on-HPC-Return-on-Investment.pdf [Accessed 24/06/2019].

[4] Gigler, B.S., Casorati A., & Verbeek, A. (2018). Financing the future of supercomputing - How to increase investments in high performance computing in Europe. https://www.eib.org/attachments/pj/financing_the_future_of_supercomputing_en.pdf [Accessed 24/6/2019].

[5] European Comission. (2017). Digital Transformation Scoreboard 2017: Evidence of positive outcomes and current opportunities for EU businesses. https://ec.europa.eu/docsroom/documents/21501/attachments/1/translations/en/renditions/nativ [Accessed 24/06/2019]

[6] CloudFlow. (2015). CloudFlow project home page: https://eu-cloudflow.eu/ [Accessed 24/06/2019]

[7] Ziegler, W., D'ippolito, R., D'Auria, M., Berends, J., Nelissen, M. & Diaz, R. (2014). Implementing a "one-stop-shop" providing SMEs with integrated HPC simulation resources using Fortissimo resources. In *eChallenges e-2014 Conference Proceedings* (pp. 1–11).

[8] ICT Innovation for SMEs. http://i4ms.eu/i4ms/i4ms.php [Accessed 24/06/2019]

[9] Kljajić Borštnar, M., Ilijaš, T., & Pucihar, A. (2015). Assessment of cloud high performance computing potential for SMES. In *Proceedings of the 13th International Symposium on Operational Research SOR 2015* (pp. 23–28).

[10] SesameNet. (2017). SesameNet Assessment tool - HPC4SME Assessment Tool. https://network.sesamenet.eu/hpc4sme-assessment-tool/ [Accessed 26/06/2019].

[11] Kljajić Borštnar, M., Ilijaš, T. (2019). Preliminarna analiza pripravljenosti malih in srednje velikih podjetij na storitve zelo zmogljivega računalništva = Preliminary Analysis of Cloud High Performance Readiness of SMEs. In *38th International Conference on Organizational Science Development: Ecosystem of Organizations in the Digital Age* (pp. 419–430).

[12] Fortissimo. (2019). Fortissimo project home page. https://www.fortissimo-project.eu [Accessed 26/06/2019].

[13] SesameNet. (2017). SesameNet project home page. https://sesamenet.eu/. [Accessed 26/06/2019].

[14] Gašperlin, B. (2019). Analysis of high-performance computing opportunities for small and medium-sized enterprises. University of Maribor. Unpublished thesys.

[15] Starc Peceny, U., Urbančič, J., Mokorel, S., Kuralt, V., & Ilijaš, T. (2019). Tourism 4.0: Challenges in Marketing a Paradigm Shift. Consumer Behavior and Marketing [Working Title], (March), 0–19. https://doi.org/10.5772/intechopen.84762

[16] European Commission. (2019). Pan-European network of Digital Innovation Hubs (DIHs). https://ec.europa.eu/digital-single-market/en/digital-innovation-hubs [Accessed 26/06/2019].

ACCELERATED ALTERNATING DIRECTION AUGMENTED LAGRANGIAN METHOD FOR SEMIDEFINITE PROGRAMS

Timotej Hrga*, **Janez Povh***,†

* University of Ljubljana, Faculty of Mechanical Engineering, Slovenia
†Institute of Mathematics, Physics and Mechanics Ljubljana, Slovenia
email: timotej.hrga@lecad.fs.uni-lj.si, janez.povh@lecad.fs.uni-lj.si

Abstract
We propose an alternating direction augmented Lagrangian method to solve semidefinite programs with additional inequality constraints. The benefit of the new approach is less computationally expensive update rule for the dual variable with respect to the inequality constraints. We provide some computational experience with this method to efficiently solve semidefinite relaxation coming from combinatorial optimization problem.

Keywords: semidefinite programming, alternating direction augmented Lagrangian method, max-cut problem

1 INTRODUCTION

For given symmetric matrices $C,\ A_i,\ i = 1, 2, \ldots, m$ and $B_j,\ j = 1, 2, \ldots, k$ of order n and vectors $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^k$ we consider the following semidefinite optimization problem (SDP)

$$\max\ \langle C, X\rangle \quad \text{such that} \quad A(X) = a,\ B(X) \leq b,\ X \succeq 0. \tag{P}$$

The dual to this problem is given by

$$\min \quad a^T y + b^T t \quad \text{such that} \quad C - A^T(y) - B^T(t) + Z = 0,\ y \in \mathbb{R}^m,\ t \in \mathbb{R}^k_+,\ Z \succeq 0. \tag{D}$$

The inner product on the space of symmetric matrices is given by $\langle X, Y\rangle = \mathrm{tr}(XY) = \sum_i \sum_j X_{ij} Y_{ij}$ and linear operator A maps symmetric matrices to $\mathbb{R}^m$ with $A(X)_i = \langle A_i, X\rangle$. Its adjoint is well known to be $A^T(y) = \sum_i y_i A_i$. Similarly for the operator B. We use the notation $X \succeq 0$ to denote that symmetric matrix X belongs to the cone of positive semidefinite matrices, and the comparison of vectors is meant componentwise.

We make the following assumption:

Assumption 1 *The matrices A and B have full ranks and the Slater condition holds, i. e. both problems* (P) *and* (D) *have strictly feasible points.*

It is well known that strong duality holds under the assumption of Slater condition, see [13].

There exist several numerical packages to solve semidefinite programs, see for example the website of Hans Mittelman [5]. Many of them are based on some variant of the primal-dual path following interior-point method, see for example [3]. These methods are only aplicable to SDP instances for which n, m and k are reasonably small, i. e. $n \leq 1000$ and $m, k \leq 10000$.

A number of algorithms have been proposed that can handle large number of constraints, for instance the boundary point method [7], the bundle method [2] and SDPAD (SDP Alternating Direction) [11]. This paper further adresses this topic by presenting accelerated augmented Lagrangian method applied to problem (D).

2 ACCELERATED AUGMENTED LAGRANGIAN SOLVER

In this section we derive the method ASDPAD (Accelerated SDP Alternating Direction) by using the augmented Lagrangian approach on the dual problem (D). Let I denote the indicator function $I\colon \mathbb{R}^n \to \mathbb{R}$ defined as $I(s) = 0$ if $s \geq 0$ and ∞ otherwise. Following the idea in [1], we first transform the problem (D) to the equivalent one with objective function $a^T y + b^T t + I(s)$ subject to additional equality constraint $t - s = 0$.

Let X and u denote the Lagrange multipliers for the equality constraints $C - A^T(y) - B^T(t) + Z = 0$ and $t - s = 0$ respectively. For fixed $\rho > 0$ consider the augmented Lagrangian L_ρ:

$$\begin{aligned} L_\rho(y, Z, t, s; X, u) &= a^T y + b^T t + I(s) + \langle Z + C - A^T(y) - B^T(t), X\rangle \, + \\ &\quad \frac{\rho}{2}\|Z + C - A^T(y) - B^T(t)\|_F^2 + \langle t - s, u\rangle + \frac{\rho}{2}\|t - s\|_2^2 \\ &= a^T y + b^T t + I(s) + \frac{\rho}{2}\|Z + C - A^T(y) - B^T(t) + X/\rho\|_F^2 \, + \\ &\quad \frac{\rho}{2}\|t - s + u/\rho\|_2^2 - \frac{1}{2\rho}\|X\|_F^2 - \frac{1}{2\rho}\|u\|_2^2, \end{aligned}$$

where $\|\cdot\|_F = \sqrt{\langle \cdot, \cdot \rangle}$ is the Frobenius norm. Alternating direction augmented Lagrangian method consists of alternating minimization of L_ρ with respect to one primal variable while keeping others fixed to get y, Z, t and s. Then the dual variables X and u are updated using the following rules, see [1]:

$$y^{k+1} = \underset{y \in \mathbb{R}^m}{\arg\min}\ L_\rho(y, Z^k, t^k, s^k; X^k, u^k), \tag{1}$$

$$Z^{k+1} = \underset{Z \succeq 0}{\arg\min}\ L_\rho(y^{k+1}, Z, t^k, s^k; X^k, u^k), \tag{2}$$

$$t^{k+1} = \underset{t \in \mathbb{R}^k}{\arg\min}\ L_\rho(y^{k+1}, Z^{k+1}, t, s^k; X^k, u^k), \tag{3}$$

$$s^{k+1} = \underset{s \in \mathbb{R}^k_+}{\arg\min}\ L_\rho(y^{k+1}, Z^{k+1}, t^{k+1}, s; X^k, u^k), \tag{4}$$

$$X^{k+1} = X^k + \rho\left(Z^{k+1} + C - A^T(y^{k+1}) - B^T(s^{k+1})\right), \tag{5}$$

$$u^{k+1} = u^k + \rho\left(t^{k+1} - s^{k+1}\right). \tag{6}$$

Let us closely look at the subproblems (1) - (4). The first order optimality conditions for problem (1) are

$$\nabla_y L_\rho = a - \rho A(Z + C - A^T(y) - B^T(t) + X/\rho) = 0.$$

By the assumption 1 the matrix AA^T is invertible and hence y is the solution of the linear system

$$AA^T y = A(Z + C - B^T t + X/\rho) - a/\rho, \tag{7}$$

where $(AA^T)(i,j) = \langle A_i, A_j \rangle$. Similarly for (3), we obtain the update rule for t:

$$(BB^T + I)t = B(Z + C - A^T y + X/\rho) + s - u/\rho - b/\rho. \tag{8}$$

Note that the matrices AA^T and $BB^T + I$ are positive definite and hence the above linear systems can be efficiently solved using Cholesky factorization.

By defining $W = A^T(t) + B^T(t) - C - X/\rho$, the subproblem (2) can be formulated as

$$\underset{Z \succeq 0}{\arg\min}\, \|W - Z\|_F^2,$$

whose solution is the projection W_+ of the matrix W onto the cone of positive semidefinite matrices. It is well known that W_+ can be computed from the spectral decomposition of $W = S\Lambda S^T$. Let $S = (S_+, S_-)$ be a partition of the eigenvectors in S according to nonnegative and negative eigenvalues $\Lambda = (\Lambda_+, \Lambda_-)$. Then

$$W = S_+\Lambda_+S_+^T + S_-\Lambda_-S_-^T = W_+ + W_-.$$

As already observed in the boundary point method [7], the update for X can be simplified and also computed from the spectral decomposition of matrix W using the formula

$$X = -\rho W_-. \tag{9}$$

The problem (4) asks for the nonnegative vector that is closest to the $t + u/\rho$. Hence the solution of the problem is

$$s = \max(0, t + u/\rho),$$

the nonnegative part of vector $t + u/\rho$.

The overall complexity of one step of the method is solving two linear systems with matrices AA^T and $BB^T + I$ and computing the eigenvalue decompostion. Compared to interior-point methods where coefficient matrix changes in each iteration, matrices AA^T and $BB^T + I$ remain constant throughout the algorithm and their factorizations can be cached at the beginning to speed up solving the systems.

The proposed method is extension of the boundary point method [7] and SDPAD [11]. The drawback of the first method is inability to handle additional inequality constraints. In the second approach the additional dual variable s for problem (D) is not introduced and the subproblem (3) is replaced with minimizing L_ρ with respect to $t \geq 0$. This leads to solving convex quadratic program of order k over the nonnegative orthant.

Including the nonnegativity constraints $X \geq 0$ on the elements of the matrix X in problem (P) can also be easily handled within this framework. Similarly to the problem (4) the Lagrange multiplier for the constraint $X \geq 0$ is computed by projecting the appropriate matrix onto the cone od nonnegative matrices. Hence there is almost no additional cost. For application, this leads to efficient algorithm for computing theta plus function $\vartheta_+(G)$ [10], associated with a graph G:

$$\max \quad \langle J, X\rangle \quad \text{such that} \;\; \operatorname{tr}(X) = 1,\; X_{ij} = 0 \;\; \forall \{i,j\} \in E(G),\; X \geq 0,\; X \succeq 0.$$

By J we denoted the matrix of all ones.

The performance of the proposed method can be significantly improved by exploiting structural properties of the matrices AA^T and BB^T, such as sparsity. Furthermore, for many SDP relaxations of combinatorial optimization problems, such as max-cut problem and maximum stable set problem, the coefficient matrix AA^T is diagonal, because of orthogonal equality constraints. Hence solving (7) is trivial.

The performance of the method is dependent on the choice of the penalty parameter ρ. Numerical tests show that for the problem we considered the starting value of $\rho = 1$ is a good choice and the value is dynamically tuned during the algorithm. The update rules (2) and (9) ensure that during the algorithm positive semidefinitness of matrices X and Z is maintained, as well as the optimality condition $ZX = 0$. Hence once the primal and dual feasibility conditions are satisfied, the method found the optimal solution. To measure the accuracy of primal and dual feasibility we use

$$r_P = \frac{\|A(X) - a\|_2 + \|\min(b - B(X), 0)\|_2}{1 + \|a\|_2} \quad \text{and} \quad r_D = \frac{\|C + Z - A^T(y) - B^T(t)\|_F}{1 + \|C\|_F}.$$

Simple strategy to adjust the value of ρ is observing the residuals:

$$\rho^{k+1} = \begin{cases} \rho^k/\tau & \text{if } r_P > \mu\, r_D \\ \tau\rho^k & \text{if } r_D > \mu\, r_P \\ \rho^k & \text{otherwise} \end{cases}$$

for some parameters μ and τ. Typical values are $\mu = 10$ and $\tau = 1.01$.

To further increase the performance of the method we use overrelaxation, see [1]. After updating the variable y in the step (1) we need to form the matrix $A^T(y)$ in order to compute spectral decomposition of W. The quantity $A^T(y)$ is replaced with

$$\alpha A^T(y) + (1-\alpha)(C + Z - B^T(t)),$$

where $\alpha \in (0, 2)$ is a relaxation parameter. Preliminary numerical results show that $\alpha = 1.8$ is a good candidate and reduces the overall number of iterations and leads to faster convergence.

3 APPLICATION: MAX-CUT WITH TRIANGLE INEQUALITIES

Max-Cut is a classical and well-known NP-hard combinatorial optimization problem that has attracted the scientific interest during the past decades, see for instance [8, 4]. For a given undirected and weighted graph $G = (V, E)$, the max-cut problem asks to find partition of the vertices into two classes such that sum of the weights of the edges connecting vertices from different classes is maximal. Encoding the partitions by vectors $x = \{-1, 1\}^n$, we obtain the following binary quadratic optimization problem

$$\max \ \frac{1}{4} x^T L x \quad \text{such that} \ \ x \in \{-1, 1\}^n, \tag{10}$$

where L is the Laplacian matrix of the graph defined as $L = \operatorname{diag}(Ae) - A$. To obtain the basic semidefinite relaxation of max-cut, observe that for any $x = \{-1, 1\}^n$, the matrix $X = xx^T$ is positive semidefinite and its diagonal is equal to vector of all ones. This leads to

$$\max \ \frac{1}{4}\langle L, X\rangle \quad \text{such that} \ \ \operatorname{diag}(X) = e, \ X \succeq 0, \tag{11}$$

which can be found in [6]. It was observed that the bound from (11) is not strong enough to be successfully used within branch and bound framework in order to solve the max-cut problem to optimality. By adding additional equality or inequality constraints known as cutting planes we strengthen the upper bound. One such class of cutting planes, called triangle inequalities, is obtained as follows. Observe that for an arbitrary triangle with vertices $i < j < k$ in the graph G, any partition of vertices cuts either 0 or 2 of its edges. Moving to our model this leads to

$$\begin{aligned} X_{ij} + X_{ik} + X_{jk} &\geq -1, \\ X_{ij} - X_{ik} - X_{jk} &\geq -1, \\ -X_{ij} + X_{ik} - X_{jk} &\geq -1, \\ -X_{ij} - X_{ik} + X_{jk} &\geq -1. \end{aligned}$$

We collect these inequalities as $B(X) \leq e$ and get the strengthened SDP relaxation

$$\max \ \frac{1}{4}\langle L, X\rangle \quad \text{such that} \ \ \operatorname{diag}(X) = e, \ B(X) \leq e, \ X \succeq 0. \tag{12}$$

Note that the matrix BB^T is sparse which can be exploited to solve (8) more efficiently.

Tab. 1 shows the comparison of computational times in seconds for an accuracy requirement of $\varepsilon = 10^{-5}$ of SDPAD and our method ASDPAD and the benefit of using the new approach. We have implemented both methods in Matlab and considered different instances of graphs with $n = 100$ nodes from Biq Mac Library [12]:

g05: unweighted graphs with edge density 0.5,

pm1d, pm1s: weighted dense and sparse graphs with edge weights chosen from the set $\{-1, 1\}$,

w05, w09: weighted dense graphs with edge weights chosen from the interval $[-10, 10]$,

as well as random unweighted graphs with edge density 0.5 and with different number of nodes produced by the graph generator Rudy [9].

Since there are $4\binom{n}{3}$ triangle inequalities, we can not include all of them, but solve the problem (12) with a limited number of them. First we compute the optimal solution of the basic semidefinite relaxation (11). Then we run through all triangle inequalities and include only the most $n \cdot 20$ violated ones. The computational times of both methods are reported in first and second column of Tab. 1.

In order to obtain tighter upper bound for the problem (10) we use the following cutting-plane procedure. After we obtain the maximizer of problem (12) with limited number of constraints, we first purge all inactive constraints. Then new violated triangle inequalites are added, the problem with new set of constraints is solved and the process is iterated until the error $\max(B(X) - e)$ is less than 0.01. The accuracy in each problem is set to $\varepsilon = 10^{-4}$. To purge the constraints that are not binding at the optimal solution we look at the values of the corresponding dual multipliers in vector s. If the value of some dual multiplier is close to zero, this indicates that the corresponding constraint is not active at the optimum and we can remove it.

In Tab. 1, we report in column 3 and 4 the timings of the above cutting-plane algorithm. Numerical results show that our method outperforms SDPAD in the case when multiple problems (12) need to be solved.

Table 1: Comparison with SDPAD method [11] on some instances of the Biq Mac Library. Computation times are in seconds.

			cutting-plane algorithm	
graph	[11]	our method	[11]	our method
g05_100.1	4.66	0.91	8.18	1.39
g05_100.2	4.57	0.83	10.45	1.57
pm1d_100.1	2.51	0.59	8.34	1.79
pm1d_100.2	4.49	0.73	8.49	1.53
pm1s_100.1	6.08	1.92	26.37	4.74
pm1s_100.2	7.87	2.12	30.15	5.09
w05_100.1	3.14	1.24	12.62	4.08
w05_100.2	2.57	1.21	16.47	4.45
w09_100.1	2.01	1.38	10.20	4.52
w09_100.2	2.62	1.42	8.09	4.24
g05_300	12.24	6.40	24.01	8.75
g05_500	30.18	20.89	45.61	30.83
g05_800	181.34	136.80	147.35	119.14

4 CONCLUSIONS

We have presented new method for solving semidefinite programs with inequality constraints. The approach is based on augmented Lagrangian method where additional variable in the dual SDP is introduced in order to eliminate solving convex quadratic program. This leads to alternating direction method where new iterates are generated as solutions of linear systems and projections onto nonnegative orthant and positive semidefinite cone. In many SDP relaxations of combinatorial optimization problems the orthogonal constraints give rise to diagonal coefficient matrix of the linear systems. In such cases the only computationally expensive step is computing the full eigenvalue decomposition of matrix variable of size $n \times n$. Thus the method is appropriate for solving large-scale problems that are out of reach for interior-point algorithms.

Since the proposed algorithm is a first order method its limitations are slow convergence to high accuracy. Furthermore, to obtain efficient variant of the method initial tuning of the penalty parameter ρ for specific classes of problems is needed. Future work will examine the incorporation of the proposed method inside branch and bound algorithm to solve max-cut problem to optimality.

References

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[2] Ilse Fischer, Gerald Gruber, Franz Rendl, and Renata Sotirov. Computational experience with a bundle approach for semidefinite cutting plane relaxations of max-cut and equipartition. *Mathematical Programming*, 105(2-3):451–469, 2006.

[3] Christoph Helmberg, Franz Rendl, Robert J Vanderbei, and Henry Wolkowicz. An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2):342–361, 1996.

[4] Monique Laurent and Franz Rendl. Semidefinite programming and integer programming. *Handbooks in Operations Research and Management Science*, 12:393–514, 2005.

[5] Hans D Mittelmann. An independent benchmarking of sdp and socp solvers. *Mathematical Programming*, 95(2):407–430, 2003.

[6] Svatopluk Poljak and Franz Rendl. Nonpolyhedral relaxations of graph-bisection problems. *SIAM Journal on Optimization*, 5(3):467–487, 1995.

[7] Janez Povh, Franz Rendl, and Angelika Wiegele. A boundary point method to solve semidefinite programs. *Computing*, 78(3):277–286, 2006.

[8] Franz Rendl, Giovanni Rinaldi, and Angelika Wiegele. Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. *Mathematical Programming*, 121(2):307, 2010.

[9] Giovanni Rinaldi. `http://www-user.tu-chemnitz.de/~helmberg/rudy.tar.gz`, 1998.

[10] Alexander Schrijver. A comparison of the delsarte and lovász bounds. *IEEE Transactions on Information Theory*, 25(4):425–429, 1979.

[11] Zaiwen Wen, Donald Goldfarb, and Wotao Yin. Alternating direction augmented lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3-4):203–230, 2010.

[12] Angelika Wiegele. Biq mac library – a collection of max-cut and quadratic 0-1 programming instances of medium size. *Preprint*, 2007.

[13] Henry Wolkowicz, Romesh Saigal, and Lieven Vandenberghe. *Handbook of semidefinite programming: theory, algorithms, and applications*, volume 27. Springer Science & Business Media, 2012.

Parallelization of BiqMac Solver

Alen Vegi Kalamar

Faculty of natural science and mathematics, University of Maribor, Koroška cesta 160, Maribor, Slovenia,

alen.vegi.kalamar1994@gmail.com

Drago Bokal

Faculty of natural science and mathematics, University of Maribor, Koroška cesta 160, Maribor, Slovenia

Janez Povh

Faculty of mechanical engineering, University of Ljubljana, Aškerčeva cesta 6, Ljubljana, Slovenia

Abstract

The max-cut problem is known to be a NP-hard problem. However, there are still possibilities and needs to solve the problem to optimality for small instances. The max-cut problem has a property which allows it to be divided into smaller subproblems (problems one dimension smaller) and if we solve these subproblems to optimality, we can get the optimal solution for the original problem. This property enabled the development of branch and bound algorithms to solve the max-cut problem.

Rendl, Rinaldi and Wiegele used semidefinite programming relaxations of the problem, and used efficient heuristics to bound the generated subproblems and efficient heuristics to generate good feasible solutions and developed an efficient branch and bound algorithm (BiqMac Solver), which can solve the max-cut problem to optimality (even for graphs with sizes up to 250). Because of their nature, branch and bound algorithms are suitable for parallelization. We implement a paralel version of the BiqMac Solver in C language using MPI library. The resulting solver was tested on the supercomputer of University of Ljubljana, Faculty of mechanical engineering. In the paper we, present details about this parallelization and numerical results obtained with the parallel BiqMac Solver.

Keywords: max-cut problem, branch and bound algorithm, parallel computation

1 Introduction

The max-cut problem is one of the many NP-hard problems in the graph theory. Beacuse of the NP hardness of the problem, researchers developed different heuristics or combinations of heuristics and exact methods to solve the problem. In [6], they present the property of the max-cut allowing the problem to be optimized by dividing the feasible solution into two disjoint sets and the subproblems over the created subsets are again max-cut problems over graphs of one dimension smaller. Using this property and the method of computing the upper bound of the subproblems in connection with eigenvalues, they present a branch and bound algorithm to solve the max-cut problem. At the same time, in [2], an effective approximation algorithm for solving the max-cut problem was presented. The algorithm provided a solution, which expected value is at least 0.87856 of the optimal one. Later, in [3], an interior point method for semidefinite relaxation of the problem was presented, where additional triangular inequalities are used. This method was extended in [4], where triangular inequalities are replaced by more general hypermetric inequalities. Using the aforementioned knowledge and improving the approaches in computing the upper bound of the problem, in [7] they presented an efficient branch and bound algorithm named BiqMac Solver.

In this paper, we focus on improving the BiqMac Solver using parallelization. We organize the paper as follows. In Section 2, we present the max-cut problem, its equivalent formulation and its semidefinite relaxation. In Section 3, BiqMac Solver algorithm is presented. In Section 4, we present the details of the parallelization of the BiqMac Solver algorithm. The Section 5 presents the test results.

2 Max-cut problem

2.1 Definition of max-cut problem

The max-cut problem is formulated as follows.

Definition 1. *Let $G = (V, E)$ be an undirected graph, where V is the set of vertices ($|V| = n$) and E is the set of edges. For each edge $ij \in E$, let w_{ij} be the edge weight. Each bipartition (V_1, V_2) of the set of vertices defines* *a cut* $(V_1 : V_2) = \{ij \in E \mid i \in V_1, j \in V_2\}$. *The problem of finding such a bipartition* (V_1, V_2)*, which maximizes*

$$w(V_1, V_2) := \sum_{ij \in (V_1:V_2)} w_{ij},$$

is called the max-cut problem.

Using the weighted adjacency matrix $A \in \mathbb{R}^{n \times n}$, $(A)_{i,j} = (A)_{j,i} = \begin{cases} w_{ij}, & ij \in E \\ 0, & ij \notin E \end{cases}$, Laplacian matrix L assocciated to A, defined as $L = Diag(Ae) - A$, where e is a vector of all ones of length n, and representation of bipartition (V_1, V_2) as a vector $x \in \{-1, 1\}^n$, where $x_i = \begin{cases} 1, & i \in V_1 \\ -1, & i \in V_2 \end{cases}$, it can be shown, that $w(V_1, V_2) = \frac{1}{4} x^T Lx$ ([5]). It can also be shown that $x^T Lx = tr(Lxx^T)$ ([5]). So, solving the max-cut problem is equivalent to solving the quadratic problem

$$z_{MC} = \max\{tr(LX) \mid x \in \{-1, 1\}^n,\ X = xx^T\}. \tag{1}$$

2.2 Semidefinite relaxation of the max-cut problem

Using the observations above and the result that the convex hull of the feasible solutions of the problem (1) is contained in the set $\{X \mid diag(X) = e,\ X \succeq 0\}$ ([5]), we can introduce the following semidefinte relaxation of the problem (1):

$$z_{SDP} = \max\{tr(LX) \mid diag(X) = e,\ X \succeq 0\}. \tag{2}$$

The relaxation can be additionaly tightened by including the triangle inequalities

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_{ij} \\ x_{ik} \\ x_{jk} \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad (1 \leq i < j < k \leq n).$$

We abberive all $4\binom{n}{3}$ of them as $\mathcal{A}(X) \leq e$. We get the following relaxation of the problem (1):

$$z_{SDPMET} = \max\{tr(LX) \mid diag(X) = e,\ X \succeq 0,\ \mathcal{A}(X) \leq e\}. \tag{3}$$

3 BiqMac Solver

BiqMac Solver [7] is an implementation of branch and bound algorithm to solve the max-cut problem. It consists of the following main steps:

- **calculating the upper bound of the (sub)problem:** Authors [7] introduce a Lagrangian dual of the relaxation problem (3) and solve it with the dynamic bundle method.
- **finding the feasible soultion of the (sub)problem:** Authors [7] use an iterative heuristic method, based on Goemans-Williamson's hyperplane rounding [2].
- **branching strategy:** It was shown (see [6, 7]), that the problem can be divided into two smaller instances of the max-cut problem. This is done by setting two vertices of the graph to either the same or the opposite partition, and constructing new graphs of one dimension smaller. For selecting vertices of graphs for branching, they use different strategies (see [3]). They use the matrix that is computed with the dynamic bundle method as a base for selecting vertices. The main idea is to find the indices of rows of the matrix to be either close or far from the $\{-1, 1\}$ vector, or indices (row and column) of the element of the matrix that is close or far from $\{-1, 1\}$.
- **strategy of selection of a (sub)problem for processing:** They use the best first strategy - the unprocessed (sub)problem is selected, of which the upper bound is closest to the best found solution.

4 Parallelization of BiqMac Solver

4.1 Master–slave communication

Master–slave communication is a model of communication protocol where one process (master) has undirectional control over one or more processes (slaves). We have two sets of processes:

- **master:** Its job is to control the state of employment of an individual slave, to send data to slaves, to obtain the processed data from slaves and then process the received data.
- **slaves:** Their job is to process the received data and return the acquired data to the master.

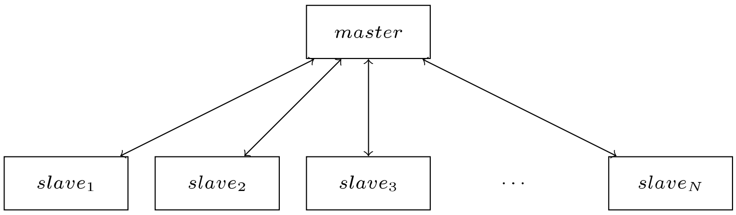


Figure 1: Master–slave communication.

4.2 Parallel implementation

The BiqMac Solver algorithm (and branch and bound algorithms in general) allows us that if we have two nodes of the rooted branch and bound tree where the first one is not an ancestor of the second one and vice versa, to independently solve each subproblem saved in the nodes. Beacuse of that, the BiqMac Solver algorithm (and branch and bound algorithms in general) is suitable for parallel implementation. For parallel implementation of the BiqMac Solver, we use the master–slave communication described above. Master and slaves have the following jobs:

- **master:** The master is the one responsible for assigning tasks. At the beginning, it creates the list of all slaves, which is then used to determine the occupancy of the slaves, and the queue, where it saves the unprocessed nodes that contain the data about the unprocessed subproblems. When the master has the data in the queue, it locates free slaves and sends them the data of subproblems (each slave gets another unprocessed subproblem), best known cuts and their value. It marks selected slaves as busy.

  When a subproblem is processed by a slave, it receives the processed data and a request for further work with it. The following requests are possible:

  - **GET_CHILDREN:** Request that the processed subproblem should be branched. In this case the slave sends the processed subproblem and the new acquired data (upper bound for a subproblem and nodes of a graph that should be branched). The master branches the subproblem and creates two new subproblems (of one dimension smaller), that are added to the queue of unprocessed subproblems.
  - **FATHOM:** Request that the processed subproblem should not be branched.
  - **NEW_VALUE:** Request that the found cut by a slave should be checked, if it is the best known cut. In this case, the slave sends a newly found cut. If its value is greater than the current best (saved in the master), then this becomes the new best cut. In response, the master sends the current best cuts to the slave.

  When the master receives a request for GET_CHILDREN or FATHOM, after the execution of the request, it continues to send the subproblems saved in the queue (it sets the request to SEND_TO_WORKERS). At this time it sets the slave, from which he got the request, as free.

- **slave:** The slave receives the data of a subproblem, the best known cuts and their value from the master. It processes the subproblem as it would be done by a non-paralel algorithm:

  - executes the bundle method to compute the upper bound of the subproblem,
  - executes the heuristic to compute a feasible solution of the subproblem,
  - selects the nodes of a graph of the subproblem for branching.

  If the value of the newly computed cut is greater than the best known by a slave, an intermediate request NEW_VALUE is sent to the master. This requst tells it to check if the new cut is the current best. The master has to verify that, because some other slave could have found a better solution in the meantime and the slave would not know about it yet. In response to this request the master sends the current best cuts. The slave then sends one of the following requests:

  - **GET_CHILDREN:** It is sent in case that the upper bound of the subproblem is greater or equal than the value of the current best cuts. The slave sends the data of the subproblem, the upper bound of the subproblem and the vertices of the graph of the subproblem to be branched.
  - **FATHOM:** It is sent in case that the upper bound of the subproblem is smaller than the value of the current best cuts.

Algorithm Pseudocode of parallel algorithm BiqMac Solver (part 1)

Input: weighted graph G
Output: best cuts $maxCut$ and their value $maxValue$

```
if is master then
    slaveList ← [0, ..., 0]
    freeSlaves ← slaveList_size
    queue ← ∅
    root ← create the root node (the original problem)
    queue ← queue ∪ {root}
    while we do not get the information about ending do
        if it is a first iteration of the loop then
            request ← SEND_TO_WORKERS
            next iteration is not the first anymore
        else
            request ← get request from a slave k
        end if
        if request = GET_CHILDREN then
            (data, i, j) ← get processed data from slave k
            left ← create left child of the subproblem data with branching i, j into the same partition
            right ← create left child of the subproblem data with branching i, j into the oposite partition
            queue ← queue ∪ {left, right}
            slaveList[k] ← 0
            freeSlaves ← freeSlaves + 1
            request ← SEND_TO_WORKERS
        else if request = FATHOM then
            slaveList[k] ← 0
            freeSlaves ← freeSlaves + 1
            request ← SEND_TO_WORKERS
        else if request = NEW_VALUE then
            cut ← get potential best cut from slave k
            cutValue ← get value of the cut cut
            if cutValue > maxValue then
                maxValue ← cutValue
                maxCut ← {cut}
            else if cutValue = maxValue then
                maxCut ← maxCut ∪ {cut}
            end if
            send maxCut to slave k
        end if
        if request = SEND_TO_WORKERS then
            if queue_size = 0 and freeSlaves = slaveList_size then
                request ← DONE
            else
                iterations ← min{queue_size, freeSlaves}
                for iteration = 1, 2, ..., iterations do
                    data ← get subproblem from queue
                    k ← get first free worker in slaveList
                    slaveList[k] ← 1
                    freeSlaves ← freeSlaves − 1
                    send (data, maxCut, maxValue) to the slave k
                    queue ← queue − {data}
                end for
            end if
        end if
```

Algorithm Pseudocode of parallel algorithm BiqMac Solver (part 2)

```
            if request = DONE then
                for all slaves k in slaveList do
                    send message to slave k that we are ending
                end for
                exit the loop
            end if
        end while
    else
        while we do not get a message from the master that we are ending do
            (data, maxCut, maxValue) ← get the data of the subproblem, current best cuts ant their value
            (data, cut, i, j) ← execute the calculating part over the subproblem data
            cutValue ← get the value of the cut cut
            if data_ub > maxValue and cutValue > maxValue then
                request ← NEW_VALUE
                send cut with request request to the master
                maxCut ← get response from the master about current best cuts
                maxValue ← get the value of current best cuts maxCut
            end if
            if data_ub < maxValue then
                request ← FATHOM
                send message to master with request request
            else
                request ← GET_CHILDREN
                send (data, i, j) to the master with request request
            end if
        end while
    end if
```

5 Testing and results

We implemented the parallel version of BiqMac (named PBM) in C language using MPI library. We tested the algorithm on the supercomputer of University Ljubljana, Faculty of mechanical engineering. For testing we used the following groups of graphs (see [8]):

- g05_100: unweighted graphs with 100 nodes and edge probability 0.50.
- pm1d_100: weighted graph with 100 nodes, edge weights chosen uniformly from $\{-1, 0, 1\}$ and density 0.99.
- w09_100: weighted graph with 100 nodes, integer edge weights chosen from $[-10, 10]$ and density 0.9.
- pw09_100: weighted graph with 100 nodes, integer edge weights chosen from $[0, 10]$ and density 0.9.

For each group of graphs, we tested the parallel algorithm PBM. We ran the algorithm with 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 slave nodes. We measured the time (in seconds) that the algorithm needed to process all the nodes of the branch and bound algorithm tree (see result tables in [5]). For comparison, we ran the $BiqMac$ algorithm over the same groups of graphs (see result tables in [5]). Using the collected data, we computed the following quotients:

$$quotient(graph, x) = \frac{\sum_{i=0}^{9} BiqMac(graph.i)}{\sum_{i=0}^{9} PBM_x(graph.i)},$$

where $x \in \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ is the number of slave nodes that are available to the parallel version PMB, $graph \in \{g05\_100, pm1d\_100, w09\_100, pw09\_100\}$ is the selected group of graphs and $i \in \{0, 1, \ldots, 9\}$ is the index of a graph instance in the selected group of graphs $graph$. This quotients shows how many times the parallel algorithm PBM is faster than the $BiqMac$ over the same group of graphs with the given number of slave nodes.

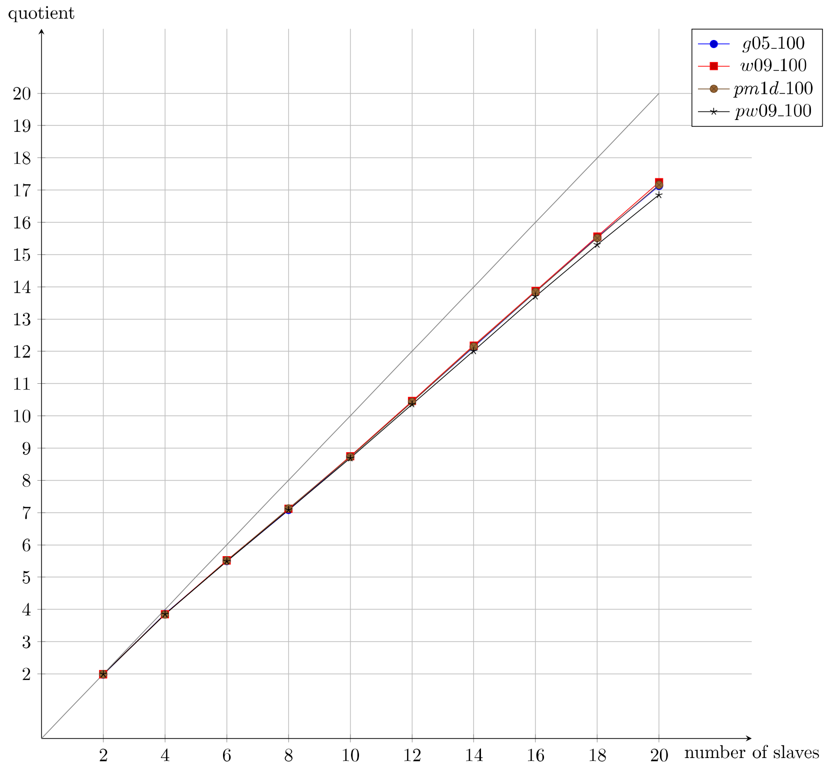


Figure 2: Comparison of quotients of parallel version PBM with different number of available slave nodes for all tested groups of graphs. Data for computing these quotients is taken from [5].

Acknowledgement

The authors would like to thank to Slovenian Research Agency (ARRS) and to Austrian Science Foundation (FWF) for support via joint Austrian-Slovenian project nr. N1-0057.

References

[1] I. Fischer, G. Gruber, F. Rendl in R. Sotirov, Computational experience with a bundle approach for semidefinite cutting plane relaxations of MaxCut and equipartition, Mathematical Programming 105 (2006): 451–469.

[2] M. X. Goemans in D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, Journal of the Association for Computing Machinery 42 (1995): 1115–1145.

[3] C. Helmberg in F. Rendl, Solving quadratic (0,1)-problems by semidefinite programs and cutting planes, Mathematical Programming 82 (1998): 291–315.

[4] C. Helmberg, F. Rendl, R. J. Vanderbei, H. Wilkowicz, An Interior-Point Method for Semidefinite Programming, SIAM Journal on Optimization 6(2) (1996): 342—361.

[5] A. V. Kalamar, Paralelni razveji in omeji algoritem BiqMac Solver, master's thesis, Faculty of natural science and mathematics, University of Maribor (2018).

[6] S. Poljak, F. Rendl, Solving the max-cut problem using eigenvalues, Discrete Applied Mathematics 62 (1995): 249–278.

[7] F. Rendl, G. Rinaldi in A. Wiegele, Solving Max-Cut to Optimality by Intersecting Semidefinite and Polyhedral Relaxations, Mathematical Programming (2010): 121–307.

[8] A. Wiegele, Biq Mac Library (2007), accessible at: http://biqmac.uni-klu.ac.at/biqmaclib.html.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 5:

Optimization in Human Environments

MULTIMEDIA-CONTENT-INDEX BASED EXPERIMENTAL CONTENT SELECTION

Evin Aslan Oğuz
Faculty of Electrical Engineering University of Ljubljana,
Tržaška cesta 25, 1000 Ljubljana, Slovenia
Nielsen, Obrtniška ulica 15, 6000, Koper, Slovenia
Email: evin.aslanoguz@nielsen.com
Andrej Košir
Faculty of Electrical Engineering University of Ljubljana,
Tržaška cesta 25, 1000 Ljubljana, Slovenia
Email: andrej.kosir@fe.uni-lj.si

Abstract: A measurement of multimedia exposure is challenging, since the impact of multimedia-content on a user is dependent on the user, the multimedia-content, the context, previous exposure, etc. We focus on measuring the impact of multimedia-content on a user's buying-decision process. We develop a novel, multimedia-exposure instrument (questionnaire) via an online crowdsourcing study, focusing on three core aspects. In our study, test users are exposed to a short film, including a commercial, and asked to answer some questions. In this paper, we describe the problem of commercial selection, and solve it as an optimization task, and then discuss the obtained results.

Keywords: optimal multimedia-content selection, experimental content selection, multimedia exposure

1 INTRODUCTION

Multimedia is the dominant media type in the information industry. As individuals and groups, we are exposed to a variety of information on different media platforms and from diverse sources, regardless of time and location [5]. The measurement of media exposure is crucial in the areas of communication science, political science, sociology, psychology, and economics. The access, composition and activities of media viewers are measured in a wide variety of content contexts, ranging from news, advertising, entertainment and health, to platforms such as television, billboards, videos, games, newspapers and social-networking sites. This penetration of multimedia into everyday life means that media exposure has far-reaching effects for individuals, businesses and industry, and is the main currency of advertising [2].

Even though the term *exposure* is usually encountered in the area of health, we refer to it in the context of experience and marketing. The dictionary definition of exposure, in the context of experience, is: *"the fact of experiencing something or being affected by it because of being in a particular situation or place"* [3], and in the context of marketing it is: *"the amount of public attention that someone or something, especially an advertisement or product, receives"* [4]. We refer to the keyword *exposure* in the sense of how the user has been attracted and affected by watching particular multimedia-content.

Media exposure was first defined by Slater as: *"the extent to which audience members have encountered specific messages or classes of messages/media content."* [6]. In the context of multimedia, this means the individual is exposed to the *"... seamless integration of data, text, images of all kinds and sound within a single, digital information environment."* [5].

People can be dramatically affected by multimedia in areas such as education, entertainment, films and advertising. For instance, students are exposed to multimedia-content in lectures, in the form of videos and animations. By employing these kinds of multimedia, teachers aim to increase the effect of exposure, which helps to hold people's attention during a lecture and finally to improve people's understanding of the presented subject and their recall of the presented facts. In the area of advertising, on the other hand, the consumer can be affected positively or negatively by being exposed to a commercial for a certain product. Specifically, this exposure can affect the five stages of the formally defined, consumer's buying-decision process in marketing, which are: problem recognition, information search, information evaluation, decision and post-purchase evaluation [1]. Because of this effect of commercials on a consumer's buying-decision process, in this paper we mainly focus on the area of advertising.

As mentioned earlier, measuring the exposure to multimedia is crucial for both science and industries such as television, advertising, news, etc. However, despite multimedia being ubiquitous in our daily lives and being crucial for both science and industry, De Vreese and Neijens conclude—based on over two hundred studies published in two leading journals (*Journal of Communication and Communication Research*)—that there is no generally accepted conceptualization and operationalization of this phenomenon in the past decade [2]. Also, to the best of our knowledge, there is still no attempt to directly measure the impact of viewed multimedia-content on a consumer's buying decisions.

The focus of this paper forms part of a wider topic, which is to research and measure the impact of multimedia commercials on an individual's buying decisions via an instrument called the Multimedia Exposure Meter (MMEM). The goal of this paper is to pose the question and solve the problem of selecting experimental multimedia-content as an optimization task. This will allow us to evaluate different experimental options in terms of different criteria of content selection. We formulate the problem of optimal experimental content selection in the form of an optimization task, provide the solution to this content selection, and evaluate its properties.

2 PROBLEM DESCRIPTION

The problem we address is how to formulate and solve the optimization task of experimental content selection to incorporate the requirements of the experiment. However, since there are several factors in the experiment, we have a multi-objective optimization problem. Also, not only the best, but the first n_a, content items need to be determined. However, as can be seen from Section 3.1, the size of the data and the multi-objective problem do not allow us to select the content manually.

2.1 Observational Study

The first phase of the multimedia exposure meter's development is conducting an observational study aimed at developing the model. Conducting an observational study is relevant to us, since we do not know the actual factors, but only investigated and decided on the relevant factors, which might affect the multimedia exposure. We decide that our observational study will be in the form of an online crowdsourcing study, since a large number of users are accessible, and with the luxury of being able to target only predefined user profiles. The aim of this online crowdsourcing study is to test the instrument gathered during the operationalization and to refine the questions that will contribute to the multimedia exposure model.

Specifying the properties of an observational study are crucial in order to eliminate biases and finally develop a more reliable measure. The factors are decided based on a literature survey on the exposure keyword, discussion with experts in the field and careful consideration of the collected information.

3 METHODOLOGY

In this section we outline the experimental data and the approach to the problem of content selection as an optimization task. We implemented our solution via a custom Python-code implementation.

3.1 Descriptive Summary of Data

The test data of the commercials together with the meta data was obtained from the Nielsen company. We have real data on the number of viewers, the total duration of the commercials, and the index derived from this information. The dataset is based on views between 01.01.2017 and 19.09.2018, it contains 1,885,203 individuals, and 2399 commercials.

As indicated in Table 1 the factors of the commercials we decided to use are: index and brand. The index of a commercial is simply the number of viewers for that commercial per second, and it is available in our dataset. We denote the index factor as F_1.

The brand can either be an internationally known brand or an unknown brand that the users are not familiar with on commercials and also on the shelves of the stores. To determine this factor, we categorize all the commercials in the range [1-5] depending on the total number of individuals and the percentage of individuals that saw the commercial. We use designed rules for categorizing each commercial. We denote the brand factor as F_2.

Commercial Factors	**Possible Values**
Index	Nonnegative reals
Brand	[Known, Unknown]

Table 1: Commercial Factors Table

3.2 Content Selection as an Optimization Task

In this section we formulate commercial selection as an optimization task. The criteria function is derived from observational study factors that we wish to take into consideration, see Table 1.

The cost (objective) function should reflect the distinct values of observational study factors F_1 and F_2. To allow the analysis of levels for these two factors on the resulting MMEM, we select commercials with a combination of extremely low and extremely high values for these two factors. Note that we do not need a single item only, but several best-fitting items. We introduce the notation $\text{argMin}(c, n_c)$ denoting a set of n_c commercial items with minimal values of the cost function c. Analogously, we define $\text{argMax}(c, n_c)$ as a set of n_c commercial items with the maximal values of the cost function c. Here, the cost function c is either factor F_1 or F_2. To select commercials with extreme values of these factors, we apply all four available combinations argMin and argMax, namely, $\text{argMin}(F_1(\text{argMin}(F_2, n_2)), n_1)$, etc. Since the results of the optimization are dependent on the order of the cost functions applied, the resulting sets are not equal, for example

$$\text{argMin}(F_1(\text{argMin}(F_2, n_2)), n_1) \neq \text{argMin}(F_2(\text{argMin}(F_1, n_1)), n_2),$$

we decide to use the intersection of these two resulting sets. Altogether, for a given set of commercials D with known values of factors F_1 and F_2, we obtain four independent optimization tasks resulting in four selected sets of commercials

$$D_{mm} = \text{argMin}(F_1(\text{argMin}(F_2, n_2)), n_1) \cap \text{argMin}(F_2(\text{argMin}(F_1, n_1)), n_2) \quad (1)$$
$$D_{mM} = \text{argMin}(F_1(\text{argMax}(F_2, n_2)), n_1) \cap \text{argMin}(F_2(\text{argMax}(F_1, n_1)), n_2) \quad (2)$$
$$D_{Mm} = \text{argMax}(F_1(\text{argMin}(F_2, n_2)), n_1) \cap \text{argMax}(F_2(\text{argMin}(F_1, n_1)), n_2) \quad (3)$$
$$D_{MM} = \text{argMax}(F_1(\text{argMin}(F_2, n_2)), n_1) \cap \text{argMax}(F_2(\text{argMax}(F_1, n_1)), n_2) \quad (4)$$

The algorithm of our solution is as follows. First, we sort a predefined set of items by the first factor and then sort again with a second factor. After that we change the order of the factors when sorting, i.e., we first sort with the second factor and then with the first factor. Finally, we intersect these two resulting sets and select the best solutions according to the parameters supplied to the optimization task.

4 EXPERIMENTAL RESULTS

In this section we report on the experimental results of the commercial selection obtained by solving optimization tasks. To implement a planned observational study, we need a list of the ten best commercials with respect to two factors F_1 and F_2 denoted by D_{mm}, D_{mM}, D_{Mm} and D_{MM}, see Subsection 3.2.

4.1 Content Selected by Optimization

Lists of the commercials selected by the optimization for all four optimization tasks are given in Tables 2, 3, 4, 5. Explanations are given in the table captions.

F1	F2	Product
0.00	1.0	Karting Btc
9.50	1.0	Saeka Naturopatski Center
10.02	1.0	Concert 50 Years On Holiday
10.51	1.0	Najlepši-Dan.Si
10.70	1.0	Expres Printing Clothes
10.87	1.0	T Rolex Honey
11.46	1.0	Petrol Hip Mobile
12.02	1.0	Osmo Dji Camera
12.55	1.0	Lutherm Cheese Ementalec
13.00	1.0	Pvg Motor Trade

Table 2: Results of min min optimization (D_{mm}). All values of $F_2 = 1$ are minimal possible.

F1	F2	Product
7.62	5.0	Hot Horse
15.55	5.0	I Feel The Film
15.99	5.0	E-Stave
17.83	5.0	Mutual Security Insurance
20.19	5.0	Film Browser 2049
20.25	5.0	New Kbm Loans
20.29	5.0	Bicycle Cinema
20.86	5.0	Baci Perugina
22.46	5.0	Www.Migimigi.Si
22.77	5.0	Wiz Automotive Insurance

Table 3: Results of min max optimization (D_{mM}). All values of $F_2 = 5$ are maximal possible.

F1	F2	Product
1550.77	1.0	Public Services Ptuj
1150.20	1.0	Husqvarna Kosilnica
1140.16	1.0	Export Windows Web Portal
1073.56	1.0	Audi
905.45	1.0	R-Moto
612.23	1.0	Concert Ambassador Good Will
585.07	1.0	Golden Lisica World Cup Maribor
556.08	1.0	Agrometal Trade With Agricultural Mechanization
539.19	1.0	Maxi Starr Vaba
535.17	1.0	Grapak
527.21	1.0	Steyr Tractors
507.39	1.0	Suzuki Vitara All Grip

Table 4: Results of max min optimization (D_{Mm}). All values of $F_2 = 1$ are minimal possible.

F1	F2	Product
987.24	5.0	Palma Tourist Agency
917.63	5.0	Olympic Committees Of Slovenia
894.34	5.0	Merkur Zavarovalnica Fit4life
769.51	5.0	Postojnska Jama Zive Jaslice
763.82	5.0	Office Of The President Of The Republic Of Slo...
750.56	5.0	Alpe Adria
707.71	5.0	Meeting Information
641.12	5.0	Fund Of Crafts And Entrepreneurs
618.19	5.0	Cankarjev Dom
599.23	5.0	Festival Ljubljana
539.93	5.0	Evening Event

Table 5: Results of max max optimization (D_{MM}). All values of $F_2 = 5$ are maximal possible.

4.2 How Close to the Best Possible Single-Criterion Items did We Get?

To verify the properties of the selected content, we graphically present the selected content in terms of the sorted factors F_1 and F_2. The vertical red lines represent the positions of the selected items.

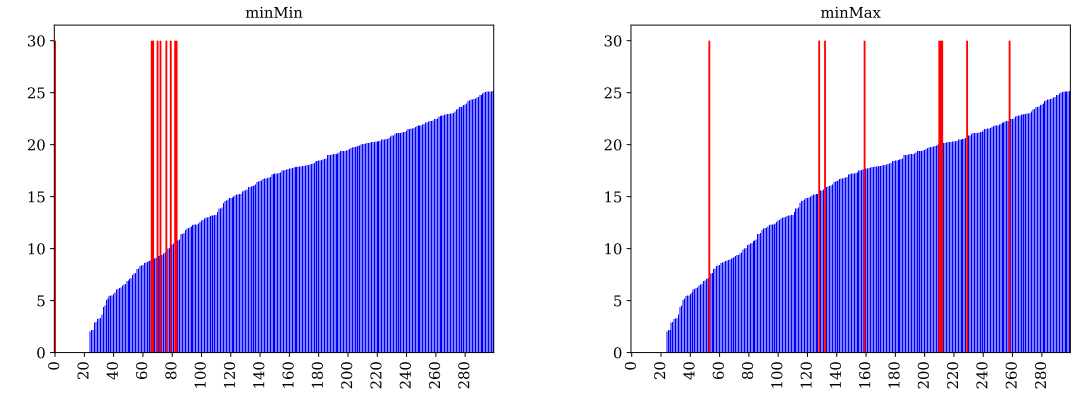


Figure 1: Vertical red lines indicate the positions of the solutions of min min (left) and min max (right) in the sorted graph of factor F_1. The values of F_2 are all optimal, i.e., $F_2 = 1$ (left) and $F_2 = 5$ (right). Only the first 300 out of 2499 are shown.

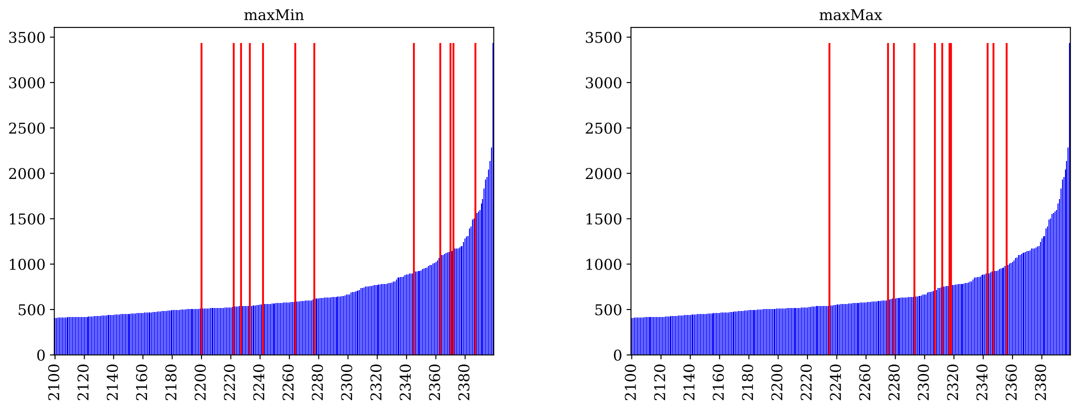


Figure 2: Vertical red lines indicate the positions of the solutions of max min (left) and max max (right) in the sorted graph of factor F_1. The values of F_2 are all optimal, i.e., $F_2 = 1$ (left) and $F_2 = 5$ (right). Only the last 300 out of 2499 are shown.

For Figs 1 and 2 we observe that the multi-objective (we optimize according to two factors F_1 and F_2) optimization results in a relatively close to best possible items regarding single-criterion optimal results.

5 DISCUSSION

In this paper we pose the question and solve the optimal experimental content selection as an optimization task. This approach proved to be useful in the design of an observational study with real test users aimed at modeling multimedia exposure.

The advantage is the resulting tool for optimal content selection, allowing us to examine the properties of the experimental content available to us.

The drawback is the unbalanced impact of factors F_1 and F_2 on the content selection. An alternative approach is to combine the cost functions as $c = a_1 F_1 + a_2 F_2$ and to control the sign and the size of the weights a_1 and a_2. We leave this as a future work.

Acknowledgement. This research was supported by the Nielsen company and project *P2-0246 ICT4QoL - Information and Communications Technologies for Quality of Life*. We thank our colleagues from Nielsen who provided data, insights and expertise that greatly helped the research.

References

[1] Frances Brassington and Stephen Pettitt. *Essentials of Marketing Management, Second Edition*. Pearson Education, Hudson in New York City, New York, 2007.

[2] Claes H. de Vreese and Peter Neijens. Measuring Media Exposure in a Changing Communications Environment. *Communication Methods and Measures*, 10(2-3):69–80, 2016.

[3] Cambridge Advanced Learner's Dictionary. *EXPOSURE — meaning in the Cambridge Advanced Learner's Dictionary*. Cambridge University Press, University Printing House Shaftesbury Road Cambridge CB2 8BS, United Kingdom, 2013.

[4] Cambridge Business English Dictionary. *EXPOSURE — meaning in the Cambridge Business English Dictionary*. Cambridge University Press, University Printing House Shaftesbury Road Cambridge CB2 8BS, United Kingdom, 2011.

[5] T. Feldman. *Multimedia*. BRBRF report. Psychology Press, Mortimer Street, London, 1994.

[6] By Michael D Slater. Operationalizing and Analyzing Exposure: The Foundation Of Media Effects Research. *Journalism and Mass Communication Quarterly*, 81(1):168–183, 2004.

OPTIMALITY OF FLIPPED LEARNING EXPERIENCE: A CASE STUDY OF USING 2-CROSSING-CRITICAL GRAPHS FOR EARLY RESEARCH EXPOSURE

Drago Bokal[1,4],Robert Repnik[1,2,3], Špela Tertinek[1], Alen Vegi Kalamar[1], Tadej Žerak[1]
[1]University of Maribor, Faculty of Natural Sciences and Mathematics, Maribor, Slovenia
[2]Faculty of Natural Sciences and Mathematics, Physics Department, University of Maribor, Koroška 160, SI-2000 Maribor, Slovenia
[3]Association for Technical Culture of Slovenia, Zaloska c. 65, 1000 Ljubljana, Slovenia
[4]Institute of Mathematics, Physics, and Mechanics, Jadranska 19, Ljubljana
[drago.bokal|robert.repnik]@um.si, [spela.tertinek|alen.vegi.kalamar|tadej.zerak]@student.um.si

This research was also supported by the Slovenian Research Agency (Grants No. P1-0403, J1-8130, and N1-0057).

Abstract: We investigate the context of flipped learning in higher education from the standpoint that academia is focused on basic knowledge production at low levels of technology readiness, but is educating students who aim to be employed in processes at higher level of technology readiness. This leads to a conflict of interest between veristically oriented perceptions, decisions, and actions of the academe, versus utilitaristically oriented perceptions, decisions, and actions of the businesses, which translates to the classroom through students' desire to apply knowledge and professor's desire to deepen it. We argue that the classroom is an ideal opportunity for combining the best of both worlds, as well as exchanging knowledge between them. Moreover, we argue that the approach of flipped classroom both synchronizes the incentives of the learner and the learned, fulfilling sufficient and necessary conditions to render teaching effectiveness the only stable mechanism for rising grades (contrary to others that yield grade inflation). Finally, we argue that flipped classroom fulfills the conditions psychologists have established for achievement of the state of flow, an optimal experience of frictionless absorption in the action a person is performing. We propose that large 2-crossing-critical graphs, a recently characterized class of graphs, is an ideal theoretical tool to test these assumptions in a laboratory setting of teaching graph theory, and test an implementation of the listed ideas.

Keywords: Learning experience, 2-crossing-critical graphs, flow, technology readiness levels.

1 Introduction

Technology readiness levels (TRL) were introduced by NASA [15] to track nine steps of progress of ideas from observed basic principles in published papers, through technological concept formulation, laboratory prototypes, demonstration in increasingly relevant contexts, to first flight missions and final proven technologies. The same principle was adopted by European Comission for determining to fund H2020 research projects [9]. Mathematically speaking, TRL1 knowledge is contained in axioms, definitions, existential theorems. TRL2 knowledge is contained in constructive theorems and accompanying mathematical structures. TRL3 knowledge is given in algorithms and corresponding proofs of correctness. Then knowledge basically leaves mathematics, and gets absorbed into other derived disciplines. For instance, operations research deals with data sets, implementations of the earlier mentioned algorithms, and management processes ensuring implementation of proven optimal solutions of stated problems. At TRL4, algorithms become libraries, at TRL5, those are implemented into functioning applications and at TRL6, they become part of information systems. At TRL7, these libraries, applications, and information systems are used for commercial purposes as part of individual, custom-made solutions. At TRL8, these custom solutions become products, tested and ready for general deployment, and at TRL9, support processes for those products are deployed and working.

It is clear from this description, that at the latest TRL5 knowledge is no longer relevant to the extent that pure mathematicians consider interesting, seemingly "just trivial technical problems still need to be solved." On the other hand, business manager who may be faced with a challenge of integrating this knowledge into his company business process would claim that "this young technology has solved no problem yet, it is too risky to be used," and he prefers to stick to the existing processes for dealing with the issues the new approach claims to solve better. This disparity between theoretical and practical skills has been termed "technological valley of death." [7].

In a parallel contribution [7], we look at the problem from theoretical biology point of view and show that in the language of theory of perceptions, one of the reasons for the technological valley of death to emerge is that utility received by agents in the environment is not a monotonous function of the amount of resources in their environment. This renders veristic perceptions that favor perceptions preserving the structure of the world inferior in comparison to utilitaristic perceptions tuned to observe the utility of the amount of resource.

Csikszentmihaly's model of emotional experience [5] claims that flow, the state of optimal experience, depends on nine conditions, upon which we elaborate in Section 3, where we also point out the ease with which a mathematically inclined person fulfills these within the well-structured field of mathematical reasoning. The first among the conditions is clarity of the goals. These are well defined within early technology readines levels, and at the higher technology readiness levels, but have less defined goals at intermediate technology readiness levels: on the first hand, perhaps a better structure, implementation could be achieved if time is invested into research and development, on the other hand, perhaps more money can be earned if time is invested into marketing, sales, and process optimization.

Similar conflict of interest occurs between students and professors of higher education. The latter have their success determined through published scientific results, and cherish the culture of knowledge seeking, but the former strive to gather the knowledge they will be able to apply at higher levels of technology readiness. As we argue in another parallel contribution [16], the result of this disparity of interest is grade inflation, which could be eliminated by synchronizing interests of both the recipient and the source agent of the knowledge exchange game.

This background is presented as the evolutionary biology and game theoretic argument to motivate flipped learning. Such approach to learning is explained in the next section, and we show in what follows, that with proper implementation, it can synchronize the incentives of both professors and students, as well as fulfill the required conditions for both to achieve flow, the state of psychologically optimal experience while performing activities. To eliminate the issues of business interest and profit dependence, as well as to demonstrate the applicability of stated concepts and processes in the setting of pure mathematics, we present a case study of 2-crossing-critical graphs, who can play a role in jump-starting research career in graph theory. We hope that this example is a good showcase to motivate similar instances at even higher levels of technology readiness.

2 Flipped learning process compared to a traditional process

The concept of flipped learning originates in 1990, when professor Mazur prepared learning materials for students so they can prepare beforehand for the group activities in the classroom [4]. In 2000, Baker first presented the "Classroom Flip" model at the Council of Independent Colleges Information Technologies Workshop [2]. It is important to point out the difference between flipped classroom and flipped learning [14, 11]. The flipped classroom is a form of blended learning and does not necessarily have to be implemented to all learning units. On the other hand, flipped learning is a pedagogical approach in which students individually study prepared materials before a lesson and group class activities are interactive and usually based on problem solving and peer discussion [8].

In the traditional learning approach, the teacher generally provides knowledge in terms of lectures. After the lesson, students do homework [1], which demands to synthesize, analyze and evaluate acquired knowledge. Students have a passive role in the classroom, which can result in low motivation, poor attitude towards the subject and lower knowledge. If complex problems are handled only in terms of homework, it could also lead to misunderstanding.

In the flipped learning, the teacher's role changes from a lecturer to a guide and facilitator. The teacher has to prepare learning materials in advance and provide them to students. It is important that learning materials are of appropriate length and available to all students [1]. Nowadays there are many options for creating and distributing learning materials of different types, for example audio and video lectures, e-textbooks, presentations, simulations, blogs, quizzes and many more. Students acquire basic level of knowledge using the provided learning materials before the class. The main advantage is that students can study learning materials in their own pace. During the class [12], the teacher leads students through activities, in which all students are actively engaging. Because students are already familiar with the learning content and have a basic understanding, time in class can be used to upgrade and deepen their knowledge. Students discuss complex problems and solve them in a group or in pairs with the help of a teacher. Commonly, learning methods are peer discussion and problem-solving. After the classroom activities, students and teacher check understanding and extend learning.

Various case studies [6, 13, 10, 18] confirm that flipped learning result in better study outcomes and increased motivation. In flipped learning approach teacher guides students through tasks that develop higher-level cognitive skills, whereas students obtain low-level skills individually with the help of provided learning materials. This approach is student-centered and demands the active engagement of students. Consequently, this can lead to problems if students are not motivated and willing to learn. The latter can emerge if the transition from the traditional approach to flipped learning and implementation of it is not properly carried out.

3 Optimality of emotional experience

In this section, we demonstrate how flipped learning presentation of 2-crossing-critical graphs fulfills conditions for establishing flow, as identified by Csikszentmihalyi [5]. The conditions he identified can be grouped into four sets. The first group contains conditions that depend on the task. They include *clear goals*, *efficient feedback*, and *control not being a problem.* All three are easily satisfied with mathematical tasks, as they are facilitated by the rigor of the mathematical logic. The second group contains conditions, depending on the environment: *deepened concentration* and *focus is on the present.* Flow inducing environment should support these two conditions, and a flipped classroom with focus on a long term goal as well as a clear route to that goal facilitates these two conditions.The last group of conditions involve the person working on the tasks. First, *challenges of the task should be balanced with the skill of the person.* The elementary set of exercisces hopes to build the skill slowly enough towards the final goal of understanding the problem well enough to show a new theorem. The final three conditions have a somewhat different nature: they are realized by devoting full attention to the task and not having it left for reflection upon it: *perception of time changes*, *the person looses her ego*, and *the experience turns autotelic*, i.e. its own award.

4 Compatibility of incentives

Smole et al. [17], identified sufficient and necessary conditions on knowledge exchange incentives so that teaching effectiveness, not grade inflation, is the mechanism behind improving grades. In this short listing, we intertwine them with arguments why they were satisfied in our case study and flipped learning in general. For the student, the conditions state that teaching effectiveness is manifested in each Nash equilibrium of the Knowledge Exchange Game if and only if (a) for the student, (a1) the utility of knowledge exchanged with a cooperating professor is higher than the utility of other work and high grade, (a2) the utility of other work must be higher than the utility of the knowledge exchanged with a professor who does not cooperate together withthe utility of high grade, and (b) for the professor, (b1) the utility of cooperating in knowledge exchang with a cooperating student is higher than the utility of other work, and (b2) the utility of cooperating in knowledge exchange with a non-cooperating student is higher than the utility of other work and high evaluation together. In Table 1 we summarize the comparison for flipped learning to be more incentive compatible than traditional learning.

Cnd.	Traditional learning	Flipped learning
a1	Focus on passive reproduction (remembering, understanding) rather than active creativity (applying, analyzing, evaluating, creating) leads to relaxation, not flow in the final outcome.	Emphasis on active skills renders higher career alignment of teaching. Student participation with cooperating professor fulfills psychological flow conditions more easily, thus (a1) is more likely fulfilled.
a2	Other courses or work outside studying on tasks directly linked by student's future career that have well defined goals likely contributes more of both emotional as well as von Neumann-Morgenstern utility.	
b1	In traditional classroom, professor derives emotional utility from student's success and vnm utility from educating a student who will be able and willing to participate in their research.	While the sources of professor's utility are the same, by introducing research into the classroom, flipped learning derives more of the professor's utility from exchanged knowledge, thus improving the likelihood of this condition being fulfilled.
b2	In absence of mechanisms deriving either emotional or vnm utility from knowledge exchange, this may easily render the condition violated in the traditional learning approach.	Attempting active objectives with non-cooperating students is a more challenging task. By Csikszentmikhaly model, to an interested professor with deep interests in the material, such students pose a challenge that renders at least a higher emotional utility if successful.

5 2-crossing-critical graphs as a case study

Our empirical confirmation of the conditions in previous sections derives from the experience of a master thesis by Žerak [19] and a paper by Bokal, Vegi Kalamar, and Žerak [3], taking the topics from the thesis to a research level. The thesis set the basis for flipped learning by preparing a list of tasks needed to be solved by a novice to gain necessary understanding of graph theory, graph drawing, crossing numbers, and c-crossing-critical graphs. Each task consists of its task statement, which could be given independently to a student able to find online sources of knowledge. To facilitate that process, each task is equipped with explanation, integrating those sources of knowledge and sufficing for the taks to be solved. Finally, each task has a solution that is the last resource a student can peek into while finding ideas to understand the material.

The tasks were prepared by Bokal and Žerak and expanded by Žerak. They were given to Vegi Kalamar, who solved them and proceeded to the research task beyond the scope of the introductory task of proving hamiltonicity of large 2-crossing-critical graphs: counting the Hamiltonian cycles in them. The challenges proved engaging, and as the bibliography search revealed significant interest not only in Hamiltonicity, which is trivial for large 2-crossing-critical graphs, but also in counting Hamiltonian cycles, we dived into this problem and managed to solve it for a more general class containing large 2-crossing-critical graphs.

We conclude with personal reflections on this experience by all three coauthors of the final paper, which we hope is the basis for a more systematic investigation of the approach.

Mentor's reflection. This research has led me to detour from traditional mathematics "converting coffee into proofs of theorems," to quote Erdős, into reflection upon why we are doing mathematics, why and how best to engage youth in it. Combining this research with the earlier quoted findings about knowledge maturing from low to high technology readiness levels opens several new directions one can dive into: while science may be emphasizing veristic aspects of understanding the world and business may be emphasizing its utilitaristic aspects, both types of understanding coevolve together. In practice, the least each side can take from this is that scientists should in part think about "why" their aspect of research is relevant to current society in general and their students in particular. Business, on the other hand, could see the importance of being able to communicate with science through their R&D departments maintaining relationship with academia through projects and scholarship.

Developing the crossing number tutorial. This approach to teach proves itself as an excellent way of splitting an objective into manageable tasks, that are also very easy to follow for both the reader and the writer. With this objective in mind, writing smaller tasks that eventually build up to the desired knowledge forces the writer to break down the topic to sufficient detail, simultaneously allowing him to check and upgrade his knowledge of the topic. In the Crossing number tutorial I started with basics. It was really fun to start with essentials, to build up the readers knowledge, as it was fun to organize the order of tasks by level of difficulty and what knowledge is required to be able to solve each one. As the difficulty progressed, I did not feel lost. Tasks were designed by a goal in mind, and if any task would not be solvable as it would appear too difficult, I would just break it even further. That way I really felt in control. And then comes the finished work. Multiple questions are raised. Are steps between questions too big for reader to understand? Or are they not big enough, and the reader will be bored? Are they understandable? And are hints really required or will the reader feel foolish that I think he will need them, but they are really not required? Feedback really helps answering those questions, and a practical view of this approach is fixing a part of the whole paper is an easy task, the only thing that must be done is either add tasks, or join few of them together, should either prove necessary in further studying of the crossing number academy.

Building research-level knowledge on top of the crossing number academy. The path to writing a paper about counting Hamiltonian cycles in 2-tiled graphs started when Bokal introduced me the new approach to acquire knowledge. With Žerak they presented the academy of 2-crossing-critical graphs, which, to a novice in this field, was a nice introduction to crossing-criticality. The academy, which is composed as a sequence of tasks introducing theoretical background of a field, has a systematic flow, which directs you to an independent dive into the field. The more the solver of the academy approaches the end of the tasks, the better idea of the field he gets, which raises new questions and problems. When I was finishing the academy it ended with tasks that dealt with the Hamiltonicity of large 2-crossing-critical graphs. A natural question appeared: Is it possible to count all Hamiltonian cycles in large 2-crossing-critical graphs? A phase when I dived into this question under the mentorship of Bokal and in cooperation with Žerak followed. The result of the research was a linear time complexity algorithm that counts all Hamiltonian cycles in large 2-crossing-critical graphs. The result naturally offered us a new question: Can the result be generalized to some generalized class of graphs? The result of the research of generalized case is an algorithm that counts Hamiltonian cycles in 2-tiled graphs. In a special case, if the set of tiles is finite, the algorithm has quadratic time complexity and if all tiles in the finite set have an additional property (as in the case of large 2-crossing-critical graphs), the algorithm has linear time complexity.

6 Conclusions

Using flipped learning approach to investigating abstract mathematical material, such as 2-crossing-critical graphs, builds upon insights from several domains of knowledge. The early experience has yielded positive feedback that confirms some theoretically predicted aspects and hence motivates further research on the topic.

References

[1] Advance HE, Flipped learning. https://www.heacademy.ac.uk/knowledge-hub/flipped-learning-0. Advance HE, Flipped learning, available at: https://www.heacademy.ac.uk/knowledge-hub/flipped-learning-0

[2] Baker, J. W. (2000). The Classroom Flip: Becoming the "Guide by the Side." Council of Independent Colleges Information Technologies Workshop, Pittsburgh, Penn. http://www.classroomflip.com/files/baker - 2000 - 03 - 31 - guide - by - the - side - CIC - workshop.pdf Advance HE, Flipped learning, available at: https://www.heacademy.ac.uk/knowledge-hub/flipped-learning-0

[3] Bokal, D., Kalamar, A. and Žerak T. (2019). Enjoying counting Hamiltonian cycles in 2-tiled graphs, article in preparation.

[4] Crouch, C.H. and Mazur, E. (2001). Peer Instruction: Ten Years of Experience and results. American Journal of Physics. 69: 970–977.

[5] Csikszentmihalyi, M. (2008) Flow: The Psychology of Optimal Experience. New York: Harper Collins.

[6] DesLauriers, L., Schelew, E. and Wieman, C. (2011) Improved Learning in a Large-Enrollment Physics Class. Science. 332: 862–4.

[7] Fic, P. and Bokal, D. (2019). Innovative veristic perceptions do have a chance: An instance of artificial technological valley of death. Submitted to SOR.

[8] Flipped Learning Network (FLN). (2014). The Four Pillars of F-L-I-P^{TM} www.flippedlearning.org/definition. [Accessed 10/04/2019]

[9] G. (2015). Technology Readiness Levels. European Commission. https:ec.europa.eu/research/participants/data/ref/h2020/wp/20142015/annexes/h2020 – wp1415 – annex – g – trlen.pdf [Accessed 10/04/2019].

[10] Gilboy, M. B., et al. (2015). Enchancing student egagement using the flipped classroom, Journal of nutrition education and behavior. 47(1): 109-114. https://doi.org/10.1016/j.jneb.2014.08.008. [Accessed 10/04/2019].

[11] Hamdan, N., McKnight, P., Mason, G., McKnight K., Arfstrom K. M. (2013) A review of flipped learning. https://flippedlearning.org/wp-content/uploads/2016/07/LitReview - FlippedLearning.pdf. [Accessed 10/04/2019].

[12] Lee J. et al. (2017). Development of an instructional design model for flipped learning in higher education. Education Tech Research Dev, 65: 427-453.

[13] Mohanty A., Parida D. (2016). Exploring the Efficacy and Suitability of Flipped Classroom Instruction at School Level in India: A Pilot Study. Creative education. 7: 768-776.

[14] O'Flaherty, J., Phillips, C. (2015). The use of flipped classrooms in higher education: A scoping review. Internet and higher education, 25: 85-95 http://dx.doi.org/10.1016/j.iheduc.2015.02.002. [Accessed 10/04/2019].

[15] Mankins, J.C. (2004). Technology readiness levels. NASA: A White Paper. 6.

[16] Smole, A., Jagrič, T., Bokal, D. (2019). Principal-Leader-Follower Model with Internal Signal. Submitted to SOR.

[17] Smole, A., Kent D., Jagrič T., and Bokal D. (2019) Social dilemma model shows valuing knowledge disincentivizes grade inflation. Submitted.

[18] Tune, J. D., et al. (2013). Flipped classroom model improves graduate student performance in cardiovascular, respiratory and renal physiology. Adv Physiol Educ, 37: 316-320.

[19] Žerak, T. (2019). Hamiltonian cycles in large 2-crossing-critical graphs (Masters thesis), in preparation.

Bounded time availability is what narrative incohesion, behavioral sink, behavioral addiction, and online social bubbles have in common.

Drago Bokal, Špela Tertinek
University of Maribor, Faculty of Natural Sciences and Math
Koroška cesta 160, 2000 Maribor, Slovenia
E-mail: drago.bokal@um.si, spela.tertinek@student.um.si

Abstract: A recent review of time allocation models by Jara-Diaz and Rodriguez-Salas exhibited a fairly neglected research area of optimal time usage that is of significant relevance to society at large and every individual in particular. Bokal and Steinbacher have closed this gap by exhibiting a task-based discrete time allocation model that features all the activity-related dimensions of existing von Neumann and Morgenstern utility based time allocation models and combines them with two determinants of Csikszentmihalys emotional utility - skill and challenge. They applied their model to a simulated process of learning to exhibit phase space of dependence of predominant emotional state on learning ability, perseverance, and passion. We build upon their research and combine it with models of Markov decision processes, used in reinforcement learning. We combine the two models, introduce a perceptive learning agent and formalize the concepts of agent's von Neumann-Morgenstern and Csikszentmihalyi utilities. We illustrate these models by showing its implications for understanding narrative incohesion, (behavioral) addiction, and for behavioral sink, showing that bounded amount of time is what these phenomena have in common.

Keywords: Time availability, Social behaviour Model of perceptive learning agent, von Neumann-Morgenstern utility, Csikszentmihalyi utility.

1 INTRODUCTION

In learning contexts, Bokal and Steinbacher [2] investigate the problem of bounded time availability in their task based discrete time allocation model (TBDTAM). TBDTAM uses time in discrete units to work on tasks which produce skills that through agent's choice affect his challenge levels and hence his emotional state. While rigorously taxonomizing the time allocation models surveyed by Jara-Diaz [18], we found out the TBDTAM model would generalize all the known models, should it account for goods in addition to skills. With this extension, we observed the model generalizes the model of an agent from interface perception theory [10], which is itself built upon models of Markov decision processes [12] that contain models of reinforcement learning, which brings us back to learning that TBDTAM started with.

We apply this general model to investigate the question of "which problems to use our attention for?" Our model allows for comparing three most common answers: 1) the interesting ones, 2) those where you can have progress, 3) those that help you sustain your career. By optimizing the time allocated to these problems, we investigate the process of sustaining appropriate levels of von Neumann-Morgenstern utility secured, as well as of Csikszentmihaly's utility experienced. Following this illustrative application of the model from the context most readers are familiar with, we continue to strengthen the case for distinction between von Neumann-Morgenstern utility and Csikszentmihalyi utility by exhibiting phenomena that could be explained in terms of Csikszentmihalyi utility of involved agents: known observations of narrative incoherence in patients with schizophrenia [11], behavioral sink, which could be explained as developing abnormal behavior due to lack of Csikszentmihalyi utility on one hand, and on the other, behavioral addiction and online social bubbles [15] which could be explained as optimizing Csikszentmihaly's utility.

The goal of this paper is, first to define a model of perceptive learning agent in Section 2, then we show results of Monte Carlo simulations of this model exhibiting the phase space of dominating emotional state affected by passion, perseverance and the number of activities practiced in Section 3. In the final Section 4, we interpret the results as narrative collapse, behavioral addiction, and behavioral sink. We also point out relationships with innovations and technological valley of death [6] and grade inflation [17].

2 A MODEL OF PERCEPTIVE LEARNING AGENT

From [10] and [2], we combine the following *perceptive learning model*. For an abstract model of an agent, let W be the measurable vector space of *all possible states of the agent's world*, regardless of agent's relationship with it. This space allows us to investigate the possible innovations an agent can introduce into his cognitive process. To avoid Russel-like paradoxes of set self-inclusion, we project W using a *focus projection* $F : W \to R$ to the measurable space of *relevant dimensions* R, which is a subspace of W. A dimension is *relevant*, if it affects agent's utility. As *utility* we identify a (small) subset of k agent's relevant dimensions, which we assume are real valued variables, and are identified through *utility projection* $U : R \to \mathbb{R}^k$. Usually, for most studied optimization problems, $k = 1$, but for multi-criterion optimization problems, where Pareto optimal solutions are sought, $k > 1$. Later we investigate relationship between von Neumann-Morgenstern utility and Csikszentmihalyi utility, so we set $k = 2$ in our model. In perception theory [10], R and W are considered the same space, which prevents seeing the possibility of innovations discussed in parallel contribution [6]. This introduces the next measurable space, the *space of agent's perceptions* X. In Markov decision processes, the spaces W, R, and X are all the same, but the distinction of W and X was introduced studying Interface theory of perception, allowing to investigate differences between various perceptive strategies [12]. Markov kernel of *perceptions* $P : R \to X$ models agent's results of perceiving the relevant dimensions. Finally, we are left with the two Markov kernels from the standard theory of Markov decision processes, the kernel of *decisions*, $D : X \to G$, where G is the semigroup of agent's actions, each of which is a Markov kernel $G : W \to W$, that acts upon the space of world states W. We use the letter A to denote the Markov kernel of actions of *implementation of a decision*, $A : G \times W \to W$. Figure 2 illustrates the diagram of these measurable spaces.

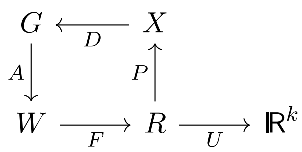


Figure 1: Illustration and definition of an universal model of a process.

By spending time on (optimizing) any dimension an agent can perceive, the agent does not spend time on other dimensions that affect his utility, hence to model self-aware agents, any dimension that is perceivable at least indirectly affects agent's utility and is hence relevant. Similarly, we assume a conscious agent is able to perceive his utility, so any relevant dimension of the space R is perceivable at least indirectly (possibly directly through some innovation, see [6]). We use this indirect relationship to justify that we do not discriminate between the space of relevant and of perceivable dimensions.

3 Von Neumann-Morgenstern and Csikszentmihalyi utility

We continue by examining two distinct sources of utility an agent can encounter in the process of learning.

The traditional, von Neumann-Morgenstern utility results from agent's activities producing commodities, goods that the agent can exchange for monetary or similar commodity in the case of an economic agent, or from agent's activities otherwise producing or securing resources needed for the agent to function. The result of those activities is primarily material, physical in the sense that when an agent is implementing its activities, it consumes and possibly produces those resources that support von Neumann-Morgenstern utility. The von Neumann-Morgenstern utility of these activities is hence monotone in the production of these commodities, which is in turn monotone in the time spent on these activities.

The Csikszentmihalyi utility we introduce as a measure of well-being, of how smoothly the agent is experiencing the activities being performed. The emotional state was originally self-reported and then taxonomized by psychologists into eight categories, which were attributed to regions of the unit square on dimensions skill (s) and challenge (c). We use this attribution to assign an emotional state to a given point (s, c) in the unit square, and we define the emotional utility to be the distance of the agents emotional state to the line of perfect balance between challenge and skill levels in the above average section, i.e. the line of $c = s \geq \frac{1}{2}$. This is not the only possible measure of emotional utility. For instance, it weights anxiety the same as relaxation, but the key difference between them is that relaxation has much attention left and is hence suboptimal due to wasted attention, which could be employed by rising challenge levels, whereas anxiety is a result of deficient skills and can only be improved through lowering challenge levels or increasing skill levels, both usually requiring (costly) external resources.

Using this approach, the total 2-dimensional utility of an agent is equal to

$$U = \sum_i (u_{v_i}(t_i, s_i, c_i) + u_{c_i}(t_i, s_i, c_i)), \tag{1}$$

where $u_{v_i}(t_i, s_i, c_i)$ is the von Neumann-Morgenstern utility and $u_{c_i}(t_i, s_i, c_i)$ is the Csikszentmihalyi utility of performing activity i for time t_i at skill level s_i and challenge level c_i.

In [2], the authors investigate the influence of learning ability, i.e. efficiency of time-attention spent learning on the emotional state of the agent. They find the following sequence of prevailing emotions that manifest by increasing the learning ability: at low learning ability, apathy prevails. By increasing the learning ability, it yields to boredom and increasing it even further, boredom yields to relaxation. All these appear first at low levels of passion (i.e. not increasing the challenge levels helps their onset). Further increase in learning ability manifests in control and finally flow, both rising from higher levels of passion (i.e. they are induced by rising challenges after success) and intermediate levels of perseverance (i.e. not dropping challenges too much after failure, but not being excessively persevere either). Thus we conclude that Csikszentmihalyi utility is for most combinations of passion and perseverance monotone in the learning ability. Note that Bokal and Steinbacher do not investigate interrelationships between emotional state and Csikszentmihalyi utility beyond stating the model.

In Figure 3, we extend these results by investigating the role of the number of tasks on the emotional state/utility of an agent. We see the effect of decreasing the number of tasks resembles the effect of increasing the learning ability. We fix learning ability $l = 0.1$ to represent a low learning ability in the setting of [2], but we vary the number of tasks. With $n = 100$, apathy prevails as in the TBDTAM setting. Decreasing down to $n = 80$ tasks, apathy is replaced by boredom, then at $n = 64$ boredom yields to relaxation, at $n = 53$, control appears and at $n = 35$, flow starts to take over, completely dominating at $n = 20$.

In this setting, the answer to the original question, Which problems to work on? has a straightforward answer: in the limited time, a researcher should work on the problems that yield sufficient von Neumann-Morgenstern utility to sustain the career the researcher desires. If there is attention left, it should be used on interesting problems where progress can be made. These will produce Csikszentmihalyi utility. Optimization would favor work on problems on which progress brings both career advancements (in form of grants, i.e. von Neumann-Morgenstern utility) as well as flow experience (i.e. Csikszentmihalyi utility); in absence of problems yielding both utilities, one would be faced with the solutions of Pareto frontier of the two criteria.

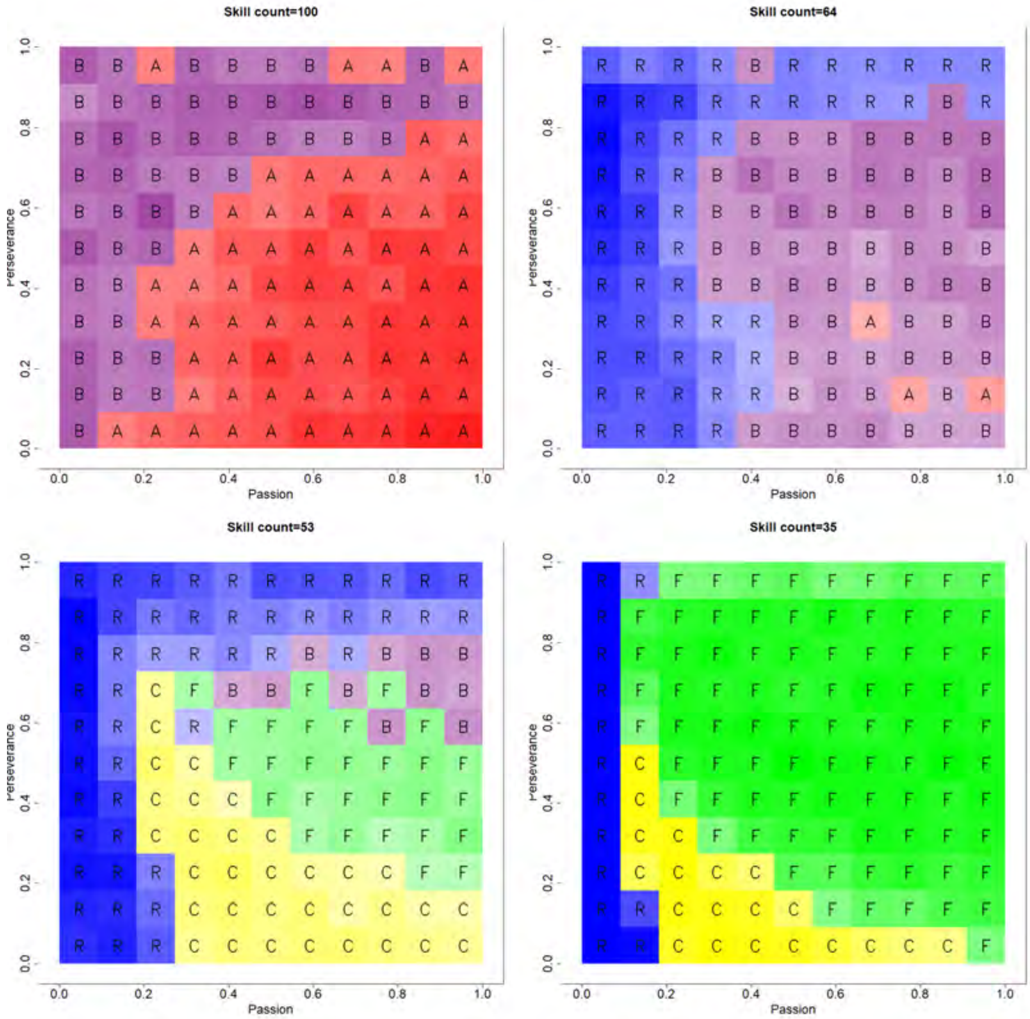


Figure 2: Phases of experience: emotional states depending on the number of tasks. Perseverance r (y-axis) and passion p (x-axis), emotional states as letters (low-F, control-C, relaxation-R, boredom-B, and apathy-A). Number of tasks: top-left $n = 100$; top-right $n = 64$; bottom left $n = 53$; bottom right $n = 35$. Learning ability in all cases, $l = 0.1$.

4 APPLICATIONS

We apply our model to interpret several less obvious situations that could be modelled by our extended model. Lysaker et al. describe four cases of narrative incoherence observed with patients of schizophrenia [11]. They observe that "each patient's story seems to be set in a world, in which he is unable to participate". They speculate that such state of apathy results from disturbed process of self-creation: the process they describe can directly be interpreted in our model as the activities of conversation between many self-positions, which is in schizophrenic patients disturbed and hence not generating emotional utility that would stimulate them to seek engagement in conversations between their self-positions. Their learning ability in this process is low, and the resulting emotional state is apathy.

A similar process explains the phenomenon observed by Calhoun, called behavioural sink, and it describes the collapse in behaviour resulted from overcrowding [3]. He experimented with rats and mice. Except for limited space, rats in the experiment were provided with optimal food, water, temperature, and hygienic conditions, which resulted in population rising from normal 32 to 50 to extreme 2200 and than declining to 0 due to abnormal behaviour. In the population prior to this collapse, nursing and care of the young was under heavy stress pressure of overcrowding, which led to infant mortality rate as high as 96 %. While the Csikszentmihalyi emotional model has not been generalized to animals, neurotransmitter models we introduce in the next illustration do offer similarity of mechanisms behind behavior and can be considered as the common source of behavior. As such, the model proposed offers mechanism insight into anthropological concept of Dunbar number [5], a cognitive limit on the number of people an individual can maintain stable social relationships with. Increasing the number of individuals decomposes this stability, resulting in apathy and conditioning for anomalous behaviors, a process similar to what we observed in schizophrenic patients.

The above two applications exhibit instances of low Csikszentmihalyi utility, but optimizing for high perception of Csikszentmihalyi utility can lead to problems as well. According to a review by [8], behavioral addictions manifest as [8] the failure to resist an impulse, drive, or temptation to perform an act that is harmful to the person or to others. A growing body of literature implicates multiple neurotransmitter systems (e.g., serotonergic, dopaminergic, noradrenergic, opioidergic) in the pathophysiology of behavioral addictions and substance use disorders. Bibliography on behavioral addiction provides insight into mechanisms of deriving Csikszentmihalyi utility. It is hereby linked to neurotransmitter systems whose levels we interpret as commodities in the space of relevant dimensions R, which are necessary for experiencing emotional utility, but not sufficient. Somewhat simplified, we can hence model behavioral addiction as spending time in activities increasing levels of neurotransmitters that contribute to perception of Csikszentmihalyi utility, yet do not actually contribute to it.

A similar, yet less damaging instance of maximizing perception of Csikszentmihalyi utility are online social bubbles [15]. Nikolov et al. state [15]: It is common for users themselves to adopt filters in their online behavior, whether they do this consciously or not. For example, on social platforms such as Facebook, a large portion of users are exposed to news shared by their friends [1, 13]. Because of the limited time and attention people possess and the large popularity of online social networks, the discovery of information is being transformed from an individual to a social endeavor. While the tendency to selectively expose ourselves to the opinion of like-minded people was present in the pre-digital world [9, 14], the ease with which we can find, follow, and focus on such people and exclude others in the online world may enhance this tendency. Regardless of whether biases in information exposure are stronger today versus in the pre-digital era, the traces of online behavior provide a valuable opportunity to quantify such biases.

Also in this case, limited time availability combined with perceived optimizing of emotional utility of users results in their selecting for a limited set of activities that can maintain their emotional utility.

5 Conclusion

We introduce a model of optimizing time allocation of a percepive learning agent who can optimize von Neumann-Morgenstern and Csikszentmihalyi utilities. We illustrate this model on several instances, where in some, Csikszentmihalyi utility is forcedly maintained at low levels, and in others, its perception is optimized. Our model indicates that these phenomena may have the common root in individual's desire to feel good within the activities they perform given the limited attention they can distribute to them. In the light of a parallel contribution on mathematical models of innovation [6], this sheds a new, possibly discouraging light on the predominating societal narrative of progress, and provides incentives for the discussion, what progress we want to devote our attention and other resources to.

References

[1] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The Role of Social Networks in Information. Proceedings of the 21st international conference on world wide web, WWW12 (pp. 512–528). New York: ACM.

[2] Bokal, D. and Steinbacher, M. (2019). Phases of Psychologically Optimal Learning Experience: Task-based Time Allocation Model. Central European Journal of Operations Research.

[3] Calhoun, J., B. (1962). Population Density and Social Pathology. Scientific American, 206(2): 139–149.

[4] Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience.

[5] Dunbar, R. (2010). How Many Friends Does One Person Need?: Dunbar's Number and Other Evolutionary Quirks. Harvard University Press.

[6] Fic, P., Bokal, D. (2019). Innovative Veristic Perceptions Do Have a Chance: An Instance of Artificial Techonogical Valley of Death. Submitted to SOR.

[7] Genesis 1,2. (2019). Bible Study Tools.
https://www.biblestudytools.com/esv/genesis/2.html [Accessed 25/05/2019].

[8] Grant, J. E., Potenza, M. N., Weinstein, A., and Gorelick, D. A. (2010). Introduction to behavioral addictions. The American Journal of Drug and Alcohol Abuse, 36(5): 233–241.

[9] Hart, W., Albarracn, D., Eagly, A.H., Brechan, I., Lindberg, M.J., and Merrill, L. (2009). Feeling Validated versus Being Correct: A Meta-analysis of Selective Exposure to Information. Psychological Bulletin, 135(4): 555-588.

[10] Hoffman, D. D., Singh, M., and Prakash, C. (2015). The Interface Theory of Perception. Psychonomic Bulletin, 22(6): 1480–1506.

[11] Lysaker, P. H., Wickett, A., M., Wilke, N. and Lysaker, J. (2005). Narrative Incoherence in Schizophrenia: The Absent Agent-Protagonist and the Collapse of Internal Dialogue. American Journal of Psychotherapy, 57(2): 153–166.

[12] Mark, J. T., Marion, B. B. and Hoffman, D. D. (2010). Natural selection and veridical perceptions. Journal of Theoretical Biology, 266: 504-515.

[13] Matsa, K., E., and Mitchell, A. (2014). 8 Key Takeaways about Social Media and News.
https://www.journalism.org/2014/03/26/8-key-takeaways-about-social-media-and-news/
[Accessed 25/05/2019].

[14] Kastenmuller, A., Greitemeyer, T., Jonas, E., Fischer, P. and Frey, D. (2010). Selective Exposure: The Impact of Collectivism and Individualism. British Journal of Social Psychology, 49(4): 745-763.

[15] Nikolov, D., Olivoera, D., F., M., Flammini, A. and Menczer, F. (2015). Meassuring Online Social Bubbles. PeerJ Computer Science, 38(1).

[16] Open Problem Garden (2007). Computer Science Institute of Charles University
http://www.openproblemgarden.org/ [Accessed 25/05/2019].

[17] Smole, A., Jagrič, T., Bokal, D. (2019). Principal-Leader-Follower Model with Internal Signal. Submitted to SOR.

[18] Tertinek, Š. and Bokal, D. (2019). Comparing Time Usage Model. In Šprajc, P., Podpregar, I., Maletič, D. and Radovanovič, M. (Eds.). 38th International Conference on Organizational Science Development: Ecosystem of Organizations in the Digital Age (pp. 1035–1045). Maribor : Univerzitetna založba Univerze v Mariboru.

INNOVATIVE VERISTIC PERCEPTIONS DO HAVE A CHANCE: AN INSTANCE OF ARTIFICIAL TECHNOLOGICAL VALLEY OF DEATH

Petra Fic,
Drago Bokal,
Faculty of natural sciences and mathematics, University of Maribor,
Koroka cesta 160, Maribor, Slovenia,
petra.ficc@gmail.com, drago.bokal@um.si

Abstract: In the paper, we build upon territorial games introduced by Mark, Marion, and Hoffman in 2010, and extend four of their territory perception and selection strategies with two novel ones that together constitute a model of technological readiness levels valley of death. We conjecture that whenever utility of a resource is not monotonous in the amount of that resource, the technological valley of death emerges. While development of the science behind these models is in its infancy, modelling and understanding the phenomenon may shed light into progress and related phenomena in society.

Keywords: evolution, perception, utility, Monte Carlo simulation, game theory.

1 Introduction

Theory of perceptions is exploring the relationship between our perception and the environment [3, 11, 12]. Several results from 2010 onwards state that veristic perceptions of the world have little if any chance of surviving evolutionary competition, which favors utilitaristic interface-like perceptions, optimized to perceive the utility derived from the actions in a given environment [4]. Subsequent proof of the Invention of symmetry theorem [5] stated that an observer who perceives the symmetry of the world, may simply be deceived by the symmetry emanating from the compositum of perceptions of the utility of results of actions upon the world, which would exhibit the stated structure, but would be just an optimal interface to the true structure of the world that could have a completely different structure. Scientific community responded grimly to this pessimistic view [6]. In this contribution, we report about further experiments in similar evolutionary contexts that have justified that attitude: the veristic perceptions do have a chance for evolutionary survival, should they prove innovative, ie. able to apply their veristic understanding of the world into an innovation that outperforms the utilitaristic perceptions of the world. However, they are outcompeted by innovative utilitaristic perceptions who take advantage of both the innovation as well as true perception of its utility.

The competition of veristic with utilitaristic perceptions is hence an abstract mathematical model behind the phenomenon of technology valley of death, which was perceived in the documents of EU commission [2] based upon knowledge progress scale introduced by NASA [7]. It is closely associated with product development [9]. Our model exhibits technology valley of death as an emergent evolutionary phenomenon in evolutionary environments in which the agents' model of the world is distinct from the true structure of the world, the utilities of resources are non-monotonous in the amount of resources, and the agents are evolving their perceptions (ie. their model), their decision process, and their actions, gaining evolutionary advantage over those who either perceive less realistically (thus unable to innovate) or less utilitaristically (thus unable to maximize evolutionary utility of the innovations).

2 Related work: territorial games

Territorial games were introduced by Mark et al [8]. We reproduce their mathematical model and adapt it to new perceptual strategies. In a territorial game, p players compete in pairs, they choose between t territories with k resources each. Each resource takes a discrete value in the set $V = \{1, 2, ..., m\}$. Let r_T be the vector of resources in territory T. The utility of territory T is defined as

$$u(T) = U(r_T) = \sum_{i=1}^{k} U_i(r_{T,i}), \tag{1}$$

where U_i is the utility of contribution of resource $r_{T,i}$ in territory T. In our investigations, it is either monotonous linear or Gaussian, as we specify later.

From interface theory of perception combined with utility theory, we take the universal model of an agent – a process, and from game theory, we take the decision tree to be the structure of decisions of each agent. The agent acting in the world W is defined with a 5-tuple $A = (X, G, P, D, M)$, where X is the space of agent's possible perceptions, G is a semigroup of agent's actions. Similarly as W, both X and G are measurable spaces. Then P, D, M are agent's perception, decision, and utility operators. In highest generality, they are modelled as Markov kernels; $P : W \to X$, $D : X \to G$, $M : W \to \mathbb{R}$. Actions $AinG$ are defined as $A : W \to W$, and they are Markov kernels as well.

For the rest of this section, we assume that $r = 1$. Mark et al. [8] introduce four agent strategies that could participate in one of two games that differ in the definition of utility. Utility is mapping $M_v : W \to \mathbb{R}$ defined as either identity in the first game or Gaussian function in the second game:

$$M_v(w) = \frac{1}{\sqrt{40\pi}} e^{-\frac{1}{2}\frac{(x-50)^2}{400}} \tag{2}$$

We apply the following proposition to determine the long-term outcome of competition between two agent strategies (A) and (B).

Proposition 1 ([1]) *Let a be the expected utility obtained by (A) competing with (A), b the expected utility of (A) competing with (B), c the expected utility of (B) competing with (A) and d the expected utility of (B) competing with (B). There are four possible long term outcomes of the evolutionary competition of the two species. If $a \geq c$ and $b \geq d$ and at least one inequality is strict, then (A) prevails and (B) goes extinct. If $c \geq a$ and $d \geq b$ and at least one inequality is strict, then (B) prevails and (A) perishes. If $a < c$ and $b > d$, they stably coexist. If $a > c$ and $b < d$, they are bistable, ie. each is asymptotically stable and which depends on the initial conditions. They are neutral, if $a = c$ and $b = d$, ie. their prevalence changes randomly.*

Furthermore, the following Lemma implicitly follows from Mark et al. [8].

Lemma 2 *Let the interaction of (A) and (B) be as described, and let p be the probability that (A) moves first when interacting with (B). Furthermore, let $\Gamma(i, X, Y)$ denote the utility of the i-th player, where $X \in \{A, B\}$ is the strategy that chooses first, and $Y \in \{A, B\}$ is the strategy that chooses second. Then, $a = \frac{1}{2}(E(\Gamma(1, A, A)) + E(\Gamma(2, A, A)))$, $b = pE(\Gamma(1, A, B)) + (1 - p)E(\Gamma(2, A, B))$, $c = pE(\Gamma(2, A, B)) + (1 - p)E(\Gamma(1, B, A))$, and $d = \frac{1}{2}(E(\Gamma(1, B, B)) + E(\Gamma(2, B, B)))$.*

A naïve realist perception faithfully and exhaustively resembles reality. On the other hand, semi-veristic strategy, the critical realist (CRn), does not perceive the true amount of resource, but categorizes it into n categories preserving the ordering. Mark et al. [8] call this a *nCat* agent. Critical realist claims that perception faithfully resembles only some aspect of reality. A *nCat* utilitaristic interface pereption (IFn) is used by Hoffman et al. [5] to argue that perceptions need not, and in general do not, resemble any aspect of reality.

In Table 1 are given details of mathematical models of a naive realist, *nCat* critical realist and *nCat* interface strategy reproduced from Mark et al. [8].

Elt	**Naive realist**	**CRn**	**IFn**
W	$\{1,\dots,m\}^t$	$\{1,\dots,m\}^t$	$\{1,\dots,m\}^t$
X	W	$\{0,\dots,n-1\}^t$	$\{0,\dots,n-1\}^t$
G	$\{1,\dots,t\}$	$\{1,\dots,t\}$	$\{1,\dots,t\}$
P	w	$\begin{cases} 0; & \text{if } 1 \leq w \leq \beta_1, \\ 1; & \text{if } \beta_1 < w \leq \beta_2, \\ ... & \\ n-1; & \text{if } \beta_{n-1} < w \leq m, \end{cases}$	$\begin{cases} 0; & \text{if } 1 \leq w \leq \beta_1, \\ 1; & \text{if } \beta_1 < w \leq \beta_2, \\ ... & \\ n-1; & \text{if } \beta_{n-1} < w \leq m, \end{cases}$
A	$u' = u + U(r_i), r'_i = 0$	$u' = u + U(r_i), r'_i = 0$	$u' = u + U(r), r'_i = 0$
D	$\text{argmax}_{p=P_n(w),i\in G_n} p[i]$	$0 < n-1 < ... < 1$	$argmax_{p=P_i(w),i\in G_i} p[i]$
M	$\{1,\dots,m\} \times \mathbb{R}$	$\{1,\dots,m\} \times \mathbb{R}$	$\{1,\dots,m\} \times \mathbb{R}$

Table 1: Mathematical models of a naive realist, CRn and IFn.

Mark et al. [8] first explore a competition between naïve realist and critical realist CR2. To percieve and handle more data, more energy is needed. Therefore, veristic strategy uses more energy to reach its goal. Furthermore, they explored a competition between $3Cat$ critical realist (CR3) and $3Cat$ interface strategy (IF3).

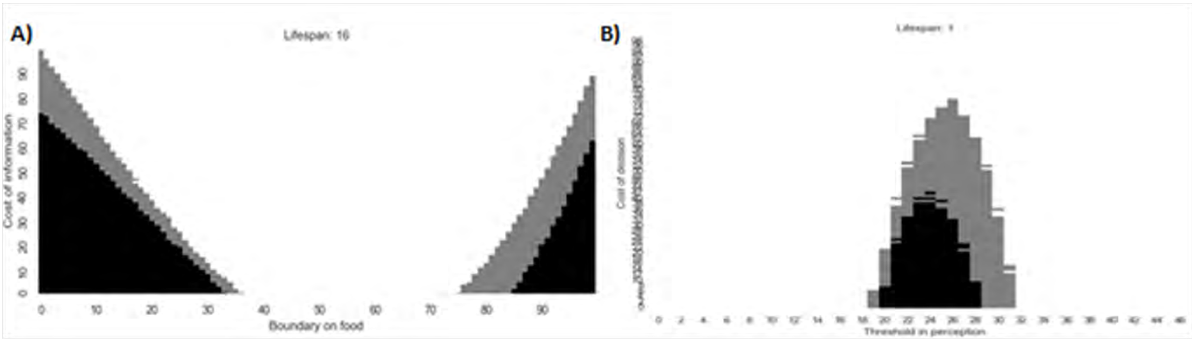


Figure 1: A) The results of an evolutionary game between naïve realist and critical realist with a single resource. B) The results of an evolutionary game between CR3 and IF3 with a single resource. [Source: A) Reproduced from [8], B) Own.]

Figure 1.A) shows how the cost of information and a threshold on food used by a critical realist in perceiving the world affects the competition of critical realist and naïve realist. A critical realist drives a naïve realist to extinction (white color) for most of the values of threshold β. Only for low cost of information, naïve realist drives critical realist to extinction (black color), and only when the threshold β is chosen poorly.

Mark et al. [8] also explore a 3-way competition between veristic, CR3 and IF3 strategies in the setting of non-monotonous Gaussian utility. They show that either veristic or IF3 strategy outperform the other two, depending on the cost of information. We reprodce their results only competing CR3 with IF3, but varying the width of the central interval that CR3 perceives as intermediate and imposing some cost on perceiving the true utility. On Figure 1.B) we can see the results: the augmented critical realist has a slight chance with narrow enough intermediate interval, (the scale of the y axis is 1:500 compared to Figure 1.A). The interface strategy drives the critical realist to extinction (white color) for most of the interval width. Only for narrow interval and very low cost of information, critical realist drives interface strategy to extinction (black color), although adjusting the parameters of the IF3 perceptive strategy may possibly compensate for that advantage.

3 New models: innovative perceptions with storage

Our contribution to the above experiments using Gaussian utility consists of two new instances of agents that have the possibility of storing surplus resources, ie. effectively they introduced a new dimension into the world. The basis for the introduction of such strategies is evolutionary: an IF3 mutation, which otherwise perceives the quantity of the resource in a utilitarian manner, can develop the possibility of storing excessive amounts of the resource. The agent using such strategy needs to decide, whether to consume the resource or store it. For such a decision, the agent needs to distinguish between too much and too little of the resource, and needs a credible perception of its amount: it is evolutionarily motivated for veristic perception. In our model, the amount of stored resource (r_s) takes value in the set $V = \{1, 2, ..., m\}$. In our model, the critical realist with storage perceives the amount of resource in the storage as 2Cat critical realist would perceive another territory, ie. there is a threshold value r^*, such that the storage is seemed empty for $r_s < r^*$, or full otherwise. In each case, the CR3SR2 agent chooses the territory with intermediate amounts of the resource first, so as to take advantage of the inherent utility of the territory first. If there is none and the storage is empty, the agent prefers the territory with too much of the resource; if the storage is perceived full, the agent prefers territory with too little of the resource.

For the storage to be fully exploited, we need to introduce another parameter into the model – lifespan l tracks the number of interactions between the two competing strategies, ie. how often can two agents compete for a territory and apply the advantage of storing excessive or supplementing insufficient amount of the resource.

In addition to CR3SR2, we also introduce a 3Cat interface strategy with storage, IF3S. This strategy has storage, but perceives the exact utility of the territory given the exact amount of resource available in the storage. The utility is perceived in three categories as with IF3, and the decision operator is the same. We hence term this to be an interface strategy, although it needs veristic inputs to produce utilitaristic perception: both territory amount and storage amount of the resource must be accounted for veristically to dsitinguish between the three categories of utility in 3Cat perceptions.

Details of our mathematical model of a $3Cat$ critical realist with storage (CR3SR2) and $3Cat$ interface strategy with storage (IF3S) are given in Table 2.

Elt	**CR3SR2**	**IF3S**
W	$\{1,\dots,m\}^{t+1}$	$\{1,\dots,m\}^{t+1}$
X	$\{0,\dots,2\}^t \times \{0,1\}$	$\{0,\dots,2\}^t \times \{0,1\}$
G	$\{1,\dots,t\}$	$\{1,\dots,t\}$
P	$P_c(t) = \begin{cases} 0; & \text{if } 1 \le t \le \beta_c, \\ 1; & \text{if } \beta_c < t \le \gamma_c, \\ 2; & \text{if } \gamma_c < t \le m. \end{cases}$	$P_u(t,s) = \begin{cases} 0; & \text{if } 0 \le U_c(t,s) \le \beta_u, \\ 1; & \text{if } \beta_u < U_c(t,s) \le \gamma_u, \\ 2; & \text{if } \gamma_u < U_c(t,s). \end{cases}$
	$P_s(s) = \begin{cases} 0; & \text{if } 1 \le s \le \beta_s, \\ 1; & \text{if } \beta_s < s \le m, \end{cases}$	
A	$s' = \begin{cases} \min(s_{max}, r_i + s - r^*); & \text{if } r_i + s > r^*, \\ \max(0, r_i + s - r^*); & \text{if } r_i + s \le r^*, \end{cases}$	$s' = \begin{cases} \min(s_{max}, r_i + s - r^*); & \text{if } r_i + s > r^*, \\ \max(0, r_i + s - r^*); & \text{if } r_i + s \le r^*, \end{cases}$
	$u' = u + \begin{cases} U(r^*); & \text{if } r_i + s > r^*, \\ U(r_i + s); & \text{if } r_i + s \le r^*, \end{cases}, r'_i = 0$	$u' = u + \begin{cases} U(r^*); & \text{if } r_i + s > r^*, \\ U(r_i + s); & \text{if } r_i + s \le r^*, \end{cases}, r'_i = 0$
D	$\begin{cases} 0 < 1 < 2; & \text{if } 1 \le s \le \beta_s, \\ 0 < 2 < 1; & \text{if } \beta_s < s \le m, \end{cases}$	$0 < 1 < 2$
M	$\{1,\dots,m\} \times \{1,\dots,m\} \times \mathbb{R}$	$\{1,\dots,m\} \times \{1,\dots,m\} \times \mathbb{R}$

Table 2: Mathematical models of CR3SR2 and IF3S.

4 Results: IF3 perishes against CR3SR2, which perishes against IF3S

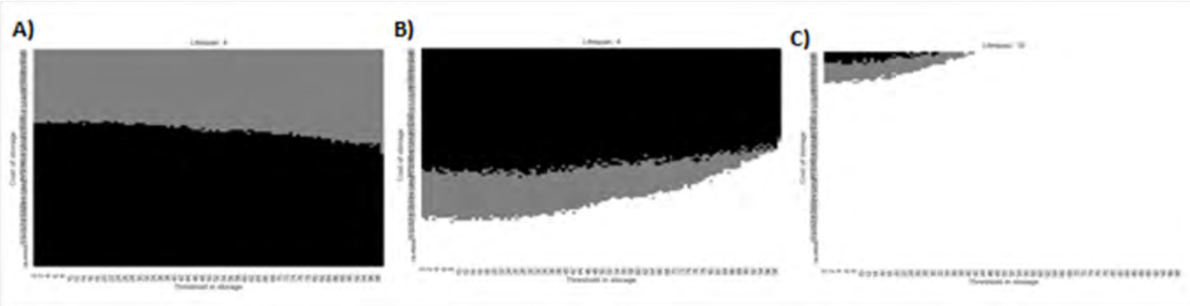


Figure 2: A) The results of an evolutionary game between CR3SR2 and IF3 with a single resource. B) The results of an evolutionary game between CR3SR2 and IF3S2 With a single resource for lifespan 4 years. C) The results of an evolutionary game between CR3SR2 and IF3S With a single resource for lifespan 16 years. [Source: Own.]

We competed IF3 against CR3SR2 for lifetimes $l \in \{1, 2, \ldots, 16, 32, 64, 128, 250, 500, 1000\}$, at which value the results converged and the last two images showed no further change. Each was competing at discrete values of $r_s^* \in \{1, 11, 21, 31, \ldots, 91, 101\}$, with the last value interpreted as always perceiving empty storage and desiring to claim territory with excessive amounts of the resource. For each of these values, we imposed a cost on the storage of $CR3SR2$ competing with $IF3$ with values $c_s \in \{1, 11, 21, 31, \ldots, 91, 101\}$, which was comparable to the maximum amount of utility M_v, taking values in the interval $[0.0039, 0.0892]$. When CR3SR2 competed with IF3S, we applied the same values for r_s^*, but imposed additional costs with the same values on $IF3S$. We assumed that IF3S would spend more energy for the additional computations in the perception mechanism, but the cost of storage would be the same for both strategies.

Figure 2.A) shows the results for given values of r_s^* (x axis) and c_s (y axis) for lifespan of $l = 4$ interactions, where the benefit of the storage is the highest. We see that CR3SR2 drives IF3 to extinction (black color) for all examined low values of storage cost and coexist for all the higher. Although CR3SR2 is slower at picking the territory, storage gives it the advantage over the utilitaristic IF3 when cost of storage is comparable to the utility of a terrain. For higher lifespans, the advantage of the storage decays and converges with $l = 1000$, but for lowest three values of c_s, CR3SR2 still dominates IF3 for all values of r_s^*, except for 101.

In Figure 2.B), we show the results of the competition between CR3SR2 and its mutation IF3S, which is able to precisely perceive the utility of combined amount of the resource in its storage and a possible territory. We assume this innovative storage imposes some additional cost to the organism. The figure shows success of the innovative utilitaristic interface perception over the original critical realist, which is emphasized with higher number of average interactions between the organisms: in comparison to rather comparable success of both strategies at four interactions in Figure 2.B), CR3SR2 barely has any viable parameters at 16 interactions in Figure 2.C), and at numbers higher than 23, the benefits of the storage outweigh any of the tested values of costs.

5 Conclusions and further research

Mark et al [8] in their research discovered that more realistic perceptions are not necessarily more successful: Natural selection can drive realistic perceptions to extinction when competing with perceptions that use specific interfaces that simplify and adapt the truth in order to better represent the utility of what is being perceived. However, we created conditions in which natural selection gives priority to (simplified) veristic perceptions. We defined strategies that store excessive amounts of the resource and studied an evolutionary game between the strategies IF3 and CR3SR2. Given the reasonably low cost of storage of the resource, innovative simplified veristic perceptions may displace the interface perceptions, even if the latter have the advantage of the first choice of the territory. Furhermore, we examined what happens when veristic strategy with storage and the interface strategy with storage compete. Our simulations show that interface strategy with storage drives the critical realist with storage to extinction.

Using these illustrative examples, we conclude with some open problems. First, we (vaguely) define four stategies A, B, C, D to constitute the valley of death, if A perishes against B, which perishes against C, which perishes against D. In addition, C uses the perceptions of A to support an innovation that cannot be supported using perceptions of B, and D uses the perceptions of B to perceive the true utility of innovation of A.

For the mathematical direction of the research, we conjecture that for each non-monotonous utility function, there exist four strategies that exploit the non-monotonicity of that function to exhibit the technological valley of death.

For microeconomic direction of further research, games with incomplete information could be defined as games where the state of the world is inaccurately perceived by both the agents. Perception strategies could be introduced into those games so as to either heuristically perceive the expected utility of the situation (modelling heuristics within the current approach to these games), or to add additional information about the state of the world (modelling cheating at such games, or research in market situations). In these cases, such perceptive games could yield improved understanding of the role of marketing, marketing research, and advertising.

For decision science and management direction of the research, the role of perceptions vs. decisions could be further explored in the stated setting. Perceptions play significant role in information systems, linked to data acquisition, data quality, data presentation. Decisions based on that data are significsnt in corporate performance management systems. The models presented here could be used for fundamental research in those settings.

References

[1] Armao, J.J. (2009). Evolutionary Game Dynamics, Cooperation and Costly Punishment. Harvard University.

[2] G. (2015). Technology Readiness Levels. European Commission. https://ec.europa.eu/research/participants/data/ref/h2020/wp/20142015/annexes/h2020wp1415 annexgtrlen.pdf [Accessed 10/04/2019].

[3] Heyer, D. and Mausfeld, R. (2002). Perception and the physical world: Psychological and philosophical issues in perception. New York: Wiley.

[4] Hoffman, D.D. (2012.) The interface theory of perception. Object Categorization: Computer and Human Vision Perspective, 148165.

[5] Hoffman, D.D., Singh, M. and Prakash, C. (2015). The interface theory of perception. Psychon Bull Rev, 22: 1480-1506.

[6] Hoffman, D.D., Singh, M. and Prakash, C. (2015). Probing the interface theory of perception: Reply to commentaries. Psychobn Bull Rev, 22: 1551 - 1576.

[7] Mankins, J.C. (1995). Technology readiness levels. White Paper, 6.

[8] Mark, J.T., Marion, B.B. and Hoffman, D.D. (2010) Natural selection and veridical perceptions. Journal of Theoretical Biology, 266: 504515.

[9] Markham, S.K., Ward, S.J., AimanSmith, L. and Kingon, A.I. (2010). The Valley of Death as Context for Role Theory in Product Innovation. The journal of product innovation management. 27(3): 402-417.

[10] Nowak, M.A. and Sigmund, K. (2004) Evolutionary dynamics of biological games. Science, 303: 793799.

[11] Radnitzky, G. and Bartley, W. W. (1987). Evolutionary Epistemology, Theory of Rationality, and the Sociology of Knowledge. LaSalle, Illinois: Open Court.

[12] Schwartz, R. (2004). Perception. Malden, MA: Blackwell Publishing.

DEMAND POINT AGGREGATION IN URBAN EMERGENCY MEDICAL SERVICE: A CASE STUDY FROM SLOVAKIA

Ľudmila Jánošíková
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 1, 010 26 Žilina, Slovak Republic
E-mail: Ludmila.Janosikova@fri.uniza.sk

Peter Jankovič
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 1, 010 26 Žilina, Slovak Republic
E-mail: Peter.Jankovic @fri.uniza.sk

Stanislav Mikolajčík
University of Žilina, Faculty of Management Science and Informatics
Univerzitná 1, 010 26 Žilina, Slovak Republic
E-mail: Stanislav.Mikolajcik@stud.uniza.sk

Abstract: The paper presents a new approach to modelling demand for emergency medical service in an urban area. We consider centres of the streets as aggregated demand points. Demand volume is the population living in a given street. Such detailed model enables to find better distribution of ambulances across the area. A case study of the city of Prešov demonstrates the approach. The *Q*-coverage *p*-median model is used to optimize ambulance locations. The performance of the original and proposed sets of locations is evaluated using computer simulation and real historical data on ambulance trips.

Keywords: emergency medical service, ambulance location, *p*-median problem, demand point aggregation, computer simulation

1 INTRODUCTION

Planning of emergency medical service (EMS) comprises all three levels of decision making. At the strategic level, the amount of ambulance stations is determined, as well as their distribution across a given area. In Slovakia, the government is in charge of these strategic decisions for the whole country. It decides, in which towns and villages the stations should be established. The term "station" has an organizational rather than a physical meaning. Currently, there are 273 stations distributed in 211 towns, city districts and villages. Large towns have multiple stations. Every station is a base for one ambulance.

At the tactical level, providers of urgent health care decide about precise addresses of those stations which they operate. They choose a proper building with a garage for the ambulance and a room where the crew waits for the rescue trip. Currently, 12 providers of ground emergency service operate in Slovakia. The largest provider is Falck Záchranná a.s. that operates 107 stations located mainly in the middle and north-eastern part of the country. If one company operates multiple stations in a town, it may decide to locate multiple stations at the same address, so the stations share the same building. For example, this is the case of Prešov, where Záchranná služba Košice operates 5 stations located at 2 addresses.

The operational level includes the dispatching of ambulances which is controlled by operators in operation centres.

This paper deals with tactical decisions about location of EMS stations in an urban area. We respect the amount of the stations in the area determined at the strategic level but look for their best locations inside the town. We do not consider the costs associated with the opening of the stations or their relocation from current positions. Instead, we adopt a patient's point of view and concentrate on the only optimization criterion that reduces the response time as much as possible.

We propose a new approach to modelling demand for emergency medical service in an urban area. We aggregate private residences in a street to one demand point and use such aggregated demand as the input for the weighted Q-coverage p-median model to optimize the locations of the stations. We use computer simulation to evaluate the proposed and current locations.

2 MODELLING DEMAND

In the literature on ambulance location, the demand for EMS is usually derived from the number of inhabitants [6, 8] or from historical data on EMS calls [1, 3, 7, 9]. The number of demand points may be quite large since every private residence may be a demand point. It is impossible to include every demand point in the location model. That is why various levels of aggregation are applied. Spatial distribution of demand is modelled either by using city zones [6], census areas [7], or by covering the urban area by a grid [1, 3, 8, 9], and aggregating the demand in the zone to its centroid. We propose a new approach to modelling the spatial distribution of demand. Thanks to open data provided by the local administration in the city of Prešov [https://egov.presov.sk/], we have the number of inhabitants in every street in the town. That is why we can use the centre of the street as an aggregated demand point. The volume of the demand is the number of inhabitants living in the particular street.

The street is represented by its centre. To calculate the centre of the street, first we need to determine the centroid (geometric centre) of the street. The centroid of the street is the arithmetic mean position of the extreme points of the polylines that form the street, i.e., it is a centroid of the envelope covering the street. The centre of the street is then the point in the polylines that is the nearest to the centroid (Fig. 1).

3 LOCATION MODEL

One of frequently used ambulance location models is the weighted p-median model that seeks for the optimal location of p facilities to minimize a demand-weighted time to access the population residing at the nodes of the network. We use a modified version of the basic model, which requires the assignment of a demand point to multiple facilities. Let us describe the model in a more formal way.

Street centres form the set I of aggregated demand points. The demand in point i denoted by b_i is the number of people living in street i. Street centres are also candidate locations, where ambulance stations can be placed. Further, t_{ij} stands for the shortest travel time of an ambulance from node $i \in I$ to node $j \in J$ of the underlying street network. The weighted p-median problem is to locate exactly p stations in points from the set I in order to minimize the total travel time needed to reach all potential patients. The amount of stations p is defined at the strategic level.

The decision on opening a station in candidate location $i \in I$ is modelled by a binary variable y_i. The value $y_i = 1$ indicates that a station will be opened in point i. Furthermore, we need to decide which stations will serve every street. In order to cope with possible unavailability of the nearest ambulance in the moment when it is needed we consider that a street can be served from two stations. It means that every street belongs to the catchment area of two stations: the nearest and the backup station. The catchment area is created using assignment binary variables. If variable x_{ij} takes value 1, then street j is in the catchment area of station i.

Figure 1: Centroid and centre of the Karpatská street

In general, the problem that requires the assignment of a customer to multiple facilities is called the Q-coverage p-median problem [5]. The model can be formulated as follows:

$$\text{minimize} \quad \sum_{i \in I} \sum_{j \in I} t_{ij} b_j x_{ij} \tag{1}$$

$$\text{subject to} \quad \sum_{i \in I} x_{ij} = Q \qquad \text{for } j \in I \tag{2}$$

$$x_{ij} \leq y_i \qquad \text{for } i \in I,\ j \in I \tag{3}$$

$$\sum_{i \in I} y_i = p \tag{4}$$

$$x_{ij}, y_i \in \{0,1\} \qquad \text{for } i \in I,\ j \in I \tag{5}$$

The objective function (1) minimizes the total travel time between stations and potential patients. This criterion reflects the main objective of the EMS system: to provide pre-hospital care as soon as possible. Constraints (2) ensure that every street j will be served and covered by exactly Q stations. Constraints (3) ensure that if a street j is assigned to a point i, then a station will be opened in the point i. Constraint (4) defines the total number of stations in the town. The remaining obligatory constraints (5) specify the definition domains of the variables.

4 RESULTS AND DISCUSSION

With the population of 85 743 (Dec 2018), Prešov is the third largest town in Slovakia. There are 356 streets in the town with the population ranging from 0 to 5246. The street network was taken from the OpenStreetMap database [https://www.openstreetmap.org], which is a freely available source of geographical data. All directional, turn and speed regulations are included within the road network model. For roads without speed limit, average travel speed of ambulances with respect to a given road category was applied to calculate the matrix $\{t_{ij}\}$ of the shortest travel times between street centres. Even if the *p*-median problem is *NP*-hard, the properties of a real road network [2] enable to solve the practical-size instances exactly in a very short time. We solved the *Q*-coverage *p*-median location model using the general solver Xpress-MP running on a personal computer equipped with the Intel Core i7 processor with 1.60 GHz and 8 GB RAM. The computational time was about 20 sec for both values of *Q*.

The locations of the stations proposed by the model with parameters $Q = 1$ and $Q = 2$ are displayed in Fig. 2, together with the current locations. The positions of two stations are the same for both parameter values. In general, the stations in the basic *p*-median model (with $Q = 1$) are more dispersed across the town area than the stations with backup coverage.

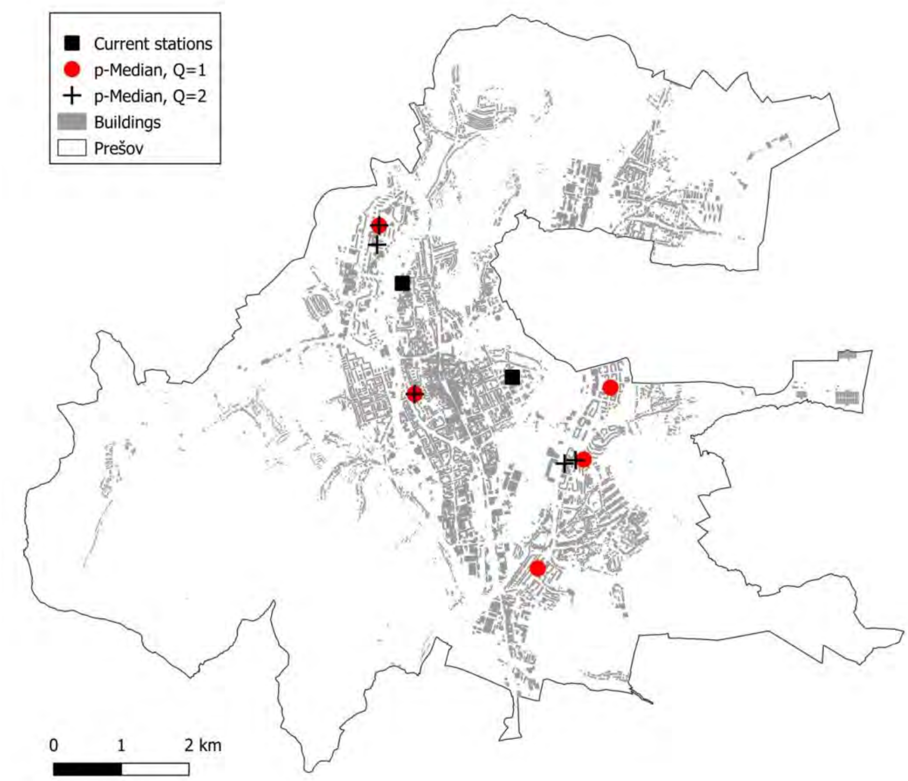


Figure 2: Current and optimal locations of stations in Prešov

The proposed and current sets of locations were evaluated using a computer simulation model described in [4]. The agent-based simulation model was implemented in AnyLogic simulation software and calibrated using a sample of EMS trips provided us by Falck Záchranná a.s.

Computer simulation enables to estimate the impact of the proposed changes in the current system. The main performance indicators evaluated by simulation are the following:

1. average response time, since it has been monitored by the Operation Centre of the EMS of the Slovak Republic;
2. percentage of calls responded to within 15 min, because a 15-minute response time is regarded as standard in Slovakia;
3. average response time for highest priority calls and percentage of these calls responded to within 8 min.

The results of the simulation study for the city of Prešov are in Table 1. Major performance indicators are compared for the current ambulance locations and the optimized locations of 5 ambulances. The best values of individual indicators are displayed in bold.

Table 1: Quality indicators for different ambulance locations in the city of Prešov

Indicator	***Current locations***	***p-median Q = 1***	***p-median Q = 2***
Average response time [min]	6:07	5:31	**5:30**
% of calls responded to within 15 min	95.1	**95.2**	95.0
Average response time for highest priority calls [min]	5:53	**4:59**	**4:59**
% of highest priority calls responded to within 8 min	88.6	89.8	**90.1**

As can be seen, the *p*-median model results in better accessibility of EMS. The average response time for all patients is reduced by about half a minute in both models. The average response time for the most critical patients is reduced even more by almost one minute. It is a significant improvement because one minute can decide about life of critical patients who are in life-threatening conditions. The amount of highest priority calls responded to within the time threshold of 8 minutes is increased in both modes. The 15-minute coverage is improved only in the basic *p*-median model. The 2-coverage *p*-median model outperforms the simple *p*-median model in two of four indicators. However, it is not possible to decide which of the two models is better considering only the town itself because the distribution of the stations in the town also affects surrounding villages. That is why it is reasonable to inspect the impact of the new station locations on the whole region.

Prešov Region is one of the eight Slovak administrative regions. It is located in the north-eastern part of the country. The city of Prešov is its administrative centre. Table 2 presents the impact of optimization of ambulance locations in Prešov on EMS provision in the Prešov Region. We can observe the improvement also at the regional level, although the changes are not so conspicuous compared to the city level. The simple *p*-median model outperforms the 2-coverage model in all indicators. So we can draw the conclusion that ensuring a backup ambulance does not improve the overall quality of the service.

Table 2: Quality indicators in the Prešov Region

Indicator	***Current locations***	***p-median Q = 1***	***p-median Q = 2***
Average response time [min]	11:34	**11:26**	11:27
% of calls responded to within 15 min	75.0	**75.3**	75.2
Average response time for highest priority calls [min]	11:21	**11:15**	11:18
% of highest priority calls responded to within 8 min	40.6	**41.0**	40.8

5 CONCLUSIONS

The paper presents a new approach to modelling demand for EMS in an urban area. In the model, aggregated demand points are individual streets and demand volume is the population in the street. Such detailed demand model has not been presented in the literature yet. It serves as an input to the *Q*-coverage *p*-median location model which seeks better distribution of ambulance stations across an urban area. A case study of the city of Prešov proves that a significant improvement of the service quality can be achieved by optimization. Better station locations in the city influence the EMS provision also in neighbouring villages. From a regional point of view follows that ensuring a backup ambulance to serve each demand point does not bring a desirable effect. We recommend to optimize the ambulance locations using the classic *p*-median model where each customer is assigned to exactly one service centre.

Acknowledgement

This research was supported by the Slovak Research and Development Agency under the project APVV-15-0179 "Reliability of emergency systems on infrastructure with uncertain functionality of critical elements" and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences under project VEGA 1/0342/18 "Optimal dimensioning of service systems".

References

[1] Aringhieri, R., Carello, G., Morale, D. (2007). Ambulance location through optimization and simulation: the case of Milano urban area. In Proceedings of the XXXVIII Annual Conference of the Italian Operations Society - Optimization and Decisions Sciences (pp. 1–29). Crema: Università degli Studi di Milano, Polo didattico e di ricerca di Crema.

[2] Czimmermann, P. (2016). Location problems in transportation networks. Communications – Scientific Letters of the University of Zilina, 18(3): 50-53.

[3] Díaz-Ramírez, J. D., Granda, E., Villarreal, B., Frutos, G. (2018). A comparison of ambulance location models in two Mexican cases. In Proceedings of the 2nd European International Conference on Industrial Engineering and Operations Management. Paris: IEOM Society.

[4] Jánošíková, Ľ., Jankovič, P., Kvet, M. (2017). Improving emergency system using simulation and optimization. In SOR ´17 : Proceedings of the 14th International Symposium on Operational Research in Slovenia (pp. 269–274). Ljubljana: Slovenian Society Informatika - Section for Operational Research.

[5] Karatas, M., Razi, N., Tozan, H. (2017). A multi-criteria assessment of the p-median, maximal coverage and p-center location models. Tehnički vjesnik, 24(2): 399–407.

[6] Maleki, M., Majlesinasab, N., Sepehri, M. M. (2014). Two new models for redeployment of ambulances. Computers & Industrial Engineering, 78: 271–284.

[7] Sasaki, S., Comber, A. J., Suzuki, H., Brunsdon, C. (2010). Using genetic algorithms to optimise current and future health planning – the example of ambulance location. International Journal of Health Geographics, 9, Article 4.

[8] Schmid, V., Doerner, K. F. (2010). Ambulance location and relocation problems with time-dependent travel times. European Journal of Operational Research, 207: 1293–1303.

[9] Zaffar, M. A., Rajagopalan, H. K., Saydam, C., Mayorga, M., Sharer, E. (2016). Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation-optimization. Operations Research for Health Care, 11: 1–12.

MINIMIZING HUMAN STRESS IN SOCIAL NETWORKS WITH TARGETED INTERVENTIONS

Dean Lipovac
InnoRenew CoE
Livade 6, 6310 Izola, Slovenia
E-mail: dean.lipovac@innorenew.eu
&
University of Primorska, Andrej Marušič Institute
Muzejski trg. 2, SI-6000 Koper, Slovenia

László Hajdu
University of Szeged, Institute of Informatics
Árpád tér 2, 6720 Szeged, Hungary
&
University of Primorska, Faculty of mathematics, Natural Sciences and information technologies
Muzejski trg. 2, SI-6000 Koper, Slovenia
&
InnoRenew CoE
Livade 6, 6310 Izola, Slovenia
E-mail: laszlo.hajdu@innorenew.eu

Sølvi Therese Strømmen Wie
NMBU, Faculty of Environmental Sciences and Natural Resource Management
Universitetstunet 3, 1432 Ås, Norway
E-mail: solvi.wie@nmbu.no

Anders Qvale Nyrud
NMBU, Faculty of Environmental Sciences and Natural Resource Management
Universitetstunet 3, 1432 Ås, Norway
E-mail: anders.qvale.nyrud@nmbu.no

Abstract: Chronic stress in humans can substantially impair functioning of organizations. As stress can spread through social networks, psychosocial interventions targeted at the most contagious people can efficiently improve organizational functioning. On the example of employees working in healthcare institutions, we show how optimization methods based on network science can guide efficient decisions to minimize human stress.

Keywords: human stress, social networks, network science, social contagion, infection model

1 INTRODUCTION

There are good reasons why stress is sometimes dubbed a silent killer; often, we are not aware of its harmful effects until it is too late. From debilitating depression to potentially fatal cardiovascular disorders, it seems like harmful effects of stress have no boundaries. Stress has even been implicated as a causal factor in cancer [3]. In this day and age, it is difficult to avoid stressful situations. We are in daily contact with stressors emerging from the environment (e.g., noise), our habits and behaviour (e.g., sleep deprivation), and the social domain (e.g., stress in the workplace).

Among the myriad of possible instigators of stress, we argue that social factors can be particularly risky. Social connections have been crucial for survival throughout the human (evolutionary) history and relationships within our families, friendships, schools, and companies are at the centre of our everyday life. It comes as no surprise that people are remarkably sensitive to social affairs; up to a point that the emotions and behaviours of one person can directly trigger similar emotions and behaviours in others [6]. While both positive and negative emotions and behaviours can spread, humans have quicker and stronger reactions towards negative stimuli. Because stress is manifested in increased unpleasant emotions (e.g., anger) and inappropriate behavioural tendencies (e.g., hostility) [9], it can spread through social interactions and decrease overall human wellbeing in any social system, including work organizations.

From the perspective of an organization, mitigating human stress typically means deploying a psychosocial intervention, such as a stress management workshop. While this approach is reasonable, companies might be reluctant to cover sizeable costs needed to provide the intervention to all employees. Even if that is not the case, their money might be better spent in providing in-depth interventions for a selected group of people instead of addressing all employees superficially. Considerably decreasing stress levels in the *most contagious* employees may mitigate stress in the entire personnel to a greater extent than slightly reducing stress in *all* employees.

Employees that will spread stress-related emotions and behaviours more readily than others will generally have more social connections and experience greater levels of stress (e.g., certain managers). Psychosocial interventions aimed directly at them might be the most efficient approach to improve the overall well-being in the workplace. Network science is an efficient tool to model the structure and the connections between any group of entities, such as companies, groups, or individuals. With this, it serves as a basis to model the infection process, in our case, the spreading of stress between people. The objective of the paper is to show how optimization methods based on network science can improve the human environment. We will demonstrate this by running network infection simulations that can provide the information on the most efficient targets (people) for stress-reducing interventions.

2 METHOD

We collected data on stress and wellbeing of nursing home personnel. The data was used to create a network representing connections between employees, employee stress levels, and the probability of stress levels to spread among employees. To model the spread of the stress between the employees, we used the Generalized Independent Cascade model in which the connection strength and the initial stress level of each employee can be simulated [1]. The most emotionally contagious employee set (and consequently the most suitable targets for stress-reducing interventions) can be computed by infection maximization [7], where the objective function minimizes the global stress level of the network. The detailed procedure is presented below.

2.1 Data collection and transformation

414 employees from 14 nursing homes in Norway completed the survey collecting demographic data, work-related information (e.g., working hours, shift work), and levels of stress and well-being. From the data, we created a network with the nodes representing employees and their overall stress levels. (Stress levels were rescaled to values between 0 and 0.8.) Connections (i.e., edges) between them were assigned if 1) they were employed at the same nursing home, 2) had the same occupation (e.g., nurse), and 3) the age difference between them was not greater than 20 years. Edge weights were higher when employees were closer in age, had matching work shifts, and worked more hours overall. (The edge weights were rescaled to values between 0 and 0.6). Employees with the missing relevant data were excluded from the study. The resulting network had 289 nodes and 731 edges.

2.2 Infection Model

The basic idea of infection models is to simulate the spread of a virus, information, or any other entity in a network. The concept was proposed by Domingos and Richardson [4] and by Granovetter [5]. Originally, it was used to improve the efficiency of viral marketing. The mathematical model of the problem was later introduced to networks by Kempe et. al. [7, 8]. In the Independent Cascade Model, the strength of the connections is given by probabilities between $\mathbf{0}$ and $\mathbf{1}$, expressing the chance of the infection or the effect spreading across the connection.

To define the model, let $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{E})$ be the network, where $\boldsymbol{V}$ is the set of the nodes and $\boldsymbol{E}$ describes the set of the edges. Let $\mathbf{0} \le \boldsymbol{p}_{\boldsymbol{v}_1, \boldsymbol{v}_2} \le \mathbf{0}$ be the edge probability between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, where $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \boldsymbol{V}$. To represent employee connections and stress spreading in the workplace, we used an extended Generalized Independent Cascade model [1]. The model defines probabilities on the nodes, so we define $\mathbf{0} \le \boldsymbol{s}_{\boldsymbol{v}} \le \mathbf{1}$ as the stress level of the node $\boldsymbol{v}$. Described from the real-life viewpoint, the probability on the node states the chance of a person becoming infectious and spreading stress to other individuals.

If the network is given with the edge and node probabilities, let $\boldsymbol{f}_{\boldsymbol{v}}$ be the final infection of the node $\boldsymbol{v}$ and the expected value $\boldsymbol{\sigma}(\boldsymbol{V})$ the sum of the final infection for each node. The final infection of the given $\boldsymbol{V}$ graph was computed by the Edge Simulation [1]. Chen et. al. [2] have proven that the simulation process is P#-complete, but the simulation can reach any precision level [7]. The code of the simulation was changed in the following way.

Algorithm 1: Edge Simulation in Generalized Independent Cascade

Input: G network, sample size k
j = 0
for all $v \in V: f_v = 0$
while $j < k$
  for all $e \in E$: set the state of the edge to active or passive based on p_e
  s=1
  Modified DFS from all $v \in V$
    $s = s(1 - p_u)$ where p_u is the stress level of the u visited node
  $f_v = f_v + 1 - s$
  $j = j + 1$
end while
for all $v \in V: f_v = \frac{f_v}{k}$

The simulation was used to compute the final stress level in one scenario. To choose the employees for the intervention, let $\boldsymbol{s}_{\boldsymbol{v}}^{\boldsymbol{intervention}}$ be the stress level of the node $\boldsymbol{v}$ after the stress-reducing intervention. If the stress of the person is decreased, we can rerun the simulation with the new stress levels, and the global infection of the network (and the overall stress levels) will also be decreased. In the case of a stress-reducing intervention, the model will decrease the stress level of a chosen employee and their local neighborhood, since this person will now have a lower probability to spread stress. If $\boldsymbol{\sigma}(\boldsymbol{V})$ is the expected value of the reference simulation and $\boldsymbol{I}$ is the set of the targeted employees, let $\boldsymbol{\sigma}(\boldsymbol{V})_{\boldsymbol{I}}$ be the expected final infection value of the simulation in which the initial stress levels of the employees in the set $\boldsymbol{I}$ are changed from $\boldsymbol{s}$ to $\boldsymbol{s}^{\boldsymbol{intervention}}$.

In the infection maximization problem, the main objective is to maximize the spread with an initial infector set. The original infection maximization problem was published by Kempe et. al [7], where they have proven the NP-hardness of the problem. The most efficient method to maximize the spread through a network is a greedy method, which can give 63% of the optimum in any case. The objective function maximizes the difference between the expected value of the initial reference simulation and in every iteration chooses the employee that minimizes the global stress level of the network (i.e., the employee undergoing a hypothetical stress-reducing intervention). A similar maximization problem was proposed in the following article [10]. The greedy method is the following:

Algorithm 2: Greedy Method to minimize the stress level of the employees

Input: G network
Output: Ordering of the employees based on stress reducing potential
$I = \emptyset$
while $|\mathbf{I}| \neq |V|$
 $I = I \cup \arg\max_{v \in G(V) \setminus I} (\sigma(V) - \sigma(V)_I)$

From a real-world point of view, the objective function that maximizes the difference between the different final infection values will minimize the global stress levels in our network. To show the optimal number of the hypothetically treated employees, it's possible to find the threshold where the global stress level will stop decreasing significantly.

3 RESULTS

A sample of the network is presented in Figure 1. Nodes are coloured based on the stress level of each employee and the width of the edges increases with the edge weight (i.e., probability of emotional (stress) contagion).

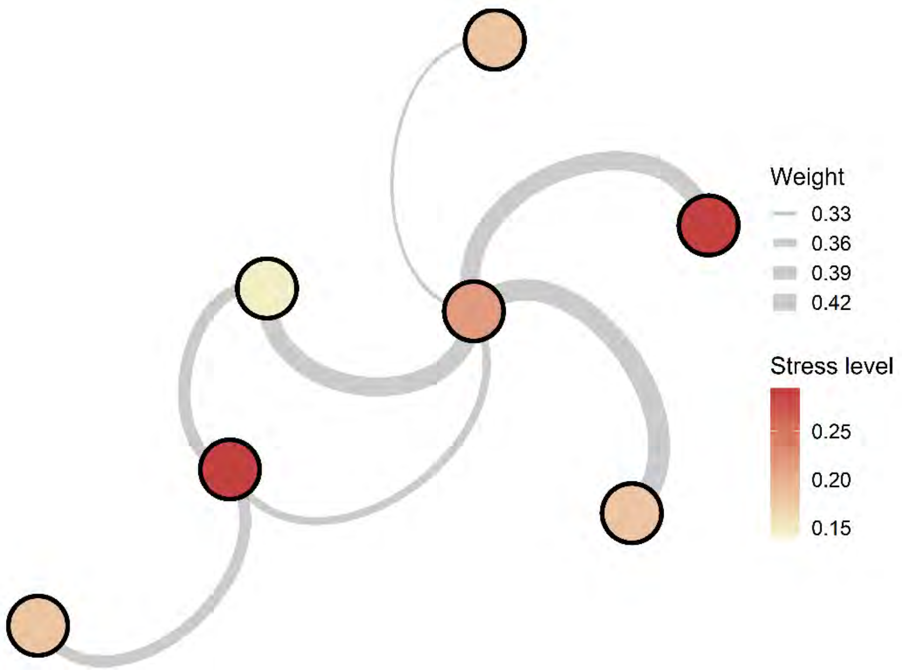


Figure 1: A sample of the network

In the network, we identified employees with the largest emotional contagion potential – people whose negative emotions are the most likely to create the largest negative overall impact by spreading to others. Table 1 displays the basic information on the top three most emotionally contagious persons. Data on such highly infectious people could be used to ascertain which qualities make a person more likely to spread emotions. (Some data have been obfuscated to protect the anonymity of respondents; each individual is at least 10-anonymous.)

Table 1: Top 3 most infectious persons (in descending order).

Gender	Age	Education	Field of work	Work hours	Work years
Female	41-50	University or more	Healthcare	21-40	0-20
Female	31-40	High school or less	Healthcare	21-40	21-40
Female	31-40	High school or less	Support staff	21-40	21-40

These highly contagious people are the prime targets for stress-reducing interventions; they are the people on whose targeted stress-reducing interventions are the most likely to have the largest positive overall impact on the entire group of people. In fact, we simulated such an intervention and the effect it has on the stress levels of the entire social network. Our hypothetical stress-reducing intervention decreased the stress levels of the employees by a randomly chosen value ranging between 0-40%. The model included the employees in order of their emotional contagion potential; those with the highest potential to impact the entire network were considered first. By summing the stress level values of all employees, we calculated the overall stress level score of the entire social network. This score was used to evaluate the effectiveness of providing hypothetical stress-reducing interventions to different numbers of employees.

Figure 2 shows the overall stress levels (of the entire social network) based on how many people received the hypothetical stress-reducing intervention. The x axis displays the number of targeted people by the hypothetical intervention, starting with the people that have the largest emotional contagion potential (on the left). The y axis represents overall stress levels in the entire group of employees (starting with the initial 100% value). As an example, the overall stress levels of all employees combined decreased to approximately 50% of the initial state when around 200 employees were targeted with the intervention.

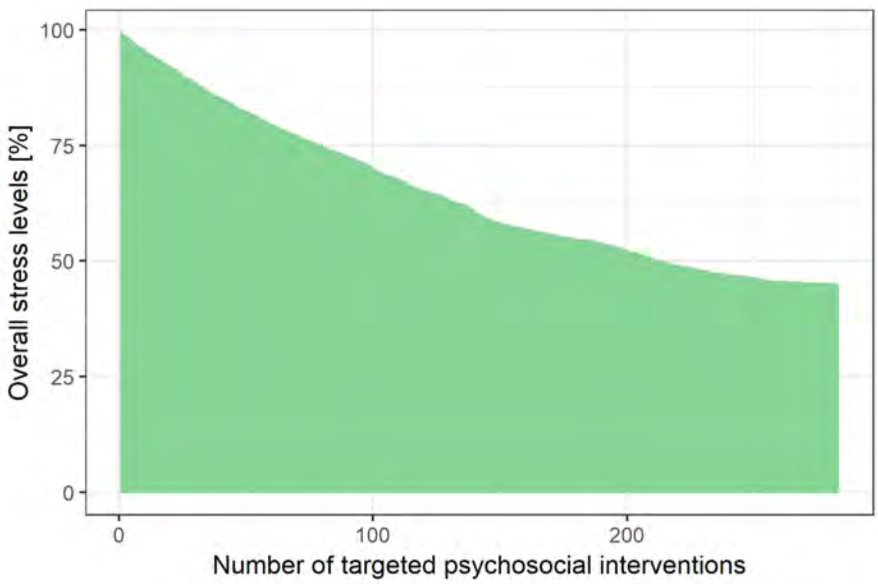


Figure 2: Overall stress levels with the growing number of psychosocial interventions.

4 CONCLUSIONS

Decreasing human stress is a challenging task. Due to the scope of the problem, stress-reducing approaches will often only scratch the surface; consequently, it is important that they are as efficient as possible. One way to increase their efficiency is to direct them at the persons with the highest potential for emotional contagion. With this, we can improve the well-being of a larger group of people despite targeting only select few. This can be achieved effectively with the application of certain network science methods.

Acknowledgement

We acknowledge the European Commission for funding the InnoRenew CoE project (Grant Agreement 739574) under the Horizon2020 Widespread-Teaming program, the Republic of Slovenia (Investment funding of the Republic of Slovenia and the EU of the European Regional Development Fund), and the National Research, Development and Innovation Office of Hungary (SNN-117879).

References

[1] Bóta, A., Pluhár, A., Krész, M. 2013. Approximations of the generalized cascade model. Acta Cybernetica, 21(1): 37–51.

[2] Chen W., Yuan Y., Zhang L. 2010. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. Proceeding ICDM ’10 Proceedings of the 2010 IEEE International Conference on Data Mining, 88–97.

[3] Cohen, S., Janicki-Deverts, D., & Miller, G. E. 2007. Psychological stress and disease. Journal of the American Medical Association, 298(14): 1685–1687.

[4] Domingos, P., Richardson, M. 2001. Mining the network value of customers, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining: 57–66.

[5] Granovetter. M. 1978. Threshold models of collective behaviour. American Journal of Sociology, 83(6): 1420–1443.

[6] Hatfield, E., Cacioppo, J. T., Rapson, R. L. 1993. Emotional contagion. Current Directions in Psychological Science, 2(3): 96–99.

[7] Kempe, D., Kleinberg, J., Tardos, E. 2003. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining: 137–146.

[8] Kempe, D., Kleinberg, J., Tardos, E. 2005. Influential nodes in a diffusion model for social networks. Proceedings of the 32nd international colloquium on automata, languages and programming: 1127–1138.

[9] Schoenewolf, G. 1990. Emotional contagion: Behavioral induction in individuals and groups. Modern Psychoanalysis, 15(1): 49–61.

[10] Tóth L., 2016. Greedy Heuristics for the Generalized Independent Cascade Model, StuCoSReC Proceedings of the 2016 3rd Student Computer Science Research Conference, 51-54.

PRINCIPAL-LEADER-FOLLOWER MODEL WITH INTERNAL SIGNAL

Andreja Smole,
Cosylab, Control System Laboratory
Ljubljana, Slovenia
E-mail: andreja.smole@cosylab.com
Timotej Jagrič,
Faculty of Economics and Business (University of Maribor)
Maribor, Slovenia
E-mail: timotej.jagric@um.si
Drago Bokal,
Faculty of Natural Sciences and Mathematics (University of Maribor)
Maribor, Slovenia
E-mail: drago.bokal@um.si

Abstract:

We develop a novel extended principal-agent model with an internal signal. The model was developed to ensure quality results in a relationship between a company's owner (principal) and two co-workers (leader and follower), whose responsibility is the realization of the project. The relationship is modeled with a two-level game. At the external level of the game, the principal establishes a contract with the leader and the follower to motivate their performance. By the internal game, we model the relationship between the leader and the follower as a two-move game. Interpretations of the conditions given in the paper may yield employment and project management policies that facilitate high-quality results.

Keywords: Game theory; Nash equilibrium; Principal-agent model; Leader-follower model; Internal signal.

1 Introduction

A leader and a follower are agents to whom the higher management outsources the task of project realization. This fits the setting of principal-agent models that study the interaction between a principal and an agent, or several agents, where the principal outsources work to the agent [1, 3, 18, 7, 15, 16, 13, 9, 10]. In those models, a contract is offered to the agent, which can be either accepted or rejected, but not negotiated. In the setting most similar to ours, an agent who accepts the contract must choose whether to cooperate or to defect. This choice is agent's private information. The principal's challenge is to design a contract that will motivate the agent to cooperate with the principal's preferences.

Related topics include using incentive design to demonstrate that a particular prize can induce two contestants to choose the efficient level of effort in a single contest [14]. A similar, but more general model was applied in [17] to design and evaluate pay schemes for educators in elementary and secondary schools, an approach not directly applicable in a company.

Our model analyses the relationship between the principal (company's owner, director-general, leader, boss, or employer) and two agents (leader and follower). We name our game *Principal-Leader-Follower Game*. The principal's goal is to ensure, with the payoff in the agreement, that both agents will deliver good results and will have conducted an excellent project. To ensure the principal's goals, we need to identify the payments that will motivate both agents. Both agents are working on the same project and to ensure high-quality results, they have to cooperate, which is why we studied the leader-follower interaction in two levels in the next step. We named it *Project-Leader-Follower game with internal signal.* In the first two moves of this game, both players choose to cooperate (C) or defect (D) and in the last two moves, they each grade the other's choice with high (HG) or low (LG) grade, thus producing an internal signal of this choice (Figure 1). We assume that the first move is done by the leader (denote $_L$), thus assuming the follower's cooperation does not affect the leader's decision on his cooperation: the leader must prepare a project plan and devote the work on the project. When the follower (denote $_F$) decides whether or not to cooperate, he already knows the leader's choice.

On a higher level, the Project-Leader-Follower game is embedded into a principal-agent type interaction, where the principal wants to induce the two agents to cooperate in the process of project realization and provide the grade as a truthful signal of each other's cooperation and work. Note that the principal has no insight into the agents' private choices regarding the cooperation, but the two agents do have information to evaluate the other's choice. In this context, our results present conditions on relative utilities of the components of the payoff for the truthful implementation of this signal in the elaborated principal-two agents setting [12] surrounding the game between the leader and the follower.

We establish sufficient and necessary conditions on these relative values that assure the pure strategy Nash equilibria of the game to occur precisely in pairs of strategies where both agents choose to cooperate and justifiably receive a high grade and evaluation. As all pure strategy Nash equilibria under these conditions have the same outcome, so do all mixed strategy Nash equilibria, implying a unique stable game-theoretic outcome of the game.

The paper is structured as follows. First, we present the newly developed game and set it into the context of related work. Then, we present the results of the deductive analysis of the game to identify the sufficient and necessary conditions on relative contributions of various utility sources for quality work to be the only equilibrium outcome of the game. We conclude with a section discussing interpretations of our deductive results into project management context and some ideas for further research.

2 Principal-Leader-Follower Game

The newly developed model is applicable under standard assumptions of incentive design [13]. In addition: (i) We focus on the interaction between a single leader and follower in a single project. (ii) We simplify the arbitrary number of choices of elusive effort quality found in other moral hazard models [8, 13], so both agents have only two choices at each move: first cooperation or defection, and then high or low grade. (iii) At the time follower is evaluating a leader's managerial work, the follower is aware of the grade that the leader has given him. (iv) By definition, cooperation yields high-quality project realization and defection yields low-quality project realization. Likewise, cooperation yields low payoff for other activities, and defection yields high payoff for other activities.

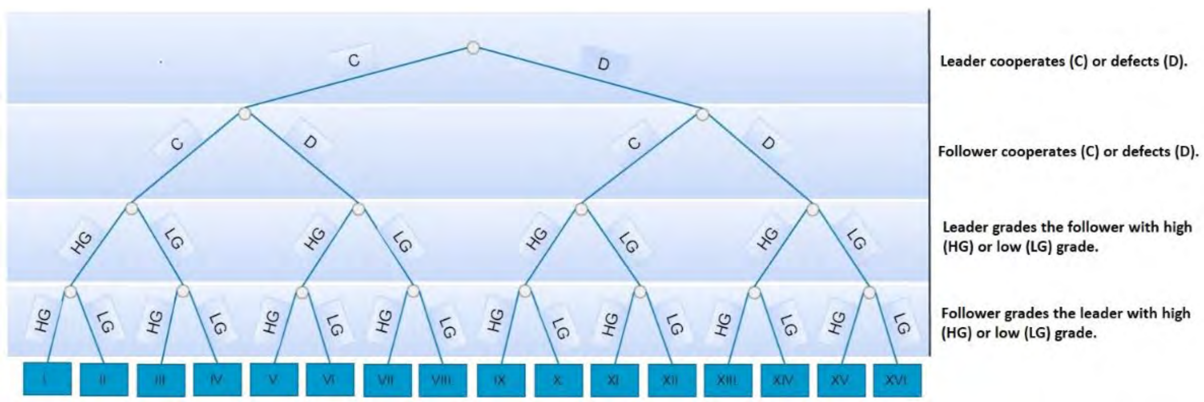


Figure 1: Game tree for the Project-Leader-Follower game with internal signal.

We address the realization of the project and the relationship between the co-workers by designing the game, analogous to the well-known prisoners' dilemma used in a similar context by [4]. We model the leader-follower interaction as a two-player four-move game with complete information. Its outcome is determined in four moves. We focused our analysis on the relationship between the leader and the follower on the execution of one project, although in some cases they are involved in more than one project. For the elementary model, we do not analyze group relationships between followers and their mutual competition or collaboration. The choice of cooperation is the agent's private information.

The combinations of agent's moves result in 16 possible outcomes, as depicted in Figure 1. There are precisely eight distinct pure strategies for the leader: two choices regarding cooperation and two independent grade levels for each of the two possible follower's cooperation choices. The second move distinguishes four types of grading:

- *Honest grading* (H) gives a high grade for cooperation and a low grade otherwise,
- *Lenient grading* (L) gives a high grade regardless of cooperation,
- *Uniform low grading* (U) gives a low grade regardless of cooperation, and finally,
- *Mean grading* (M) gives a low grade for cooperation and a high grade otherwise.

Contrary to the rather small number of leader's pure strategies, there are 64 distinct pure strategies for the follower. As a leader's payoff does not depend on the follower's evaluation of the leader, the follower is completely at liberty, without explicit incentives for either choice in the evaluation. Relying on the warm glow principle [2], we model the follower's last move by the possible characters:

- The warm glow principle inclines a *Partially honest follower* (PH) to give a good evaluation in case that the leader cooperated and has given a fair grade (a grade actually reflecting the follower's cooperation).
- A *Partially biased follower* (PB) is inclined to reflect the received grade: he gives a high evaluation for a high grade and low for a low grade.
- An *Attributive honest follower* (AH) strategy models grade attribution theory: the follower assigns low evaluation when receiving a low grade (attributes failure to others), but behaves as a partially honest follower when receiving a high grade.
- An *Attributive biased follower* (AB) assigns a low evaluation after receiving a low grade, but his evaluation in the case of high grade reflects only the leader's cooperation choice, not the grading.

We list all the leader's strategies as columns and the follower's strategies as rows of Table 1. For the leader, we denote the choice of cooperation with C and defection with D. The leader's strategy is fully described by adding C or D to the letter corresponding to the leader's grading strategy. For the follower, however, we first need to state his cooperation choices for each cooperation choice of the leader, with two labels, where the first label represents the follower's response to the leader's choice of C and the second one the response to D. Supplementing this with the two-letter notation for the follower's evaluation strategy denotes each follower strategy with four letters.

The payoffs of this game depend on the execution of the project, which is more dependent on follower's work, the total payoff from any other activities, and the grade or evaluation received by each agent. Relations between these payoffs are in practice determined by career policies and the company's award policy. Both the leader and the follower need to contribute their effort to the execution of the project, which improves the quality of the executed project.

Given the discrete choices of the agents, the 16 outcomes exhibit four different levels of quality of the executed project - excellent (e), resulting from both agents choosing to cooperate; good (g), resulting from only the follower cooperating; sufficient (s), resulting from only the leader cooperating, and insignificant (i), resulting from both agents opting to defect. It is clear that the utility of excellent quality of the project is highest and the utility of insignificant quality of the project is lowest. We assume the utility of the good is more than of the sufficient, thus acknowledging that the project's quality depends more on the cooperation choice of the follower than of the leader, who is assumed to hold a more managerial, less content related role. For the other activities and grade evaluation, we postulate two levels of the utility: there is either a high amount of other activities (f) or low amount (f'), and similarly for utility for a high grade (g) or a low grade (g').

		LEADER							
	Strategy	CH	CL	CU	CM	DH	DL	DU	DM
FOLLOWER	CCPH	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	CCPB	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L + o_L$ $f_F' + g_F + o_F$	$f_L + g_L + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	CCAH	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	CCAB	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	CDPH	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$
	CDPB	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L + i_L$ $f_F + g_F + i_F$
	CDAH	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$
	CDAB	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L + e_L$ $f_F' + g_F + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L' + g_L' + e_L$ $f_F' + g_F' + e_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$
	DCPH	$f_L' + g_L + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L' + g_L + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	DCPB	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L + g_L + o_L$ $f_F' + g_F + o_F$	$f_L + g_L + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	DCAH	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	DCAB	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$	$f_L + g_L' + o_L$ $f_F' + g_F' + o_F$
	DDPH	$f_L' + g_L + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L' + g_L + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$
	DDPB	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L + i_L$ $f_F + g_F + i_F$
	DDAH	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$
	DDAB	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L' + g_L' + u_L$ $f_F + g_F' + u_F$	$f_L' + g_L + u_L$ $f_F + g_F + u_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$	$f_L + g_L' + i_L$ $f_F + g_F' + i_F$	$f_L + g_L' + i_L$ $f_F + g_F + i_F$

Table 1: Table of utilities (payoffs) for the Principal-Leader-Follower Game.

Table 1 presents the leader's strategies as columns and the follower's strategies as rows. The outcome of the agents' strategies appears at the intersection of the corresponding row and column. The outcomes corresponding to them are colored, marking them as possible Nash equilibria of the Project-Leader-Follower Game, which can simultaneously occur due to the many leader-follower pairs it involved. The Nash equilibrium that is particularly favoured is the outcome I, which we name *desired outcome*. For each player, we assume the differences between components for high and low value are strictly positive numbers:

Definition 2.1 Rational assumptions *state that for each* $x \in \{L, F\}$, df_x *and* dg_x *are strictly positive numbers and that* $e_x > o_x > u_x > i_x$.

We assume that each player's utility function is the sum of utilities for each of the quantities. The total utility is set for the pair of moves chosen by the leader and the follower, respectively. Note that only the differences ($d_x = x - x'$ for $x \in f, g$) between the utilities of discrete choices play a role in the analysis. We also define the differences $d_{eo} = e - o$, $d_{eu} = e - u$, $d_{oi} = o - i$, and $d_{ui} = u - i$ (with appropriate agent subscripts).

3 Results

We emphasize the Nash equilibrium [as] a self-enforcing agreement, that is, an (implicit or explicit) agreement that, once reached by the players, does not need any external means of enforcement, because it is in the self-interest of each player to follow the agreement if others do [11]. Thus, even if a certain behavior is socially unacceptable, rational individuals may enact it when faced with a choice, thus putting their own self-interest ahead of outcomes desired by society (the principal in our model). The policies in place at the company and in the hiring process of the employers affect the relationships between these three components, thus determining the Nash equilibria of the internal game.

We identify pairs of strategies that yield the principal's desired outcome. We then use the approach of mechanism design to deductively identify the award structures of the two players that are sufficient and necessary conditions for these pairs to be the only Nash equilibria of the internal game. In Theorem 3.1, we define the sufficient and the necessary conditions for these pairs to be the only Nash equilibria, hence eliminating any rational desire for other pairs of strategies to become stable behavior.

Theorem 3.1 *If the rational assumptions hold, then the following is true: The pairs of strategies* $(\{C\ \ H, C\ \ L\}, \{C\ \ \cdot\})$ *are the only pure strategy Nash equilibria of the Project-Leader-Follower Game with internal signal, if and only if the following conditions hold:*

$$deo_L \geq df_L; \tag{1}$$

$$dui_L > df_L + dg_L; \tag{2}$$

$$deu_F > df_F + dg_F; \tag{3}$$

$$df_F > doi_F + dg_F. \tag{4}$$

The sufficient and necessary conditions for a pair of strategies manifesting the desired outcome to be a Nash equilibrium can be interpreted as follows:

- for the leader: the difference between the utility for excellent and good project quality is larger than the difference between utility of a high or low amount of other activities, i.e. the payoff contribution of cooperation in the presence of cooperating follower is greater than the contribution of other activities;
- for the follower: the difference between the utility for excellent and sufficient project quality is larger than the difference between utility of high and low amount of other activities, i.e. the payoff contribution of cooperation in the presence of cooperating leader needs to be higher than the contribution of other activities.

4 Conclusions

In the paper, we developed Principal-Leader-Follower Game, a theoretical model of interaction between an individual principal, an individual leader, and an individual follower aiming to understand their incentives in the project realization process.

The model represents a novel extension over the currently separate principal-agent and leader-follower models of effective project management, since it proposes to the principal the conditions that ensure that the leader and the follower will realize a high-quality project. The role of the leader is also to mentor the follower, which affects the increase in the follower's qualifications and his satisfaction through personal growth.

Analyzing the model, we provide deductive evidence for what experience suggests: motivation stemming directly from quality, not from opinions, is the key factor in providing long term cooperation, company's growth, and personal satisfaction.

While the model analyzes the phenomenon under the assumption of leader's and follower's rationality, we are aware that this assumption is strong and known not to hold in several social contexts [6]. For practical research, hence, we propose to model the relationship between ethical and utilitaristic approaches to mechanism design aiming to understand the behavior of agents who are incentivized for or against their ethical imperatives. The other aspect, neglected by the rationality assumption of the two agents, is the emotional well-being of the agents. In this direction, we propose to continue the research by including the emotional aspects of the two agents into the decision evaluation process. A basis for combining the utility of well-being has already been developed in [5], and the next step is to evaluate the strategic aspect of the Principal-Leader-Follower Game in the context of the emotions experienced.

References

[1] Aliprantis, C. D., and Chakrabarti, S. K. (2000). Games and decision making. *Oxford University Press.*

[2] Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal, Volume 100, Issue 401.*

[3] Bebchuk, L., and Fried, J. (2004). Pay Without Performance. *Harvard University Press.*

[4] Bhattacharyya, N. (2004). Student Evaluations and Moral Hazard. *Journal of Academic Ethics*, 2.

[5] Bokal, D., and Steinbacher, M. (2019). Phases of psychologically optimal learning experience: task-based time allocation model. *Central European Journal of Operations Research*, Volume 27, Issue 3, pp. 863–885.

[6] Bowles, S. (2016). The moral economy: why good incentives are no substitute for good citizens. *Yale University Press.*

[7] Chang, C. (2013). Principal-Agent Model of Risk Allocation in Construction Contracts and Its Critique. *Journal of Construction Engineering and Management.*

[8] Dembe, A., and Boden, L. (2000). Moral Hazard: A Question of Morality? NEW SOLUTIONS. *A Journal of Environmental and Occupational Health Policy, Volume 10 , Issue 3, pp. 257-279.*

[9] Harris, M., and Raviv, A. (1978). Some results on incentive contracts with application to education and employment, health insurance, and law enforcement. *American Economic Review, pp. 20–30.*

[10] Holmstrom, B., and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization. Volume 7, Special Issue: Papers from the Conference on the New Science of Organization, pp. 24–52.*

[11] Holt, C., and Roth, A. (2004). The Nash Equilibrium: A Perspective. *Proceedings of the National Academy of Sciences of the United States of America, Volume 101.*

[12] Krautmann, A. C., and Sander, W. (1999). Grades and student evaluation of researchers. *Economics of Education Review, Issue 18, pp. 59–63.*

[13] Laffont, J., and Martimort, D. (2009). The Theory of Incentives: The Principal-Agent Model. *Princeton University Press.*

[14] Lazear, E. P., and Rosen, S. (1981). Rank-Order Tournaments as Optimum Labor Contracts. *The Journal of Political Economy, pp. 841–864.*

[15] Liu, H., and Liu, S. (2013). Salesperson compensation contract design based on multi-agent principal agent model. *Nankai Business Review International, Volume 4, Issue 1, pp. 78–86.*

[16] McAfee, R. P., and McMillan, J. (1996). Competition and Game Theory. *Journal of Marketing Research, Volume 33, Number 3, pp. 263–267.*

[17] Neal,D.(2011). The design of performance pay in education. *Handbook of the Economics of Education, Volume 4, pp. 495–550.*

[18] Waterman, R. W., and Meier, K. J. (1998). Principal-Agent Models: An Expansion? Journal of Public Admin. *Research and Theory, Volume 8, Issue 2, pp. 173–202.*

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 6:

System Modelling & Soft Operational Research

A UNIFIED ENVIRONMENT FOR QUANTITATIVE AND QUALITATIVE MODELLING OF DYNAMIC SYSTEMS

Dariusz Banas
University of Economics, 1 Maja 50, 40-287 Katowice, Poland

Abstract: In analysing the possibility of practically applied decision-making process, according to our comprehension and business practice, a one in thinking process shall take into account not only the essential properties of economic entities characterising themselves by complexity, dynamics, nonlinearity, but also the knowledge of experts. Use of expert knowledge in the process of supporting decisions in the enterprise is complementary to quantitative models, forming the modelling approach that reflects surrounding business reality. Yet, to address above practically in the context of Operations Research, there is lack of fully dynamic approach that simultaneously represents quantitative modelling (e.g. System Dynamics (SD)[1]) and qualitative (e.g. Soft System Dynamics (SSD)[2]) to outline relationships and build a coherent model with use of exact mathematical equations and tacit knowledge of experts. In this manuscript we tend to find an answer and possibly 'pave out' novelty, seen as a concrete solution based on analysis of the real business use case with business implications to the hypothesis we raised: whether a uniform computation approach of system dynamics methodology can be established for quantitative and qualitative constituents of the economic entity (description by means of state equations). Results obtained in such solution are promising, however still to be verified on different sets of models, data and configurations.

Keywords: System Dynamics, Soft System Dynamics, Weighted Influence Non-linear Gauge System (WINGS), Fuzzy Cognitive Maps (FCM), decision making, qualitative data, quantitative data

1 THEORETICAL AND PRACTICAL SETTLEMENTS

In the practice of Operations Research, decision making techniques refer to the statics or quasi-dynamics of complex structures that result from the initial description of the system, thus imposing simplifications that linearise the native structure of the system, and consequently, narrow the field of analysis and the results obtained. There is no fully dynamic[3] approach, including non-zero initial conditions, and being in this way equivalent to native[4] features of such system, and capturing the aspects of interdependence, mutual interactions, feedback, cause-and-effect relations, but also randomness and determinism, which are the features of complex systems [1]. To complement above as for decision making there is neither the approach[5] that simultaneously represents quantitative (considered as SD [5]) nor qualitative (considered as SSD [8]) modellings. Entire manuscript creates an exemplary path, as excerpt from the overall fully dynamic approach, as for analysing the system build with the use of expert knowledge[6] in the environment of system dynamics[7]. Important to note is that SD [4] is predominately used as for quantitative data modelling what for qualitative ones delivers ambiguity and unclarity, i.e

[1]System Dynamics (SD) - according to Forrester
[2]Soft System Dynamics (SSD) - according to Mendoza, Prabhu
[3]in a sense of time (not only with the roots of present time, but also with the past and characteristics acquired in previous competitive episodes that may form the context for future ones)
[4]native features are "innate" features that characterize the real nature of system
[5]currently there is no outlook that combines the strengths of both modelling methods / compensates for the disadvantages of each of them
[6]expert's knowledge cannot be captured well by system dynamics, i.e., the framework of J. Forrester's System Dynamics - applied in this manuscript
[7]other approaches (similar in the context and applicability) of system dynamics were developed by Lukaszewicz, Coyle

assessment of knowledge of personnel (qualitative data) engaged to the quantitative system [3]. For the purpose of qualitative analysis usually soft system dynamics is applied as better fitting method dealing with qualitative data. For the latter ones as for SSD, the WINGS [9] method is used in the manuscript to compare results against SD, both fed by the same sets of randomly generated data set to step apart for testing purposes from the real world data (enterprise and expert data) and drive conclusions (see in Section 2). WINGS computation results are comparable to the other well known SSD approach used by practitioners for modelling called as FCM methods [6]. Those ones however (also as per details in [2]) may deliver inconclusive results, thus FCM's require special attention[8] and real practice of use depending on the analysed system properties. The case study considered in this manuscript comes from the evaluation of strategic offers at enterprise level with aim to select the most appropriative one for the submission to the client against enterprise financial and strategic imperatives and its conditions (details in [2]). In the outcome conclusions [2] it can be drawn attention that the quality of decisions may directly contribute to financial and strategic health of the company. In practice it turns out to be dependent on procedures and methods of assessment, what leads to different results and implications, whilst financial and strategic implications may range billions of Euros in international corporation operations (for strategic offers - examples in [2]). In summary as for analysis performed with WINGS [9] and FCM [6], the assessment of the impact of strategic offers does remain susceptible to the method used for analysis and computations, with implication for verification of the usefulness of qualitative models in relation to the system studied and the managerial problem being solved.

1.1 Methodological approach

In methodological approach the aim of the manuscript is find out a universal solution that supports qualitative and quantitative models, in a possibly simple manner, therefore initial testing of the hypothesis[9] takes place in the Section 2 with use of simulated data. Note: although there is general agreement on the importance of qualitative data when developing a SD model, there is no 'clear path' of usage (when, how [7]), given that coexistence of 'qualitative and quantitative' is indispensable in systems practice. The latter one has been pointed out by J. Forrester (Fig. 1), without however a clear practical-computational solution proposed. This gap is more noticeable, especially when the model requires the use of so-called soft variables such as 'customer satisfaction', 'product quality', 'price pressure', 'engagement' or 'perceived productivity'. Although SD and SSD pertain with the same goal of decision-making, both in SD and SSD models (similarly MCDA models) only a few attempts[10] have been described to merge these methods and provide a tool for better understanding and control of systems [10]. To transform these observations and the anticipated approach (and continue towards Section 2), let's outline first some specifics of system dynamics computation approach (enclosed in 1.2).

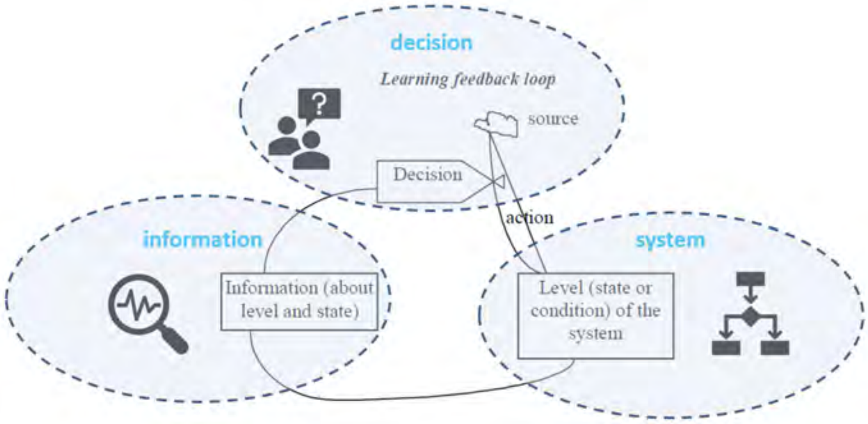
1.2 Specifics of SD computation approach and initial calculations with use of real model data

In a general view we consider a basics of SD, seen as the *stock* x with incoming *input* and outgoing *output* flows, where its mathematical representations in the time t domain are presented below (1),(2),(3),(4) as for the preparation of the general code for computer-aided calculations

[8]comparisons WINGS-FCM with use of real data (enterprise, expert) are described in [2]

[9]can system dynamics methodology (description by means of state equations) be used as a computing environment for both quantitative (SD) and qualitative (SSD) models?

[10]identified tangible (for decision-making) examples of method integration are the combined SD and PROMETHEE procedures (SD: scenario simulation, strategic options and long-term test; MCDA: Ranking and selection using the PROMETHEE procedure) among 675 articles in the topic of combining SD with another method for a sake of homogeneous patch combining quantitative and qualitative views)



Source : Forrester, 1968

Figure 1: A general example of a system, information and decision - model (compilation from the original J. Forrester's picture published in 1968)

with use of equations (5),(6):

$$\frac{dx(t)}{dt} = input - output \tag{1}$$

$$\int \frac{dx(t)}{dt} dt = \int (input - output) dt \tag{2}$$

$$x(t) = \int (input - output) dt \tag{3}$$

$$\frac{x(t) - x(t - \Delta t)}{\Delta t} = input - output \tag{4}$$

$$x(t) - x(t - \Delta t) = (input - output)\Delta t \tag{5}$$

$$x(t) = x(t - \Delta t) + (input - output)\Delta t \tag{6}$$

Important to note: all relationships, influences, weights between elements of the model (system) in SD are provided by the experts based on their assessment, whilst the entire state of the enterprise is supplied by the business support systems (both normalised to the values from 0 to 9 - Fig. 2 [11]). As for computations with use of differential equations the RK4 (Runge-Kutta 4) calculation method for system dynamics has been applied (with step dt = 1/4 which is a standard / universal value). This method is extrapolative estimation of inter-ocular value, equivalent to the approximation of the trajectory between steps (k1, k2, k3, k4) for the higher order polynomial curve[12], what returns more precise results [11]. Computation in line with the above were performed in Stella software environment[13] for the three independent offers A, B, C that impact the enterprise conditions (state), thus may result in its financial health [2]. Obtained results of computations are very close to each other as to compare FCM, WINGS, SD (Fig. 3), being independent to the method used (exact values naturally differ between FCM, WINGS and SD). The analogy of the SD calculations to the SSD methods (FCM, WINGS) comes from the matrix balance. In the SD interpretation, the matrix balance is the state(s) of the concept (node), taking into account the input and output, influences, with the behaviour of graphs logic (note: the same logic of directed graphs is applied for FCM and WINGS).

[11] in this data there are no mentioned influences between network elements (however all data details can be found in [2])

[12] with auxiliary values k1, k2, k3, k4 and using these values to calculate x(n+1). The sizes k1, k4 describe the slope of the trajectory at the beginning and end of the intervals, and k2, k3 - at the center, where the final result is x(n+1)=x(n)+dt/6(k1+2k2+2k3+k4) - according to [11]

[13] https://www.iseesystems.com/

Current status in the firm (values in parenthesis)	Offer A	Offer B	Offer C
Net cash flow (per year) (4)	2	1	3
Revenue (6)	3	5	4
Quoted sales margin per BG (2)	1	0	0
Quoted sales margin per unit (4)	2	2	2
Capital Cost (1)	0	0	0
Delivery costs (per year) (9)	2	5	5
Further upsell (per year) (9)	6	2	4
Expected upsell (9)	4	8	2
Deal upsell NPV (7)	1	2	3
Strategic deal quality (6)	4	1	4
Strategic context (3)	1	0	1
Committed upsell (per year) (9)	1	2	6
Market unit share impact (1)	1	0	1
Price erosion (0)	0	0	0
Revenue risk (2)	1	0	2
Sales risk (2)	1	1	0
Discount (per year) (2)	1	1	0
Cost risk (1)	0	2	2
Technical risk (3)	2	0	1
Customer Credit risk (1)	0	0	1
Margin risk (2)	1	0	1
Customer risk (1)	1	2	1
Delivery risk (3)	3	1	2
Total risk provision (4)	3	2	3

Figure 2: Model input data: status of the firm and impact of the offer

2 RESULTS OF SD COMPUTATIONS USING RANDOM DATA - AND CONCLUSIONS

2.1 Computations using random data

As to test the hypothesis whether system dynamics methodology with use of state equations can serve as for a computing environment for both SD and SSD models it has been assumed that randomly generated data for all the parameters related to the offers of A,B,C and enterprise state (conditions) feed the same model-structure (with conditions of Sub-section 1.2 in Section 1 in this manuscript). Moreover at the entire step of research, all the data has been generated as positive integers (without negative ones), to narrow complexity of SD model and computations[14]. All together 32670 data points have been randomly generated (with above conditions and logic of the system(networks)) to feed 10 independent 'generations' of calculations with use of SD modelling. Each 'generation' contained 168 stocks/differential equations[15] for offers (A,B,C) and enterprise status, imitating SSD networks (SSD computations to compare vs. SD were obtained with the use of WINGS method). Example of results are shown on the Fig. 4.

2.2 Conclusions

Despite some initial promising results, not for all 'generations' SD vs. WINGS obtained results were identical (placement on the axis), what highly drives additional attention and analysis of root causes of this numerical computations (planned as further works). It also has been found that SD method returns results very fast - in the 3rd step of computations, what is equivalent to 3*dt=3*(1/4)=0.75 with use of RK4 integration method. Other methods, like Euler or RK2 have been tested too, returning similar results to RK4, but in a longer horizon of computations

[14]note: through SD modelling it is possible to mimic negative influences too

[15]and other relevant SD building blocks

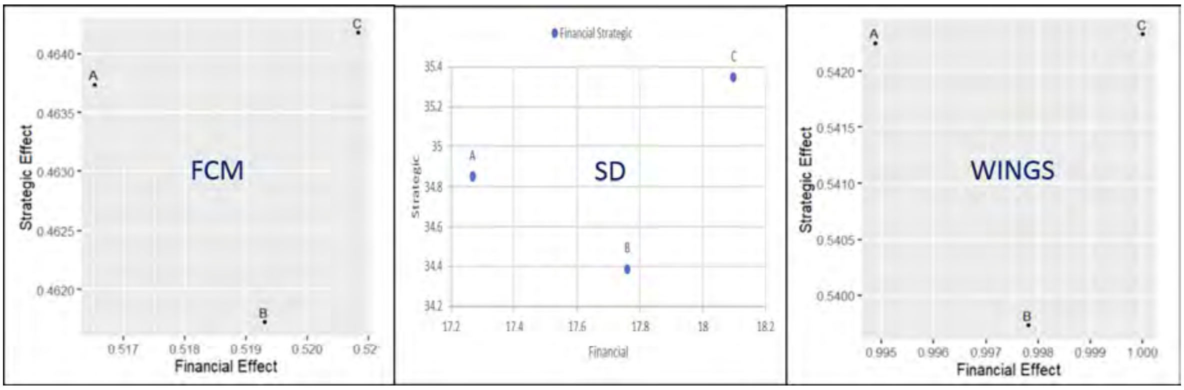


Figure 3: Calculation results in the SD environment compared to FCM and WINGS - total relative involvement (placement on the axis) of Financial and Strategic Effects (caused by offers A, B, and C onto enterprise status) is very similar irrespective to the calculation method

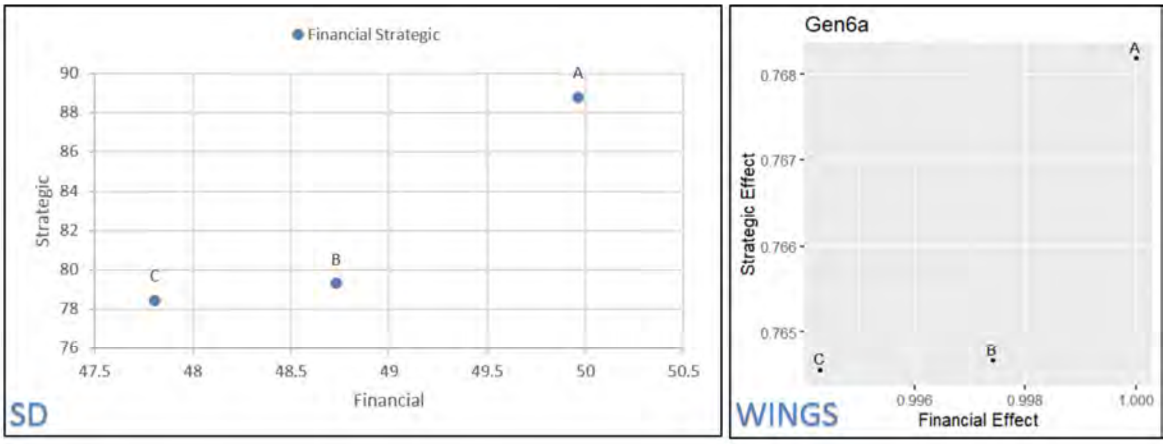


Figure 4: Calculation results in the SD environment compared to WINGS - Financial and Strategic Effects caused by offers A, B, and C onto enterprise status (example)

(greater number of 'dt' steps required) to reach equilibrium of the model. The 'dt' (as the step) stands only for a technical factor resulting from the calculations, so in the approach presented in this manuscript it doesn't convey the time elapsed in case of dynamics per se. Use of differential equations in SD (to reflect and simulate SSD) is nothing more than a computational approach describing the state of an object(s) with inputs and outputs, whilst 'dt' is just the step in calculations. If 'dt' is small enough (used 1/4 in calc. for RK4), so that it does not interfere with the shortest impulse in the modelled system, then 'dt' does not affect the calculated results (acc. to Forrester, Coyle). Practically obtained results during conducted experiments may testify to this, as the confirmation of the statement about required character of 'dt'. In it there is as well an analogy to the Nyquist theorem in the theory of signals, stating that the sampling frequency f_p should be at least twice as high as the frequency f_s of the sampled signal ($f_p \geq 2f_s$). In conclusion however, conducted experiments with the use of system dynamics methodology and its specifics[16], can't yet fully confirm the usefulness and applicability of SD for SSD [17] [18]. Presented approach in the manuscript needs some further research and tests

[16]SD is above all intended (with assumptions about 'dt' step) for modelling dynamic behaviours, which is essentially the 'eigen value' of system dynamics

[17]for FCM / WINGS calculations in fact nothing is modelled in time

[18]for SD - time strictly matters in dynamic relations to observe their evolution over time

with different sets of models and data, to strive to create potentially a new coherent, universal and trustworthy 'environment-method', both for qualitative and quantitative calculations.

3 ACKNOWLEDGEMENTS

The author of this system dynamics would like to express his gratitude to the Professor Jerzy Michnik, the author of the WINGS method, for WINGS calculations and very insightful, and valuable comments that significantly helped to improve this manuscript.

References

[1] Abraham-Frois G. (1998): Non-linear Dynamics and Endogenous Cycles. Springer, Lecture Notes in Economics and Mathematical Systems; 463, str 168,169; https://www.springer.com/la/book/9783540643210

[2] Banas, D., Michnik, J. (2019): Evaluation of the Impact of Strategic Offers on the Financial and Strategic Health of the Company - A Soft System Dynamics Approach; Mathematics 7(2), 208; doi.org/10.3390/math7020208

[3] Banas, D. Michnik, J. Targiel K. (2016): System modeling and control of organization business processes with balance scorecard and system dynamics, Control and Cybernetics Journal of Polish Academy of Sciences, Control and Cybernetics vol. 45 No. 2; http://control.ibspan.waw.pl:3000/contents/export?filename=Banas-Michnik-Targiel..pdf

[4] Coyle, G. (2000): Qualitative and quantitative modelling in system dynamics: some research questions. System Dynamics Review, 16(3), 225-244; doi.org/10.1002/1099-1727(200023)16:3¡225::AID-SDR195¿3.0.CO;2-D

[5] Forrester, J.W. (2013): Industrial Dynamics; Martino Fine Books; https://www.amazon.com/Industrial-Dynamics-Jay-Wright-Forrester/dp/1614275335

[6] Kosko, B. (1986): Fuzzy cognitive maps. International journal of Man-Machine Studies, 24, 65-75; doi: 10.1016/ S0020-7373(86)80040-2.

[7] Luna-Reyes, L. F. Andersen, D. L. (2003): Collecting and Analyzing Qualitative Data: methods and models System Dynamic Review 19, 271-296. doi.org/10.1002/sdr.280

[8] Mendoza, G.A.; Prabhu, R. (2006): Participatory modeling and analysis for sustainable forest management: Overview of soft system dynamics models and applications. Forest Policy and Economics 9,179-196.1; doi.org/10.1016/j.forpol.2005.06.006

[9] Michnik, J. (2013): Weighted Influence Non-linear Gauge System (WINGS) - An analysis method for the systems of interrelated components. European Journal of Operational Research 2013,228, 536-544; doi.org/10.1016/j.ejor.2013.02.007

[10] Mohammadreza, Z., Georges, A., Romme, L., Walrave, B. (2018): Why, when, and how to combine system dynamics with other methods: Towards an evidence-based framework; doi.org/10.1080/17477778.2017.1418639

[11] Karczmarczuk, J. (2012): Rownania rozniczkowe i komputery; FOTON119; http://www.foton.if.uj.edu.pl/documents/12579485/e5c1c244-5564-4eef-9252-5528eeb86e93

MODELLING THE KIDNEY TRANSPLANT WAITING LIST

Katarína Cechlárová
Institute of Mathematics, Faculty of Science, P.J. Šafárik University, Jesenná 5, Košice, Slovakia
E-mail: katarina.cechlarova@upjs.sk

Diana Plačková
Institute of Mathematics, Faculty of Science, P.J. Šafárik University, Jesenná 5, Košice, Slovakia
E-mail: diana.plackova@student.upjs.sk

Tatiana Baltesová
L. Pasteur University Hospital, Trieda SNP 1, Košice, Slovakia
E-mail: tatiana.baltesova@unlp.sk

Abstract: We created a dynamic stochastic model to evaluate the performance of a kidney transplantation system with respect to various parameters that include different rates of deceased kidneys harvesting, the proportion of patients with a willing living donor and different allocation policies used. Our model is applicable in the context of a small country and clearly demonstrates the efficiency gains brought about by inclusions of kidney exchange programs.

Keywords: transplantation, waiting list, kidney exchange, stochastic modelling

1 INTRODUCTION

Chronic kidney disease is a world-wide problem. It is estimated that approximately 0,1% of the world population suffer from end stage renal disease (ESRD). For ESRD patients, two treatment options are available: dialysis or transplantation. The country of the authors has the population of about 5.5 million and in 2017 there were around 4 500 patients receiving regular dialysis treatment [9].

While dialysis is associated with a low life quality and many undesired side effects, a patient after a successful kidney transplantation can lead a practically normal life, apart from life-long immunosupressive treatment.

Transplantation is impossible without donors. An organ from a deceased donor (DD) can be used, but availability of such an organ is very much unpredictable. Patients are listed in a transplant waiting list for deceased donors (DDWL) and in each country there is a national authority responsible for its management and allocation of kidneys. Further, as it is proved that one kidney is perfectly sufficient for life, a healthy person can donate one of his/her kidneys to a patient in need. Moreover, the long term function of kidneys from living donor is significantly better comparing to cadaveric kidneys. With the improvement of surgical techniques that minimize the risk for a donor, living kidney transplantation has become a treatment of choice.

Table 1 depicts the numbers of patients registered in the DDWL and the numbers of transplantations in Slovakia for the last 10 years [10].

Table 1: Statistics on the DDWL and transplantations in Slovakia

	2009	***2010***	***2011***	***2012***	***2013***	***2014***	***2015***	***2016***	***2017***	***2018***
Patients in the DDWL	589	493	478	508	471	504	426	386	377	257
Transplantations										
deceased donors	153	156	116	130	108	110	165	124	142	135
living donors	19	7	13	3	10	15	19	19	11	11

To prevent graft rejection, some immunological prerequisites have to be fulfilled. First is the AB0 blood type compatibility: a donor of blood group 0 can donate to anybody, a donor with blood type A only to recipients with blood group A or AB, a B-donor only to B or AB recipients and an AB donor only to recipients of the same blood group. The second condition for transplantation is negative crossmatch. This test confirms that there are no clinically relevant antibodies against donor's HLA antigens in the recipient's serum. Although various desensitization protocols allowing transplantation across the blood group and HLA barrier are available, they are very expensive and the transplantation results are less favourable for the patients.

The aim of this paper is to model how the size of the DDWL and waiting times for transplantation could evolve under various assumptions. We modelled a stationary patients population and different intensities of deceased donors arrival. Then we assumed a certain proportion of patients bringing their willing donors – a genetically or emotional relative. To increase the utilization of living donors, we also employed the idea of kidney paired donation [2]. The results can help to estimate the influence of various parameters and policies on the donation process in a small country.

2 RELATED WORK

There are many scientific papers that analyze the dynamic process of organ donation. We would like to draw the reader's attention only to a selection of them.

Among the first works to use queueing models in connection with DDWL was paper [7] by Zenios. The author assumed that there are several classes of patients, several classes of organs, and patients reneging due to death. Focusing on randomized organ allocation policies he developed closed-form asymptotic expressions for the stationary waiting time, stationary waiting time until transplantation, and fraction of patients who receive transplantation for each patient class.

Zenios in [8] considers a question of how to dynamically manage the mix of direct (between two incompatible patient-donor pairs) and indirect exchanges (the living donor donates to DDWL and the associated recipient receives a kidney fom a DD – these are assumed to be available immediately) to maximize the expected total discounted quality years of the candidates in the participating pairs. Direct exchanges are preferable because the candidate receives a living-donor organ instead of the inferior cadaveric organ an indirect exchange provides. However, the latter involves a shorter waiting. To address this question, Zenios developed a double-ended queueing model for a simple exchange system with two types of donor-recipient pairs. The optimal policy takes the form of a two-sided regulator with newly arrived pairs enrolling in the exchange system to wait for a direct exchange as long as the double-ended queueing process is between the two barriers.

Stanford et al. [5] address the 'blood type 0 problem', which means that recipients of blood type 0 experience longer waiting times than those of other blood types, partly due to cross-transplantation of too many 0 organs to compatible donors of other blood types. The authors show that AB0 identical transplantation cannot achieve equity either, so they present a model for restricted cross-transplantations to achieve comparable waiting times for all blood types.

The model developed in [1] combines the information of the arrival of patients to the DDWL and the process of donation, while the authors assumed that a DD can provide either one or two kidneys. The authors used Bayesian inference and compared the results obtained by the modelling with the real evolution of the DDWL in País Valenciá (Spain).

Segev et al. in [4] followed a cohort of simulated patients through a kidney paired donation (KPD) program for several years to predict median waiting times for various blood types and sensitization subgroups of recipients. Their simulations suggest that the most cost-efficient modality for transplanting incompatible patient-donor pairs is a national kidney paired donation program (KPD) utilizing an optimized matching algorithm. In a recent work, Santos et al. [3] took into consideration various policies of KPD programs found in practice (incompatible pairs, altruistic donors, and compatible pairs) and various matching policies. Their final results show that shorter time intervals between matches lead to higher number of effective transplants and to shorter waiting times for patients. Furthermore, the inclusion of compatible pairs can lead to greater benefit for 0-blood type patients.

3 MODEL

To start our dynamic model in time zero, we used the data of real patients that were registered in the national DDWL on May 18, 2019 in the following structure: the date of registration on the DDWL, blood group and the panel reactive antibody (PRA%) level. The number of these patients was 278.

We modelled the arrival of patients and DDs as two independent Poisson processes (see [6]) with parameters λ_p and λ_d; the used time unit was one month. Parameter λ_p=153/12 has been estimated on the basis of clinical experience. For the flow of DDs we used two different values λ_d=108/24 and λ_d=165/24 to capture the most pesimistic as well as the most optimistic scenario based on the National Transplant Organization [10] data from the last 10 years. For each arriving patient and DD we randomly generated his/her blood group according to the blood groups distribution in the Slovak population, taken from [11]. In Slovakia, the most prevalent blood group is A with 42 %, then 0 with 32 %, blood groups B and AB correspond to 18 % and 8%, respectively. For each arriving patient we also randomly generated the level of his/her sensitization. H (high sensitization) patients are those with the PRA level above 80, the other patiens are considered to be L (patients with low sensitization). The proportion of H patients in our starting sample was 8%, so we used this probability.We generated 50 samples of patients flows and each one was paired with one sample of DD flow with a low and one sample of DD flow with a high arrival rate as specified above.

To be able to model living donation, for each patients arrival sample we also randomly generated a willing living donor for 20% and 40% of arriving patients. The donor's blood group was generated in the same manner as for patients, and we also randomly generated a positive crossmatch of the donor with his/her intended recipient, different probabilities were used for H and L patients.

We allocated DDs using the FIFO (first-in-first-out) principle. That means that when a deceased donor *d* arrives, we process the patients on the DDWL that have not yet received a transplant in the order of their entry date to the DDWL. First we check the blood group compatibility of the patient with the arriving donor. If they turn out not to be compatible, we do not proceed with transplantation. Otherwise we randomly generate a catch-all variable β with a uniform distribution in [0,1] interval that captures tissue type compatibility, possible cross-match and other immunological and health associated factors in the patient. If β was higher than 20 for a H patient and higher than 80 for a L patient then no transplantation is possible. The first patient *p* for whom both tests give green light, is marked as transplanted, she/he is remowed from DDWL and we record the time of the arrival of *d* as the time when *p* was transplanted.

To assess the influence of various policies on the performance of the kidney transplantation system we considered 4 different models.

Model 1 assumes only DDs and it is described above.

Model 2 considers living donors in addition to DDs. If a patient *p* with his/her willing living donor *d(p)* arrives then donation of *d(p)* to *p* is attempted in the same manner as with a DD. If this is not possible, donor *d(p)* is lost and patient *p* is put into DDWL. By contrast, if this donation is not excluded, we assume that it is performed immediately, i.e., the waiting time of patient *p* is 0.

Model 3 uses the idea of kidney exchange, first suggested by Rapaport in 1986 [2] and now a part of kidney transplantation programs in many countries [12]. Incompatible patient-donor pair X is kept in the database and they wait for another incompatible patient-donor pair. When another such pair Y arrives, kidney exchange is attempted. This means that we check whether the donor of pair X can donate to the patient of pair Y and simultaneously whether the donor of pair Y can donate to the patient of pair X. If the compatibility tests are positive, paired donation is performed at the time of the arrival of the second pair and both patients are removed from DDWL. In this model, the patient of an incompatible pair X is kept in DDWL and he/she is considered as a candidate for a cadaveric transplantation in accordance with Model 1 for each arriving DD.

Model 4 evaluates what happens, if patients of incompatible patient-donor pairs are not considered for cadaveric transplantation for some time, as is the case in many countries with a functioning kidney paired donation program. We set this time to be 3 months. After that time, the incompatible pair still wait for another incompatible pair but the patient is treated in the same way as other patients in DDWL.

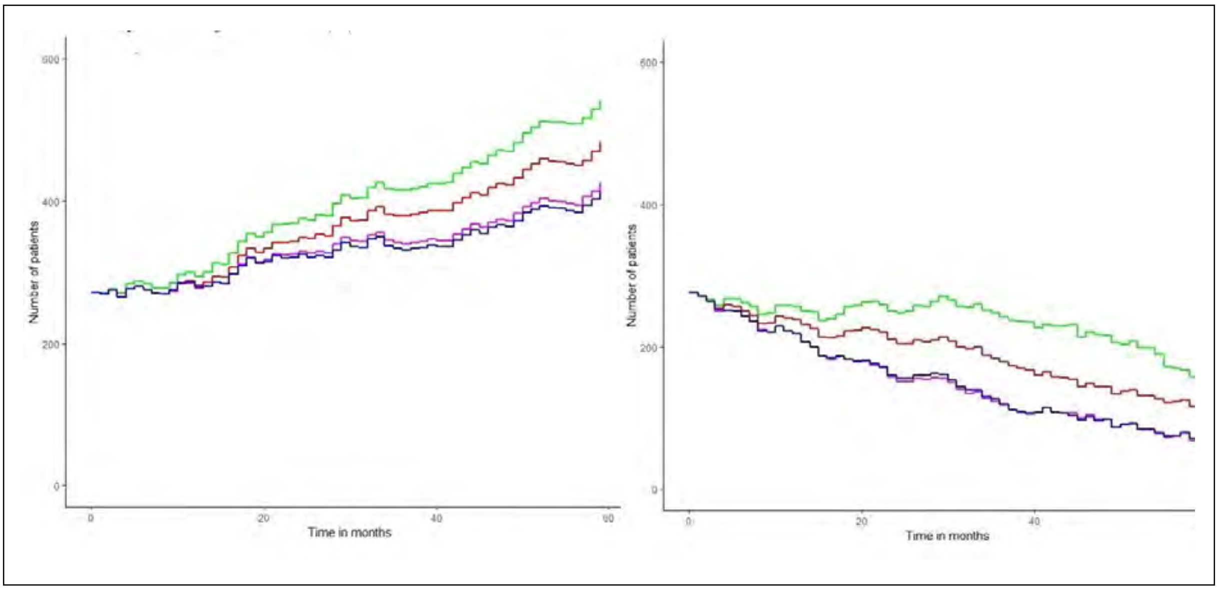


Figure 1: The evolution of the number of patients in the DDWL in time. Left : 108 DDs yearly and 20% patients with a LD. Right :165 DDs yearly and 40% patients with a LD. The green line correspond to DD only, the red line to Model 2, the pink line to Model 3 and the blue line to Model 4.

4 RESULTS

We generated 50 samples of patients arrival flows and 50 samples of DD arrival flows with two different intensities. We assumed that each DD provides both kidneys for transplantation, and we chose these parameters so as to the obtained the average monthly flows of organs correspond to the minimum and maximum number of cadaveric transplantations that were performed in the last 10 years. We simulated the evolution of the process in the timespan of 5 years. Figure 1 depicts the evolution of the DDWL over the considered time span for one of the generated samples in the most pesimistic (left) and the most optimistic case (right).

The number of patients in the generated samples varied from 991 to 1095 and the number of DDs between 473 and 605 for λ_d=108/24 and between 715 and 924 for λ_d=165/24. The average number of cadaveric transplantations in the former case was 533.48 and in the latter case 834.96 (when no living donation was considered). Living donors contributed on average 68.14 transplantations which increased to 117.7 with the introduction of KPD (20% patients with LD); for 40% patients with LD these figures are 132.4 and 242.34 respectively. The average and median waiting times obtained in our simulations are given in Tables 2 and 3.

Table 2: Waiting times of patients on the DDWL in the simulations for 108 deceased donors yearly

		20% patients with LD			*40% patients with LD*		
	DDs only	***Model 2***	***Model 3***	***Model 4***	***Model 2***	***Model 3***	***Model 4***
All patients average	41.18	35.85	33.47	33.50	31.78	27.28	27.41
median	36.60	30.31	27.42	24.41	26.00	21.11	21.26
L patients average	39.13	33.81	33.47	31.61	29.77	25.62	25.75
median	35.44	25.75	27.07	26.53	25.52	20.84	20.50
H patients average	64.86	60.78	55.90	56.18	57.20	47.46	47.49
median	43.58	40.87	37.22	38.56	38.57	27.54	26.88

Table 3: Waiting times of patients on the DDWL in the simulations for 165 deceased donors yearly

		20% patients with LD			*40% patients with LD*		
	DDs only	***Model 2***	***Model 3***	***Model 4***	***Model 2***	***Model 3***	***Model 4***
All patients average	25.70	22.71	21.22	21.21	20.46	18.108	18.11
median	20.27	16.63	14.71	14.65	14.31	11.30	11.21
L patients average	24.30	21.33	19.89	19.88	19.10	16.82	16.83
median	19.95	16.24	14.53	14.38	14.12	11.02	11.05
H patients average	41.60	38.61	36.40	36.47	36.06	32.64	32.78
median	23.21	21.00	18.64	18.28	18.99	14.08	13.76

5 DISCUSSION AND CONCLUSION

Besides of medical and logistic issues associated with transplantations that have to be solved, there are ethical, religious and legal questions which have to be answered to achieve positive attitudes in the society towards transplantations and organ donation. Mathematical modelling can help in weighing various pros and cons in that it has the power to visualize the possible outcomes of various scenarios before they have been observed in reality.

We have presented a 'what if' analysis to see what happens with the evolution of the numbers of patients in the waiting list and the waiting times for transplantation if various parameters change. Specifically, we considered the average number of deceased donors and the average percentage of patients with living donors. We also estimated the outcomes of various allocations policies.

As expected, the increased number of DDs has shortened the waiting time for transplantation by almost 40%. Living donors have a great potential in cutting down waiting times further: if there is a shortage of DDs, the average waiting time of all patients shrinks by almost 6 months if approximately one in 5 patients has a willing living donors. If the number of willing living donors doubles, even by astonishing 10 months, which is one quarter of the original waiting time. Even in societies when the supply of DDs is sufficient, the average waiting time decreases by 2 months in the former case and by more than 5 months in the latter (20% decrease).

Our simulations also showed a discouraging phenomenon: less than one half of the wiling living donors can be used. Notice that implementing the policy of kidney paired donations the utilization rate increased to 74%-79%. In most countries with a developed KPD program, the recipients who register in this program are usually temporarily off listed from DDWL [12]. Our simulations, when one compares Model 4 to Model 3, show very small increase in the number of transplantations and practically no shortening of waiting times, rather its slight increase in some cases.

Notice also the large gap in average as well as median waiting times between highly sensitized and other patients. This calls for a modification of the used allocation policy, that was in our case the simple FIFO. In a follow up work we shall try to see the effects of some prioritization rules for highly sensitized patients.

Acknowledgement

This work has been supported by VEGA grant 1/0311/18, APVV grant APVV-17-0568 and COST action CA15210 ENCKEP. The authors would like to thank Daniel Kuba and Magdaléna Krátka, National Transplant Organization, for providing the details of the composition of the Slovak DDWL.

References

[1] Abellán, J., Armero, C., Conesa, D., Pérez-Panadés, J. Martinez-Beneito, M., Zurriaga, O., Garcia-Blasco, M. J., Vanaclocha, H. 2016. Analysis of the renal transplant waiting list in the paísValenciá (Spain). Stat.Med., 25: 345-358..

[2] Rapaport, F. 1986. The case for a living emotionally related international kidney donor exchange registry. Transplant Proceedings 18(3): 5-9.

[3] Santos, N., Tubertini, P., Viana, A., Pedroso, J. 2017. Kidney exchange simulation and optimization. Journal of the Operational Research Society 68:1521-1532.

[4] Segev, D. L., Gentry, S. E., Melancon, J.K., Montgomery, R.A. 2005 Characterization of waiting times in a simulation of kidney-paireddonation.American Journal of Transplantation, 5: 2248-2455.

[5] Stanford, D.A., Lee, J. M., Chandok, N., McAlister.V.C. 2014. A queueing model to address wait time inconsistency in solid-organ transplantation. Oper. Res. Health Care 3: 40-45.

[6] Tijms, H.C. 2003. A First Course in Stochastic Models.Wiley. Chichester.

[7] Zenios, S.A 1999. Modeling the transplant waiting list: A queueing model with reneging. Queueing Systems 31: 239-251.

[8] Zenios, S.A. 2002. Optimal Control of a Paired-Kidney Exchange Program. Management Science 48(3): 328-342.

[9] Health Statistics Yearbook 2017. National Health Information Center. Web document available at http://www.nczisk.sk/. (accessed 15 June 2019).

[10] Statistics on transplantations in Slovak Republic. National Transplant Organization.Web document available at http://www.nto.sk/. (accessed 15 June 2019).

[11] Národná transfúzna služba (National blood service). Web document available at http://www.ntssr.sk/. (accessed 15 June 2019).

[12] European Network for Collaboration on Kidney Exchange Programmes. Web document available at http://www.enckep-cost.eu. (accessed 15 June 2019).

FRAMEWORK FOR DISCRETE-EVENT SIMULATION MODELING SUPPORTED BY LMS DATA AND PROCESS MINING

Associate professor Mario Jadrić, PhD
University of Split, Faculty of Economics, Business and Tourism, Department of Business Informatics
Cvite Fiskovića 5, Split, Croatia
mario.jadric@efst.hr

Abstract: The use of Learning Management Systems (LMS) generates a large amount of data that could, among other purposes, be used to uncover hidden processes reflecting student learning behaviour. To gain valuable insights from the collected data and activities that occurred in the process, techniques such as data and process mining need to be employed. Apart from presenting the advantages of the advanced detection of student behaviour and processes based on the data from LMSs, this paper proposes a conceptual model for using the techniques to support the development of discrete-event simulation models in an educational environment. Additionally, research and practical challenges for successful integration of the presented concepts and techniques are elaborated.

Keywords: learning management systems, learning analytics, educational data mining, educational process mining, discrete-event simulation model

1 INTRODUCTION

Studies confirm the potential of educational data analysis to anticipate existing and emerging challenges for educational institutions and students, provide a basis for decision-making on limited resources and pedagogical models, and propose a structure for improving academic results [8]. Knowledge discovery from the databases can result in anticipating student enrolment on a particular course, discovering cheating in an online assessment, revealing extreme values in achieved student results, and similar examples, all leading to increased quality in educational environments [2]. It is obvious that an increasing amount of data can be found in university databases, so the main goal of the educational data mining (EDM) is to use a large amount of collected educational data that comes from different sources, in different formats and at different levels of detail, in order to better understand the learning process and improve its outcomes [9]. It is becoming a common practice that standard e-learning systems such as Moodle have already built-in mechanisms for tracking log files for teachers, administrators and students [20]. Furthermore, in exploring student behavior in higher education, one research area is specially focused on the analysis of the data logs from information systems using tools for process mining [6]. On the other hand, simulation is often used for understanding, analyzing and predicting the behavior of complex and large systems. It is used to better understand existing systems and to facilitate the design of new systems by predicting their future behavior.

In the paper, first, an overview of existing literature through the perspective of LMS data analysis, educational process mining and discrete-event simulation modeling is provided. Then the results of the LMS educational data analysis are presented using the process mining. This is followed by a discussion of the potential of using the obtained process models to support simulation modeling.

2 THEORETICAL BACKGROUND

2.1 LMS Data Analysis

LMS supports delivery and management of content, identification, and evaluation of learning objectives, communication, and providing all the information needed for learning. Knowledge hidden in educational data sets can be extracted with data mining techniques. For example, in an earlier study in order to better understand the problem of student dropout, we processed the data available from the higher education information system using several methods of data mining: logistic regression, decision trees and neural networks [10]. The models were built following the Sample, Explore, Modify, Model, Assess (SEMMA) methodology and then compared in order to select the method that best predicts student dropout.

A wide range of available educational information that can be analyzed and used to make the best decisions about different learning issues highlights learning analytics as a new research area. Recent trends in this field are the application of data mining tools and techniques used in large data analytics projects for more rational data-driven decision-making in educational contexts [21]. The ability to extract useful patterns from educational databases is extremely useful, but one of the limitations of such studies is that they do not analyze the offline activities, because a large number of online materials will be printed and used in a paper version. In addition, we can expect that a certain number of students will attempt to exploit the system failures or applied methodological approaches in an uncontrolled environment. For example, in a study that analyzed student behavior in the LMS by focusing on the aspect of online self-assessment, it became apparent that students under the conditions of possible multiple attempts to access the test largely opted for the method of "trial and error" instead of learning the course content [11]. The results obtained in another experimental study [20] are very useful for classifying problems related to students as well as for discovering other interesting patterns in the Moodle log data. For example, another study [5] shows several approaches for the educational data mining including K-means clustering, multiple regression, and classification, that was used to investigate and predict final grades and completion rates. In summation, LMSs archive detailed logs of students' activities in e-learning courses and represent a rich-data environment for the use of educational data mining methods with many practical implications.

2.2 Educational Process Mining

Educational process mining (EPM) is a relatively new research area within educational data mining (EDM). The aim is to make the hidden knowledge explicit using logs collected from educational environments in order to analyze and provide a visual representation and better understanding of the educational processes [3]. The first type of process mining is a *discovery* that uses event logs and creates a process model without using a priori information. The process algorithm constructs a process model that reflects the behavior observed in the event log. Another type of process mining is *conformance*. In this technique, the existing process model is compared with the event log of the same process. Compliance check can be used to check whether the actual data are recorded in the logs, according to the model and vice versa. The third type of process mining is an *enhancement*. This technique is used for expanding or improving the existing process model using information about the actual process recorded in an event log, with a goal to change or extend a priori model [6]. As can be seen, the basic idea of process mining is to detect, track and improve the actual processes by extracting knowledge from logs that are automatically collected in various information systems. The main goals of process mining in educational environments are [9]: (1) extracting knowledge related to educational processes from large event logs, such as process models that tracks key educational performance indicators; (2) analysis of educational processes and their compliance with the curriculum; (3) improvement of the educational process model; (4) personalization of educational processes through the recommendation of the best learning paths depending on students profiles, their preferences or target skills.

EPM is used in a wide range of educational settings. As the process mining is getting matured, and the information becomes more accessible, organizations are becoming more interested in comparative mining to understand how processes can be improved. Therefore, there have been suggestions [19] to use process cubes as a way to organize event data in a multidimensional data structure tailored toward process mining. This allows comparison of different process variants or different case groups. The initial step is usually to visualize the students' behavior using process mining and to determine the probability of their graduation [7]. The process mining has been successfully used to detect, track and improve processes, based on the information from event logs and other traces. A study [6] provided an insight into the potential of process mining techniques in the context of higher education systems such as detection of extreme values in student achievements, the anticipation of students' success, drop-out identification, and detecting students that need special attention. There are other examples of using process mining techniques: to monitor and analyze learning behaviors based on data from MOOC [15], and similarly to improve the participants' experience in MOOC [18].

2.3 From Process Mining to Discrete-Event Simulation Modelling

The framework presented in a recent study [1] integrates process mining techniques into the conceptual modeling phase as one of the steps in developing the simulation model. The proposed hybrid framework was used in the context of an emergency department in order to determine performance bottlenecks and explore possible improvement strategies. The hybrid approach overcomes issues of traditional conceptual modeling using process mining techniques in order to reveal valuable process knowledge from the event log analysis. Developing conceptual models that use traditional data sources such as expert interviewing is definitely the longest stage in the simulation modeling process. However, the use of process mining does not necessarily exclude traditional techniques, as the results of the process mining can be verified by experts to determine any mistakes or inaccuracies in data sets that were the basis for process mining [1]. Similarly, other authors [17] use a combination of process mining techniques to detect control-flow, data, performance and process resources based on historical data, and integrate them into a comprehensive simulation model presented as a Coloured Petri Network (CPN) to be used for process analysis and performance evaluation of different models.

3 FRAMEWORK FOR DISCRETE-EVENT SIMULATION MODELLING

For the purpose of this study, student behavior data was collected at University of Split, Faculty of Economics, Business and Tourism from the Moodle LMS. To prove the feasibility of the approach, only a segment of collected data from the course “Information Technology” in the academic year 2018/19 was used. In the course, students access resources and activities (read texts, watch video materials, complete surveys, and so on) in the order and dynamics that best suited them. Selected data from Moodle reports were then exported to a .xls file that is appropriate for further analysis in the data and process mining tools.

Process mining as a technique can be used to analyze all data that consists of sequences of different types of activity. Activities that occur by re-engaging different students during the semester in the learning process, such as browsing, deleting, modifying, adding content to the Moodle course, were the basis for conducting the analysis. In this study, the process begins with enrolling students in the course and continues with different activities while students’ progress through the course. For the purpose of the analysis, Disco process mining software has been used to detect the patterns of student behavior in the form of process maps. The tool enables automatic discovery of a process model based on imported data using an optimized high-speed process discovery algorithm (available at https://fluxicon.com/disco). The resulting process maps were relatively intuitive and easy to use, while the software enables a dynamic view of the process. Due to a sizable report, a large number of process maps that can be generated and the impracticality of displaying them within this paper, below only a small sample of the processed data is provided for an arbitrarily selected scenario for the first two days of the course “Information Technology” (Figure 1). Notably, it is possible to apply numerous algorithms for process mining to visualize the actual students’ behaviors. Disco provides a complete set of process metrics for activities and paths, like case frequency, maximum number of repetitions, total duration, mean duration, maximum duration, and more.

The model in Figure 1 shows performance in terms of activity duration and transition between them. In this scenario, the relative frequencies for the resources are: System = 50%, File = 20%, Test = 10%, Task = 10%, Forum = 10%), while for activities “View e-Course - Information Technology” 50% and for the remaining five activities 10%. In addition, the software automatically computes statistics for activities, resources, context and activity attribute name, for example for activity “View e-course - Information Technology”: max repetition = 5, min = 1, median = 1, mean = 1.67, standard dev. = 1.63.

A detailed list of contexts, activity attribute names, and resources, together with the timestamp for this arbitrary case is listed in Table 1.

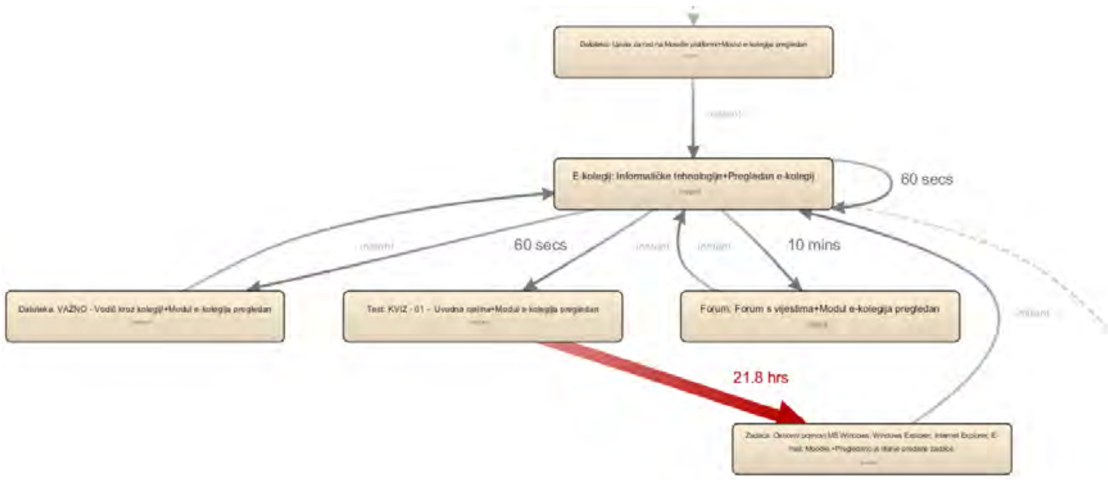


Figure 1: Arbitrarily selected scenario for the first 2 days of accessing the course

Table 1: First 5 activities of arbitrary scenario with resources, date and time associated

Context	*Activity*	*Resource*	*Date*	*Time*
Instructions for Moodle	Module viewed	File	09.10.2018	21:46:00
Information Technology	E-course viewed	System	09.10.2018	21:46:00
Course Guide	Module viewed	File	09.10.2018	21:47:00
Information Technology	E-course viewed	System	09.10.2018	21:47:00
Quiz – Introduction	Module viewed	Test	09.10.2018	21:47:00

The information presented in the process map (Figure 1) and in Table 1 represents input data for the discrete-event simulation modeling. Table 2 shows descriptive statistics and distribution fit for this segment of selected data with a number of accessed activities in the LMS course "Information technology" for the first two days of the semester. Attention should be drawn to the fact that the distribution presented here with respect to the smaller processing segment and the large variations in the observed data are not statistically significant and only serve to indicate the potential of their usage in simulation modeling, as explained in the next paragraphs.

Table 2: Input data analysis of the LMS e-course access data

Context	*Count*	*Min*	*Max*	*Mean*	*StD*	*Distribution*	*Expression*
Instructions	48	1	3	1.26	0.503	Erlang	0.5 + ERLA(0.191, 4)
Course Guide	147	1	8	1.79	1.22	Exponential	0.5 + EXPO(1.29)
Infor. Techn.	1404	1	75	6.08	6.99	Weibull	0.5 + WEIB(5.41, 0.942)
News forum	186	1	13	2.04	1.73	Exponential	0.5 + EXPO(1.54)
Quiz – Intro.	955	1	121	9.85	21.4	Weibull	0.999 + WEIB(0.354, 0.224)

The diagram in Figure 2 presents a framework for discrete-event simulation modeling supported by data and process mining in an educational environment. A brief explanation of the framework is given in the paragraphs below. A random number generator is used to generate a random number set. The input model is most often determined by the analysis of the *Goodness of Fit Tests* and is used to transform the set of random numbers into a set of input data. This transformation process is called "*generating random variables*". Data collection procedures should follow standard sampling practices in order to collect data on the real system (educational environment) in a representative manner. Input data is used to determine the distribution of random input variables (as presented in table 2), but also to create a conceptual and simulation model, and finally for validation purposes. Apart from the LMS system data, sources in this example can include other databases related to students, manually recorded data, or sampling studies.

The conceptual model is created based on process discovery results and on some other assumptions about the modeled system. However, the conceptual modeling step can be avoided in elementary models and the development of a simulation model can be started directly from the process map. In the next step, output data is generated by applying a simulation model to a set of input data, while, statistical analysis is performed on the output data, which includes, for example, calculation of the arithmetic mean, median or variation and confidence intervals. It is also necessary to conduct a simulation model validation by comparing the output data of a simulation model with the output of the real system and the results of the process discovery.

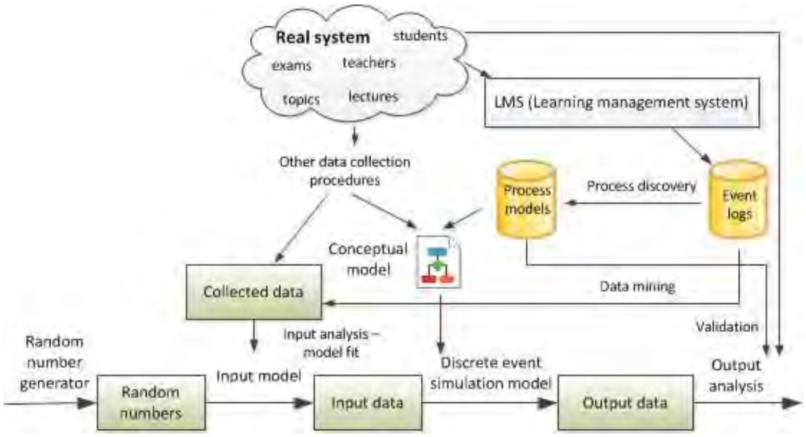


Figure 2: Framework for discrete-event simulation modeling supported by LMS data and process mining

The list of resources and activities for the simulation entity "student" together with their timestamps, descriptive statistics, and distributions that can be used in the phases of the input model development, the conceptual and discrete-event simulation model development, as well as in the validation procedure are shown in Table 1 and Table 2. The key challenges in this process of connecting several research areas [14] are associated to: (1) entity modelling – selection of relevant attributes, types and distribution; (2) activities modelling – defining activities, mining the duration of activities, determining the conditions of branching, defining the discipline of the order and closing events; (3) control flow modelling; (4) resource modelling – rules for assigning activities, retrieving a predefined resource schedule, defining unavailability, etc. From all of the above, it is evident that input data is a key element of simulation modeling and is used to set simulation parameters and variables, from conceptual modeling to final validation of the simulation model.

4 CONCLUSION

Data mining brings about the ability to extract useful patterns from educational databases while process mining detects control-flow, data, performance and process resources based on historical data. However, the standard process mining and its results are not enough for improving the process in question since there is little opportunity for experimentation with parameters, what-if analyses, forecasting, leading to the need for simulation modeling and value added by the proposed conceptual model. Despite the potential of processes mining to support the creation of discrete-event simulation models, there are still many research challenges to successfully integrate these two areas. It is also necessary to add that the use of learning analytics and process mining on educational data opens up numerous ethical, legal and security issues [16].

By implementing the proposed conceptual framework using a larger dataset. It would be possible to generalize and extrapolate the results for a specific group or generation of students for the purpose of analysis and predicting of students' success, drop-out identification, and detecting students that need special attention. The limitation of this paper is certainly the simplification of the process discovery step by presenting only one learning scenario based on Moodle LMS data. Concerning future research, it would be particularly interesting to operationalize the presented conceptual model and even explore the possibilities of using process mining in some other contexts. Bearing in mind the findings [21] highlighting the potential of integrating learning analytics with smart education and smart library services in the context of emerging concepts such as smart cities, the potential of using process mining as a support to simulation modelling will certainly become increasingly significant, especially in the context of big data and the growing need for data-driven decision making. At the moment, the most important area for the use of discrete-event simulation in the context of smart cities is mobility [12], however, it is to be expected that other data-rich and smart environments (like healthcare, security, water, waste, education) will become more suitable cases for application of process mining coupled with discrete-event simulation modeling.

Acknowledgment
This work has been supported by the Croatian Science Foundation under the project UIP-2017-05-7625.

References

[1] Abohamad, W., Ramy, A., & Amr, A. 2017. A hybrid process-mining approach for simulation modeling. In Proceedings of the 2017 Winter Simulation Conference: 1527–1538.
[2] Baradwaj, B. K., & Pa, S. 2011. Mining Educational Data to Analyze Students‟ Performance Brijesh. International Journal of Advanced Computer Science and Applications, 2(6):63–69.
[3] Bogarín, A., Cerezo, R., & Romero, C. 2018a. A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(1).
[4] Bogarín, A., Cerezo, R., & Romero, C. 2018b. Discovering learning processes using inductive miner: A case study with learning management systems (LMSs). Psicothema, 30(3): 322–329.
[5] Choi, H., Lee, J. E., Hong, W., Lee, K., Recker, M., & Walker, A. 2018. Exploring Learning Management System Interaction Data: Combining Data- driven and Theory-driven Approaches. Proceedings of the 9th International Conference on Educational Data Mining, 324–329.
[6] Devi, K. L., M.Phil., M.Suryakala, & M.Phil. 2014. Educational Process Mining-Different Perspectives, 16(1): 57–60.
[7] Douzali, E., & Darabi, H. 2016. A case study for the application of data and process mining in intervention program assessment. ASEE Annual Co., 2016.
[8] Grant, M. R. 2012. University of Missouri- St. Louis: Data-Driven Online Course Design and Effective Practices Mary. Continuing Higher Education Review, 76: 183–192.
[9] Grigorova, K., Malysheva, E., & Bobrovskiy, S. 2017. Application of Data Mining and Process Mining approaches for improving e-Learning Processes. 3rd International Conference on Information Technology and Nanotechnology (ITNT-2017): 115–121.
[10] Jadrić, M., Garača, Ž., Ćukušić, M. 2010. Student Dropout Analysis with Application of Data Mining Methods. Management : Journal of Contemporary Management Issues, 15(1): 31–46.
[11] Jadrić, M., Ćukušić, M., & Garača, Ž. 2017. Analysing Students' Behavior Patterns in Online Assessment. In CECIIS, Varaždin, Croatia: 83–90.
[12] Jadrić, M., Ćukušić, M., Pavlić, D. 2019. Review of discrete simulation modelling use in the context of smart cities, SSRCI - Smart, Sustainable and Resilient Cities and Infrastructures, Opatija, Croatia
[13] Juhaňák, L., Zounek, J., & Rohlíková, L. 2017. Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. Computers in Human Behavior, 92: 496-506.
[14] Martin, N., Depaire, B., & Caris, A. 2015. The use of process mining in business process simulation model construction structuring the field. Business and Information Systems Engineering, 58(1): 73–87.
[15] Mukala, P., Buijs, J., Leemans, M., & Van Der Aalst, W. 2015. Learning analytics on coursera event data: A process mining approach. CEUR Workshop Proceedings, 1527: 18–32.
[16] Nunn, S., Avella, J. T., Kanai, T., Kebritchi, M. 2016. Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. Online Learning, 20(2): 13–29.
[17] Rozinat, A., Mans, R. S., Song, M., & W.M.P. van der Aalst. 2009. Discovering simulation models. Information Systems, 34(3): 305–327.
[18] Umer, R., Susnjak, T., Mathrani, A., Suriadi, S. 2017. On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative Teaching & Learning, 10(2):160–176.
[19] Van der Aalst, W. M. P., Guo, S., & Gorissen, P. 2015. Comparative process mining in education: An approach based on process cubes. Lecture Notes in Business Information Processing, 203: 110–134.
[20] Verma, A., Rathore, S., Vishwakarma, S. K., & Goswami, S. 2019. Mining CMS Log Data for Students' Feedback Analysis. In Third International Congress on Information and Communication Technology, Advances in Intelligent Systems, Vol. 797: 417–425.
[21] Waheed, H., Hassan, S. U., Aljohani, N. R., & Wasif, M. 2018. A bibliometric perspective of learning analytics research landscape. Behavior and Information Technology, 37(10–11): 941–957.

IT SERVICE BUSINESS ANALYSIS WITH BALANCED SCORECARD AND WEIGHTED INFLUENCE NON-LINEAR GAUGE SYSTEM

Jerzy Michnik
University of Economics in Katowice, ul. 1 Maja 50, 40-287 Katowice, Poland
jerzy.michnik@ue.katowice.pl

Abstract: The paper proposes to apply the extended Weighted Influence Non-linear Gauge System method to support building strategy and decision making. Ability to combine in a uniform way qualitative and quantitative information derived from various sources is the main advantage of this approach. The model of firm's performance has been built within the Balance Scorecard framework. The case of information technology service provider has been used to illustrate the proposed methodology.

Keywords: balanced scorecard, cognitive maps, decision making, system dynamics, Weighted Influence Non-linear Gauge System

1 INTRODUCTION

The Balanced Scorecard (BSC) has been introduced to motivate, measure and evaluate company performance by using four perspectives: 'financial', 'customer', 'internal business process' and 'learning and growth' [4]. Since it introduction BSC becomes very popular among scholars and practitioners. The original paper by Kaplan and Norton counts more than 20 000 citations in Google Scholar.

The long term consequences of various initial states were studied using the system dynamics (SD) approach which was originated in 60s of XX century by Forrester [3] and later widely applied to study many social systems (see eg. [6, 7]).

BSC and SD served as a framework for developing strategic policy in company providing IT services [2]. In their work authors found that some variables in the model like knowledge growth caused by professional trainings or influence of sales and marketing expenses on growth of number of customers were difficult to estimate in strict functional form, what is necessary condition for evaluating the SD model. In view of the problems mentioned above we propose to apply an approach that is better designed to cope with data that are extracted from expert knowledge and have semi-quantitative character. We propose to use WINGS method as a semi-quantitative tool, complementary to SD – the strict quantitative method. Based on BSC, WINGS can help to set up the policy directions and support decision making. The case study of IT service provider from [2] is used to illustrate the potentiality of the WINGS method.

In the following section we present shortly the WINGS method. Sec. 3 contain the description of the model and discuss its results. Conclusions close the main part of the paper.

2 THE WINGS METHOD

The original form of the WINGS method has been introduced in 2013 [5]. The extended (dynamic) version of it appeared recently [1]. Due to lack of enough space we present here only very brief description of it. The WINGS method considers studied system as a set of interacting concepts that makes system something more than the merely sum of its components. In an analysis led with WINGS two kinds of variables are considered as important attributes of concepts: internal strength and influence.

Three main stages of the WINGS method are: (1) qualitative analysis, (2) quantitative evaluation, and (3) calculations (technical stage).

2.1 Qualitative analysis

During this first stage a user, who can be both a single person and a group of people, develops a list of concepts that according to his view play an important role in the studied system behaviour. Then, the user tries to recognise direct relations between concepts. These relations describe the positive or negative impact that may appear between a pair of concepts. The effect of this initial step is summarised in the form of square matrix (containing zero – no impact, plus or minus sign) and corresponding map (digraph).

In the second step the user provides a linguistic scale for the evaluation of influences in the system and the internal strength (power, importance) of each concept. The number of points on the scale depends on the given problem and on the preferences of the user.

2.2 Quantitative evaluation

Using his best knowledge about the problem the user assess the relative power of concepts and strength of non-zero influences. This allows him to assign numerical values to the verbal evaluations that were set up in the previous stage. The values are inserted in the corresponding places in to the matrix and digraph.

2.3 The WINGS output

In this final stage the total values of internal strength and impact are calculated. They are the sum of values of all direct and indirect impacts calculated along all paths of any length that can appear in a WINGS digraph. Technically it means to add together all powers of the matrix of direct impacts (which is in fact the adjacency matrix of a digraph).

We denote the initial matrix containing the evaluated variables (internal strengths and direct impacts) as $\mathbf{D}$, It is an $n \times n$ matrix with elements d_{ij}, where n is the number of components in the system. To provide convergence of the infinite sum we scale matrix $\mathbf{D}$ according to the following formula:

$$\mathbf{S} = \frac{1}{s}\mathbf{D}, \tag{1}$$

where scaling s is defined as a sum of all elements of absolute values of elements d_{ij}:

$$s = \sum_{i=1}^{n} \sum_{j=1}^{n} |d_{ij}| \tag{2}$$

The total strength-influence matrix $\mathbf{T}$ is given by

$$\mathbf{T} = \mathbf{S} + \mathbf{S}^2 + \mathbf{S}^3 + \ldots = \mathbf{S}(\mathbf{I} - \mathbf{S})^{-1}, \tag{3}$$

For each element in the system, the row sum I_i^r and the column sum R_j^r of $\mathbf{T}$ are calculated:

$$I_i^r = \sum_{j=1}^{n} t_{ij}, \quad R_j^r = \sum_{i=1}^{n} t_{ij}, \tag{4}$$

For each element in the system, the row sum I_i^a and the column sum R_j^a of $|\mathbf{T}|$ are calculated:

$$I_i^a = \sum_{j=1}^{n} |t_{ij}|, \quad R_j^a = \sum_{i=1}^{n} |t_{ij}|, \tag{5}$$

where t_{ij} are the elements of $\mathbf{T}$.

The core WINGS output for each component i consists of two groups of numerical indicators. The first group comprises the relative (net) impact, receptivity, involvement and role:

I_i^r – *relative total impact* represents the net (algebraic sum of positive and negative items) influence of component i on all other components in the system; R_i^r – *relative total receptivity* represents the net influence of all other components in the system on component i; $|I_i^r| + |R_i^r|$ – *relative total involvement* represents the sum of all net influences exerted on and received by component i; and $|I_i^r| - |R_i^r|$ indicates the *net role (position)* of component i in the system; if it is positive, component i belongs to the *net influencing (cause)* group; if it is negative, component i belongs to the *net influenced (result)* group.

Each of the item from the first group has its counterpart in the second group: I_i^a – *absolute total impact* represents the absolute influence of component i on all other components in the system; R_i^a – *absolute total receptivity* represents the absolute influence of all other components in the system on component i; $I_i^a + R_i^a$ – *absolute total involvement* represents the sum of all absolute influences exerted on and received by component i; $I_i^a - R_i^a$ – *absolute role (position)*; if it is positive, component i belongs to the *absolute influencing (cause)* group; if it is negative, component i belongs to the *absolute influenced (result)* group.

3 MODEL

The SD model, developed in [2] served as a point of departure to construct the system of concepts and their relations. Our ultimate goal is to analyse the main strategic issues concerning the company development and this goal determined the content and scope of our model, in particular which concepts have been included in analysis. Partly based on the case study presented in [2], the levels of internal strengths and interrelations have been estimated. The resulting map is presented on Fig. 1.

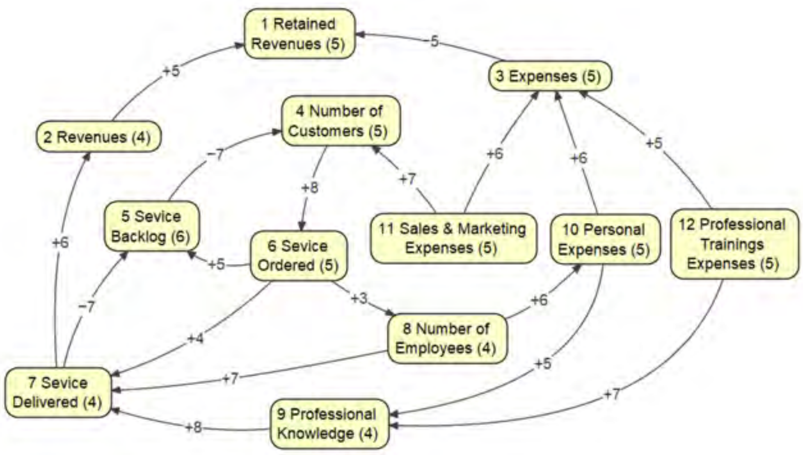


Figure 1: WINGS map of the system (internal strength of each concept is given after its name in parentheses).

Drawing on the BSC, in financial perspective we distinguished Retained Revenues as the ultimate financial measure. The retained Revenues consist of Profits and Expenses. In turn, Sales and Marketing Expenses, Personal Expenses and Professional Training Expenses contribute to Expenses. Customer perspective is represented by a single component – Number of Customers. In internal processes perspective there is Service Backlog which is positively influenced by Service Ordered (measure of how much work results from orders) and negatively by Service Delivered (measure of how much work is done). Finally, in learning and growth perspective we distinguished two concepts: Number of Employees and Professional Knowledge, both positively influencing Service Delivered.

3.1 Results

Results of calculations done with the WINGS method are presented in Tables 1, 2 and on Fig. 2. Service Delivered is the strongest concept in terms of absolute involvement what pointed out the central role of this concept in studied system. Its position in relative involvement is weaker because one of its impacts (on Service Backlog) is negative. On both figures Service Delivered belongs to influenced group. Expenses is the other most influenced concept. When we look at relative total influence (which takes into account positive and negative relations), we see that Service Ordered and Personal Expenses are strongly involved and have positive Role. That means that these concepts are crucial to the system evolvement. The strongly involved concept – Professional Knowledge – belongs to influenced concepts.

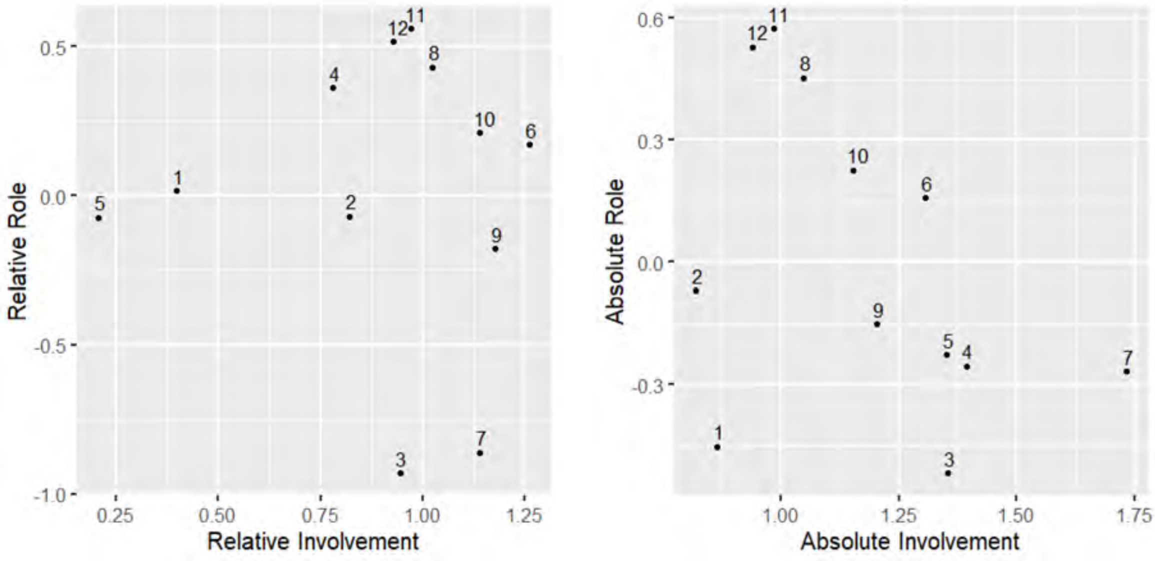
Table 1: WINGS relative output for basic data

Evaluated concept	Impact	Receptivity	Involvement	Role
1 Retained Revenues	0.2062	0.1917	0.3979	0.0146
2 Revenues	0.3751	0.4453	0.8204	-0.0702
3 Expences	-0.0061	0.9373	0.9434	-0.9312
4 Number of Customers	0.5700	0.2089	0.7790	0.3611
5 Service Backlog	-0.0652	0.1405	0.2058	-0.0753
6 Service Ordered	0.7154	0.5461	1.2615	0.1693
7 Service Delivered	0.1389	1.0000	1.1389	-0.8611
8 Number of Employees	0.7265	0.2966	1.0232	0.4299
9 Professional Knowledge	0.4985	0.6782	1.1767	-0.1796
10 Personal Expenses	0.6744	0.4642	1.1386	0.2102
11 Sales and Marketing Expenses	0.7658	0.2062	0.9720	0.5596
12 Professional Training Expenses	0.7216	0.2062	0.9278	0.5154

Table 2: WINGS absolute output for basic data

Evaluated concept	Impact	Receptivity	Involvement	Role
1 Retained Revenues	0.2061	0.6585	0.8646	-0.4524
2 Revenues	0.3748	0.4450	0.8198	-0.0702
3 Expences	0.4183	0.9366	1.3549	-0.5184
4 Number of Customers	0.5696	0.8235	1.3931	-0.2538
5 Service Backlog	0.5619	0.7890	1.3509	-0.2270
6 Service Ordered	0.7320	0.5740	1.3060	0.1580
7 Service Delivered	0.7329	1.0000	1.7329	-0.2671
8 Number of Employees	0.7505	0.2969	1.0475	0.4536
9 Professional Knowledge	0.5262	0.6777	1.2039	-0.1515
10 Personal Expenses	0.6898	0.4639	1.1538	0.2259
11 Sales and Marketing Expenses	0.7803	0.2061	0.9864	0.5742
12 Professional Training Expenses	0.7348	0.2061	0.9409	0.5287

Six concepts have strong relative involvement and belong to cause group (in a decreasing order of involvement): Service Ordered, Personal Expenses, Number of Employees, Sales and Marketing Expenses, Professional Training Expenses, Number of Customers. Among them Service Ordered and Number of Customers cannot be directly controlled by the firm. The other four can be used to improve firm's performance. On both figures - relative and absolute these four concepts have similar positions. Only Professional Training Expenses are slightly dominated by Sales and Marketing Expenses. The Personal Expenses, Number of Employees, Sales and Marketing Expenses form a kind of "efficient border" on the involvement-role plane.



(a) Relative (b) Absolute

Figure 2: Involvement vs. Role for basic data (the number of each dot identify the concept according to numbering given in Tables 1 and 2)

We can conclude that these factors should have the first priority in making the firm's strategic decisions.

4 Conclusions

The BSC approach provides a useful general framework for formulating the firm's strategy and selecting the best options for achieving long-term goals. The strict methodology, which is SD, allows for dynamic simulations of future states of main performance indicators. The numerical output of SD approach gives deeper insight into firm development. But the strict mathematical framework of SD is also its weakness. It is difficult to put into mathematical (functional) form many important, but semi-quantitative or even qualitative, factors which cannot be measured precisely in numbers. The information about them is limited and most often can be obtained from experts' assessments.

In view of the above we propose to apply the WINGS method which is able to combine in a uniform way quantitative and semi-quantitative data. The WINGS model can be considered as a complementary approach that can help management to understand the relations among main factors influencing the firm's performance, and in consequence, to shape the strategic decisions. The case study of IT service company shows how our proposed approach works and what conclusions can be drawn from WINGS results.

As this is one of the first attempts to apply the extended WINGS method to study firm's evolution and its future performance, still many research questions are opened. More experiments with various case studies and different models may validate the method usefulness and indicate its limits of applicability. The WINGS output is quite rich and further research can reveal other ways of using it for the best results.

References

[1] Banaś, D., Michnik, J. (2019). Evaluation of the Impact of Strategic Offers on the Financial and Strategic Health of the Company—A Soft System Dynamics Approach. Mathematics 7, 208.

[2] Banaś, D., Michnik, J., Targiel, K. (2016). System modeling and control of organization business processes by use of balanced scorecard and system dynamics. Control and Cybernetics 45, 185-205.

[3] Forrester J. W. (1961). Industrial Dynamics. M.I.T. Press, Cambridge, Mass.; Wiley, New York.

[4] Kaplan R., Norton D. (1992). The Balanced Scorecard – Measures that Drive Performance. Harvard Business Review, January-February 1992.

[5] Michnik, J. (2013). Weighted Influence Non-linear Gauge System (WINGS) – An analysis method for the systems of interrelated components. European Journal of Operational Research 228, 536-544.

[6] Morecroft, J.D.W. (2015). Strategic Modelling and Business Dynamics, + Website: A feedback systems approach, 2 edition. ed. Wiley, Hoboken, New Jersey.

[7] Sterman, J. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill.

PARTICLE SWARM OPTIMIZATION IN GEODETIC DATUM TRANSFORMATION

Polona Pavlovčič Prešeren and Aleš Marjetič
University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova cesta 2, Ljubljana, Slovenia
polona.pavlovcic-preseren@fgg.uni-lj.si; ales.marjetic@fgg.uni-lj.si

Abstract: Transformations between spatial coordinate systems have become more important due to the utilization of different sensors in several practical solutions. In geodesy, transformations are essential in maintaining the connection between elderly local and modern global positioning, which is based on global navigation satellite systems (GNSS). In this study, the performance of transformation by particle swarm optimization (PSO) is presented and analyzed according to the – in geodesy – traditional least squares adjustment (LSA). The aim is to propose a different solution in setting the connections between several coordinate systems/geodetic datums. For this, local coordinates were gained from topographies in non-geocentric coordinate system, while the global coordinates were determined by the use of GNSS. Simulation results show a good agreement of the adopted PSO with the traditional LSA. Thus, PSO gives promising results for further use in geodetic datum transformation.

Keywords: transformation, geodetic coordinate system/geodetic datum, GNSS, particle swarm optimization (PSO), least squares adjustment (LSA)

1 INTRODUCTION

The issue of coordinate transformations from local non-geocentric to the global geocentric coordinate system and vice versa is a crucial geodetic dilemma worldwide. The problem has become more advent due to the utilization of GNSS (*Global Navigation Satellite System*) in geodetic applications. Since previously used local geodetic reference networks contain in-homogeneities [10], the use of regional transformation parameters is not always an optimal solution. For the specific geodetic tasks, for example, cadaster surveying, it is necessary to present local transformation models.

Local transformation can be established by the use of common points with coordinates in both coordinate systems/geodetic datums. In geodesy, the similarity datum transformations are the most commonly used models in setting the relation between two coordinate systems [15]. Usually, redundant points are required for the best transformation model of the area under the consideration. However, it is not always the case that more common points yield to better results. This comes from the fact that unknown irregularities and biases in the terrestrial networks, established prior the GNSS use, exist [10].

Traditionally, local transformation parameters are determined by least squares adjustment [8]. Determination of unknown parameters follows the minimization of the error cost function. In this, the objective function presents the basis in setting optimal parameters. Most of the optimization techniques for non-linear problems follow two basic concepts. The first is an expansion of a non-linear model as a Taylor-series, further use of the linearized model around approximate values and iterations to reach the final goal. The second one is based on a gradient steepest descent method. However, some other optimization methods exist as well. El-Habiby et al. [7] presented several non-linear least squares optimizations and claimed that they differ in their effectiveness of solving the specific tasks.

The main objective of current research can be stated as follows: the investigation and testing the metaheuristic algorithm of particle swarm optimization to acquire transformation model between geocentric and non-geocentric geodetic datums when redundant common points are available. The second objective is the evaluation of the proposed method according to the least squares adjustment results.

2 METHODOLOGY

2.1 An overview of geodetic datum transformations

There are several classes of transformations, namely Euclidean, affine, projective, similarity and polynomial, which can be used in geodetic applications. Each of them contains specific characteristics. The Euclidean transformations do not change the length of the measures and preserve the shape of a geometric object. Affine transformations preserve collinearity and ratios of distances. The main difference between Euclidean and affine transformation comes from the fact, that all the Euclidean spaces are affine, but affine spaces can be also non-Euclidean [16]. Projective transformations are widely used in remote sensing and geographic information systems. They follow the rule that certain characteristics (for example collinearity, concurrency, tangency, and incidence) remain invariant after transformations [2].

Several classical similarity transformations, for example, *Bursa-Wolf*, *Molodensky-Badekas* and *Veis*, which are based on the only one set of rotations, are suitable for 3D geodetic datum transformations [4]. Transformation models differ in operation, however, they all should lead to the same transformed coordinates. The *Molodensky–Badekas* uses a centroid, but the *Bursa-Wolf* does not. The *Bursa–Wolf* model is more popular due to its simplicity, but the *Molodensky–Badekas* transformation is according to Dawod and Alnagar [3] often superior to the *Bursa-Wolf*. This was confirmed also by Kutoglu et al. [11], who showed that the mathematical model of the *Bursa-Wolf* causes high correlations between transformation parameters, hence the *Molodensky-Badekas* model determines the translations better. More complicated models, like *Vanicek-Wells* or *Krakiwsky-Thompson,* follow more than one set of rotations [16]. Consequently, they result in non-unique solutions, which come from the errors and biases in the terrestrial networks. This is the main reason that they are often not used as much as others. Due to the distortions, non-geocentric terrestrial models often do not contain a uniform accuracy. Even more, coordinates of points in the local national coordinate systems do not contain global accuracy. Consequently, it is sometimes beneficial to present more complex transformations techniques in the datum transformations, for example, stepwise multiple regression, already presented by Dawod and Alnagar [3].

However, in this study, the *Bursa-Wolf* model has been used, given by expression [14]:

$$\begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix}_{global} = \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} + (1+k) \cdot \begin{bmatrix} 1 & \omega_z & -\omega_y \\ -\omega_z & 1 & \omega_x \\ \omega_y & -\omega_x & 1 \end{bmatrix} \cdot \begin{bmatrix} X_P \\ Y_P \\ Z_P \end{bmatrix}_{local} \quad (1)$$

with seven unknowns, namely: translations ΔX, ΔY and ΔZ, the difference in scale k and three rotation angles $\omega_x, \omega_y, \omega_z$, where the coordinates of the point P in global and local coordinate systems are recognized as known. In this mathematical expression, the three-step product of rotational matrices is presented only by the matrix R (the only matrix in the Eq. 1) and is due to small rotation angles simplified with the above form.

2.2 Least squares adjustment

The similarity transformation model can be described through a set of continuous and/or discrete vectors of unknowns Δ. Let Δ_0 contain *a priori* values of unknowns, and let Σ_{ll} be the corresponding variance-covariance matrix of observations (in this case coordinates in a local coordinate system) in the vector l. From Σ_{ll} and *a priori* reference variance σ_0^2 (in the first iteration process) matrix of weights P can be defined as $P = \sigma_0^2 {\Sigma_{ll}}^{-1}$. A physical theory imposes a non-linear relationship of the form $f(\Delta) = 0$ on the possible values of Δ, where $f()$

can be any non-linear differentiable operator acting on Δ. The least squares adjustment is stated as searching the vector $\widehat{\Delta}$ as well as the vector of observation residuals v, which is based on the minimization of the quadratic form $v^T P v$. The adjustment follows the linearized equations, given in the vector-matrix form [12]:

$$v + B \cdot \Delta = \mathrm{f}, \tag{2}$$

where the matrix B contains first order partial derivatives of the nonlinear operator $f()$ and the vector f contains differences between definitive and approximate function values. Since the linearization, the non-linear problem has to be solved iteratively:

$$\widehat{\Delta} = (B^T P B)^{-1}(B^T P \mathrm{f}) \\ v = \mathrm{f} - B \cdot \widehat{\Delta} \text{ and } \hat{l} = l + v. \tag{3}$$

When the number of common points is not less than three points, the optimal probability values of seven transformation parameters can be obtained by least squares adjustment. Hereinafter, a different approach of particle swarm optimization will be presented, which does not need to linearize the non-linear equations.

2.3 Particle swarm optimization

Eberhard and Kennedy proposed the particle swarm optimization in 1995 [6]. The algorithm is inspired by social behavior and dynamics of bird flocking, fish schooling, and swarm theory, where a population of the individuals adapt their behavior to the environment according to the rule of survival. The initialization follows a population of random solutions, called particles, and searches for optima by updating generations. Particles as potential solutions fly through the problem space by following the previous optimum particles. In the search process, they have a tendency to fly towards the gradually better search area.

Particle swarm optimizer is insensitive to scaling of design variables, derivative free, and follows simple implementation based on a few parameters. It is not always the case that PSO is successful in finding the global minimum, but it does a great job in high dimensional, non-convex, non-continuous environments [13], [14].

The original PSO algorithm can be described as follows. In PSO, each point is called a particle. For the N-dimensional search space, the i^{th} swarm particle is represented by the column vector, $x_i = [x_{i,1} \quad x_{i,2} \quad \cdots \quad x_{i,n}]^T, i = 1, 2, \ldots, N_{part}$, where N_{part} stands for the number of particles, usually defined by the user. This particular particle contains the velocity, represented by the N-dimensional vector $v_i = [v_{i,1} \quad v_{i,2} \quad \cdots \quad v_{i,n}]^T$. At each iteration step, indexed by $t = 0, 1, \ldots$, the fitness function $f(x)$ (i.e. a function that is under the consideration in the optimization process) is evaluated in the current particle position. The particles move in a loop according to special rules. Let $x_i(t)$ denote the position of the i^{th} swarm particle at the t^{th} iteration. Each specific particle is a record of position, where the best fitness value is found previously. Such a location is denoted as $p_i(t)$ and is called "*particle best*" (*pbest*):

$$f(p_i(t)) = \min_{j=t,t-1,\ldots,0} f(x_i(t)). \tag{4}$$

The vector $p_i(t) = [p_{i,1}(t) \quad p_{i,2}(t) \quad \ldots \quad p_{i,n}(t)]^T$ contains previously best visited positions of the i^{th} particle. Best fitness result found in the history is called "*global best*", $g(t)$, and is evaluated by the fitness function as:

$$f(g(t)) = \min_{j=1,2,\ldots,N_{part}} f\left(p_j(t)\right). \tag{5}$$

The particle, whose current "*particle best*" is also "*global best*" of the swarm wins and is denoted as "*best particle*" in the swarm. The next value in the iteration step $x_i(t+1)$ is acquired from $x_i(t)$, $p_i(t)$ and $g(t)$ as follows:

$$x_i(t+1) = x_i(t) + v_i(t) \tag{6}$$
$$v_i(t+1) = v_i(t) + c_1 \cdot r_1() \cdot \left(p_i(t) - x_i(t)\right) + c_2 \cdot r_2() \cdot \left(g(t) - x_i(t)\right).$$

In the Eq. 6, c_1 and c_2 are constants, also known as cognitive and social scaling parameters. The second and the third term in the equation for the velocity represent accelerations, which change the velocity randomly. $r_1()$ and $r_2()$ are diagonal matrices with random numbers drawn from a uniform distribution. Another constant, the user defined maximum velocity v_{max}, controls the global exploration ability of the swarm.

A step forward from the traditional PSO algorithm is the presentation of the inertia weight coefficient, which is even more successful in global minimum finding comparing to the traditional PSO [6]. The motivation to introduce the inertia weight coefficient is the elimination of user-defined coefficient v_{max}. Moreover, inertia weight significantly fastens the convergence and has a prominent role in balancing the search through the cases that tend to have a critical success of an optimization. The modified equation for velocity follows the form [12]:

$$v_i(t+1) = w \cdot v_i(t) + c_1 \cdot r_1() \cdot \left(p_i(t) - x_i(t)\right) + c_2 \cdot r_2() \cdot \left(g(t) - x_i(t)\right). \tag{7}$$

In such representation, the weight w controls the velocity. Shi and Eberhart [13] proposed a constant value of inertia weight and showed the benefits of its use in global optimization by reducing the number of iterations. They claimed that a large inertia weight allows better global search, while the small inertia weight allows a local minima finding.

The original procedure for implementing the PSO is as follows [14]:

1. initialization of a population of particles with random positions and velocities in the problem space;
2. evaluation of the desired fitness function (in this particular case Eq. 1) for each particle;
3. comparison of particle's fitness evaluation with its *pbest*. If the value is better than *pbest*, then the set *pbest* will be equal to the current value, and previously best visited position $p_i(t)$ will be equal to the current location $x_i(t)$;
4. identification of the particle in the neighborhood with the best success in history, and assignment its index to the variable "*global best*", $g(t)$;
5. change the position of the particle according to Eq. 6 and velocity according to Eq. 7;
6. loop to step 2 until a criterion (a good fitness function or a maximum number of iterations) is met.

3 DATA AND RESULTS

For this study, local coordinates were gained from the topographies of points in the previously used non-geocentric Slovenian coordinate system. GNSS carrier-phase positioning was used to determine global coordinates (Table 1).

Table 1: Coordinates in two coordinate systems, D48/GK and D96/TM.

	Non-geocentric national coordinate system D48/GK			***Geocentric national coordinate system ETRS89 (D96/TM)***		
Point	y_{GK} [m]	x_{GK} [m]	H [m]	*e* [m]	*n* [m]	h [m]
1	461,849.93	99,989.22	339.46	461,478.88	100,475.72	385.88
2	459,984.02	99,868.19	340.50	459,613.03	100,354.67	386.94
3	459,937.19	101,340.75	425.80	459,566.22	101,827.28	472.23
4	461,403.90	99,880.86	291.83	461,032.91	100,367.32	338.30
5	461,159.60	100,963.57	301.04	460,788.62	101,450.02	347.33
6	460,699.53	100,536.99	294.99	460,329.10	101,023.32	341.34
7	460,619.60	100,413.60	295.00	460,248.61	100,900.06	341.48
8	460,919.91	101,001.67	300.441	460,548.98	101,488.10	346.86
9	461,137.50	100,197.46	292.95	460,766.68	100,683.87	339.30
10	460,870.72	100,438.41	294.28	460,499.72	100,924.89	340.62
11	461,284.66	99,750.99	292.55	460,913.66	100,237.43	338.89

For the transformation, data were transformed from both coordinate systems, namely *D48/GK* and *D96/TM*, into Cartesian coordinates $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Equations for the Gauβ-Krüger projection can be found in [5] and the rest of the equations in [9]. From the available set of data, several smaller sets of common points in both coordinate systems were chosen randomly to determine parameters of the transformation model.

To avoid different types of errors in available set of data, affine triangle-based 2D transformation [1] was used. Two points, namely 6 and 9, claimed to have *gross errors*, which were probably due to the physically displaced points due to road reconstructions. Those points were excluded from further investigation.

According to the model in the Eq. 1, determination of the *Bursa-Wolf* coefficients was carried out for both algorithms under consideration, namely, LSA and PSO. The performance tests were carried out for twelve sample sets with different distribution and number of common points. In each determination, the rest of the points, which were not used in parameters' determination, were used for the validation. This study used the *RMSE* index for each specific position component to evaluate the accuracy of each fitted model (Table 2).

Table 2: Validation results according to different number of common-points as well as different points.

	Least Squares Adjustment			***Particle swarm optimization***		
Common points	*RMSE of e* [cm]	*RMSE of n* [cm]	*RMSE of h* [cm]	*RMSE of e* [cm]	*RMSE of n* [cm]	*RMSE of h* [cm]
1, 2, 3	*3.0*	*3.8*	*5.5*	*3.2*	*4.5*	*4.8*
2, 4, 5	*2.6*	*2.7*	*7.6*	*2.8*	*3.0*	*4.9*
4, 8, 11	*2.9*	*4.4*	*5.5*	*3.1*	*5.0*	*4.3*
1, 3, 5, 7	*3.6*	*3.6*	*5.5*	*3.7*	*3.7*	*6.0*
2, 4, 7, 10	*5.6*	*8.6*	*5.2*	*5.8*	*9.0*	*5.1*
4, 7, 8, 10	*9.7*	*6.6*	*5.2*	*9.8*	*6.5*	*9.9*
1, 3, 5, 7, 10	*3.6*	*6.7*	*5.5*	*3.7*	*7.2*	*5.4*
2, 4, 7, 10, 11	*4.2*	*4.5*	*5.2*	*4.2*	*4.7*	*5.3*
4, 5, 7, 8, 10	*7.1*	*4.6*	*5.2*	*7.3*	*5.2*	*7.6*
1, 3, 5, 7, 8, 10	*2.7*	*4.6*	*7.0*	*3.0*	*4.6*	*7.0*
2, 4, 5, 7, 8, 11	*2.3*	*2.8*	*4.6*	*2.6*	*2.8*	*4.9*
4, 5, 7, 8, 10, 11	*7.5*	*4.4*	*6.5*	*7.4*	*4.5*	*6.6*

As can be seen from the results, several centimeters differences between sets of common points under consideration exist. From this, we can confirm the problem of distortions in the area under consideration. Finally, PSO results show a good agreement to the LSA for this particular case.

4 CONCLUSIONS

This paper proposes PSO for the geodetic datum transformation. The PSO results agree with the traditional LSA. Since PSO follows simple implementation, which is derivative free, as well as execution time is short the fitting process can be rapidly and easily performed, However, results depend on the iterations and prior desired values of final goal achievement.

The experimental analysis shows that PSO is feasible and brings a different aspect of the solution of datum transformations with promising results. However, research by using a bigger sample size of data might be used in further experiments to check the robustness of the method.

Acknowledgement

This research was partially funded by the Slovenian Research Agency within the research program P2-0227 (A), *Geoinformation Infrastructure and Sustainable Spatial Development in Slovenia*.

References

[1] Berk, S., Komadina, Ž. 2013. Local to ETRS89 Datum Transformation for Slovenia: Triangle-Based Transformation Using Virtual Tie Points. Survey Review, 45(3): 25–34.

[2] Chang, N.−B., Bai, K. 2018. Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing. Milton: Taylor and Francis.

[3] Dawod, G., Alnagar, D. 2000. Optimum geodetic datum transformation techniques for GPS surveys in Egypt. Proceedings of Al-Azhar Engineering Sixth International Conference, Cairo, Egypt, (4): 709–718.

[4] Deakin, R. E. 2006. A note on the Bursa-Wolf and Molodensky-Badekas transformations. School of Mathematical and Geospatial Sciences, RMIT University: 1–21.

[5] Deakin, R. E., Hunter, M. N., Karney, C. F. F. 2010. The Gauss-Krűger Projection. Presented at the 23rd Victorian Regional Survey Conference, Warrnambool 10–12 September 2010, The Institution of Surveyors, Victoria.

[6] Eberhart, R. C., Kennedy, J. 1995. Particle Swarm Optimization. Proceedings IEEE International Conference on Neural Networks. Perth, Australia: 1942–1948.

[7] El-Habiby, M., Gao, Y., Sideris, M. G. 2009. Comparison and analysis of non-linear least squares methods for 3-D coordinates transformation. Survey Review, 41(311): 26–43.

[8] Grafarend, E., Awange, L. J. 2003. Nonlinear analysis of the three-dimensional datum transformation (conformal group C_7). Journal of Geodesy, 77(1): 66–76.

[9] Ligas, M., Banasik, P. 2011. Conversion between Cartesian and geodetic coordinates on a rotational ellipsoid by solving a system of nonlinear equations. Geodesy and cartography, 60(2):145–159.

[10] Kohli, A., Jenni, L. 2008. Transformation of Cadastral Data between Geodetic Reference Frames using Finite Element Method Integrating the Generations. Working Week 2008 Stockholm, Sweden 14–19 June 2008.

[11] Kutoglu, H. S., Mekik, C., Akcin, H. 2002. A Comparison of Two Well Known Models for 7-Parameter Transformation. Australian Surveyor, 47(1): 24–30.

[12] Mikhail, E. M., Ackerman, F. 1976. Observations and least squares. New York: IEP.

[13] Shi, Y., Eberhart, R.C., 1998. A modified Particle Swarm Optimizers. Proceedings on the IEEE International Conference on Evolutionary Computation: 69–73.

[14] Shi, Y., 2004. Particle Swarm Optimization. IEEE connections, 2(1): 8–13.

[15] Vanicek, P., Steeves, R. R. 1996. Transformation of coordinates between two horizontal geodetic datums. Journal of Geodesy, 70(11): 740–745.

[16] Weisstein, E. W. 2017. Affine Transformation. From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/AffineTransformation.html. Accessed 2019.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Special Session 7:

Towards Industry 4.0

DIGITAL TWIN OF UNIQUE TYPE OF PRODUCTION FOR INNOVATIVE TRAINING OF PRODUCTION SPECIALISTS

Mihael Debevec
University of Ljubljana, Faculty of Mechanical Engineering, Chair of Manufacturing Technologies and Systems, Laboratory for Handling, Assembly and Pneumatics
Askerceva 6, SI-1000 Ljubljana, Slovenia
E-mail: mihael.debevec@fs.uni-lj.si

Niko Herakovič
University of Ljubljana, Faculty of Mechanical Engineering, Chair of Manufacturing Technologies and Systems, Laboratory for Handling, Assembly and Pneumatics
Askerceva 6, SI-1000 Ljubljana, Slovenia
E-mail: niko.herakovic@fs.uni-lj.si

Abstract: The most important goal of management in every company is the production process with as few as possible deadlocks. And an effective production process can be reached with well-trained production specialists. In this contribution, a new strategy on how to train production specialists for unique type of production by using digital twin is presented. Also, a specially developed digital twin for typical unique type of production is explained and describes the real production process in detail so that it enables the observation of responses to the different input data. The main goal of the presented strategy covers the training of personnel so that they learn how varied input data reflect in the output results. The digital twin at the same time clearly and responsibly indicates the response of the production system to the changes in parameter settings.

Keywords: digital twin, unique type of production, simulation, schedule plan, training, virtual factory.

1 INTRODUCTION

Modern manufacturing requires well trained personnel who must respond actively to different disturbances in the production process. To achieve this goal, companies should introduce modern and innovative tools to train their employees and prepare them as much as possible for situations in the manufacturing process where circumstances change and disturbances occur. Training should be performed on developed digital twins of production process that enables performing different simulations in a dynamic production environment by using computers. Traditional training approaches do not enable that sort of upgrading skills.

Existing research addresses with a number of useful examples [1-4] which are usable for training in specialized areas. Also a number of related studies have been made and different approaches proposed in other areas with the goal to improve the existing processes.

Our research work has proven that digital twin (DT) [5-8] is a very useful tool for the training of production specialists, usually these are the production process planners. A major advantage of the concept is that the digital twin does not consume any material, resources or energy – digital twin operate only with data. During the training process, the real production process is not interrupted and consequently the equipment is not occupied and cannot be damaged.

2 THE BASICS OF DIGITAL TWIN

A digital twin is defined as a virtual representation of a physical product or process, used to understand and predict the physical counterpart's performance characteristics. Digital twins are used throughout the product lifecycle to simulate, predict, and optimize the product and production system before investing in physical prototypes and resources [7].

On the field of production processes the digital twin enables execution of the production or schedule plan in a virtual computer environment or, in other words, the performing of the production process in a DT [5]. Just like for a real factory, it is considered that one of the most important parts in the DT is the resource management.

In the field of the production processes, research is generally oriented in several specialized orientations:

- scheduling of the production operations in time scale, and that is the most frequently addressed issue in practice,
- improvement of the production process at the level of the performing of operations,
- studies about reducing the costs and reducing power consumption,
- some special principles to improve the efficiency of the production process.

For the development of the concepts and models for training, the researchers have been focused on the improvement of the production process at the level of the performing of operations.

For both concepts and models of the production processes, logical models were designed in the first step, and they were created based on the fundamental principle of the DT (see Fig. 1), the theoretical assumptions [3] and the recent results of our research work [5]. DT basically consists of two basic parts: a virtual physical subsystem (VPS) and a real information subsystem (RIS). The virtual physical subsystem describes the logical dependencies between processes, material flow and the flow of resources in the production system [5]. For the real information subsystem, the existing information system of the production system is taken into account where the data flow is controlled by the integral information system. So the RIS covers the appropriate data structure that is needed to control the physical part of the DT. Relevant connections are established between the VPS and RIS to ensure the correct functioning of the DT.

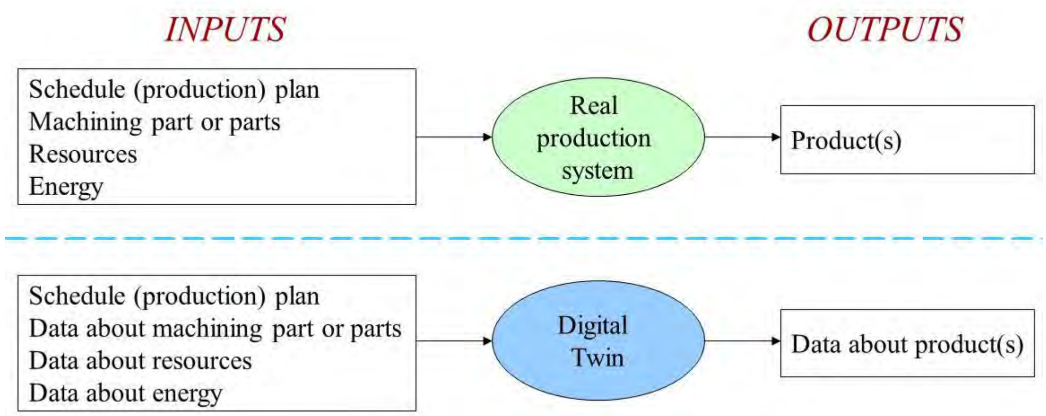


Figure 1: The basic principle of the digital twin.

Then, computer simulation models, which represent the central logical part of the digital twin, were designed on the basis of the logical models of the production process, and also the appropriate input and output data structures were created. The basic and main objective of the research was the creation of the DT which enables and ensures the execution of a schedule plan in the same way as it is carried out in the real production process (see Fig. 1). This means that all logical dependencies, which are characteristic of a real production system, are described in detail in the developed digital twin, including all necessary simulation models.

In order to make the digital twin easy for use, it is developed in the standard programming tool Tecnomatix Plant Simulation [9, 10] for several important reasons and advantages:

- the simulation tool is standard and generally applicable,
- the tool is based on discrete events execution,
- the program is object-oriented,
- complex logical dependencies from the production system can be modelled through programming in the programming language SimTalk [10],
- data exchanges between the digital twin and databases are relatively easy to establish,
- the tool enables an easy way to create different graphical or numerical presentations about the progress of the production process in the digital twin.

In DT, it is assumed that the energy for the operating of the production system is always available. In the case of a power failure, the manufacturing process cannot be carried out.

3 DIGITAL TWIN OF UNIQUE TYPE OF PRODUCTION

The concept for the model of a unique type of production (UTP) treats the operation as the elementary unit of the production process. The concept includes a logical rule that states that the performing of an operation can be started only when all the required resources, the data sets and the materials are available and present at the place where the planned operation will be performed. And the sequence of operations represents the production process (see Fig. 2).

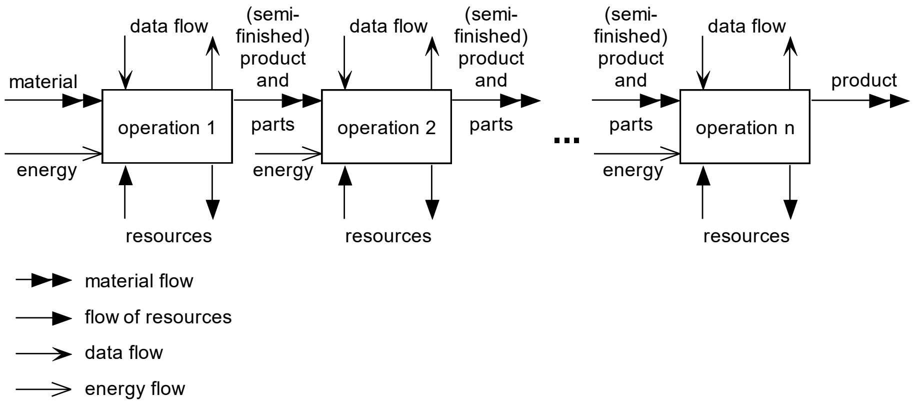


Figure 2: The logical scheme of the simulation model for a unique type of production.

In the production process model for the UTP, first the data for the production process and then the presence of resources has been taken into account because due to the unavailability of resources deadlocks frequently occur. Among the most important resources are the transportation equipment, clamping equipment, cutting tools, measuring devices, machining centres, special tools and equipment, and as a very important influential factor human resources.

Based on the formed logical model (see Fig. 2), a digital twin of the production process for a UTP was built in the computer environment Tecnomatix Plant Simulation (see Fig. 3, left part).

In the digital twin are used standard components of the tool Plant Simulation and they are mutually logically connected. The meaning of individual types of objects is explained separately (see Tab. 1).

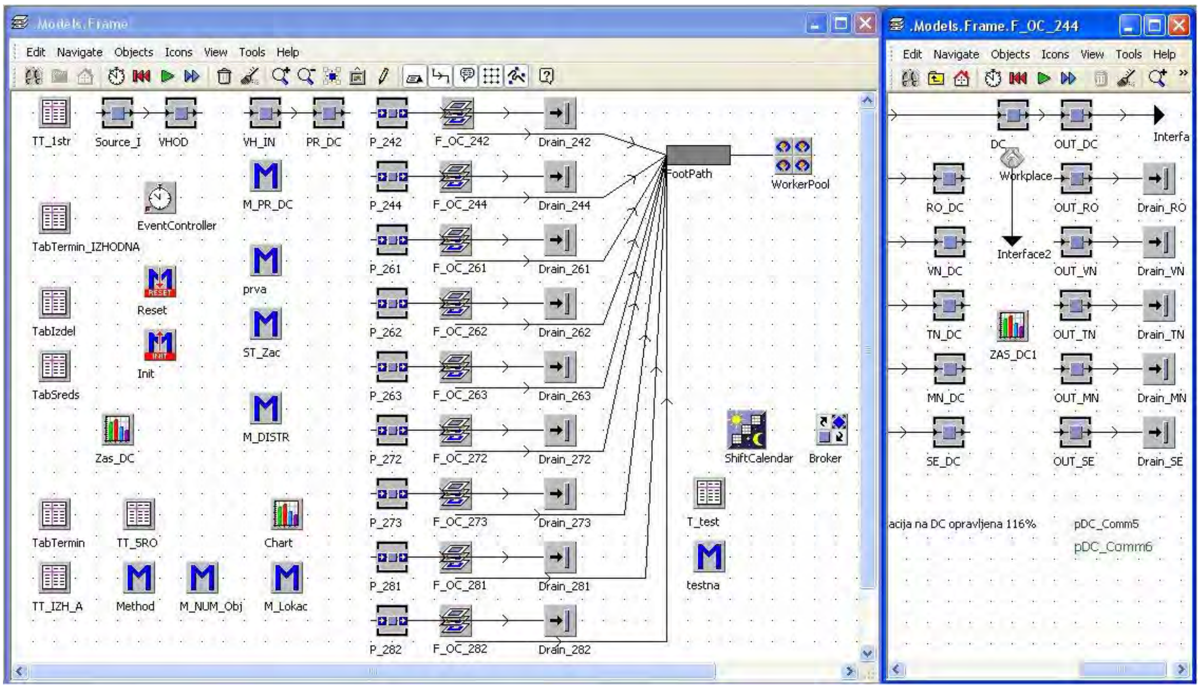
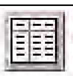
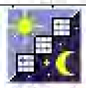
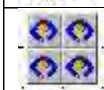


Figure 3: An example of the digital twin for a UTP in Plant Simulation.

Table 1: Important logical elements in the simulation model for a UTP.

Icon on Fig. 3	*Description*
 TT_1str	table type objects: TT_1str: the table with input data – schedule data TabSredst: the table with data about resources TabTermin_IZHODNA: the table with output data – the results of the simulation
 EventController	EventController is the object for controlling the simulation process
 F_OC_244	process type object: F_OC_244 … 282: objects that represent machining centres
 Zas_DC	objects of the chart type are used to display the efficiency of a production unit
 ShiftCalendar	the object of the ShiftCalendar type is used to manage the working calendar and shifts
 WorkerPool	the object of the WorkerPool type is used to manage workers
	the object of the Method type contains programs for the logical actions execution

Digital twin is developed in level mode. The basic model in DT is developed on the first level (see Fig. 3, left part) and covers logical dependencies between segments of UTP process. Some individual segments, such as CNC machines, are modelled more detail in sublevels (see Fig. 3, right part – example of one CNC machine) in order to ensure transparency of the DT.

The DT for UTP is parametric so that the user inserts into the DT the input data in table format that is intended for the real production process for the observed period. Among the input data is the schedule plan, the list of available resources, work calendar and the number of available workers. After the initial setup, the user performs a simulation in DT for the desired observation period or for the production of the desired number of pieces. During the execution of the simulation, the speed of the simulation execution can be set to stepless, the simulation can be carried out step by step or it can be stopped at any time. The start date and time of every individual simulation run can be set to an arbitrary time. It is also possible to configure the production parameters for every object of production process.

In the presented DT, there are included pre-designed illustrative displays for real-time tracking of important indicators such as the number of the finished parts, the consumed production time, the occupancy analysis of individual resources and time course of operations on CNC machines (see Fig. 4). The indicators can be displayed in text format or graphically on the screen or structured in table form. With simple additions, any indicator can be installed or any calculation and analysis can be performed online.

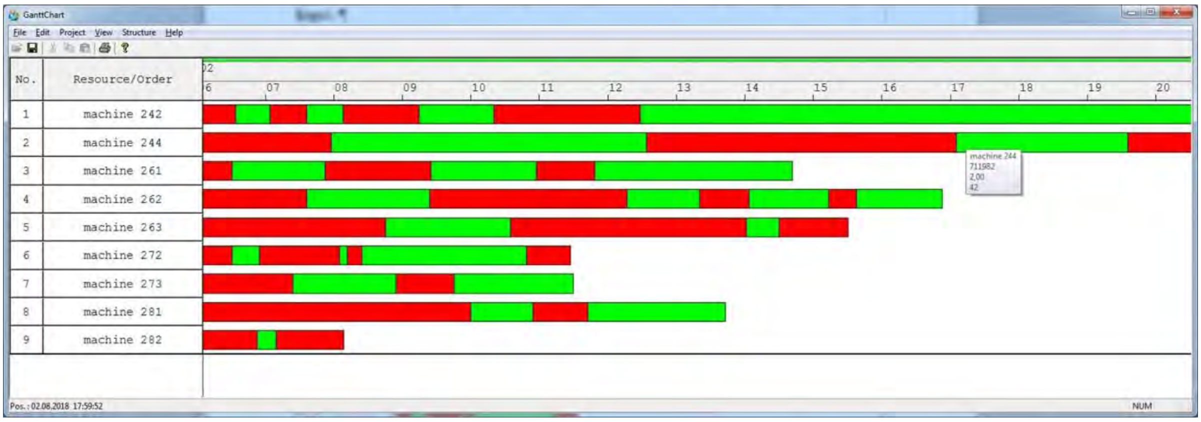


Figure 4: An example of time course of operations on CNC machines.

The UTP model has been tested and verified using known data from a company for an already performed production process, so that the input data and results of the production process were well known. To prove the adequacy of the model several testing simulation runs were carried out, and the acquired data showed that the developed model satisfactorily describes the real production process.

4 CONCLUSION

The use of digital twin for unique type of production for the purposes of training brings a number of advantages. The first advantage is obtaining the simulation results about the estimated execution of the schedule plan quickly. Testing has proven that the execution of a production process in the DT for an entire work shift takes only a few minutes. Furthermore, with the testing of production plans in a DT we do not intervene in the real production system and thereby we do not cause any disturbances. Because of this, we have practically unlimited possibilities of testing different schedule plans where we observe the behaviour of the production system as a function of time for an individual plan, or only observe the outputs of the production system. The DT is designed parametrically, so for the purposes of training, we can easily test different production plans as input data. The developed DT is user-friendly so that the user inserts into the DT the input data, sets the process parameters, performs the simulation and evaluates the results of the simulation. So the presented DT is very useful tool for training of the production specialists since it offers them practically unlimited possibilities for testing different production scenarios.

In general, DT offers an option for the users to perform a simulation for an existing or known production system where they observe the outputs of the model according to different rearrangements in the DT. Also they can study a planned production system where they test various configurations of the production system, or test the response of the planned system to different settings, among which we included the working calendar, number of shifts, different break times, number of employees, number of machines, variants of parallel processes, different process times, transport times, different transport routes and manners and transport strategy for components. Based on the different settings and acquired responses the users can determine most suitable production parameters.

Acknowledgement

The work was carried out in the framework of the GOSTOP programme (OP20.00361), which is partially financed by the Republic of Slovenia – Ministry of Education, Science and Sport, and the European Union – European Regional Development Fund.

The authors also acknowledge the financial support from the Slovenian Research Agency (research core funding No. (P2-0248)).

References

[1] Padhi, S.S., Wagner, S.M., Niranjan, T.T., Aggarwal, V. 2013. A simulation-based methodology to analyse production line disruptions. International Journal of Production Research, 51(6): 1885-1897. https://doi.org/10.1080/00207543.2012.720389.

[2] Uzam, M., Li, Z. 2014. On an iterative deadlock prevention approach for automated manufacturing systems. The International Journal of Advanced Manufacturing Technology, 74(1-4): 503-507. https://doi.org/10.1007/s00170-014-6018-x.

[3] Negahban, A., Smith, J.S. 2014. Simulation for manufacturing system design and operation: Literature review and analysis. Journal of Manufacturing Systems, 33(2): 241-261. https://doi.org/10.1016/j.jmsy.2013.12.007.

[4] Thürer, M., Fernandes, N.O., Carmo-Silva, S., Stevenson, M. 2018. Lot splitting under load-limiting order release in high-variety shops: An assessment by simulation. Journal of Manufacturing Systems. 48(Part A): 63-72. https://doi.org/10.1016/j.jmsy.2018.06.003.

[5] Debevec, M., Simic, M., Herakovic, N. 2014. Virtual factory as an advanced approach for production process optimization. International Journal of Simulation Modelling, 13(1): 66-78. https://doi.org/10.2507/IJSIMM13(1)6.260.

[6] Mikell, M., Clark, J. 2018. Cheat sheet: What is Digital Twin? IBM. https://www.ibm.com/blogs/internet-of-things/iot-cheat-sheet-digital-twin/ [Accessed 07/06/2019].

[7] Digital Twin. 2019. SIEMENS. https://www.plm.automation.siemens.com/global/en/our-story/glossary/digital-twin/24465 [Accessed 10/06/2019].

[8] Digital Twin. 2019. Fraunhofer IOSB. https://www.iosb.fraunhofer.de/servlet/is/81047/ [Accessed 10/06/2019].

[9] Plant Simulation. 2018. SIEMENS. https://www.plm.automation.siemens.com/store/en-us/trial/plant-simulation.html [Accessed 08/05/2019].

[10] Bangsow, S. 2015. Tecnomatix Plant Simulation: Modeling and programming by means of examples. Switzerland: Springer International Publishing; 2015. https://doi.org/10.1007/978-3-319-19503-2.

PHYSICAL TESTING OF A TRAILING ARM BY DISCRETE OPTIMIZATION

Matic Muc
TPV d.o.o., Kandijska cesta 60, 8000 Novo mesto, Slovenia
m.muc@tpv.si
Vili Malnarič
TPV d.o.o., Kandijska cesta 60, 8000 Novo mesto, Slovenia
v.malnaric@tpv.si
Jernej Klemenc
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
jernej.klemenc@fs.uni-lj.si
Janez Žerovnik
FS, University of Ljubljana, Aškerčeva 6, 1000 Ljubljana, Slovenia
Institute of Mathematics, Physics and Mechanics, Jadranska 19, Ljubljana, Slovenia
janez.zerovnik@fs.uni-lj.si

Abstract: In design of car parts weight, price and time of production process are very important. That is why engineers use softwares for CAD modelling and FEM analyses on a daily basis. When we are dealing with a complex product with lots of different loading cases proper job scheduling is of significant importance. In this paper we optimize the testing of the trailing arm FEM in real life testing settings by application of heuristics for the travelling salesman problem.

Keywords: Travelling salesman problem, minimum spanning tree, distance matrix, Hamiltonian cycle, discrete optimization, asymmetric travelling salesman problem

1 INTRODUCTION

The two most important questions in modern car industry are "How to reduce CO_2 emissions of a car in order to make it more eco-friendly?" and "How can we make our product as cheap as possible without losing its function?". It turns out that there is a similar answer to both questions - we have to make our car parts lighter and consequently the whole car assembly is going to be lighter, eco-friendlier and cheaper. This type of design is called *lightweight design.*

One of the most important car chassis parts is trailing arm, which connects car body and tyre. Assume the situation in which a costumer wants the team of engineers to develop the next generation trailing arm which could withstand higher loads and be as light as possible.

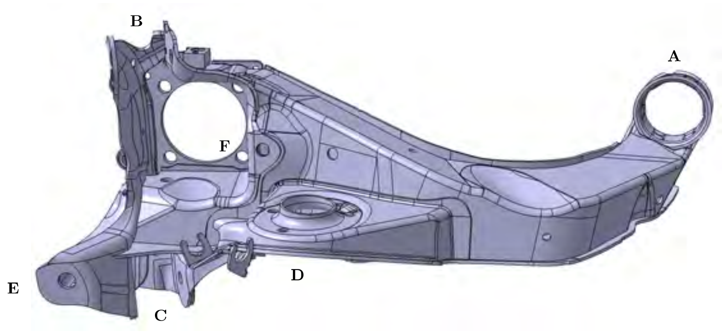


Figure 1: CAD model of trailing arm with kinematic point labels

In Figure 1, CAD model of old generation trailing arm can be seen. Its geometry is very complex due to limited space and very important function of the part. Trailing arm is made out of different sheet metal parts which are later welded together into final design.

Modern product development is a tight collaboration between various engineering fields. In order to make a product as light as possible a lot of different virtual tests (FEM analyses) have to be performed during the development phase. Even though FEM solvers are becoming more efficient and accurate in the prototype phase real tests have to be performed as well.

In this paper we first consider ways to find the best sequence of virtual analyses settings in order to set all of the jobs as fast as possible. We are going to tackle this problem with graph theory, in particular we model our problem as the travelling salesman problem (TSP). TSP is a well known computationaly hard problem in combinatorial optimization [1, 2]. In the next step, we will be looking for best sequence of physical testing. In this case, the appropriate model is the asymmetric travelling salesman problem (ATSP). For the moderate size realistic problems of symmetric TSP, we find that even optimal solutions are found by classical approach for TSP [4]. For the general case the heuristic algorithm for asymmetric travelling salesman problem (ATSP) [3] was implemented.

2 TRAVELLING SALESMAN PROBLEM

In this section we first explain our optimization problem, and observe its natural relation to the travelling salesman problem. An example with symmetric distance matrix is solved to optimality. The symmetric versions of TSP appear when the order of virtual simulations is optimized.

2.1 LOAD CASES EXPLANATION

For better understanding of this paper, we will first explain how many different load scenarios there are and what they mean.

Trailing arm shown in Figure 1 is attached to car chassis in six kinematic points (labelled with letters A-F). The trailing arm is attached to the chassis directly through point "A". Points "B" and "C" represent connections for upper and lower control arm, point "D" denotes connection between trailing arm and spring support. Points "E" and "F" are connected to damper and stabilizer. In each kinematic point we have the same loads. We have to check dynamic durability when the trailing arm is loaded in each of them. Additionally we have to control plastic deformation of each point when loaded in tension or compression.

Denotations in Tables 1 and 2 represent kinematic point and current loading case, e.g. *Adyn* when dynamic load is applied in point A, *AzF* when point A is loaded with force in tension, *AdF* when point A is loaded with force in compression and AdD when point A is loaded with displacement in compression.

There is also seventh kinematic point which is not shown on Figure 1 - "G" which simulates car braking. In this case we only simulate dynamic durability (In Tables 1 and 2 denoted as *Gdyn*). All together we have 29 different loading cases which have to be calculated every time.

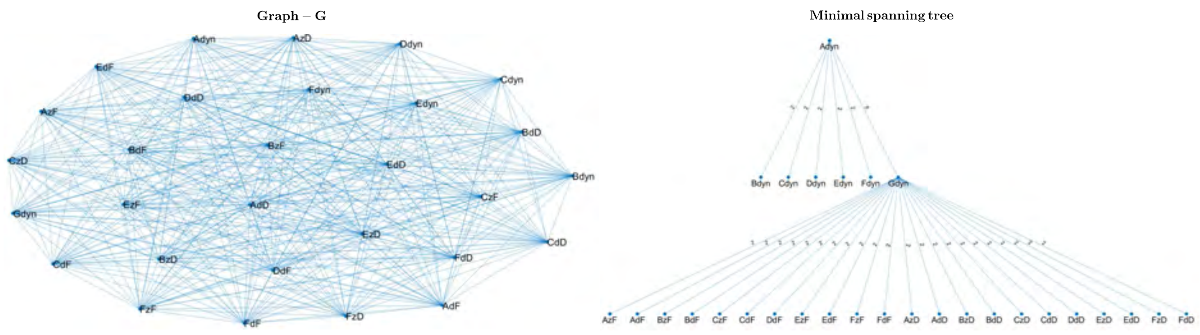
2.2 SOLVING TRAVELLING SALESMAN PROBLEM

In this case every task which has to be set up for FEM analysis is going to present one vertex - $n = 29$. The edges between vertices are weighted by time in minutes, which is needed for transition from one setting to another. Table 1 is showing the weighted adjacency matrix of size 29 x 29. As the graph is complete, the weighted adjacency matrix is at the same time the distance matrix and is thus denoted by $D(G)$.

Table 1: Distance matrix for FEM simulation [5].

	Adyn	Bdyn	Cdyn	Ddyn	Edyn	Fdyn	Gdyn	AzF	AdF		FdD
Adyn	0	2	2	2	2	2	5	10	10		10
Bdyn	2	0	2	2	2	2	5	10	10		10
Cdyn	2	2	0	2	2	2	5	2	2		2
Ddyn	2	2	2	0	2	2	5	10	10		10
Edyn	2	2	2	2	0	2	5	10	10	⋯	10
Fdyn	2	2	2	2	2	0	5	10	10		10
Gdyn	5	5	5	5	5	5	0	2	2		2
AzF	10	10	2	10	10	10	2	0	2		2
AdF	10	10	2	10	10	10	2	2	0		2
					⋮					⋱	2
FdD	10	10	2	10	10	10	2	2	2	2	0

Figure 2a shows graph G according to distance matrix 1. We can see that we are dealing with full graph which has 406 connections (According to *Handsaking Lemma* [4]).



(a) (b)

Figure 2: (a) Graph G and (b) Minimum spanning tree T of graph G.

In order to find a Hamiltonian cycle which represents solution of TSP, we have to construct minimal spanning tree (MST) which is shown in Figure 2b. Weight of T - $W(T) = 59\,\text{min}$ which means that TSP solution lies between 59 min and 118 min. In the next step we find one of the solutions for TSP, which is shown in Figure 3.

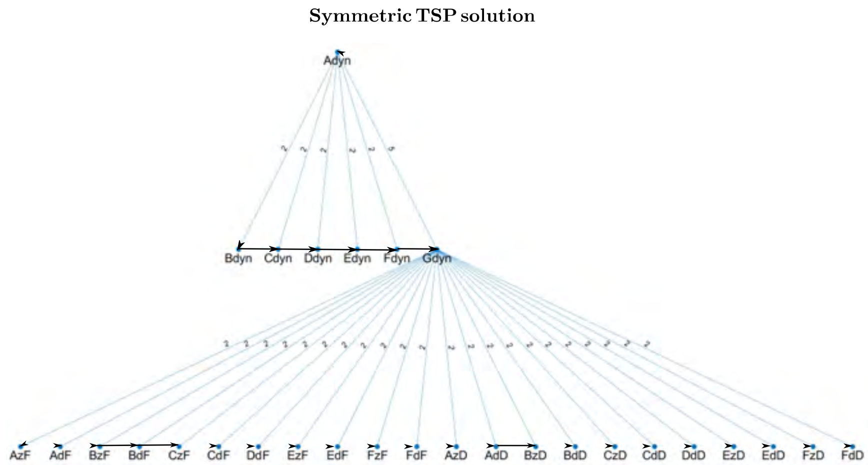


Figure 3: Solution for Symmetric TSP.

Hamiltonian cycle shown in Figure 3 weights 69 min, which means that we can start FEM calculations for all load cases in 69 minutes. The solution can be shown to be optimal (the proof is ommited due to space limitations).

3 ASYMMETRIC TRAVELLING SALESMAN PROBLEM

As explained before, there are many different ways to construct Hamiltonian cycle which represents near-optimal solution for symmetric TSP. Solving TSP becomes even more problematic when the distance matrix is not symmetric. The distance matrix $D = (d_{ij})$ of dimension $n \times n$ contains information about distances between vertex i and vertex j. If distance $d_{ij} = d_{ji}$ we are dealing with symmetric TSP, otherwise it is asymmetric (ATSP) [3]. There are few algorithms which can help us solve ATSP. We are going to use heuristic presented by Brest and Žerovnik [3].

3.1 THE HEURISTIC

The heuristic is based on *arbitrary insertion algorithm [6]* which is improved by optimization in a later phase. Used heuristic is presented below:

1. In the beginning we have to specify number of vertices - n_{start} which will be used for calculating partial solution ($n_{start} < n$).
2. Randomly select n_{start} vertices and write temporary distance matrix D_{temp}.
3. Find a optimal solution for TSP with distance matrix D_{temp}.
4. Randomly select one of the remaining vertices (which is still not included in cycle) and insert it into solution from step 3 in cheapest way possible.
5. Repeat step 4 until all of the vertices are used.
6. Keep this tour solution, say S.
7. Repeat n - times steps 1 through 6.
8. Compare the current solution with solution S. Keep the better one.

3.2 USAGE OF HEURISTIC ON REAL EXAMPLE

The heuristic was first applied to the symmetric example elaborated in section 2.2 and the optimal solution was found instantly. The next example, corresponding to physical tests in the workshop, is asymmetric. In this case, distance matrix (shown in Table 2) is not symmetric because we have to use different tools in order to mount our specimen into testing rig. The values in Table 2 denote time in minutes which is needed to set up testing rig from one test to another.

Table 2: Distance matrix for real physical testing [5].

	Adyn	Bdyn	Cdyn	Ddyn	Edyn	Fdyn	Gdyn	AzF	AdF		FdD
Adyn	0	20	20	30	20	20	30	5	5		20
Bdyn	30	0	20	30	30	30	30	30	30		30
Cdyn	30	20	0	30	30	30	30	30	30		30
Ddyn	40	30	30	0	10	10	30	40	40		10
Edyn	40	30	30	20	0	20	30	40	40	…	20
Fdyn	20	30	30	10	20	0	30	20	20		5
Gdyn	30	30	30	30	30	30	0	30	30		30
AzF	5	20	30	20	30	30	30	0	5		30
AdF	5	20	30	20	30	30	30	5	0		30
					⋮					⋱	5
FdD	35	30	30	15	35	5	30	35	35	5	0

The proposed heuristic was applied on distance matrix showed in Table 2. The calculated solution for ATSP is shown in Figure 4. In this case, we have no proof of optimality. Further analysis is needed to provide either optimality proof or the approximation error.

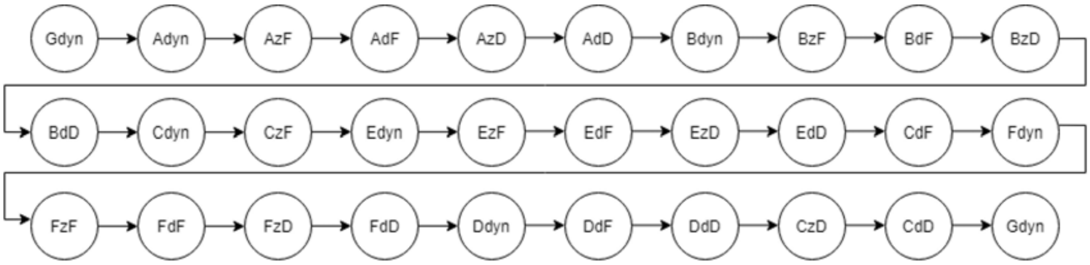


Figure 4: Solution for ATSP

With proposed solution we could set up tests for all loading cases in 280 minutes.

4 CONCLUSIONS AND FUTURE WORK

Within this article we have described the usage of methods for discrete optimisation on a real engineering problem. The optimal testing sequence in FEM calculation for valuation of trailing arm and the optimal sequence of physical testing for all loading cases of trailing arm were observed to correspond to optimal solutions of the corresponding travelling salesman problems. The well-known method based on minimal spanning tree and teh heuristic proposed by Žerovnik in [3] with few minor modifications were used. The heuristic was applied to both symmetric and asymmetric TSP. It is interesting that proposed algorithm finds solution faster than the one we have used for solving TSP. With our heuristic we determined that all of the tests can be set up in 280 minutes.

In the future it would be interesting to further test and improve our heuristic. In reality, several pieces of the same product may be tested at different testing positions i n the workshop yielding a more complex problem where the optimal sequence my be replaced by a set of subsequences. For good job scheduling it would be essential that test duration for each loading case is also included in the calculation.

References

[1] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnoy Kan and D. B. Shmoys (1985). The Travelling Salesmen Problem. John Wiley & Sons, New York.

[2] G.Reinelt (1991), TSPLIB - Traveling Salesman Problem Library. In: European Journal of Operations Research 52, p.125.

[3] J. Brest, J. Žerovnik (2005), A Heuristic for the Asymmetric Travelling Salesman Problem. MIC2005, The 6th Metaheuristics International Conference, p. 145 - 150.

[4] J. Žerovnik (2011), Osnove teorije grafov in diskretne optimizacije. Univerza v Mariboru, Fakulteta za strojništvo, Maribor.

[5] TPV d.o.o. (2018), TPV d.o.o. company archives, Novo mesto.

[6] D. J. Rosenkrantz, R. R. Stearns and P. M. Lewis (1997), An Analysis of Several Heuristics for the Traveling Salesman Problem, In: SIAM J. Comput. 6, 563-581.

DISTRIBUTED MANUFACTURING NODE CONTROL WITH DIGITAL TWIN

Miha Pipan
University of Ljubljana, Faculty of Mechanical Engineering
1000 Ljubljana, Slovenia
E-mail: miha.pipan@fs.uni-lj.si

Jernej Protner
University of Ljubljana, Faculty of Mechanical Engineering
1000 Ljubljana, Slovenia
E-mail: jernej.protner@fs.uni-lj.si

Niko Herakovič
University of Ljubljana, Faculty of Mechanical Engineering
1000 Ljubljana, Slovenia
E-mail: niko.herakovic@fs.uni-lj.si

Abstract: In this paper, we present a new and tested approach for integration of distributed manufacturing network in Smart Factory where the system is controlled using Digital Twin and Digital Agents instead of MES system. Standard MES and ERP systems are in general based on the centralization of all data and execution of decision-making algorithms in cloud server. However, the newest IT trends use new decentralised management systems to store data and execution of decision making algorithms to shorten response time of network and increase security and data change traceability.

Keywords: Manufacturing, Distributed Systems, Local Services, Digital twin

1 INTRODUCTION

Smart Factory that incorporates guidelines of Industry 4.0 must be based on following concepts: Internet of Things, Industrial Internet, Smart Manufacturing and Cloud based Manufacturing. The definition of Smart factory suggests use of [1]: Horizontal and vertical system integration, Big/Smart data and its Analytics, Collaborative and autonomous robots, Simulation models and digital twins, Cybersecurity, Additive manufacturing, Industrial IoT (IIoT) and Augmented Reality. The goal of Smart Factory must not only be customisable manufacturing (customisable product) but also customisable manufacturing process (production process that can be upgraded, downgraded and changed) [2].

2 SMART FACTORY

Smart Factory has to have distributed hierarchical structure, which enables customisable manufacturing process [3]. The distributed manufacturing nodes use the internal services to enable its general connectability/functionality and have faster response times since all decision-making, process optimisation, quality checking and error solving is done locally without communication lag. The nodes can communicate between them and with Digital Twin to control and optimise production.

2.1 SMART FACTORY AND DIGITAL TWIN

The backbone of any Smart Factory is Digital Twin of all logistics and processes upgraded with Digital Agents (Artificial Intelligence). Digital Twin must be connected to every manufacturing Node in Smart Factory (Figure 1). This way optimisation of global processes (decisions of production order based on times, warehouse and current production status) can be done. This Digital Twin and Digital Agents can use white or black box methods for optimisation and therefore enable local optimisation.

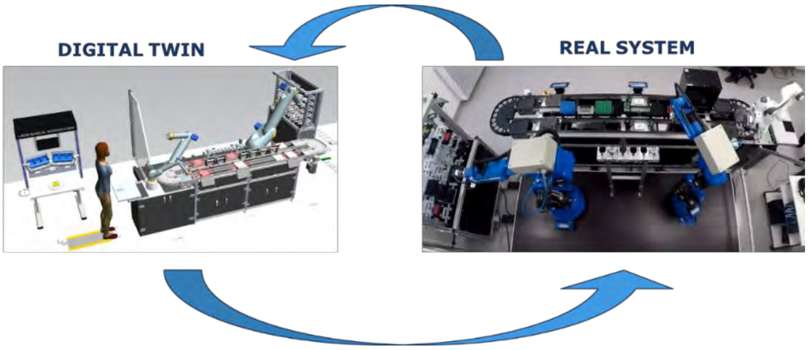


Figure 1: Real-time communication between Digital Twin and real system is backbone of Smart Factory.

2.2 DIGITAL TWIN AS MES SYSTEM

Digital Twin can replace classical MES system if it incorporates all logistics and processes presented in real system, is integrated in manufacturing IT structure and gathers data from ERP system (orders, warehouse status, workers and their schedule, …) and manufacturing Nodes (connected manufacturing units, assembly status, automatic inspection units…) as shown on Figure 2.

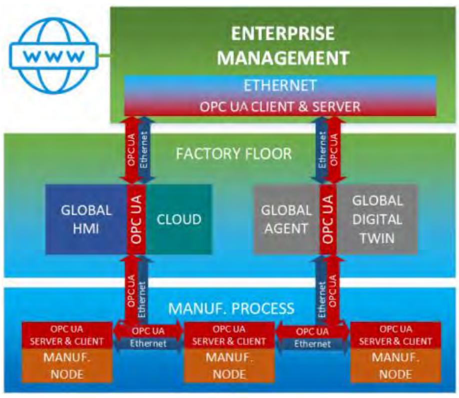


Figure 2: Proposed distributed manufacturing system with Digital Twin.

2.3 DATA EXCHANGE

The used (Figure 3) model uses directed communication between enterprise management systems to manufacturing/assembly process. All customised orders are structured based on current manufacturing network structure and optimal manufacturing order obtained from Digital Twin with Digital Agents. The manufacturing orders are send to genesis module, that acts as input for production process and encrypts manufacturing/assembly instructions and planned times for every module that is needed for successful complement of the customised production order. The manufacturing data is located on every product (RFID tag, microchip…). These readable/writable data modules on the product serve as data transfer protocol between manufacturing modules and are used to conduct local piloting of manufacturing process.

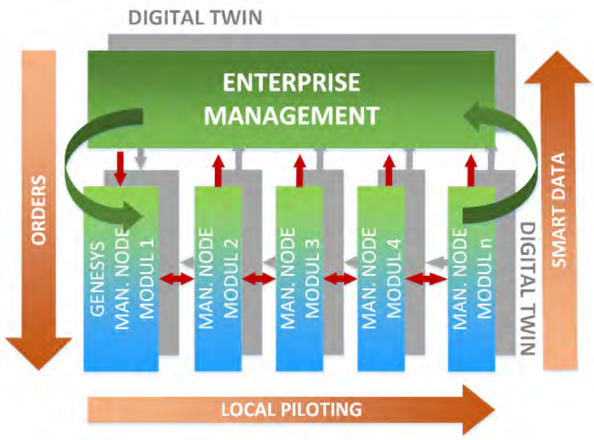


Figure 3: Proposed distributed manufacturing system with digital twin.

During the manufacturing process Smart Data is sent via local network to enterprise management system to enable tracking of orders and status of all Nodes. Smart data represents data filtered with local agent. The production/assembly faults are locally and globally handled by manufacturing Nodes with help of the Digital Twin that finds the optimal action and reorders manufacturing if needed.

3 MANUFACTURING ORDER EXECUTION

In this section we describe the information flow from the initial selection of the desired products, to the process of assembly of the desired products in an actual demo smart factory that was built in our laboratory LASIM. As stated before in Section 2, the framework of smart factory consists of distributed Nodes which are interconnected over local area network. For data transmission between the Nodes we used OPC UA communication protocol, since it is an open source protocol, it can be used cross-platform (it is not tied to a specific programming language or operating system), it is secure and it was a common link between the existing equipment and newly acquired hard- and soft-ware [4]. It is based on the client-server concept, where the server is an application that provides its data to other applications, and the client is an application that requires or acquires data.

As can be seen on Figure 4, our smart factory consists of ten distributed Nodes. Nodes 1 to 6 represent controllers on the assembly stations on the production line, Node 7 is a controller of a warehouse, Node 8 is a PLC controller responsible for the information flow between Nodes 1 to 6 and sensors, position units and stoppers on the line, Node 10 represents the controller on a smart manual workstation and Node 9 represents a computer which contains the Digital Twin of the smart factory. Each of the ten Nodes is equipped with either an OPC UA server, client, both or a number of different OPC UA clients. Information flow between the appropriate server and client is represented with a specific colour, for example, the communication between the server on Node 3, appropriate clients on Node 9 and a client on Node 4 is represented with light blue colour. It should be noted that even though the Nodes on the Figure 4 seem centralized around Node 8, this is not the case. Node 8 serves only as a broker between the Node 9, Nodes 1 to 6 and hardware on the production line. It is not equipped with advanced algorithms and it merely executes simple tasks like counting the number of pallets on the line between the assembly stations.

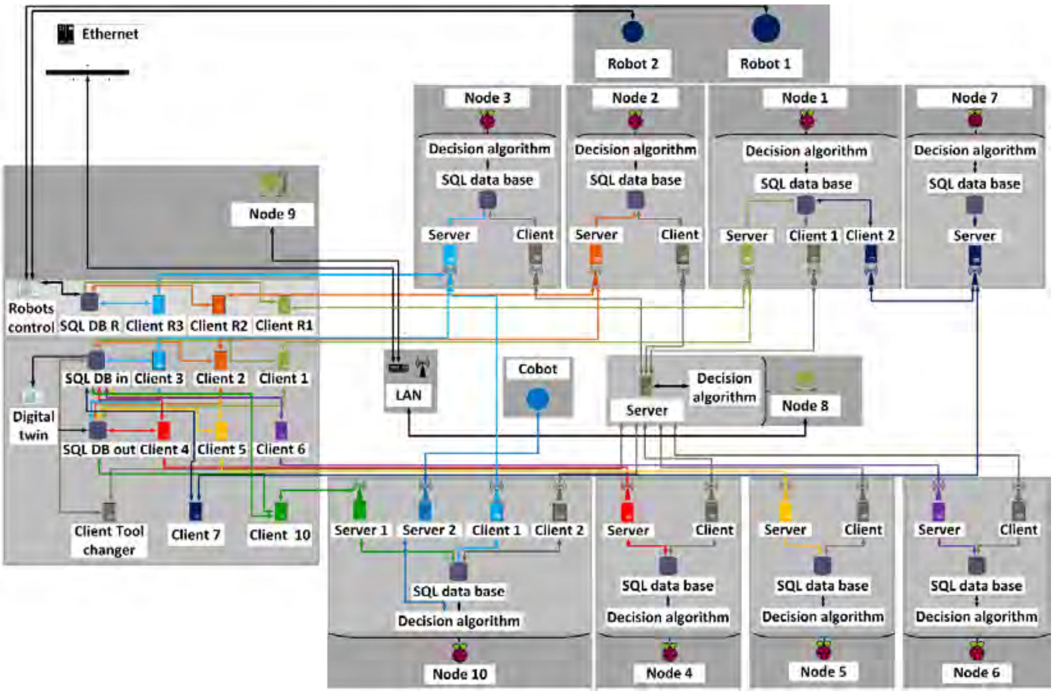


Figure 4: Scheme of connections between the distributed Nodes in demo smart factory LASIM

3.1 DIGITAL TWIN and ORDER OPTIMISATION

Let's assume that we would like to assemble five products (three mosaics of different colours and shapes (letters) and two raspberry pi cases). Through use of the GUI (Graphical User Interface) of the Digital Twin we select the desired products. Digital Twin first checks the state of the buffers and then reports which orders are accepted as shown on Figure 5.

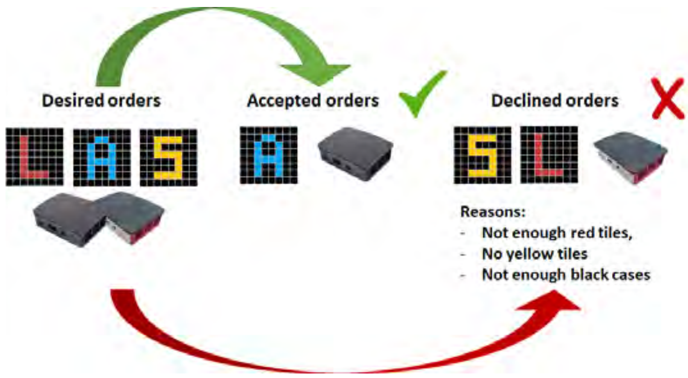


Figure 5: Overview of desired orders in digital twin

When either some or all desired orders are approved, Digital Twin optimizes the sequence of orders using intelligent algorithms developed in our department, generates the necessary data and writes it to the appropriate SQL database ("SQL DB out" on Figure 4). It first generates the data which is going to be recorded on the RFID tag on each of the pallets (manufacturing data) and then appropriate data for each of the assembly stations (nodes) separately. Manufacturing data (for example STAT 000000000000RPI-OH-RN00002214100) contains status of the order (STAT), empty space (zeroes) for the start time and date, name of the product (RPI-OH-RN), serial number of the product (N00002) and lastly six-digit assembly code which defines what protocol should be executed on each of the assembly stations (214100).

3.2 MANUFACTURING ORDERS EXECUTION

In this section the communication between the nodes in our smart factory, during manufacturing order execution is presented. It should be noted that this communication depends on the six-digit assembly code written to the RFID tag. As stated before in Section 2.3 RFID technology is used as a data transmission protocol between the assembly stations (nodes 1 to 6). When the production line is started and the first pallet with an empty RFID tag arrives to the first assembly station or Node 1 (Figure 4) the following happens:

1. Server on Node 1 sends a request to the Client 1 on Node 9 (light green colour) to send the manufacturing data (STAT 000000000000RPI-OH-RN00002204100) to be written on the RFID tag.
2. Decision algorithm on Node 1 checks the six-digit assembly code (204100) and follows the appropriate protocol. Since the first digit is “2” a task is to be executed on the Node 1. Depending on the digit in the first place of the code, the pallet could just move forward (if no task is to be executed).
3. Decision algorithm on Node 1 writes the actual start date and time of the operation in its local SQL database and on the RFID tag. Then it reads the serial number written on the RFID tag (N00002), and saves it in its local SQL database.
4. Server on Node 1 then sends the request to Client 1 on Node 9 (light green colour) to send the appropriate manufacturing order data (Table 1) according to the acquired serial number. Node 9 sends information about start and finish time of the assembly operation and task 1. Task 1 is a job sequence to be sent to robot 1.

Table 1: Manufacturing order data according to serial number

Serial	Start time	Finish time	Task 1
N00002	14:30:12 03.05.2019	14:34:03 03.05.2019	X1T6H111

5. Since the task is about to be executed on Node 1, Client 1 (light brown colour) requests the Server on Node 8 to lift the positioning unit.
6. Client 2 (dark blue colour) on Node 1 then asks Server on Node 7 whether the warehouse unit is ready for the robot to acquire the base part for the first order.
7. Server on Node 1 sends Task 1 to Client R1 on Node 9 (light green colour) which then communicates with robots.
8. The robot executes the first task (job sequence) and puts the base part on the pallet. Decision algorithm writes the actual finish time to its local database and the Server on Node 1 sends the information about the actual start and finish times back to the Node 9, for analysis in Digital Twin.
9. Server on Node 1 requests the lowering of the position unit from Server on Node 8 and pallet moves forward.

Since the assembly code on RFID tag has number “0” in the second place, the pallet with the base part skips Node 2 and stops on Node 3 since the assembly code has number “4” on third place. This means that Node 3 only sends the serial number of this product to Node 10 (smart manual station which is connected to production line over collaborative robot):

1. Server on Node 3 sends information about serial number to Node 10 (light blue colour). Client on Node 3 requests lowering of the position unit from Server on Node 8 (light brown colour).
2. Server on Node 10 then sends the request to Client 10 on Node 9 (dark green colour) to send the appropriate manufacturing order data according to the acquired serial

number. Node 9 sends information about start time, finish time of the assembly operation.
3. When the manual assembly is finished, Server on Node 10 sends the information about the actual start and finish times of the assembly operation back to the Node 9, for analysis in digital twin.

Pallet next stops at Node 4 (this station performs quality check with machine vision), since the fourth number in the assembly code is "1".

1. Decision algorithm on Node 4 first performs the same task a decision algorithm on Node 1 does in step 3 (except writing data to RFID tag).
2. Then it performs the same task a decision algorithm on Node 1 does in step 4. Node 9 sends information about start time, finish time and correct machine vision analysis.
3. After the quality check on Node 4 is finished, its server sends the information about actual start and finish times of the operation and the analysis of the machine vision back to Client 4 on Node 9 (red colour) for digital twin to analyse whether the product is correctly assembled or if it needs to be repaired. If digital twin decides that the product is correctly assembled, then the first digit of the assembly code on the RFID tag is changed to "6", which means that the product should be put in the warehouse. If not different digit is written and the repair protocol is executed.
4. Client on Node 4 requests lowering of the position unit from Server on Node 8 (light brown colour).

When the pallet arrives back to Node 1, its decision algorithm first checks the first digit in the assembly code on the RFID tag. If the digit is "6", the product is finished and should be taken to the warehouse.

1. Server on Node 1 sends the request to Client 1 on Node 9 (light green colour) to send the manufacturing order data (Table 3) according to the acquired serial number. Node 9 sends information about start and finish time of the final operation and task 2. Task 2 is a job sequences to be sent to robot 1.
2. Steps 5 to 8 from the first arrival of the pallet to Node 1 are repeated.
3. When the data is sent back to Node 9, server on Node 1 asks the Client 1 on Node 9 if there are any more orders. If not the decision algorithm deletes the data on the RFID tag and send the pallet away, if orders are available the steps are repeated from the start.

4 CONCLUSIONS

The distributed manufacturing nodes with Digital Twin present backbone for Smart Factory if the hierarchy and communication channels are correctly defined and organised. This way we can achieve tradability of product and minimise the data that needs to be exchanged (Smart Data vs. Big Data).

References

[1] Vaidya, S., Ambad, P. and Bhosle, S. (2018). Industry 4.0 – A Glimpse. *Procedia Manufacturing*, 20, 233 – 238.

[2] Fernández-Miranda, S.S., Marcos, M., Peralta, M.E. and Aguayo, F. (2017). The challenge of integrating Industry 4.0 in the degree of Mechanical Engineering. *Procedia Manufacturing*, 13, 1229 – 1236.

[3] Thames, L. and Schaefer, D. (2016). Software-defined cloud manufacturing for Industry 4.0. *Procedia CIRP*, 52, 12 – 17.

[4] Schleipen, M., Gilani, S-S., Bischoff, T. and Pfrommer, J. (2016). OPC UA & Industry 4.0 – enabling technology with high diversity and variability. *Procedia CIRP*, 57, 315 – 320.

SCRAP DETERMINATION WITH PROCESS MINING – LITERATURE REVIEW

Jaka Toman
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, 4000 Kranj
E-mail: jaka.toman@gmail.com

Uroš Rajkovič
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, 4000 Kranj
E-mail: uros.rajkovic@um.si

Mirjana Kljajić Borštnar
University of Maribor, Faculty of Organizational Sciences
Kidričeva 55a, 4000 Kranj
E-mail: mirjana.kljajic@um.si

Abstract: In this paper we address the cause for scrap on a production line. Usually, the root cause analysis is a method of choice, but sometimes yields no results. Based on the assumption that a data driven approach could be used to address this problem, we propose to combine the root cause analysis with process mining approach. For this purpose, we have conducted a literature review, in which 4981 relevant articles regarding root cause analysis and process mining were found. The review of articles provides better understanding of root cause analysis and process mining and overview publications to date.

Keywords: Root-Cause Analysis, Process Mining, Manufacturing, Scrap

1 INTRODUCTION

Manufacturing companies are constantly striving to improve their processes. Thomas, Hayes and Wheelwright [1] introduced the term "world-class manufacturing" in 1985, defining organizations which achieve a global competitive advantage through use of their manufacturing capabilities as a strategic weapon. Effective Root Cause Analysis (RCA) is arguably one of the most valuable management tools in any organization [2]. When looking at a large context, RCA is a part of a problem-solving process. RCA is therefore one of the core building blocks in an organization's continuous improvement efforts, also on scrap management topic [3]. However, most of these RCA approaches are data oriented. Some authors argue that one needs to carefully combine process-centric and data-centric approaches [4]. Therefore, process mining emerged as a new scientific discipline on the interface between process models and event data [5].

The purpose of this article is to bridge this existing literature gap and review the works published on root cause analysis and process mining topics combined. Research objectives of this paper are to:

- clarify the definition of root cause analysis and process mining,
- classify and summarize all relevant articles and
- propose a research framework to answer the following research question:

"Can reasons for scrap be determined with combining approaches of RCA and process mining?"

1.1 Root Cause Analysis

RCA is a collective term used to describe a wide range of approaches, tools, and techniques for uncovering causes of problems [3]. Consequently, the definition of a root cause and root cause analysis varies between authors and root causes methodologies, with different levels of causation being adopted by different systems [6]. Some of the approaches are focused more on identifying the true root causes than others and some are more general problem-solving techniques. Others simply offer support for the core activity of root cause analysis. Some tools are characterized by a structured approach, while others are more creative in nature [3].

Reid and Smyth-Renshaw [7] provided examples of RCA use in various contexts: manufacturing improvements, software projects, crime reports and client incidents. Despite this versatility we can find one basic definition that prevails in literature and will be used also in our paper. Paradise and Butch [8] defined a root cause as the most basic cause that can reasonably be identified and that management has control to fix. Root cause analysis is the task of identifying root causes. This definition contains three key elements. Root causes must be so basic that one can fix them. On the other hand, given that fixing them is the whole point, it is not reasonable to further split root causes into more basic causes [9].

Through history different RCA methodologies developed. Some of more known are Ford's 8D method, Six Sigma [10] and A3 thinking [11]. The majority of root causes analysis methodologies reviewed were essentially checklists of potential root cause factors to stimulate thought [6]. Supporting these methodologies, various tools and techniques were developed. Some of more known tools and techniques are Pareto Chart, Six Sigma, Five Why's and Fishbone diagram. Due to vast number of tools, Andersen and Fagerhaug [3] suggested grouping them according to their purpose: Problem understanding, Problem cause brainstorming, Problem cause data collection, Problem cause data analysis, Root cause identification, Root cause elimination, and Solution implementation.

These tools contribute in their own way to the RCA. Some are best applied sequentially while others can be applied at many different points in the analysis. In general, these tools are simple and easy to use. But with growing complexity of organizational environments, new tools have emerged, based on data mining software, leveraging computer capabilities and vast amount of available data. An extensive literature review of data mining in quality improvement field has been made by Köksal, Batmaz and Testik [12].

1.2 Process Mining

Modern RCA approaches (machine learning, data mining…) are strongly data-oriented techniques. They typically focus on classification, clustering, regression, or rule-learning problems. Process aspect is usually not included. Recent developments in the field of process mining are filling that gap.

Process mining is positioned as the missing link between process model analysis and data-oriented analysis [13]. It represents a technique of obtaining useful process related information from event logs and extends the approaches generally found within Business Process Management [14].

It aims to identify, monitor and improve real processes by acquiring knowledge from archived data in the contemporary information systems. "The basic idea of process mining is to diagnose processes by mining event logs for knowledge" [15]. It allows to analyse these event logs, sometimes also referred to as 'audit trail', 'trans- action log' or 'history'. Records in these logs are called events, or 'audit trail entries'. In process mining, each event needs to refer to an activity for a specific case or process instance [16]. Preferably, each event also refers to the performer, the originator of the event, and a time stamp. For each process under investigation these are the constraining assumptions. If available data fulfils these assumptions, process mining can be applied on that particular process. Event logs are the starting point of process mining. The data of the event log can be mined and different aspects about the underlying process can be analysed.

However, it is important to be aware that an event log contains only example behaviour. We cannot assume that all possible runs have been observed. In fact, an event log often contains only a fraction of the possible behaviour [17]. Often event logs store additional information about events and these additional data attributes may be used during analysis. Therefore, it is important to confront existing tools and techniques with event logs taken from real-life applications [13].

When dealing with process mining techniques we can distinguish between three perspectives of process models:

- The control-flow perspective that deals with both the existence of certain process elements and the ordering in which these process elements can occur. The goal of mining this perspective is to find a good characterization of all possible paths.
- The organizational perspective that focuses on the analysis of the organization behind the business process. Consequently, the key question is ‘‘who performed which process elements’’. The goal is to either structure the organization by classifying people in terms of roles and organizational units or to show relations between individual performers.
- The case data perspective that represents the information elements which are produced, used or manipulated during the process. Cases can be characterized by their path in the process or by the originators working on a case.

2 RESEARCH METHODOLOGY

A comprehensive literature review of journal articles dealing with "Root Cause Analysis" and "Process Mining" topics was conducted. This review was based on a similar approach used in [18] for big data, [19] for RFiD, and in [20] for review of e-commerce related topics. Three key characteristics relevant for our paper are: conduction of literature review, classification of relevant journal articles, and identification of research gaps.

A comprehensive search using the descriptor "Root Cause Analysis" and "Process Mining" was conducted within following databases: Web of Science, Scopus, Emerald and ProQuest. We started our review on May 2nd, 2019 and ended on May 30th, 2019. Search was made within topics, titles, abstracts and full texts. We excluded irrelevant articles considering titles and category. From research perspective, mainly articles with Medicine research area were excluded. The review resulted in total of 4981 results. All the results were then exported to Mendeley, a reference management software package, for further analysis. Export included articles references, abstracts and full texts in pdf format. These articles were then a subject of further, more detailed review. Each article was then reviewed to assess relevance to our research objectives and identify duplicated ones. Due to our filed of interest, articles were then categorized according to their topic.

3 RESULTS

Figure 1 shows a distribution of RCA relevant articles over the years. Separately are show articles that address RCA topic relating to manufacturing; scrap, yield, defects and with data mining and machine learning topics. From first publications in year 1981 we can notice steady increase of publicized articles. In year 2017 12% of all articles were published. From 2013 to 2017, in average 115 articles per year were published. 52 % of all articles were published in this last 5 years. This continuous increase of numbers of publications is showing an increase of interest about RCA topic.

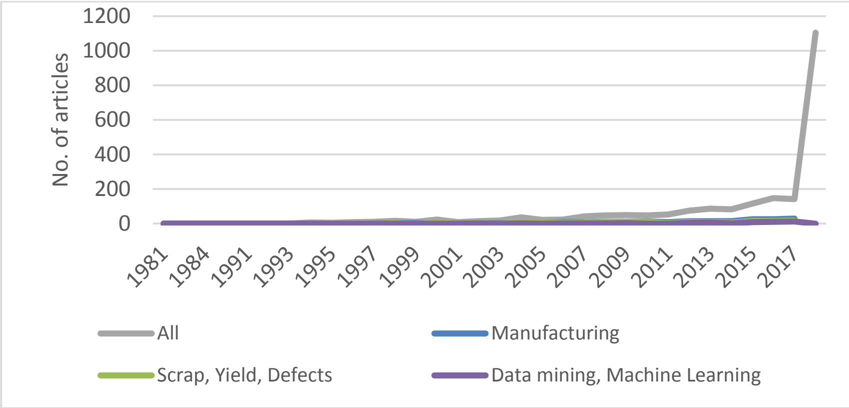


Figure 1: Distribution of articles by the year of publication (RCA related topics)

Figure 2 shows a distribution of Process Mining relevant articles. Like RCA topic we can notice continuous increase in number of publications. Number of publications are rising from the year 1985. However, from 2011 we can notice a sharp incline of articles, representing 80% of all published articles. Figure 2 also shows the number of Process Mining articles in combination with Manufacturing and Root Cause topics. However, there are less than 6% of such articles.

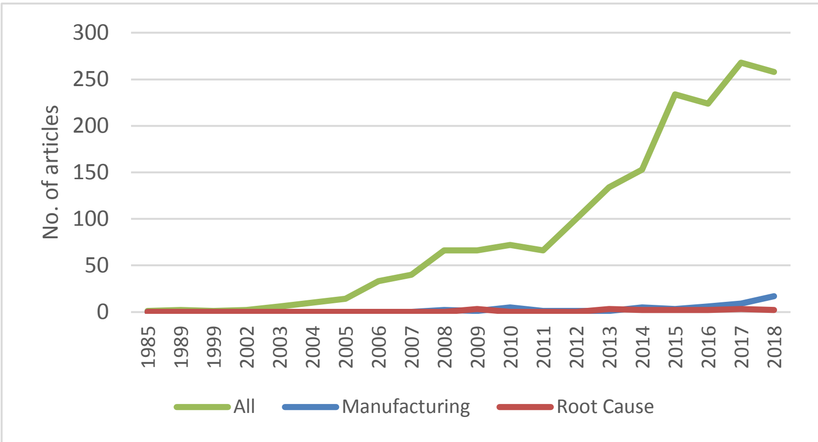


Figure 2: Distribution of articles by the year of publication (Process Mining related topics)

In these years process mining has received a fair amount of research attention resulting in a range of techniques. Such as process discovery techniques [21, 22], techniques for the analysis of event log data [16, 23] techniques for trace classifications [24], process metrics [25] and specific application areas [16, 26, 27].

In Table 1 we present the matrix of published articles by research topics. Matrix shows there have been researches made in the manufacturing and scrap management area. However, these topics were generally addressed separately from root cause analysis perspective and from process mining perspective. Matrix therefore shows a research gap in addressing this topic in combination with RCA and process mining. Managing scrap in combination with process mining is poorly researched.

Table 1: Matrix of published articles by research topics, highlighting research gap.

	Manufacturing	***Scrap***
Root Cause Analysis	181	18
Process Mining	53	**0**
Root Cause Analysis + Process Mining	**0**	**0**

4 CONCLUSIONS

The problem of scrap management in a production line was addressed by identifying a research framework combining root cause analysis and process mining. For this purpose, the literature review was conducted In the process of the literature search, we found 4981 relevant articles, referring to RCA and Process Mining. The results show a constant growth and interest in the topics (RCA related articles addressing manufacturing, production, scrap, yield, defects and data mining, machine learning) since more than 50% of published articles are dating in past 5 years. This growth is also seen in subcategories of manufacturing, scrap and data mining. We found similar observations in reviewing articles on Process Mining topic. From 2011 we can notice a sharp increase of articles, representing 80% of all published articles. However, we found little Process Mining articles in combination with Manufacturing and Root Cause topics (less than 6%). We continue the review of articles by their keywords. Review showed that for each topic separately (root cause analysis and process mining) there are some papers in manufacturing environment, however no article was found combining root cause analysis and process mining topics in terms of manufacturing.

Our review showed a growing interest in root cause analysis and process mining topics. Accordingly, the number of publications on these topics are growing. Despite many articles in manufacturing environments for both root cause analysis as process mining we did not recognize any paper combing these approaches. Especially in terms of scrap management in manufacturing environments. Therefore, we recognize a potential for case study on this topic.

Acknowledgement
Mirjana Kljajić Borštnar and Uroš Rajkovič were supported by the Slovenian Research Agency, ARRS, through research programme P5-0018.

References

[1] Thomas, L. J., Hayes, R. H., & Wheelwright, S. C. (1985). Restoring Our Competitive Edge: Competing through Manufacturing. Administrative Science Quarterly, 30(2), 305. https://doi.org/10.2307/2393118

[2] Latino, R. J. (2015). How is the effectiveness of root cause analysis measured in healthcare? Journal of Healthcare Risk Management, 35(2), 21–30. https://doi.org/10.1002/jhrm.21198

[3] Andersen, B., & Fagerhaug, T. (2006). Root cause analysis: simplified tools and techniques (2nd ed.). Retrieved from https://www.amazon.com/Root-Cause-Analysis-Simplified-Techniques/dp/0873896920

[4] Azzini, A., & Damiani, E. (2014). Process Mining in Big Data Scenario. Retrieved from https://pdfs.semanticscholar.org/3e41/a4da19832a4338b363d0590551a447121fa0.pdf

[5] Van Der Aalst, W. (2013). Association for Information Systems AIS Electronic Library (AISeL) ECIS 2013 Completed Research ECIS 2013 Proceedings "Mine your own business": using process mining to turn big data into real value. Retrieved from http://aisel.aisnet.org/ecis2013_cr

[6] Livingston, A. D., Jackson, G., & Priestley, K. (2001). Safety Executive Root causes analysis: Literature review Root causes analysis: Literature review. Retrieved from http://www.hse.gov.uk/research/crr_pdf/2001/crr01325.pdf

[7] Reid, I., & Smyth-Renshaw, J. (2012). Exploring the Fundamentals of Root Cause Analysis: Are We Asking the Right Questions in Defining the Problem? QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL, 28(5, SI), 535–545. https://doi.org/10.1002/qre.1435

[8] Paradies, M., & Busch, D. (1988). Root Cause Analysis at Savannah River. In Proceedings of the IEEE Conference on Human Factors and Power Plants (pp. 479–483).

[9] Klaus, J. (2003). Using Root Cause Analysis to Handle Intrusion Detection Alarms. Universit¨ at Dortmund am Fachbereich Informatik. Retrieved from https://d-nb.info/969432127/34

[10] Gamal Aboelmaged, M. (2010). Six Sigma quality: a structured review and implications for future research. International Journal of Quality & Reliability Management, 27(3), 268–317. https://doi.org/10.1108/02656711011023294

[11] Sobek, D. K., & Smalley, A. (2008). Understanding A3 thinking: a critical component of Toyota's PDCA management system.

[12] Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. https://doi.org/10.1016/j.eswa.2011.04.063

[13] Van der Aalst, W.M.P., Reijers, H. A., Weijters, A. J. M. M., van Dongen, B. F., Alves de Medeiros, A. K., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. Information Systems, 32(5), 713–732. https://doi.org/10.1016/j.is.2006.05.003

[14] Mušič, G., & Rojec, P. (2012). PROCESS MINING OF PRODUCTION MANAGEMENT DATA FOR IMPROVEMENT OF PRODUCTION PLANNING AND MANUFACTURING EXECUTION. Retrieved from http://www.msc-les.org/proceedings/emss/2012/EMSS2012_483.pdf

[15] Van der Aalst, W. M. P., de Medeiros, A. K. A., & Weijters, A. J. M. M. (2005). Genetic Process Mining (pp. 48–69). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11494744_5

[16] Jans, M., Martijn Van Der Werf, J., Lybaert, N., & Vanhoof, K. (2011). A business process mining application for internal transaction fraud mitigation. https://doi.org/10.1016/j.eswa.2011.04.159

[17] Rozinat, A., & Van der Aalst, W. M. P. (2008). Conformance checking of processes based on monitoring real behavior. Information Systems, 33(1), 64–95. https://doi.org/10.1016/j.is.2007.07.001

[18] Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. Intern. Journal of Production Economics, 165, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

[19] Lim, M. K., Bahr, W., & Leung, S. C. H. (2013). RFID in the warehouse: A literature analysis (1995–2010) of its applications, benefits, challenges and future trends. International Journal of Production Economics, 145(1), 409–430. https://doi.org/10.1016/j.ijpe.2013.05.006

[20] Ngai, E. W. T., & Wat, F. K. T. (2002). A literature review and classification of electronic commerce research. Information & Management, 39(5), 415–429. https://doi.org/10.1016/S0378-7206(01)00107-0

[21] Goedertier, S., Martens, D., Vanthienen, J., Baesens, B., Frasconi, P., Kersting, K., … Tsuda, K. (2009). Robust Process Discovery with Artificial Negative Events. Journal of Machine Learning Research (Vol. 10). Retrieved from http://jmlr.csail.mit.edu/papers/volume10/goedertier09a/goedertier09a.pdf

[22] Van Der Aalst, Wil M P, & Song, M. (2004). Mining Social Networks: Uncovering Interaction Patterns in Business Processes. Retrieved from http://www.buddygraph.com/

[23] Van der Aalst, W. M. P., & van Dongen, B. F. (2002). Discovering Workflow Performance Models from Timed Logs (pp. 45–63). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45785-2_4

[24] Ferreira, D. R. (2009). Applied Sequence Clustering Techniques for Process Mining. In Handbook of Research on Business Process Modeling (pp. 481–502). IGI Global. https://doi.org/10.4018/978-1-60566-288-6.ch022

[25] Dijkman, R., Dumas, M., van Dongen, B., Käärik, R., & Mendling, J. (2011). Similarity of business process models: Metrics and evaluation. Information Systems, 36(2), 498–516. https://doi.org/10.1016/j.is.2010.09.006

[26] Günther, C. W., & van der Aalst, W. M. P. (2007). Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics (pp. 328–343). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75183-0_24

[27] Wang, Y., Caron, F., Vanthienen, J., Huang, L., & Guo, Y. (2014). Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port. Expert Systems with Applications, 41(1), 195–209. https://doi.org/10.1016/j.eswa.2013.07.021

USE OF A STANDARD RISK MODEL AND A RISK MAP FOR PRODUCT DEVELOPMENT PROJECT PLANNING AND MANAGEMENT

Tena Žužek, Lidija Rihar, Tomaž Berlec, Janez Kušar
University of Ljubljana, Faculty of Mechanical Engineering
Aškerčeva 6, 1000 Ljubljana, Slovenia
E-mails: tena.zuzek@fs.uni-lj.si, lidija.rihar@fs.uni-lj.si, tomaz.berlec@fs.uni-lj.si, janez.kusar@fs.uni-lj.si

Abstract: Today's market is very unpredictable, products are becoming increasingly complex, and customer demands are constantly changing. The key to company's competitiveness and project's success is an effective risk management. This paper outlines the use of a standard risk model and a risk map for planning and managing a product development project. A risk map is a simple and demonstrative tool that graphically displays project risks with their levels of seriousness and enables a quick analysis of the impact of preventive and corrective measures on the level of risk. The use of the proposed approach is illustrated using the example of a project relating to a die-cast tool development.

Keywords: project management, risk management, standard risk model, risk map

1 INTRODUCTION

Today, companies are facing more and more complex products, rapid technology changes, customer's changing demands and unpredictable markets. To survive and remain competitive, companies must be able to adapt to the competitive environment and to manage uncertainties. In this regard, it is crucial that the risks are effectively managed and both preventive and corrective measures planned to eliminate or mitigate the consequences of risk events.

There are several models for project risk management: standard, simple, cascade and Ishikawa risk model. The paper will illustrate the use of the standard risk management model upgraded with the use of a risk map. The standard risk model separately examines a risk event and the consequences of the risk event, making it easier to plan the preventive measures on one hand and corrective measures on the other.

A risk map graphically shows the identified risks. The x-axis in the risk map represents the total loss (it may be expressed in units of time, money or quality), while the y-axis represents a risk probability. The graph also shows the threshold line of expected losses which represents still acceptable losses for a company. The graph allows an instant read-out of the critical risks and observation of how the preventive and corrective measures impact the risk level.

The aim of the paper is to present the risk management process using the standard risk model and risk map, and to present an analysis of the effects of the preventive and corrective measures using the example of a project relating to a die-cast tool development.

2 RISK MANAGEMENT

Efficient risk management is essential for the successful management of a product development project. Fundamentally, a risk is defined as a probability of the occurrence of an undesired event or a probability of the non-occurrence of a desired event [5]. Risks can only be prevented by identifying their origins and by systematically eliminating them [2]. Companies must prepare for the identified risks and formulate adequate preventive and corrective measures. Both, external risks, on which the companies have no influence (for instance changes on the market), and internal risks, which originate in the company and can be influenced upon must be taken into consideration [7]. An increasing emphasis is put on cognitive factors that can have a significant impact on the success of a project.

Risks are one of the key factors that impact the success of a project [3]. PMBOK® (*Project Management Body of Knowledge*) lists risk management as one of the basic areas of successful project management skills [1]. Project managers need to be trained to use appropriate risk management tools throughout the project and not only when adverse effects occur.

2.1 Standard risk model

Risk models are a tool for systematic risk management. They facilitate communication between stakeholders and allow for an easier risk identification and analysis. The models clearly show the identified risks and their origins [6].

There are several models for project risk management: standard, simple, cascade and Ishikawa risk model [6]. In practice, the Ishikawa model [4] proved to be the most useful for the identification of risks, while the use of the standard model shown in Figure 1 is proposed for a further analysis of the identified risks. The standard risk model relates the risk event, the impact of the risk event, and the total loss. The left side of the model refers to the very risk event, while the impact of the risk event is shown in the centre. This allows for a separate treatment of and reduction in risk occurrence reasons on one hand and consequence occurrence reasons on the other [6].

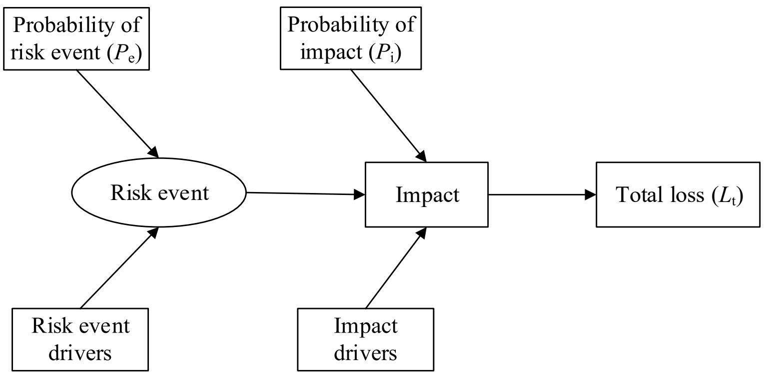


Figure 1: Standard risk model [6].

A project team assesses a probability of risk event occurrence P_e and a probability of the occurrence of its impact/consequence P_i. The probabilities are a subjective assessment and are normally based on experience from previous similar projects or on prepared decision-making tables prepared by an expert group. It is also necessary to assess the total loss L_t in the event of the occurrence of a risk event. The evaluated probabilities P_e and P_i and the total loss L_t serve as a basis to calculate the expected loss L_e (equation 1) which indicates a risk level of a certain event.

$$L_e = P_e \cdot P_i \cdot L_t \tag{1}$$

2.2 Risk map

The number of identified risks is usually high, while the available resources for dealing with the risks are limited. So, the risks need to be classified and attention and resources should be focused on those that pose the greatest danger to the project.

The risks are classified as to the expected loss L_e. The calculated values serve as a basis to decide which risks will be actively dealt with (active risks) and which will not be given more attention (inactive risks). It is advisable to create a list of 10 critical risk events (the list may also be shorter or longer, depending on the nature and objectives of the project), which are entered into a risk map [4]. A risk map is a graph, in which the x-axis represents the total loss L_t (it can be expressed in units of time, money or quality) and the y-axis represents a probability of a risk [6]. A probability of a risk is a product of the probability of risk event occurrence P_e and the probability of risk impact Pi. The graph also shows the threshold line of expected losses that represents still acceptable losses. The threshold line is defined by equation 2. L_e in the equation represents the expected loss that is still acceptable for the company.

$$\frac{L_e}{L_t} = P_e \cdot P_i \tag{2}$$

The threshold line divides the graph into two parts. The upper part (above the threshold line) represents the field of critical risks (Risk 1 in Figure 2), while the risks below the threshold line are non-critical risks (Risk 2 in Figure 2). The first ones will have to be addressed with more attention and adequate measures will be needed to reduce the risk. The risk level can be reduced by lowering the probability of the occurrence of a risk event P_e, by lowering the probability of the occurrence of a risk impact P_i or by lowering the total loss L_t. The goal is to have all risks below the threshold line.

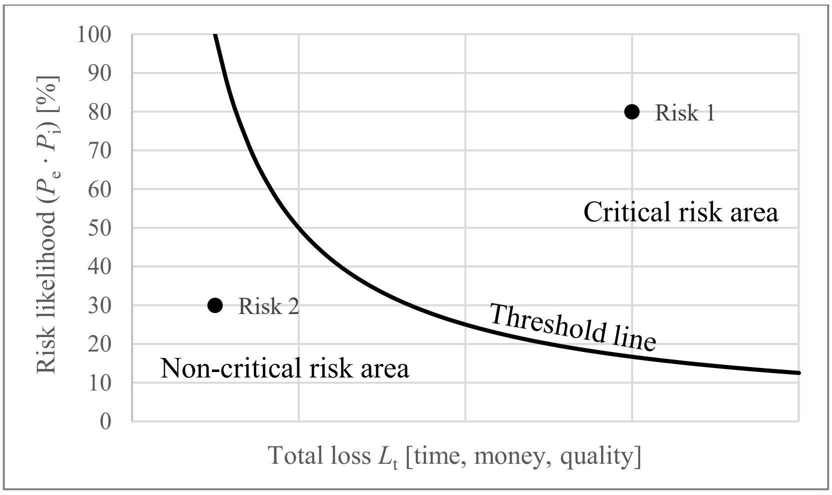


Figure 2: Risk map [6]

3 EXAMPLE CASE

In the following, an example of a project relating to the development of a die-cast tool is presented. Using the Ishikawa model, the project team first defined the major risk factors (project team, buyer, suppliers, development and technology, manufacture, quality control) and looked for all possible risks of the project within the individual factors. The project team studied the WBS (Work Breakdown Structure) of the project and assigned the identified risks to individual activities.

The project team also defined possible consequences of each risk event. Using the standard model, the team assessed a probability of the occurrence of a risk event P_e and a probability of the occurrence of a risk event impact P_i. The total loss L_t that would appear in the event of the occurrence of a risk event was also assessed. Then, the expected losses L_e for each risk event were calculated. The team classified the risks by the value of the expected losses. The losses can refer to monetary losses (extra costs) or to time losses (delays).

To illustrate the use of the standard risk model and the risk map, four most critical monetary-loss related risks of the project were selected. These risks are summarized in Table 1.

Table 1: Critical risks of the project for the execution of the order for the die-cast tool.

Risk	***Activity***	***Risk description***	P_e	***Impact description***	P_i	$P_e \cdot P_i$	L_t [€]	L_e [€]
T1	Design freeze	Late confirmation of documents by the buyer (delay: 1 month)	0.9	Plan for the tool not prepared in time	0.8	0.72	5 000	3 600
T2	Confirmation of first pieces	Rejection by the buyer	0.3	Corrections of the method under consideration of buyer's remarks	1.0	0.30	5 000	1 500
T3	Manufacture of first pieces	Poor quality of pieces	0.5	Corrections of the method and new manufacture of pieces	0.8	0.40	3 000	1 200
T4	Delivery of special tool	Late delivery	0.6	Manufacture of special parts not in time	0.7	0.42	500	210

The company had also defined the expected loss L_e, up to which they are prepared to risk. The maximum tolerable value of the expected losses was set at 1 000 €. The risk map from Table 1 is shown in Figure 3, the threshold line is plotted as well. We can see that risks T1, T2 and T3 are critical (they lie above the threshold line), while risk T4 is not critical (it lies below the threshold line).

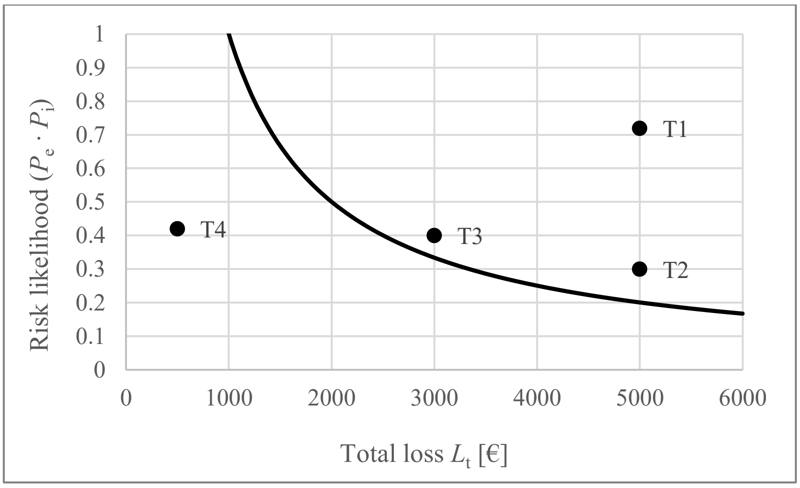


Figure 3: Map of monetary-loss related project risks

For critical risks, the project team must carry out a detailed analysis and prepare both preventive and corrective measures. The entire process of risk management will be illustrated on the example of the risk T1, i.e. design freeze. The risk T1 is the most critical among the identified risks (it has the highest expected loss) and it also appears as first in the project's timeline.

It is a known fact that the buyer often confirms the documentation later than originally agreed. Since the project team cannot complete the tool plan without the documentation being confirmed and continue its work according to the foreseen schedule, high extra costs are incurred. If the buyer is one month late with the confirmation, the total loss of the company amounts to 5 000 €. The expected loss calculated by the equation 1 equals 3 600 €, which is higher than the predetermined still acceptable value of 1 000 €. The project team must therefore prepare adequate preventive and corrective measures to lower the expected loss.

The main reason (risk event driver) for the buyer not to have confirmed the documentation in time lies in the fact that the timelines of the buyer are not harmonised with those of the company. The company therefore decided to coordinate the dates before signing the contract and to have the dates fixed in the contract. According to the project team's assessment, the probability for the documentation not being confirmed in time is reduced to 70% and at the same time half of the costs are passed on to the buyer in case of a delay. In this case, the total loss of the company is 2 500 €, and the expected loss is 1 400 €.

As the expected loss is still above the threshold value, the company decided to reduce the probability of an impact, by means of corrective measures, that is, the probability of late completion of plan for the tool in the event the documentation is not confirmed in time. If, despite the harmonized timelines, the buyer is still late in confirming the documentation, the company sends a written request to the buyer once a delay of one week has passed. The company assessed that the buyer provides the confirmation quite rapidly in this case and the probability for the tool plan not being prepared in time is thus reduced by 50%. The expected loss is reduced to 875 €, which is within the safe area.

The results of the envisaged measures are summarized in Table 2. The basic risk is designated T1, once a preventive measure was adopted, the risk was designated T1.1, and once a corrective measure was adopted, it was designated T1.2. The impact of the measures is also evident on the risk map shown in Figure 4.

Table 2: Results of the calculated impacts of the proposed measures to mitigate the risk T1.

Risk	P_e	P_i	$P_e \cdot P_i$	L_t [€]	L_e [€]
T1	0.9	0.8	0.72	5 000	3 600
T1.1	0.7	0.8	0.56	2 500	1 400
T1.2	0.7	0.5	0.35	2 500	875

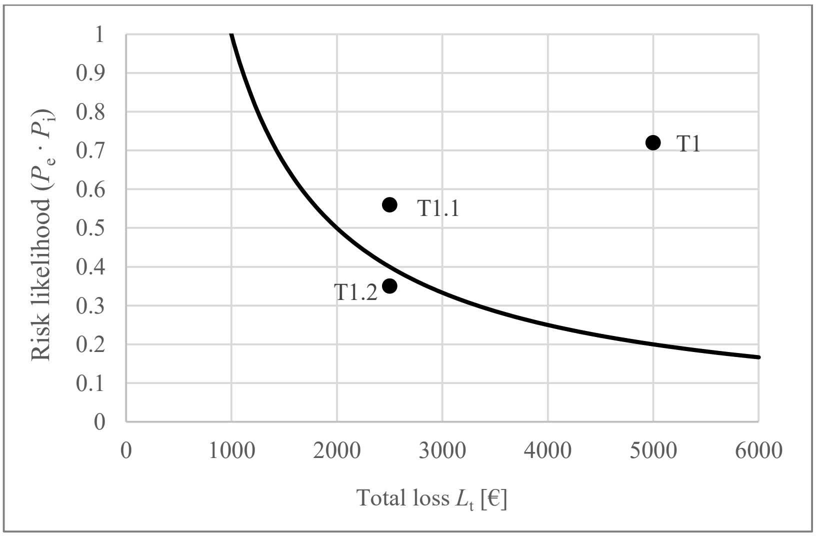


Figure 4: Map of project risk T1 (T1: the initial very critical risk; T1.1: a somewhat lower risk after the adoption of a preventive measure; T1.2: a non-critical risk after the adoption of corrective measure)

4 CONCLUSIONS

Efficient risk management is crucial for a successful business operation and project management. The paper proposes a project risk management model by using a standard risk model and a risk map.

The use of the proposed model has been successfully tested on the example of a development of a die-cast tool. The project team identified all potential project risks, assigned them to individual activities and determined the expected losses for each risk event using the standard model. Based on the calculated expected losses, the project team classified the risks and entered them into the risk map. A plan of preventive and corrective measures was prepared for the critical risks. The paper precisely outlines the risk management for the event of a late confirmation of documents by the buyer, which may occur in the design freeze activity. By adopting preventive and corrective measures, the project team managed to lower the risk level below the threshold value. The impact of individual measures is clearly evident from the risk map.

The risk management model by using the standard model and the risk map proved to be very efficient. The main advantage of using the proposed model is its simplicity and clarity. The graphical presentation of risks on the risk map provides for an easier risk analysis and monitoring of the impact of both preventive and corrective measures on the risk level.

Acknowledgement

The authors acknowledge the financial support of the Slovenian Research Agency (research core funding No. P2-0270).

References

[1] A Guide to the Project Management Body of Knowledge (PMBOK® Guide), 6th edition. 2017. Pennsylvania, USA: Project Management Institute, Inc.

[2] Elkjaer, M., Felding, F. 1999. Applied Project Risk Management – Introducing the Project Risk Management Loop of Control. Project Management, 5(1): 16–25.

[3] Krane, H. P., Rolstadås, A., Olsson, N. O. E. 2010. Categorizing Risks in Seven Large Projects – Which Risks Do the Projects Focus On? Project Management Journal, 41(1): 81–86.

[4] Kušar, J., Rihar, L., Žargi, U., Starbek, M. 2013. Extended Risk-Analysis Model for Activities of the Project. SpringerPlus, 2(1):227.

[5] Merritt, G. M., Smith, P. G. 2004. Techniques for Managing Project Risk. In: Cleland, D. I. Field Guide to Project Management, Second Edition (pp. 202–218). New Jersey, USA: John Wiley & Sons, Inc.

[6] Smith, P. G., Merritt, G. M. 2002. Proactive Risk Management: Controlling Uncertainty in Product Development. New York, USA: CRC Press, Taylor & Francis Group.

[7] Škec, S., Štorga, M., Marjanović, D. 2013. Mapping Risks on Various Product Development Process Types. Transactions of FAMENA, 37(3): 1–16.

The 15th International Symposium on
Operational Research in Slovenia

SOR ’19

Bled, SLOVENIA
September 25 - 27, 2019

Session 1:

Econometric Models and Statistics

THE ROLE OF LOCAL ACTION GROUPS FOR THE OPTIMAL ALLOCATION OF INVESTMENTS IN THE LONG-TERM CARE

Samo Drobne[1] and **Marija Bogataj**[2]
[2] University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova cesta 2, Ljubljana, Slovenia
[1] University of Ljubljana, Faculty of Economics, SEB & INRISK, Slovenia
marija.bogataj@ef.uni-lj.si, samo.drobne@fgg.uni-lj.si

Abstract: European states are responsible for planning, funding, and administration of deinstitutionalization processes in the long-term care. In this paper we analyse the possibilities for the local action groups (LAGs) of municipalities in Slovenia to participate in the optimal allocation of these investments and services. We present the method for optimal coverage of Slovenian territory with functional regions where LAGs are centres of health-care activities for the elderly and jobs for most commuters in a functional region. Sets of hierarchical functional regions were modelled using the Intramax method and the existing position of LAGs in Slovenia is compared with the optimal results.

Keywords: local action groups, LAG, functional region, central place, optimal allocation, eldercare.

1 INTRODUCTION

In this paper, an application of a method for evaluating functional regions (FRs) for servicing the elderly, like that developed by Drobne and Bogataj [12, 13], is expanded and the results are compared with the delineation determined in the structure of Slovenian local action groups (LAGs). We are looking for such allocation of investments in the social infrastructure for seniors in LAGs and delineation of FRs where the costs of communication between a city as a central place and other areas in the FR, and services (including investments in service centres) would be minimal.

The remainder of this paper is structured as follows. In the following three subsections, definitions and explanations about functional regions, local action groups, and smart villages are given. Section 2 provides a methodology description. This is followed by a short results presentation in section 3. Section 4 discusses the results and concludes the paper.

1.1 Functional regions

Functional regions (FRs) are spatial units on different levels of spatial hierarchical structures, which are increasingly being considered when analysing economic, social, environmental, and spatial development and when making development-related decisions on investments and operations of spatial networks [10]. Related to demographic and technological changes, these regions are a generalization of changeable social and economic functional interactions in a territory [8]. In analyses of FRs we are faced very often with the problem of defining their number and their delineation regarding the functions and actions under the analysis. Drobne [11] proposed an evaluation model of the areas and the number of FRs. But the first consideration regarding the social infrastructure for Long-Term Care (LTC) was presented by [12, 13]. Their model is based on the Intramax procedure as a hierarchical method to model a regionalization using selected criteria.

Various services in a central place of a region have a different optimal size of the territory extending around them. While a formal region is understood as an area having a well-defined border, the FR extends around an activity and its border can change many times in a time horizon. FR is an area made up of different basic spatial units (BSUs, e.g. communities, municipalities) that are linked and function as a unit in a higher level of hierarchical spatial structures. The participating BSUs could change their membership in a time horizon.

In the past, many researchers (e.g., [1, 2, 8, 9, 15, 16, 21]) showed that the existing administrative regions covering the territory of European Member States used as basic entities for policy making, resource allocation, and research do not provide meaningful information, as they are too heterogeneous inside and have many similarities. Therefore, a FR was defined as a region characterised by its agglomeration of activities and by its intra-regional intensity of flows of people, cargo, information, and financial flows, facilitating production and services, also enabling services within its interaction borders, which is rarely well defined. In the industrial society the basic characteristic of a FR is the integrated labour market, being much more intensive inside the region than across its borders. Consequently, the delineation of FRs is based on the conditions of the local labour market. Based on this perception, OECD [21] as well as many authors, e.g. Cörvers et al. [10] and Casado-Diaz and Coombes [8], accepted FRs' approach as a labour-market approach. But economic shocks, as we have witnessed over a decade ago, make us reconsider the labour market perception. Several applications to determine functional areas have been developed and used for statistical purposes to analyse various aspects of labour market performances [8, 11] and other socio-economic aspects, to evaluate and/or define administrative regions [1, 10], or to analyse housing market areas for housing policy [5] and commodity market areas [6]. Like the concepts and their applications of FRs, location-allocation models and applications are not new in research community. The challenge of location-allocation models trying to answer the question about the best site to locate the facilities for older inhabitants and how to allocate resources is multifaceted and adds to the ever-growing body of literature.

1.2 Local Action Groups in Slovenia

To enable the local inhabitants to participate actively in decision-making processes regarding priority tasks and development actions, the Community-Led Local Development (CLLD) approach aims to promote a comprehensive and well-balanced development of municipalities according to the "bottom-up" approach, by forming local action groups (LAGs) as functional areas of two or more municipalities. In the 2014–2020 programming period, 37 local action groups were created in Slovenian territory, homogeneously covering the entire territory; see Figure 1.

Figure 1: Local action groups in Slovenia in programming period 2014–2020 (source [20])

Each LAG was formed as a group of communities with common local needs and challenges, with a common goal of achieving a set of local development goals. The area of each LAG must be connected in a homogeneous geographic and functional unit, and the area of each municipality cannot be divided among two or more LAGs. The CLLD financial sources in the programming period 2014–2020 include three funds, namely the European Agricultural Fund for Rural Development (EAFRD), the European Regional Development Fund (ERDF), and the European Maritime and Fisheries Fund (EMFF). In the 2014–2020 programming period, a total of 96 million EUR from European and national funds were allocated to the implementation of CLLD, following four priorities: (a) Preparatory support – funds intended for the formation of local partnerships, strengthening institutional capacity, and training and networking during the preparation and implementation of the local development strategy; (b) Support for the implementation of operations within the community-led local development strategy, like funds intended to co-finance expenditure incurred in the implementation of operations of LAG or local actors, the results of which contribute to the achievement of objectives set out in the local development strategies; (c) Preparation and implementation of the cooperation activities of the LAG; (d) Support for running management, operation, and animation costs.

1.3 Smart villages

In Europe, the number of people aged 65 or older is about to grow from 85 million today to more than 151 million in 2060 [14]. Slovenian inhabitants are among the fastest ageing population. New programs should be put in place, not only in production [7], but also in elder care [3, 4]. To keep public finances and pension funds sustainable due to the aging of the population, contributions to pension funds relative to wages decrease, while the purchasing power of pensions also declines [22]. This situation requires additional national and community support to older inhabitants [3].

How to provide adequate services and housing for an increasing number of people that are dependent on the help of others is a crucial question in the EU and also in Slovenian municipalities. The housing stock in LAGs is not fit to support the shift from institutional care to home-based independent living. The majority of houses in Slovenian municipalities are not adequately built, as they contain accessibility barriers for people with emerging functional impairments. The financial sources are not provided in general to improve the living standard for seniors. Retirement communities are not developed at all [23]. How to finance the adaptation of housing units and to build new facilities is discussed. It is assumed that some facilities are optimal to be constructed on the level of LAGs. The management of these spatial units should consider how to build smart villages where the centres of facilities and services for seniors would be constructed and would operate in the future. In our approach, we want to determine the investments in LAGs, which are nearly optimally allocated in dependence of the forecasting level of fixed costs of investments and operations.

Higher investments and other fixed costs of public services in a central place decrease the number of central places and the broader territory of a country where they are placed. Therefore, some “expensive” activities will find their optimal location in less central places and some “cheaper” ones in more central ones. If some central places of a certain level are stronger than others, their market area grows and attracts customers from other central places. Activity cells need to find an optimal level of central places to benefit the appropriate structure of human resources, subventions, and other fiscal policies, and lower production or distribution costs. Their influence on LAGs will be studied in the following chapters.

2 DELINEATION OF FUNCTIONAL REGIONS

2.1 Intramax procedure

Functional regions at local level can be considered as local labour market (LLM) areas [2]. Let us consider the labour commuter as a person in employment whose territorial unit (BSU, e.g. community, municipality) of workplace j is not the same as territorial unit of residence i. To analyse functionally delineated areas as LLM areas, the groupings can been arranged using the hierarchical clustering method Intramax, initially based on the ideas of Markov chain techniques of Masser and Brown [17], later further developed in [18, 19]. The objective of the Intramax procedure is to maximise the proportion within the group interaction at each stage of the grouping process, while considering the variations in the row and column totals of the matrix. In the grouping process, two BSUs (in our case municipalities) are grouped together for which the objective function (1) is maximised [6, 17, 18, 19]:

$$\max I; \quad I = \frac{I_{ij}}{O_i D_j} + \frac{I_{ji}}{O_j D_i}; \quad i,j = 1,2, \dots N \tag{1}$$

where I_{ij} is the flow from home i to working area j, $O_i = \sum_j I_{ij}$ is the total of flows originating from origin i, $D_j = \sum_i I_{ij}$ is the total of flows coming to destination j, and $O_i, O_j, D_i, D_j > 0$. The Intramax analysis is a stepwise procedure. In each step two BSUs are grouped together and the interaction between them becomes the internal interaction for the new resulting BSU. This new BSU takes the place of the two parent BSUs at the next step of the analyses. So with N elementary spatial units after $N-1$ steps all BSUs are united into a new (B)SU and all interactions become internal. So, after $N-k$ steps we get k functional regions in which inner flows together have maximal intensity $I_{ij}^{\cdot k} + I_{ji}^{\cdot k}$.

2.2 The optimal allocation of public service centres

For flow of workers ($\cdot\, w$) the transportation costs to regional central places $c_{ij}^{\cdot w} = c_{ji}^{\cdot w}$ have been calculated for each potential FR. The costs of services needed for a population or part of the members in a household, in an area determined in this way, have been calculated and added to the travel costs.

Notation**:** a_j are fixed costs of daily activity for activity (service) A in central place area j; k is the potential number of central places j where services A will be located or from where they will originate; c_{ij}^h are the transportation costs regarding daily servicing a person living at BSU i if service A is located at central place j, if the service is available at her/his home (h), and c_{ij}^s, if it is available in central place j; n_i is the number of inhabitants in BSU$_i$, p_{ij}^h, p_{ij}^s ... is the percentage of potential users of A living in BSU i needing service A per day at home or in the service centre respectively.

To the criteria of optimal regionalisation on the bases of the number of daily commuters F_1 to work (1) we shall add the criterion of costs of services F_A:

$$F = \alpha F_1 + \beta F_A = \alpha \sum_{i \in \Gamma_j} c_{ij}\left(I_{ij}^{\cdot k} + I_{ji}^{\cdot k}\right) + \beta \left(\sum_{j=1}^{k} a_j + \sum_{i \in \Gamma_j} \left(\left(c_{ij}^s + c_j\right) p_{ij}^s + \left(c_{ij}^h + c_j\right) p_{ij}^h \right) n_i \right) \tag{2}$$

In general, when we plan public services in a region, α and β are subject to budget availability and negotiation in the process of governance of regions. Therefore we are looking for such allocation of activities A in k functional regions where the costs of communication and services between the central place and other areas in the FR are minimal:

$$\min_k F = \min_k \left(\alpha \sum_{i \in \Gamma_j} c_{ij} \left(I_{ij}^{:k} + I_{ji}^{:k} \right) + \beta \left(\sum_{j=1}^{k} a_j + \sum_{i \in \Gamma_j} \left(\left(c_{ij}^s + c_j \right) p_{ij}^s + \left(c_{ij}^h + c_j \right) p_{ij}^h \right) n_i \right) \right) \tag{3}$$

3 NUMERICAL EXAMPLE: LOCAL ACTION GROUPS IN THE CONTEXT OF SLOVENIAN FUNCTIONAL REGIONS

In the numerical example, we considered FRs modelled by Intramax procedure as a proxy for LAGs in Slovenia. Hierarchically aggregated FRs in Slovenia have been modelled by using data on inter-municipal labour commuters in 2010 and our own programme code in *Mathematica 12* [11]. The territory has been divided into k sets of FRs stepwise, where k =1,2,…,20 in this application. For each set of FRs the value of the criterion function (2) has been calculated for the case of servicing the elderly. Here we used: $\alpha = \beta = 1, p_{ij}^h = 0, p_{ij}^s = 0.001, c_j = 25€, n_i$ was calculated at 16.5% of the total number of inhabitants BSU. The transportation costs were calculated as the shortest road distance by car (using ArcGIS software) multiplied by 0.26 €. The investments and other costs of such activities were calculated by [24–26]. Fixed costs a_j of daily activity A were assumed as a variable taking a value between 2,000 EUR and 200,000 EUR: Figure 2 shows the optimal numbers of FRs in dependence of fixed costs of daily activity for services in the FR. We can see that at higher a_j the optimal k is lower.

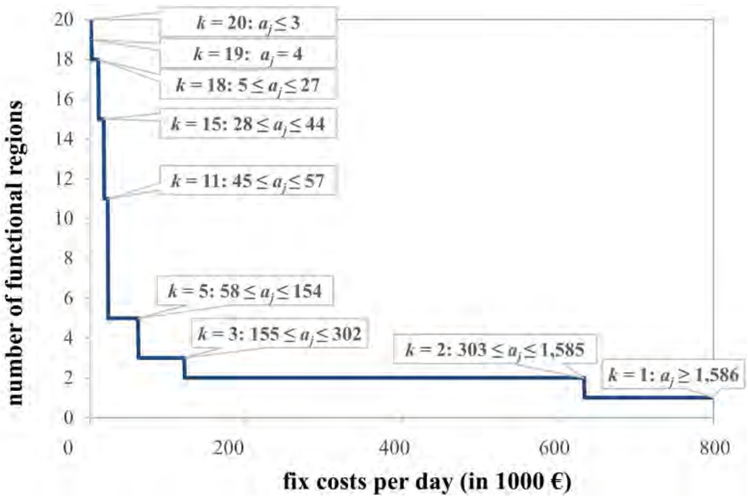


Figure 2: Optimal number of functional regions at given fixed costs per day for activity *A*

4 DISCUSSION AND CONCLUSION

We have seen that the Rural Development in Slovenia which is managed nationally through the Rural Development Programme (RDP), funded under the European Agricultural Fund for Rural Development and national contributions, consists of 37 LAGs. The RDP sets out priority approaches and actions to meet the needs of the LAGs. One of their needs is the social infrastructure for the inhabitants aged 65+. Among the basic needs for achieving this goal is the construction of an intergenerational center, including a day-care center for the elderly. Each LAG wants to build such a center with strong medical, nursing, and ICT support. We have calculated that the repayment of investment costs and fixed operating costs would require between 5000 and 10,000 € daily, which means that the optimal number of LAGs with such investments would be $k=18$. This means that two LAGs should be combined for such an investment. Figure 3 shows existent LAGs and the analyzed 18 FRs modeled by the Intramax method. We can clearly see the LAGs that can be simply aggregated into one FR and others where a new analysis of the territory should be performed.

The improved size rule and the hierarchical system of settlements has been explained by answering the question where to locate services, especially elder-care centres, in the process of deinstitutionalization. The answer has been found by looking for such an allocation of activities in their LAGs as functional regions, when the LAGs cover the complete national territory, in case of the optimality of allocation at given fixed costs.

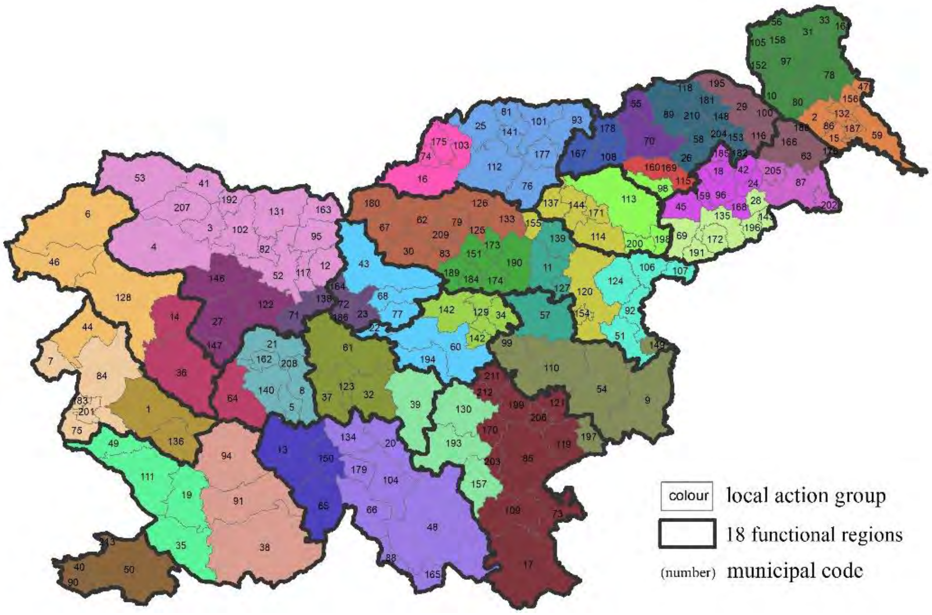


Figure 3: Local action groups in Slovenia in programming period 2014–2020 (source [22]) and 18 functional regions modelled by Intramax method

If each existing LAG wishes to have their own intergenerational center, including nursing and other more sophisticated day care for the elderly, the fixed cost must be much lower. If fixed costs a_j of daily activity A take a value between 3000 € and 4000 € then it is advisable that two LAGs merge in one service network on average.

As a possible future work, calculations should be made for each investment separately for each LAG or a group of LAGs.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research project J6-9396 Development of Social Infrastructure and Services for Community Based Long-Term Care, research project J5-1784 Creating Social Value with Age-Friendly Housing Stock Management in Lifetime Neighborhoods and research core funding P2-0406 Earth Observation and Geoinformatics).

References

[1] Andersen, A.K. (2002). Are commuting areas relevant for the delimitation of administrative regions in Denmark?, Regional Studies, 36(8): 833–844.

[2] Ball, R.M. (1980). The use and definition of Travel-to-Work areas in Great Britain: Some problems. Regional Studies, 14(2): 125–139.

[3] Bogataj, D. 2013. Pensions and home ownership in the welfare mix for older persons. In Zadnik Stirn, Lidija (ed.), et al. SOR'13 proceedings, 12th International Symposium on Operational Research in Slovenia Ljubljana: SDI_SOR pp. 281–286.

[4] Bogataj, D., Battini, D., Calzavara, M., Persona, A. 2019. The ageing workforce challenge: investments in collaborative robots or contribution to pension schemes, from the multi-echelon perspective. International Journal of Production Economics, 210: 97–106.

[5] Brown, P. J. B., Hincks, S. 2008. A framework for housing market area delineation: Principles and application. Urban Studies, 45,11: 2225–2247.

[6] Brown, P. J. B., Pitfield, D. E. 1990. The Intramax derivation of commodity market structures from freight flow data. Transportation Planning and Technology, 15(1): 59–81.

[7] Calzavara, M., Battinia, D., Bogataj, D., Sgarbossa, F., Zennaro, I. 2019. Ageing workforce management in manufacturing systems: state of the art and future research agenda. Int. Journal of Production Research, doi: 10.1080/00207543.2019.1600759.

[8] Casado-Díaz, J. and Coombes, M. (2011). The Delineation of 21st Century Local Labour Market Areas. Boletín de la Asociación de Geógrafos Españoles, (57): 7–32.

[9] Coombes, M.G., and Bond, S. (2008): Travel-to-Work Areas: the 2007 review. Office for National Statistics, London.

[10] Cörvers, F., Hensen, M., and Bongaerts, D. (2009). Delimitation and coherence of functional and administrative regions. Regional Studies, 43(1), 19–31.

[11] Drobne, S. (2016). Model vrednotenja števila in območij funkcionalnih regij: doktorska disertacija Ljubljana: University of Ljubljana, Faculty of Civil and Geodetic Engineering.

[12] Drobne, S., Bogataj, M. 2015. Optimal allocation of public service centres in the central places of functional regions. IFAC- PapersOnLine, 48(3): 2362–2367.

[13] Drobne, S., Bogataj, M. 2017. The impact of public investments in facilities on the potential housing market for older persons. Facilities, 35(7/8): 422–435.

[14] EC, 2015. The 2015 ageing report, Economic and budgetary projections for the 28 EU Member States (2013-2060). European Economy 3|2015, European Commission, Brussels.

[15] Karlsson C., and Olsson, M. (2006). The identification of functional regions: theory, methods, and applications. Ann Reg Sci, 40(1): 1–18.

[16] Laan, L.V.D., and Schalke, R. (2001). Reality versus policy: The delineation and testing of local labour market and spatial policy areas. European Planning Studies, 9(2): 201–221.

[17] Masser, I., and Brown, P.J.B. (1975). Hierarchical aggregation procedures for interaction data. Environment and Planning A, 7(5): 509–523.

[18] Masser, I., and Brown, P.J.B. (1977). Spatial representation and spatial interaction. Papers of the Regional Science Association, 38(1): 71–92.

[19] Masser, I., and Scheurwater, J. (1980). Functional regionalisation of spatial interaction data: an evaluation of some suggested strategies. Environment and Planning A, 12(12): 1357–1382.

[20] Ministry of Agriculture, Forestry and Food (2019). European Agricultural Fund for Rural Development. https://www.program-podezelja.si/en/43-news/593-30-9-2016-first-decisions-on-the-approval-of-local-action-groups-for-2014-2020-in-slovenia-issued.

[21] OECD (2002). Redefining territories – The functional regions. Organisation for Economic Co-operation and Development. Paris, France.

[22] Rogelj, V., Bogataj, D. 2018a. Planning and financing the home and facility-based care using the multiple decrement approach. Journal of decision systems, 27(3): 132–143.

[23] Rogelj, V., Bogataj, D. 2018b. Planning the home and facility-based care dynamics using the multiple decrement approach: the case study for Slovenia. IFAC-PapersOnLine, 51(11): 1004–1009.

MULTI-CONSTRAINED GRAVITY MODEL OF LABOUR COMMUTING: CASE STUDY OF SLOVENIA

Samo Drobne[a] and **Metka Mesojedec**[b]
[a] University of Ljubljana / Faculty of Civil and Geodetic Engineering, Jamova cesta 2, 1000 Ljubljana, Slovenia
b Dolenje Mokro Polje 22, 8310 Šentjernej, Slovenia
samo.drobne@fgg.uni-lj.si, metka.mesojedec@gmail.com

Abstract: In this paper, we analyse inter-municipal commuting flows in Slovenia by using an adopted multi-constrained gravity model as suggested by Olsson (2016). The model considers 428 unique spatial constraints defined by a spatial structure, i.e. municipalities and regions of origin and destination. The model analyses separately the measures for commuting inside a municipality, between municipalities inside a region, and between regions. In general, the results for Slovenia are in line with the results for Sweden.

Keywords: spatial interaction, gravity model, entropy, constrained optimization, commuting, Slovenia.

1 INTRODUCTION

A considerable amount of empirical and theoretical literature has been published on gravity models during the past half a century. The conception of gravity models was originally introduced by Tinbergen (1962) and Pöyhönen (1963) in order to explain the differences in the cross-country variation of trade flows. The gravity model became very popular because of its quite simple usage combined with a substantial power of explaining very different flows in general (e.g., international trade flows, tourism flows, migration flows, commuting flows). In spite of its popularity, the literature points out that the gravity model has several limitations. One of them is the calibration of all the parameters of a gravity model (Gargiulo et al., 2012), which is not always easy. However, gravity models continue to be a popular tool among economists, as they explain flows between regional and other spatial units.

In this paper, we applied the concept of a spatial multi-constrained gravity model as suggested by Olsson (2016) to the inter-municipal labour commuting flows in Slovenia. In the gravity model, we calibrated accessibility measures at three different levels: inside basic spatial units (BSUs), between them in a region, and between regions. The differentiation between these three spatial levels is important in spatial studies, because of the different drives, whose impacts are quite different for labour commuting over short, mid, and long distances (Evers and Van der Veen, 1985; Champion et al., 2009; Lundholm, 2010). This is the first time that a spatial multi-constrained gravity model is calculated for Slovenia.

The remainder of this paper is structured as follows. In section 2, we describe the material and methods. This is followed by results, presentation, and discussion in section 3. Section 4 summarizes and concludes the paper.

2 METHODOLOGY

In this paper, we analysed accessibility measures in the multi-constrained gravity model at three different spatial levels of Slovenia: inside basic spatial units (BSU), between BSUs in a region, and between regions. In the case study, we applied the model, whose conception had been suggested by Olsson (2006), to the inter-municipal labour commuting flows as registered for 2017. 212 Slovenian municipalities were considered as BSUs, and the regional level was defined by twelve statistical regions that can be considered as functional regions, inside which most of the gravitational interactions exist (Drobne, 2016).

A functional region (FR) is a region characterised by its agglomeration of activities and by its intra-regional transport infrastructure. The basic characteristic of a FR is the integrated labour market, in which intra-regional commuting as well as intra-regional job search and search for labour demand is much more intensive than the inter-regional counterparts (OECD, 2002; Karlsson and Olsson, 2006; Cörvers et al., 2009). Statistical regions of Slovenia were established for the purpose of regional planning and cooperation in various sectors in the mid-1970s. At that time, the regionalization of statistical regions was supported by exhaustive gravity analysis of labour markets, education areas and supply markets in twelve regional, and their sub-regional, centres. Later, the borders of statistical regions were modified several times to accommodate to changing gravity interrelationships (SORS, 2011).

Let us denote with i municipalities of origin, $i = 1,2, \dots, n$, and with j municipalities of destination, $j = 1,2, \dots, n$; $n = 212$. The analysed interaction data on labour commuting are collected in $n \times n$ commuting matrix, $\mathbf{c} = \{c_{ij}\}$. After calibrating the gravity model, we obtain the estimated commuting matrix, $\tilde{\mathbf{c}}$. Commuting-time matrix of dimension $n \times n$ is denoted by $\mathbf{t} = \{t_{ij}\}$. Let us define a $1 \times n$ unit row vector, $\mathbf{u}$. $\mathbf{o} = \mathbf{c}\mathbf{u}^{\mathrm{T}}$ denotes $n \times 1$ vector with the number of active population in the municipality, i.e. sums of commuting matrix's rows, and $\mathbf{d} = \mathbf{u}\mathbf{c}$ denotes $1 \times n$ vector with the number of jobs in the municipality, i.e. sums of commuting matrix's columns.

The spatial structure of Slovenian territory was captured in three $n \times n$ dummy matrices, $\mathbf{k}$, $\mathbf{l}$, and $\mathbf{m}$, expressing whether commuting ends in the origin municipality ($k_{ij} = 1$, otherwise $k_{ij} = 0$), in another municipality in the same FR ($l_{ij} = 1$, otherwise $l_{ij} = 0$), or in another FR ($m_{ij} = 1$, otherwise $m_{ij} = 0$).

The multi-constrained gravity modelling approach allows an analysis of only selected interactions as well. The selection can be done regarding the commuting time. To identify all the links that are included in an analysis zones of interest can be created and collected into the $(n \times n)$ zone matrix, $\mathbf{z}$. However, in our case study for Slovenia, $z_{ij} = 1$ for all t_{ij}.

In the constrained gravity models, constraints normally consider commuters and commuting time under consideration. Considering the analysed spatial structure in the model, the observed number of commuters within a home municipality is equal to $p_1 = \mathbf{u}(\mathbf{k} \circ \mathbf{c})\mathbf{u}^{\mathrm{T}}$ (where the Hadamard product sign, $\circ$, is used for entrywise multiplication of matrices), the observed number of commuters between municipalities within a home FR is equal to $p_2 = \mathbf{u}(\mathbf{l} \circ \mathbf{c})\mathbf{u}^{\mathrm{T}}$, and the observed number of commuters between FRs is equal to $p_3 = \mathbf{u}(\mathbf{m} \circ \mathbf{c})\mathbf{u}^{\mathrm{T}}$. They are collected in the column vector $\mathbf{p}$. The constraints on commuting time are: the observed total commuting time for commutes within a municipality equals $r_1 = \mathbf{u}(\mathbf{k} \circ \mathbf{c} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}$, the observed total commuting time for commutes between municipalities within the home FR is $r_2 = \mathbf{u}(\mathbf{l} \circ \mathbf{c} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}$, and the observed total commuting time for commutes between FRs is $r_3 = \mathbf{u}(\mathbf{m} \circ \mathbf{c} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}$. They are collected in the column vector $\mathbf{r}$.

In constrained gravity models the objective is to maximise the system entropy, $\sum_i \sum_j c_{ij} \ln(c_{ij}) - c_{ij} = -\mathbf{u}(\tilde{\mathbf{c}} \circ \ln(\tilde{\mathbf{c}}) - \tilde{\mathbf{c}})\mathbf{u}^{\mathrm{T}}$, subject to the constraints (Olsson, 2016), where ln()is the matrix natural logarithm. The formulation of the problem is then

$$\max \mathrm{L}(\tilde{\mathbf{c}}, \alpha, \beta, \gamma, \delta) = \sum_{s=0}^{8} \mathrm{L}_s \tag{1}$$

where the Lagrangian parts, L_S, are

$$\mathrm{L}_0 = -\mathbf{u}(\tilde{\mathbf{c}} \circ \ln(\tilde{\mathbf{c}}) - \tilde{\mathbf{c}})\mathbf{u}^{\mathrm{T}}, \tag{2}$$

$$\mathrm{L}_1 = \mathbf{u}(\alpha \circ (\tilde{\mathbf{c}}\mathbf{u}^{\mathrm{T}} - \mathbf{o})), \tag{3}$$

$$\mathrm{L}_2 = (\beta \circ (\mathbf{u}\tilde{\mathbf{c}} - \mathbf{d}))\mathbf{u}^{\mathrm{T}}, \quad (4)$$

$$\mathrm{L}_3 = \delta_1(\mathbf{u}(\mathbf{k} \circ \tilde{\mathbf{c}})\mathbf{u}^{\mathrm{T}} - p_1), \quad (5)$$

$$\mathrm{L}_4 = \delta_2(\mathbf{u}(\mathbf{l} \circ \tilde{\mathbf{c}})\mathbf{u}^{\mathrm{T}} - p_2), \quad (6)$$

$$\mathrm{L}_5 = \delta_3(\mathbf{u}(\mathbf{m} \circ \tilde{\mathbf{c}})\mathbf{u}^{\mathrm{T}} - p_3), \quad (7)$$

$$\mathrm{L}_6 = \gamma_1(r_1 - \mathbf{u}(\mathbf{k} \circ \tilde{\mathbf{c}} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}), \quad (8)$$

$$\mathrm{L}_7 = \gamma_2(r_2 - \mathbf{u}(\mathbf{l} \circ \tilde{\mathbf{c}} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}) \text{ and} \quad (9)$$

$$\mathrm{L}_8 = \gamma_3(r_3 - \mathbf{u}(\mathbf{m} \circ \tilde{\mathbf{c}} \circ \mathbf{t})\mathbf{u}^{\mathrm{T}}). \quad (10)$$

Let us denote the Lagrangian multipliers, δ and γ, the spatial parameters, where δ is the proximity-preference parameter and γ is the distance-friction parameter. The gravity model has three proximity-preference parameters and three distance-friction parameters, considering spatial interactions inside a municipality (δ_1 and γ_1), spatial interactions between municipalities in the same FR (δ_2 and γ_2), and spatial interactions between FRs (δ_3 and γ_3).

Constraints and spatial parameters are collected in the column vectors: p_1, p_2 and p_3 are collected in the column vector $\mathbf{p}$, r_1, r_2 and r_3 are collected in the column vector $\mathbf{r}$, δ_1, δ_2 and δ_3 are collected in the column vector $\boldsymbol{\delta}$, and γ_1, γ_2 and γ_3 are collected in the column vector $\boldsymbol{\gamma}$.

In addition to the three constraints for the amount on commuting, (5)–(7), and three time constraints (8)–(10), the model has also commuting origin and destination constraints. It also enforces that the estimated number of workers that live in each municipality is equal to the observed number, $\mathbf{o} = \mathbf{c}\mathbf{u}^{\mathrm{T}} = \tilde{\mathbf{c}}\mathbf{u}^{\mathrm{T}}$. This adds 212 origin constraints, (3). However, only 211 origin constraints provide new information. The three constraints on the number of commuters together enforce that the estimated number of commuters is equal to the observed working population. This makes the 212th origin constraint redundant, since it will be enforced by the other constraints. To each origin constraint there is a Lagrangian multiplier which is called a push factor. They are collected in the column vector $\boldsymbol{\alpha}$. As suggested by Olsson (2016), for programming convenience all 212 destination constraints are used, but one origin is used as base, here $\alpha_1 = 0$. The model also enforces that the estimated number of jobs in each municipality is equal to the observed number of jobs, $\mathbf{d} = \mathbf{u}\mathbf{c} = \mathbf{u}\tilde{\mathbf{c}}$. This adds 212 destination constraints, (4). As for the origin constraints, one of the destination constraints is redundant, since only 211 destination constraints provide information. To each destination constraint there is a Lagrangian multiplier which is called a pull factor. They are collected in the row vector $\boldsymbol{\beta}$. Again, because of programming convenience all 212 destination constraints are used, but one pull factor is used as base, here $\beta_1 = 0$. So, the model has 428 (3+3+211+211) constraints.

For programming purpose, it is assumed that all distance-friction parameters, push and pull factors are zero ($\gamma_1 = 0,\ \gamma_2 = 0,\ \gamma_3 = 0,\ \alpha = 0,\ \beta = 0$). The start values for the proximity preferences are $\delta_1 = \ln(p_1/\mathbf{u}(\mathbf{k} \circ \mathbf{z})\mathbf{u}^{\mathrm{T}})$, $\delta_2 = \ln(p_2/\mathbf{u}(\mathbf{l} \circ \mathbf{z})\mathbf{u}^{\mathrm{T}})$, and $\delta_3 = \ln(p_3/\mathbf{u}(\mathbf{m} \circ \mathbf{z})\mathbf{u}^{\mathrm{T}})$, respectively.

The partial derivative of the Lagrangian with respect to commuting gives the estimated commuting matrix $\tilde{\mathbf{c}} = \exp(\alpha\mathbf{u} + \mathbf{u}^{\mathrm{T}}\beta + \delta_1\mathbf{k} + \delta_2\mathbf{l} + \delta_3\mathbf{m} - (\gamma_1\mathbf{k} + \gamma_2\mathbf{l} + \gamma_3\mathbf{m}) \circ \mathbf{t})$, where exp() is matrix exponential. Inserting this into the Lagrangian gives the dual form, $\min \mathrm{D}(\alpha, \beta, \gamma, \delta)$ where $\mathrm{D}(\alpha, \beta, \gamma, \delta) = \mathbf{u} \exp(\alpha\mathbf{u} + \mathbf{u}^{\mathrm{T}}\beta + \delta_1\mathbf{k} + \delta_2\mathbf{l} + \delta_3\mathbf{m} - (\gamma_1\mathbf{k} + \gamma_2\mathbf{l} + \gamma_3\mathbf{m}) \circ \mathbf{t})\, \mathbf{u}^{\mathrm{T}} - \alpha^{T}\mathbf{o} - \beta\mathbf{d}^{\mathrm{T}} + \gamma^{\mathrm{T}}\mathbf{r} - \delta^{\mathrm{T}}\mathbf{p}.$

In the case study of Slovenia, the Matlab program (Olsson, 2016), which uses the Newton-Raphson iterative procedure, was adopted and used to find the optimum. In the program, each group of the parameters is adjusted separately. The push factors are adjusted first. They are adjusted using $\alpha_{(n+1)} = \alpha_n - \rho(\tilde{o}_n - o)./\tilde{o}_n$. ./ is the symbol for piecewise division. After recalculating the estimated commuting flows the pull factors are adjusted using $\beta_{(n+1)} = \beta_n - \rho(\tilde{d}_n - d)./\tilde{d}_n$. So, the estimated commuting flows are recalculated before the distance-friction and proximity-preference vectors are adjusted. At the end, the proximity-preference vector is adjusted and the estimated commuting flows are recalculated once more. The distance-friction vector and the proximity-preference vector are adjusted as described above. As suggested by Olsson (2016), the relative adjustment factor was set to $\rho = 0.2$.

3 RESULTS AND DISCUSSION

Fig. 1 shows distance-friction parameter convergence (left), proximity-preference parameter convergence (middle), and paths to solution (right) in the Newton-Raphson iterative procedure of adjustment of parameters in the multi-constrained gravity model of inter-municipal labour commuting in Slovenia in 2017.

The results on distance-friction parameters are: $\gamma_1^{(SI)} = 0.0792$, $\gamma_2^{(SI)} = 0.1043$, and $\gamma_3^{(SI)} = 0.0458$, where (SI) denotes the results for Slovenia. These results are in line with the results for Sweden for 1998 (Olsson, 2016), where distance-friction parameter for commuting within a municipality $\gamma_1^{(SE)} = 0.0248$, the distance-friction parameter for commuting between municipalities within a region $\gamma_2^{(SE)} = 0.0958$, and the distance-friction parameter for commuting between regions $\gamma_3^{(SE)} = 0.0514$. The highest difference between the results, i.e. for commuting inside the municipality, γ_1, can be, most probably, explained because of the different size of municipalities in the two compared countries. The size of the municipalities in Sweden is in general bigger than in Slovenia. You find the convergence process for the distance-friction parameters in Fig. 1 (left).

The results on proximity-preference parameters are: $\delta_1^{(SI)} = 6.8977$, $\delta_2^{(SI)} = 6.0444$, and $\delta_3^{(SI)} = 3.7707$. Regarding the order of the values, these results are again in line with the results for Sweden (Olsson, 2016). However, the differences are higher than for distance-friction parameters: distance-friction parameter for commuting within a municipality $\delta_1^{(SE)} = 8.5147$, the distance- friction parameter for commuting between municipalities within a region $\delta_2^{(SE)} = 7.4679$, and the distance-friction parameter for commuting between regions $\delta_3^{(SE)} = 5.4938$. Again, the differences can be, most probably, explained with the different size of municipalities. One can find the convergence process for the proximity-preference parameters in Fig. 1 (middle).

The right side of Fig. 1 shows the proximity-preference parameter and distance-friction parameter pairs from the start (along the x-axis) to the solution. In the background, the 211 push and 211 pull factors are adjusted as well.

Fig. 2 shows the value of the dual function per iteration. As in the case study for Sweden (Olsson, 2016), little happens to the parameter values (Fig. 1) and the value of the dual function (Fig. 2) after 200 iterations.

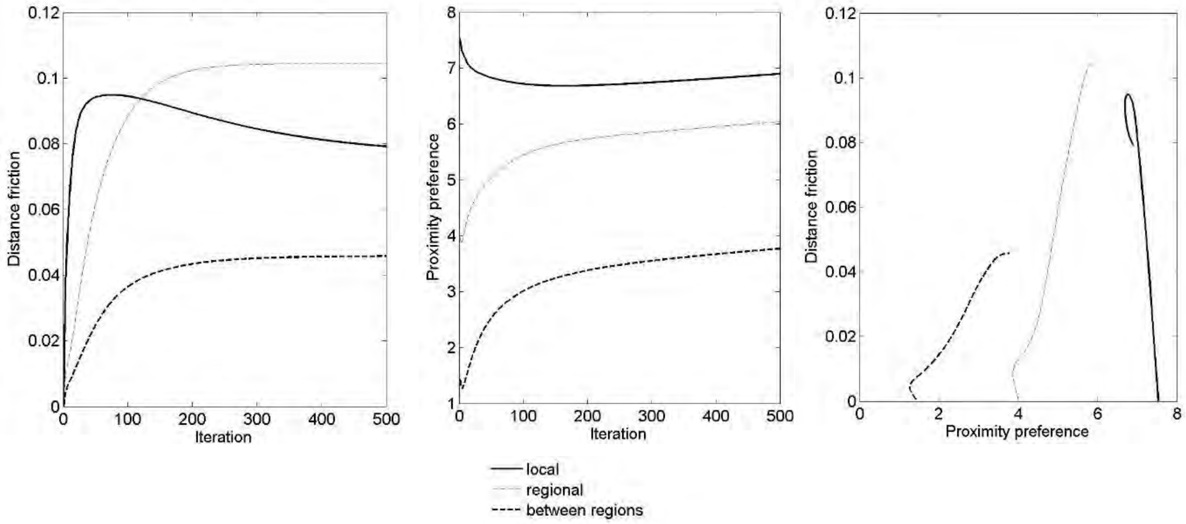


Figure 1: Distance-friction parameter convergence (left), proximity-preference parameter convergence (middle), and paths to solution (right); the multi-constrained gravity model of inter-municipal labour commuting in Slovenia, 2017.

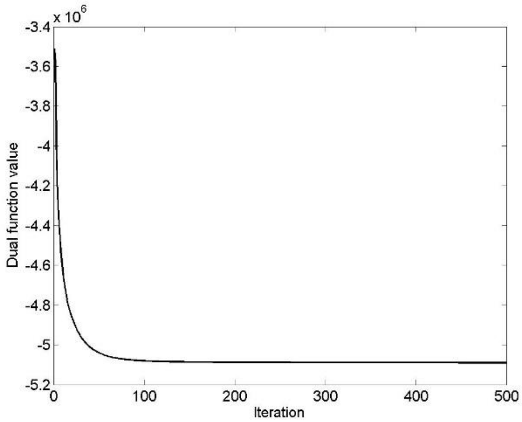


Figure 2: Dual function; the multi-constrained gravity model of inter-municipal labour commuting in Slovenia, 2017.

4 CONCLUSIONS

In this paper, we analysed a multi-constrained gravity model of inter-municipal commuting flows in Slovenia. The approach, as suggested by Olsson (2016), was adopted for the case study of Slovenia. The model for Slovenia considers 428 unique spatial constraints defined by a spatial structure of municipalities and statistical regions. This was the first time that the spatial multi-constrained gravity model was calculated for Slovenia.

In general, the results for Slovenia (for 2017 in this study) are in line with the results for Sweden (for 1998; Olsson, 2016): the highest distance-friction parameter is for commuting between municipalities inside a region, followed by the distance-friction parameter for commuting within a municipality, and by the distance-friction parameter for commuting between regions.

As a possible direction for future work, we see the analysis of the impact of the municipal area size on the adjusted parameters.

Acknowledgements
The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0406 Earth Observation and Geoinformatics).

References

[1] Champion, T., Coombes, M., Brown, D. L. E. 2009. Migration and Longer-Distance Commuting in Rural England. Regional Studies, 43(10): 1245–1259.

[2] Cörvers, F., Hensen, M., Bongaerts, D. 2009. Delimitation and coherence of functional and administrative regions. Regional Studies, 43(1): 19–31.

[3] Drobne, S. 2016. Model vrednotenja števila in območij funkcionalnih regij (=A model evaluating the number and areas of functional regions). PhD thesis. Ljubljana.

[4] Evers, G. H. M., Van der Veen, A. 1985. A Simultaneous Non-Linear Model for Labour Migration and Commuting, Regional Studies, 19(3): 217–229.

[5] Gargiulo, F., Lenormand, M., Huet, S., Espinosa, O. B., 2012. A commuting network model: Getting the Essentials. Journal of Artificial Societies and Social Simulation, 15 (2), DOI: 10.18564/jasss.1964.

[6] Karlsson C., Olsson, M. 2006. The identification of functional regions: theory, methods, and applications. Ann Reg Sci, 40: 1–18.

[7] Lundholm, E. 2010. Interregional Migration Propensity and Labour Market Size in Sweden, 1970–2001. Regional Studies, 44(4): 455–464.

[8] Pöyhönen, P. 1963. A tentative model for the flows of trade between countries. Weltwirtschaftliches Archiv, 90(1): 93–99.

[9] OECD 2002. Redefining territories – The functional regions. Organisation for Economic Co-operation and Development. Paris, France.

[10] Olsson, M. 2016. Functional Regions in Gravity Models and Accessibility Measures. Moravian Geographical Reports, 24(2): 60–70.

[11] SORS (2011). Administrative-territorial division = "Upravno-teritorialna razdelitev", Statistical Office of the Republic of Slovenia, Ljubljana. http://www.stat.si [Accessed: 25/03/2011].

[12] Tinbergen, J. 1962. Shaping the World Economy, Twentieth Century Fund, New York.

CHALLENGES OF BENFORD'S LAW GOODNESS-OF-FIT TESTING IN DISCOVERING THE DISTRIBUTION OF FIRST DIGITS: COMPARISON OF TWO INDUSTRIES

Ksenija Dumičić
University of Zagreb, Faculty of Economics and Business – Zagreb
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: kdumicic@net.efzg.hr

Ivana Cunjak Mataković
Centar revizija d.o.o.,
V. Mačeka 28, HR–47000 Karlovac, Croatia
E-mail: ivana.cunjak@gmail.com

Abstract: Benford's Law is a logarithmic distribution that gives the expected patterns of digits in numerical data. Digital analysis based on Benford's Law is used by auditors and forensic accountants to detect anomalies in financial reports. This paper describes the first digit law that predicts the appearance of expected data as to respect Benford's distributions. The aim of this paper was to test whether the financial statements from selected companies of two industries from Croatia comply with Benford's Law of first digit. The paper demonstrates the difference in performance of two compared industries' results.

Keywords: Benford's Law, first digit distribution, audit risk, fraud detection, financial statements, significance level, Croatia

JEL classification: C1, C2, G3, M4

1 INTRODUCTION

The game with financial number is unfortunately alive and doing very well. There are regular reports of occupational fraud which indicate that already detected fraud cases might be just the tip of iceberg [6]. The 2018 *Reports to the Nations on Occupational Fraud and Abuse*, [1], assert that the total loss, caused by the cases from their study, exceeded USD 7.1 billion. In fraud cases presented, the leading detection methods were tips, and more than half of tips were provided by victim-organization employees. The fraud in financial statements was detected by external auditors in 4 percent of cases [1]. The above-mentioned fraud presents the significant risk for external auditors. The risk exists due to auditors who are examining less than 100 percent of account balances or classes of transaction. Nigrini and Miller state that usage of computer-assisted audit techniques (CAAT-s) within the audit has twofold benefits: risk reduction and increase of audit effectiveness, and audit economy and efficiency [7]. The digitalization of data sets allows the usage of statistical techniques and methods. One of the statistical methods which could be applied in audit is Benford's Law. The Benford's Law is based on logarithmic distribution and describes the distribution of the first digit. Forensic accounting techniques are designed on the basis of the Benford's Law and their aim is to identify the existence of unusual transaction, events and trends. If the first digit in financial reports do not comply with Benford's Law this could be the indication of some "cosmetic earning management", according to [13].

This paper analyses the financial statements of companies from two industries in Croatia, with a goal to determine whether they comply with Benford's Law. What has been analysed is the financial reports of companies from food processing industry and companies from tourism sector, and in our research, companies were grouped in two categories: 1) the companies that are not in process of pre-bankruptcy settlement, and 2) the companies that are in process of pre-bankruptcy settlement. The companies that are not in process of pre-bankruptcy settlement are those, whose financial reports are prepared on assumption of "going concern". In this analysis the Benford's Law settings are used to detect the possible manipulation in financial statements, and are based on the null hypotheses H_0: "The financial statements of companies comply with the Benford's Law of first digit distribution", and the alternative, being H_1: "The financial statements of companies do not comply with Benford's Law of first digit distribution". The research hypothesis of this research would be that Benford's Law Goodness-of-Fit (GoF) testing is a powerful tool in fraud detection, i.e. the auditors may get meaningful results simply.

2 PREVIOUS RESEARCH

In 1995 the Wall Street Journal published an article entitled "He's got their number: Scholar uses math to foil financial fraud" by Mark Nigrini, [6]. Nigrini examined compliance of financial reports with Benford's Law. The basic results of his research are the following: if the frequencies of first digits don't follow the Benford's distribution, it might be the risk of fraud or error. Nonconformity does not signal fraud or error with certainty [6]. The Benford's Law has been adopted and applied in other scopes of economy. For instance, the paper [2] analysed the overall probability distribution of the first significant digit on the whole set of prices and on the whole set of returns of 361 stocks belonging to the S&P 500 market with Benford's Law. The paper [11] analysed the suitability of Benford's Law to check the quality of macroeconomics data which are relevant to the euro deficit criteria, where the dataset consisted of all relevant data from the 27 EU states for the period from 1999 to 2009. In [12] the financial statements of twenty tech companies from Fortune 500 list are studied. This research includes, inter alia, some of the following companies: Amazon, Google, IBM, Apple, eBay, Intel, Microsoft and Oracle. It was found that the first leading digit of financial statements of observed companies almost followed the Benford's Law. The similar research was carried out in Slovenia in 2018, [8], which investigated whether the financial statements of companies listed on the Ljubljana Stock Exchange could pass the Benford's Law test, using sample survey containing 44 companies, covering the period of three years (2011, 2012 and 2013), where the analysed data passed the Benford's Law test. In Croatia, the first research based on Benford's law was carried out in 2008, [5]. The authors observed the parameters on Zagreb Stock Exchange in the period from January 1st 1998 to February 26th 2008. The reached results shown that the closing daily stock price in the observed period do not fit the Benford's Law, but the total daily stock turnovers completely do. In [13] statistical fraud detection tools to the financial reports of 7 large Croatian public and state-owned companies were examined. By analysing available data for the period from 2010 to 2011, they found that the 7 annual reports are deviated from the Benford's Law. The paper [9] investigated the usage of the Benford's Law on financial statements of companies listed on Zagreb Stock Exchange in the period from 2011 to 2016, and it reached the results, which show that financial reports of listed companies did not follow Benford's Law. In [4], Benford's Law application in psychological pricing detection is studied, where the results of the analysis shown discrepancy from Benford's Law.

3 RESEARCH METHODOLOGY

Simeon Newcomb, the Canadian-American astronomer/mathematician, observing logarithmic tables, noted that the first pages wear out much faster than the last ones. He discovered that not all digits (1, 2, … 9) occur with the same frequency in the first place of such numbers. In 1881 he wrote two-page article "Note on the Frequency of Use of the Different Digits in Natural Numbers" [10] in the *American Journal of Mathematics* [3]. On the basis of this findings Newcomb formulated a law, which stated:" The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally likely" [3]. According to this law if a number has the leading significant digit 1 with probability $log_{10}2 \cong 0.301$, the leading significant digit 2 with probability log_{10} (3/2) ≅0.176, and so on to monotonically down to probability 0.046 for leading digit 9 [3]. In 1938 physicist Frank Benford rediscovered this law established by Simeon Newcomb. Benford analysed data from as many sources as possible and tried to include a variety of different types of data sets in his analysis. He analysed numbers from the front pages of newspapers and all numbers in an issue of Reader's Digest. It was shown in data that he collected that a varied random number had no relationship to each other. Results of Benford's analysis showed that on average 30.6 percent of the numbers had a first digit 1, and 18.5 percent of the numbers had a first digit 2. "Benford then saw a pattern to his results. The actual proportion for the first digit 1 was almost equal to the common logarithm of 2 (or 2/1), and the actual proportion for the first digit 2 was almost equal to the common logarithm of 3/2. The logarithmic pattern continued through to the number 9, with a proportion for the first digit 9 approximating the common logarithm of 10/9. Benford derived the expected frequencies of digits in the lists of number (1), which comprise Benford's Law" [6, p. 87].

$$P(D_1 = d_1) = \log\left(1 + \frac{1}{d_1}\right) d_1 \in \{1, 2, \dots 9\} \tag{1}$$

Table 1: Benford's distribution of first digit

D	1	2	3	4	5	6	7	8	9
π(d)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

In 1995, after almost 60 years, [3, p. 360], T. P. Hill gave an explanation why Benford's Law is found in many empirical contexts. He proved that "if probability distributions are selected at random, and random samples are then taken from each of these distributions in any way so that the overall process is scale (base) neutral, then the significant-digit frequencies of the combine sample converge to the logarithmic distribution" This explanation is named by Hill's theorem. Hill's theorem presented the new form of a Central Limit Theorem and explains why the significant digits of the combined sample converge to Benford's distributions. For the conduction of the data sets analysis and testing of compliance with Benford's Law, the usually standard criteria that should be obeyed are: 1) the recommended minimum scope of statistical set is 1000 data, 2) median is smaller than the arithmetic mean and skewness is positive. The various statistical tests measure discrepancy of data sets from Benford's distribution. The Pearson χ^2 statistic with 8 degrees of freedom is often used to compare an actual set of results with an expected set of results:

$$\chi^2 = n \sum_{i=1}^{9} \frac{\left(\pi(d) - r(d)\right)^2}{\pi(d)}, \tag{2}$$

where n is the sample size, $\pi(d)$ is the Benford probability of occurrence of digit d and $r(d)$ is the actual relative frequency of occurrence of the digit d in the sample. The null hypothesis is rejected at 5% significance if the χ^2statistic exceeds 15.51% and at 1% significance if χ^2 exceeds 20.09 [13].

Accounting data presents events and transactions that are contained in a form of financial statements. This data sets have various forms of distributions. Hill's theorem explains why combined sample from different distributions converge to Benford's (logarithmic) distribution. Benford's Law is used as statistical tool for detecting a "red flags" or anomalies in financial statements. If analysed data sets do not comply with Benford's Law, the technical reasons for that could be: the median is larger than arithmetic mean, skewness is negative, the sample is small, negative numbers or the digit 0 appeared in the first place and totals and subtotals should be ignored because they are results of arithmetic operation and they could not be manipulated, [6]. If hypothesis tested by Pearson χ^2 , statistically this coefficient is typically larger for larger samples, [13]. The existing indicators could be analysed when these technical standard criteria are satisfied.

4 EMPIRICAL RESULTS

The aim of empirical research was to test weather financial statements of companies from two industries comply with Benford's first digit distributions. The companies observed were from food processing industry and tourism sector in Croatia.

This study includes financial statements of "*companies that are in process of pre-bankruptcy settlement" (*further on*: companies in process)* and "*companies that are not in process of pre-bankruptcy settlement" (*further*: companies that are not in process)*, for each category of industry. For the benefit of our research interest financial statements of companies, *companies in process* and *companies that are not in process*, were compared for every industry. Research features are shown in Table 2.

Table 2: Research Matrix

Status of companies	**Categories of industry (economic activity or sector)**	
	Food Processing Industry	*Tourism Sector*
Companies that are not in process of the pre-bankruptcy settlement	No. of companies: four Observed period: 2013 - 2018	No. of companies: four Observed period: 2013 - 2018
Companies that are in process of pre-bankruptcy settlement	No. of companies: four Observed period: 2013 - 2018	No. of companies: four Observed period: 2013 - 2018

From the Zagreb Stock Exchange web site, which is a data source of all Croatian companies listed on the stock exchange, not consolidated financial statements for company sample were taken over to Excel.

Data imported in Excel were grouped into four clusters according to categories of industry and status of companies. For each company, in the observed period, there was a separated balance sheet from income statement. Before the process of analysing the first digit was started, all negative numbers were removed, as well as digit 0, total and subtotals from financial statements.

For obtained data sets, assumptions for Benford distribution GoF of digits were checked, as given in Table 3. The empirical results of goodness-fit-to testing regarding Benford's Law distribution of the first digit for compared categories is shown by Figure 1.

Table 3: Assumptions for Benford distribution goodness-of-fit of digits

Assumptions	Categories of industry (economic activity or sector)			
	Food Processing Industry		Tourism Sector	
	Companies that are not in process	Companies in process	Companies that are not in process	Companies in process
Skewness	2.20	1.74	1.90	1.98
Median	107	122	92	90
Mean	170.78	163.11	156.33	116.89
Sample size n	1,537	1,468	1,407	1,052

The obtained data sets obeyed the assumptions of Benford's Law distribution of first digit. A statistical method of Benford's Law was applied on observed numerical data group by two categories. For each category of industry and financial statement of *companies that are not in process* and *companies in process,* χ^2 test was performed for GoF, measuring significance of discrepancy from Benford's distribution, as given in Table 4.

Table 4: The empirical results of Benford's Law distribution of first digit

Digit	Food Processing Industry						Tourism Sector					
	Companies that are not in process			Companies in process			Companies that are not in process			Companies in process		
d	O*	B*	χ^2	O*	B*	χ^2	O*	B*	χ^2	O*	B*	χ^2
1	513	463	5.47	411	442	2.16	431	424	0.13	327	317	0.34
2	225	271	7.70	276	259	1.18	242	248	0.13	169	185	1.43
3	208	192	1.33	165	183	1.85	182	176	0.22	130	131	0.02
4	142	149	0.32	122	142	2.89	136	136	0.00	96	102	0.35
5	102	122	3.19	135	116	3.03	89	111	4.51	78	83	0.34
6	69	103	11.17	92	98	0.40	92	94	0.05	76	70	0.44
7	107	89	3.58	105	85	4.64	74	82	0.71	90	61	13.78
8	96	79	3.84	91	75	3.37	87	72	3.14	45	54	1.44
9	75	70	0.31	71	67	0.22	74	64	1.44	41	48	1.06
Total	1,537	1,537	**36.91**	1,468	1,468	**19.74**	1,407	1,407	**10.33**	1,052	1,052	**19.18**

*O*Observed, B*Benford's Law expected*

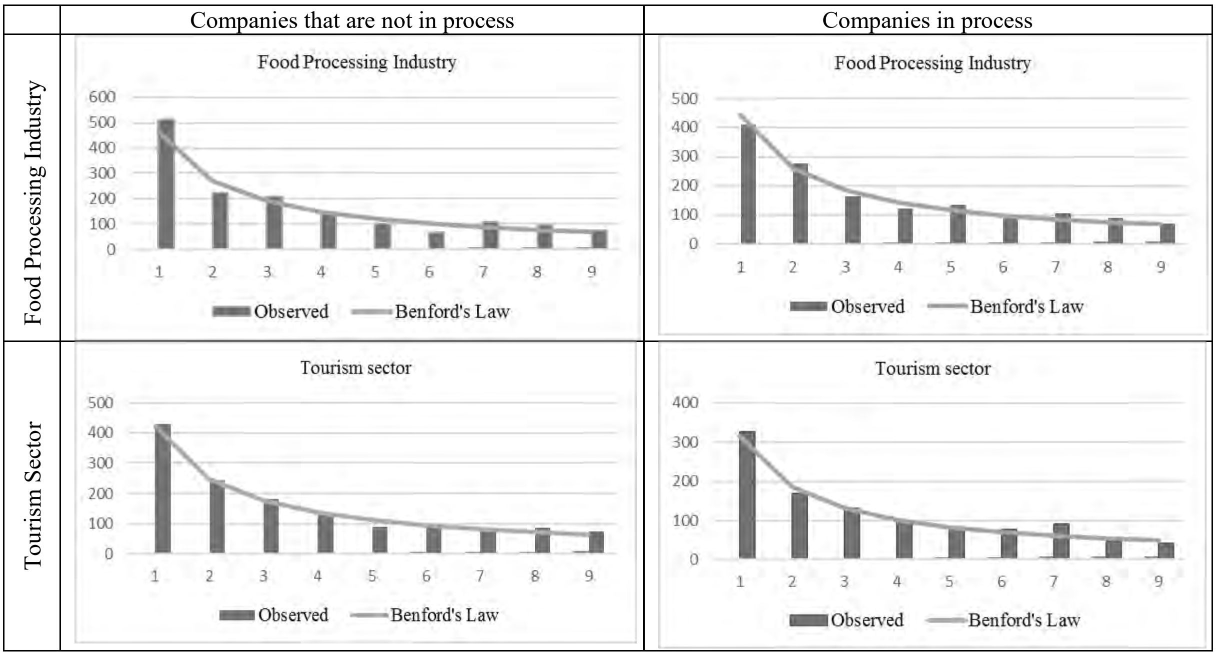


Figure 1: Goodness-of-fit to Benford's Law distribution for companies regarding two situations of being or not being in the process and by two industry types

To test H_0 and H_1 χ^2 test was used to compare the observed distribution of data with Benford's Law first digit distribution. The null hypothesis is rejected at 5% significance if χ^2exceed 15.51 and at 1% significance if χ^2exceeds 20.09. The hypotheses are verified, as it is shown in the Table 5, when applying the χ^2 test with 8 degrees of freedom.

Table 5: Hypothesis Matrix

α	Food Processing Industry		Tourism Sector	
	Companies that are not in process	Companies in process	Companies that are not in process	Companies in process
	χ^2=36.91	**χ^2=19.74**	**χ^2=10.33**	**χ^2=19.18**
0.05	Adopt H_1 Hypothesis	Adopt H_1 Hypothesis	Not reject H_0 Hypothesis	Adopt H_1 Hypothesis
0.01	Adopt H_1 Hypothesis	Not reject H_0 Hypothesis	Not reject H_0 Hypothesis	Not reject H_0 Hypothesis

The testing of the hypothesis at 5% significance has shown that only the financial statements of *companies that are not in process* of pre-bankruptcy settlement from *Tourism Sector* category complied with Benford's Law first digit distribution. At 1% significance the financial statements of *companies that are not in process* of pre-bankruptcy settlement from *Food Processing Industry* category did not comply with Benford's Law first digit distribution.

5 CONCLUSION

In this paper, the financial statements' data distribution fit of included Croatian companies from *Food Processing Industry* and from *Tourism Sector* as regard to Benford's Law first digit distribution was tested using Goodness-of-Fit χ^2-test. The obtained results showed that financial statements of *companies that are not in process of pre-bankruptcy-settlement* from *Tourism Sector*, at significance level of 5%, comply with Benford's Law first digit distribution. The financial statements of *Food Processing Industry* companies deviate from Benford' distribution in a statistically significant way and therefore they should be investigated further.

Benford's Law is useful statistical technique for digital analysis of financial statements but it is necessary to emphasize limitations of this Law, proofing the main research hypothesis set in this paper, but the results of digital analysis should be interpreted carefully, since the Benford's Law of the first digit distribution can be applied only to data sets converging to a logarithmic distribution, respectively.

6 REFERENCES

[1] ACFE (Association of Certified Fraud Examiners, Inc.). 2018. *Reports to the Nations on Occupational Fraud and Abuse.* Retrieved from https://www.acfe.com/report-to-the-nations/2018/.

[2] Corazza, M., Ellero, A., Zorzi, A. 2010. Checking financial markets via Benford's Law the S&P 500 case. In: Corazza M., Pizzi C. (eds.): *Mathematical and Statistical Methods for Actuarial Sciences and Finance*. Springer, Milano, pp. 93-102.

[3] Hill, P. T. 1995. A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, 10(4), 354-363.

[4] Jošić, H., Žmuk B. 2018. The Application of Benford's Law in psychological pricing detection. *Zbornik radova Ekonomskog fakulteta Sveučilišta u Mostaru*, 24, 37-57.

[5] Krakar, Z., Žgela, M. (2009): Evaluation of Benford's Low application in Stock Prices and Stock Turnover. *Informatologia*, 42(3), 158-165.

[6] Nigrini, M. J. 2011. *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations.* New Jersey: John Wiley& Sons, Inc. Hoboken.

[7] Nigrini, M. J., Miller, S. J. 2009. Data Diagnostics Using Second-Order Test of Benford's Law. *Auditing: A Journal of Practice & Theory*, 28(2), 305-324.

[8] Omerzu, N., Kolar, I. 2019. Do the Financial Statements of Listed Companies on the Ljubljana Stock Exchange Pass the Benford's Law Test? *International Business Research*, 12(1), 54-64.

[9] Papić, M., Vudrić, N., Jerin, K. 2017. Benfordov zakon i njegova primjena u forenzičnom računovodstvu. *Zbornik sveučilišta Libertas*, 1-2(1-2), 153-172.

[10] Phillips, T. 2009. Simon Newcomb and "Natural Numbers" (Benford's Law). *American Mathematical Society Feature Column*, Monthly Essays on Mathematical Topics. Retrieved from: http://www.ams.org/publicoutreach/feature-column/fcarc-newcomb.

[11] Rauch, B., Göttsche, M. 2011. Fact and Fiction in EU-Governmental Economic Data. *German Economic Review*, 12(3), 243-255.

[12] Shrestha, I. 2016. *Validity of Financial Statements: Benford's Law*. Indiana. United States.

[13] Sljepčević, S., Blašković, B. (2014): Statistical detection of fraud in the reporting of Croatian public companies. *Financial Theory and Practice*, 38(1), 81-96.

[14] Zagreb Stock Exchange. 2013-2018. *List of securities*. Retrieved from: http://www.zse.hr/default.aspx?id=26474.

CLUSTERS OF EUROPEAN COUNTRIES REGARDING RECENT CHANGES IN BUSINESS DEMOGRAPHY STATISTICS

Ksenija Dumičić, PhD
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: kdumicic@net.efzg.hr

Berislav Žmuk, PhD
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: bzmuk@efzg.hr

Anita Harmina, MA
VERN' University of Applied Sciences
Trg Drage Iblera 10, HR-10000 Zagreb, Croatia
E-mail: anita.harmina@vern.hr

Abstract: This paper analyses the dynamics of business demography statistics variables: number of active enterprises, number of enterprises' births and number of enterprises' deaths, and an additional variable, named the rate of natural increase of enterprises ($RNI_{ent.}$), for 28 European countries in 2016 compared to 2012. Among analysed, 1/3 of countries decreased the $RNI_{ent.}$ indicator values and 2/3 of them increased them in 2016 compared to 2012. Hierearchical clustering performed on three demography variables in both years resulted with three clusters of countries having low, medium or high demography statistics intensity changes, indicating different directions of movements of the $RNI_{ent.}$ indicator values.

Keywords: business demography statistics, hierarchical cluster analysis, rate of natural increase of enterprises, squared Euclidean distances, Eurostat.

1 INTRODUCTION

According to [9], the business demography statistics describes the characteristics and demography of the businesses' population which has been composed of almost 27 million active enterprises for 28 European Union (EU) countries in 2016. About 2.6 million enterprises were newly born in EU-28 that year, which was 90,000 (or 3.5%) newly born enterprises more in 2016 than in the previous year.

Based on data on business demography statistics and through the methods of descriptive statistics, this paper aims to analyse the changes that have occurred between 2012 and 2016 in the number of active enterprises, the birth of enterprises and the death of enterprises for 28 European countries (25 EU members, highly developed Iceland and Norway, and North Macedonia, which is the EU candidate since 2005). Additionally, for the same time and geographical frame, this paper presents the calculation and the analyses of changes in the natural growth of enterprises' population, given as the rate of natural increase of enterprises, denoted here by $RNI_{ent.}$, an indicator which is usually considered in the context of a population of a country. Furthermore, by applying the hierarchical cluster analysis approach for 28 European countries, this research presents and compares the clusters for 2012 and 2016, and discusses the relations between the two cluster solutions and the $RNI_{ent.}$ indicator. When applying the business demography concepts, the population demography formulas are respected, as given in [14].

The birth of new enterprises and the death of unproductive ones are the main events of business dynamism. The newly born enterprises are considered as the innovators, as drivers of economic productivity and as new employment creators, since they stimulate the competitiveness and efficiency of an economy [2, 4, 5, 13]. The death of enterprises is, on the other hand, crucial for the process of "creative destruction" [12]. Even though for most EU countries about 80% of newly born enterprises survive after one year and less than 50% of them are still active after five years, the employment created by enterprises' births still overcomes the jobs lost due to the death of enterprises [4, 9]. Larger enterprises contribute less to job creation than smaller enterprises do [5], so policies which restrict large enterprises competitiveness also support more enterprises births [11]. In 2015 the EU-28 businesses' population recorded an estimated birth rate of 9.6% and an estimated death rate of 8.3% [7]. In recent years (up to the first quarter of 2017) the number of new enterprises in most OECD countries returns to, and in many cases even exceeds the pre-crisis levels [12]. By applying the cluster analysis, the author [10] found four clusters of EU countries characterized by different entrepreneurial dynamics, income level and cultural values, the authors [15] clustered transition EU countries according to their productive entrepreneurial performance and institutions and the authors [1] found five entrepreneurial types of European countries explained by innovation, employment, formal institutions, entrepreneurship and governance.

2 DATA AND METHODS

In the paper, 28 European countries have been observed. Regarding the geographical scope of the analysis, it should be noted that due to official data availability 25 EU member states were included but Cyprus, Greece and Poland were excluded from the analysis. However, 3 additional countries: the EU candidate country North Macedonia and two highly developed countries, Iceland and Norway have been included. In the analysis, the main focus is given to the following three variables: number of active enterprises, number of enterprises' births and number of enterprises' deaths. Data for all variables are taken from the Eurostat database [8]. Onwards, due to missing data, analysed data are related only to 2012 and to 2016. In addition, the variable rate of natural increase of enterprises indicator ($RNI_{ent.}$) has been calculated for the purpose of the analysis by using the following equation, according to [14]:

$$RNI_{ent.} = \frac{\left(\text{No. of births - No. of deaths}\right)}{\text{No. of active enterprises}} \cdot 100\% . \tag{1}$$

The analysis was first conducted by using descriptive statistics methods and afterwards the hierarchical cluster analysis based on the Ward's linkage and squared Euclidean distances was performed. In the cluster analysis, the variables number of active enterprises, number of enterprises' births and number of enterprises' deaths were used, whereas the variable rate of natural increase of enterprises ($RNI_{ent.}$) was used as a control variable for comparisons. In order to determine the most appropriate number of clusters, the Calinski and Harabasz pseudo-F index, according to [3], and the Duda-Hart Je(2)/Je(1) index F, according to [6], were used. The Calinski and Harabasz pseudo-F index is defined as follows:

$$\text{C - H pseudo F index} = \frac{\text{trace}(B)/(c-1)}{\text{trace}(W)/(n-c)}, \tag{2}$$

where B is the between-cluster sum of squares and cross-products matrix, c is the number of clusters, W is the within-cluster sum of squares and cross-products matrix, n is the number of observations. On the other hand, the Duda-Hart Je(2)/Je(1) index F is defined as a ratio of the sum of squared errors within the group and the sum of squared errors in the two resulting subgroups. The better distinct cluster structure is achieved the larger both indices are. Pseudo T-squared measure is closely connected with the Duda-Hart Je(2)/Je(1) index F and it will be observed as well. However, lower pseudo T-squared measure values are preferable.

3 DESCRIPTIVE STATISTICS ANALYSIS

In Table 1 main descriptive statistics results for all four observed variables are given. The results are shown separately for 2012 and for 2016.

Table 1: Descriptive statistics of business statistics demography and rate of natural increase of enterprises, for selected 28 European countries, in 2012 and 2016

Statistics	***No. of active enterprises***		***No. of enterprises' births***		***No. of enterprises' deaths***		RNI$_{ent.}$	
	2012	***2016***	***2012***	***2012***	***2016***	***2016***	***2012***	***2016***
Mean	832,055	875,897	72,034	70,185	75,861	85,197	0.6	2.7
Stand. Dev.	1,103,549	1,137,177	89,190	91,717	93,616	110,555	2.7	3.9
Coeff. of var.	133	130	124	131	123	130	470	145
Minimum	24,164	27,653	2,133	1,101	1,947	2,891	0.6	2.7
1st quartile	142,871	146,249	12,144	10,349	12,828	17,827	2.7	3.9
Median	364,057	377,021	28,607	23,048	38,057	39,145	470	145
3rd quartile	852,421	892,241	92,758	85,051	89,323	95,617	0.6	2.7
Maximum	3,953,714	3,849,594	294,961	316,786	308,326	371,365	2.7	3.9

At all four analysed variables, Table 1, the differences between the countries are remarkable. In 2012, the lowest number of active enterprises had Iceland (24,164) whereas the highest number of active enterprises had Italy (3,953,714). These two countries had the same position in 2016, as well. In 2012, the number of enterprises' births was the lowest for Malta (1,947), whereas the largest was for France (308,326). In 2016, the lowest number of enterprises' births had Luxembourg (2,891) and the largest had United Kingdom (371,365). In 2012, if the number of enterprises' deaths is observed, it can be noticed that the lowest value had Luxembourg (2,133), whereas the largest had Spain (294,961). In 2016, the lowest number of enterprises' births had Malta (1,101), and the largest one was noticed for Italy (316,786). Finally, regarding the rate of natural increase of enterprises, in 2012, the lowest RNI$_{ent.}$ had Croatia (-3.6%), whereas the largest RNI$_{ent.}$ had Lithuania (6.2%). If situation in 2016 is observed, it can be noticed that the lowest RNI$_{ent.}$ had Bulgaria (-4.3%) and that the largest RNI$_{ent.}$ was found for Malta (13.3%).

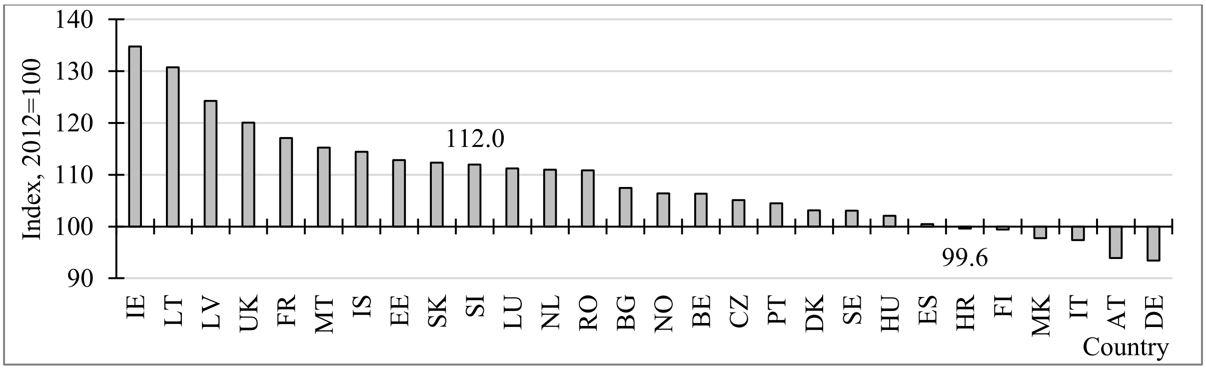


Figure 1: Indices of active enterprises in 28 European countries in 2016, 2012=100 (ISO country codes)

When comparing Croatia and Slovenia, it can be noticed that Croatia had more active enterprises in both observed years (2012: Croatia 147,798 vs. Slovenia 128,088; 2016: Croatia 147,181 vs. Slovenia 143,451). However, Figure 1 reveals that the difference between two countries in the number of active enterprises was reduced. Slovenia had better trends at the other variables than Croatia, as well. In both years Slovenia had higher number of enterprises' births (2012: Croatia 12,123 vs. Slovenia 12,920; 2016: Croatia 12,856 vs. Slovenia 14,884), lower number of enterprises' deaths (2012: Croatia 17,448 vs. Slovenia 11,615; 2016: Croatia 11,832 vs. Slovenia 10,390) and higher $RNI_{ent.}$ (2012: Croatia -3.6% vs. Slovenia 1.0%; 2016: Croatia 0.7% vs. Slovenia 3.1%).

4 CLUSTER ANALYSIS AND DISCUSSION

The hierarchical cluster analyses were conducted by using data for all observed European countries at three variables: number of active enterprises, number of enterprises' births and number of enterprises' deaths. Two cluster analyses were conducted by separately using data from 2012 and from 2016. In both cases the Ward's clustering method and squared Euclidean distances, as a distance measure, were used. In order to select the optimal number of clusters the results of the Calinski and Harabasz pseudo-F index, the Duda-Hart Je(2)/Je(1) index F and pseudo T-squared measure values up to ten-cluster-solution are given in Table 2 for both years.

Table 2: Calinski and Harabasz pseudo-F index and the Duda-Hart Je(2)/Je(1) index F values, data related to 2012 and 2016

Number of clusters	***2012***			***2016***		
	Calinski and Harabasz pseudo-F index	***Duda-Hart***		***Calinski and Harabasz pseudo-F index***	***Duda-Hart***	
		Je(2)/Je(1) index F	***Pseudo T-squared***		***Je(2)/Je(1) index F***	***Pseudo T-squared***
2	171.71	0.2746	55.49	163.24	0.3308	42.47
3	184.74	0.4905	3.12	145.24	0.6253	1.80
4	172.90	0.3873	26.90	124.62	0.2656	19.36
5	182.39	0.0000	-	130.31	0.4884	2.09
6	209.58	0.1618	5.18	164.14	0.2649	2.78
7	255.96	0.4281	10.69	233.50	0.0000	-
8	310.41	0.4592	2.36	264.59	0.1107	8.03
9	325.99	0.4841	6.39	288.06	0.4226	16.40
10	349.23	0.0000	-	343.96	0.6075	2.58

It arose, Table 2, that the optimum number of clusters should be three. In Table 3 and Table 4, the members of clusters in 2012 and 2016 are given, respectively.

Table 3: Clusters' members, data related to 2012

Cluster	***Countries (ISO country code)***
A	AT, BE, BG, HR, DK, EE, FI, HU, IS, IE, LV, LT, LU, MT, MK, NO, SK, SI, SE
B	CZ, NL, PT, RO
C	FR, DE, IT, ES, UK

Table 4: Clusters' members, data related to 2016

Cluster	***Countries (ISO country code)***
A	AT, HR, DK, EE, FI, IS, IE, LV, LT, LU, MT, MK, NO, SI
B	BE, BG, CZ, HU, NL, PT, RO, SK, SE
C	FR, DE, IT, ES, UK

According to Table 3 and Table 4, the cluster A contains 19 countries in 2012 and 14 in 2016, cluster B has 4 countries in 2012 and 9 in 2016, and cluster C has the same 5 countries in 2012 and 2016. The cluster A has the lowest average value of all three business demography variables in 2012 and 2016. The cluster averages for the number of enterprises' births (23,136 in 2012; 16,289 in 2016), the number of enterprises' deaths (20,980 in 2012; 10,277 in 2016) and the number of active enterprises (250,335 in 2012; 165,161 in 2016) are much lower in 2016 than in 2012 and are also much lower than the sample averages in each year. For the countries grouped in this cluster in 2012, the $RNI_{ent.}$ indicator, however, varies greatly from -3.6% in Croatia to 6.2% in Lithuania, with an average of 0.9% (close to the sample average of 0.8% for 2012). Seven countries in this cluster in 2012 have negative $RNI_{ent.}$: Sweden, Denmark, Ireland, Hungary, Malta, North Macedonia and Croatia. The average of the $RNI_{ent.}$ indicator in cluster A in 2016 equals 4.3%, which is much higher than the sample average in 2016, and is also the highest average value for this variable among all clusters in 2016. Only Finland, among 14 countries grouped in this cluster, has a negative $RNI_{ent.}$ indicator, and only Finland, Luxembourg and Estonia have lower $RNI_{ent.}$ indicator in 2012. Specially, despite the earlier observed differences between Croatia and Slovenia regarding all four variables, these two EU members are both found in cluster A in 2012 and in 2016. The cluster B is described by mostly medium values in business demography variables in 2012 and in 2016. The average number of enterprises' births for countries in this cluster in 2012 equals 93,120, the cluster's B average number of enterprises' deaths equals 100,510 and the average number of active enterprises equals 870,761. All these averages are higher than the corresponding sample averages in 2012. Only the Netherlands has a positive value of the $RNI_{ent.}$ indicator in 2012. Cluster B in 2016 grouped all countries that were formerly in this cluster in 2012 plus Belgium, Bulgaria, Hungary, Slovakia and Sweden. This cluster averages are the closest to the sample averages in 2016. The $RNI_{ent.}$ indicator for countries in this cluster ranges from -4.3% in Bulgaria (which is also the only country in this cluster with negative $RNI_{ent.}$ indicator and the only country who’s $RNI_{ent.}$ indicator severely declined in comparison to 2012) to 2.9% in Belgium, and has an average equal to 0.7%, which is much lower compared to the sample average in 2016. The cluster C is characterized by very high values of all clustering variables in 2012 and 2016. The cluster's C averages in 2012 for the number of enterprises' births (262,410), the number of enterprises' deaths (243,257) and the number of active enterprises (3,011,625) are all much higher than the corresponding sample averages in 2012. In this cluster, in 2012, the countries are very different regarding the $RNI_{ent.}$ indicator, which varies from -1.6% in Spain to 4.5% in France, and has an average of 0.7%. This cluster's averages of business demography variables in 2016 are, however, higher than averages in cluster C in 2012. The average $RNI_{ent.}$ indicator in cluster C, in 2016, equals 1.8%, and is higher than the average for these same countries in 2012. Unlike the other countries in this cluster in 2016, the $RNI_{ent.}$ indicator for Italy and Germany continued with further decrease.

5 CONCLUSIONS

The performed hierarchical cluster analyses for 28 European countries in 2012 and 2016, were based on three basic business demography variables: number of active enterprises, number of enterprises’ births and number of enterprises’ deaths, and resulted with three clusters of countries: the demography statistics lowest intensity changing cluster A; the demography statistics medium intensity changing cluster B; and the demography statistics highest intensity changing cluster C. This resulted with clusters of countries having obviously different demography statistics intensity changes, low, medium or high. In the year 2016 regarding the 2012, within the resulting clusters, obvious differences in dynamics based on indicator for rate of natural increase of enterprises, RNIent., were indicated. Among the analysed countries, 9 decreased the $RNI_{ent.}$ indicator values and 19 of them increased them in 2016 compared to 2012. Malta, Ireland, Lithuania, Hungary, Portugal and Croatia showed obvious positive tendencies in improving the $RNI_{ent.}$ indicator values in 2016 compared to 2012, while Bulgaria, Finland, Estonia, Luxembourg, Italy and Germany shown a surprising decrease of the rate of natural increase of enterprises indicator, $RNI_{ent.}$ indicator, in the same period. In general

Slovenia shown better values for all three basic business demography variables compared to Croatia. The number of active enterprises in Croatia dropped by 0.4%, whereas the number of active enterprises in Slovenia rose by 12% in 2016 compared to 2012, but Croatia gave better improvement in the $RNI_{ent.}$ indicator by the absolute difference amount in that indicator of 4.3 in 2016 compared to 2012, and Slovenia had an absolute difference in the $RNI_{ent.}$ indicator extent of 2.1.

References

[1] Abdesselam, R., Bonnet, J., Renou-Maissant, P. 2017. The Demography of Enterprises and Employment in the European Union Countries. Center for Research in Economics and Management (CREM), University of Rennes 1, University of Caen and CNRS. No. 2017-10.

[2] Audretsch, D. B. 2012. Entrepreneurship research. Management Decision, 50(5): 755-764.

[3] Calinski, T., Harabasz. J. 1974. A dendrite method for cluster analysis. Communications in Statistics, 3(1): 1–27.

[4] de Kok, J., Vroonhof, P., Verhoeven, W., Timmermans, N., Kwaak, T., Snijders, J., Westhof, F. 2011. Do SMEs create more and better jobs. Brussels: European Commission.

[5] de Wit, G., de Kok, J. 2014. Do small businesses create more jobs? New evidence for Europe. Small Business Economics, 42(2): 283-295.

[6] Duda, R. O., Hart, P. E. Stork, D. G. 2000. Pattern Classification. New York: Wiley.

[7] Eurostat. 2018. Structural business statistics at regional level. Statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Structural_business_statistics_at_regional_level#Enterprise_demography [Accessed 12/06/2019].

[8] Eurostat. 2019. Business demography by legal form (from 2004 onwards, NACE Rev. 2). Eurostat. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=bd_9ac_l_form_r2&lang=en [Accessed 07/06/2019].

[9] Eurostat. 2019a. Business demography statistics. Statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php/Business_demography_statistics [Accessed 28/4/2019].

[10] Linan, F., Fernandez-Serrano, J. 2014. National culture, entrepreneurship and economic development: different patterns across the European Union. Small Business Economics, 42(4): 685-701.

[11] Nofsinger, J., Zykaj, B. B. 2014. Business Policies and New Firm Birth Rates Internationally. Accounting and Finance Research, 3(4): 1-14.

[12] OECD. 2017. Entrepreneurship at a Glance 2017. Paris: OECD Publishing.

[13] Pinillos, M. J., Reyes, L. 2011. Relationship between individualist–collectivist culture and entrepreneurial activity: evidence from Global Entrepreneurship Monitor data. Small Business Economics, 37(1): 23-37.

[14] Pol, L. G., Thomas, R. K. 1997. Demography for business decision making. Westport, CT.: Quorum Books.

[15] Szabo, Z. K., Herman, E. 2014. Productive entrepreneurship in the EU and its barriers in transition economies: A cluster analysis. Acta Polytechnica Hungarica, 11(6): 73-94.

DEEP LEARNING PREDICTIVE MODELS FOR TERMINAL CALL RATE PREDICTION DURING THE WARRANTY PERIOD

Aljaž Ferencek
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
aljaz.ferencek@student.um.si

Mirjana Kljajić Borštnar
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
mirjana.kljajic@um.si

Davorin Kofjač
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
davorin.kofjac@um.si

Andrej Škraba
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
andrej.skraba@um.si

Blaž Sašek
University of Maribor, Faculty of Organizational Sciences
Kidričeva cesta 55a, 4000 Kranj, Slovenia
blaz.sasek@student.um.si

Abstract: The problem of products' terminal call rate (TCR) prediction during the warranty period is addressed in this paper. TCR represents a key information for a quality management department to reserve the necessary funds for product repair during the warranty period. Various methods have been used to address this problem, from discrete even simulation, time series, to machine learning predictive models. We have developed a deep learning predictive models and analysed their quality and performance. Results suggest deep learning is an approach worth further exploring but require large volumes of quality data.

Keywords: manufacturing, product lifecycle, management product failure, machine learning, prediction

1 INTRODUCTION

Todays' business environment is highly competitive thus businesses need to optimize their costs and improve their profit and/or market share. Warranty claim control as a part of quality control department is one of the most important departments since servicing warranties involves additional costs to the manufacturer. Warranty is an important element of marketing new products as better warranty signals higher product quality and provides greater assurance to customers [8]. Higher quality of the product is related with product reliability which is, in a more technical definition, the probability that the product (system) will perform its intended function for a specified time period when operating under normal (or stated) environmental conditions [1].

Many quality and reliability engineers who are involved in the warranty claims predictions use empirical models based on past data of products with similar design and complexity adjusted by certain, experience-based correction factors [5], [10]. In this paper we set out to develop and validate a prediction model, using deep learning models on a case of a production-oriented company on the field of home appliances.

1.1 Previous research

When addressing problems that aim to forecast the future, we are talking about machine learning and more specifically predictive analytics. Predictive analytics has been used in many business applications such as customer relationship management, predicting crime fighting in law enforcement, predicting warranty problems for automobile manufacturers, predicting change in stock price, etc. [3].

The problem addressed in this paper deals with predicting warranty call rates into the future given up some initial data for each production series, which was previously addressed by using a combined model of market absorption and failure process [6], where forecasting activity for current products were done by using warranty claims for the first few months of service. Their objective was to anticipate the final numbers of warranty returns at the end of the warranty cycle. In general, there were two types of data. First type was focusing on products from sales to failure, and the other type from production to failure, where the market absorption time was considered. For claims prediction, authors proposed the Markov Modulated Fluid Model. The model was verified and validated against the real-world data, authors also proposed an application of grid computing due to costs of prediction in means of computer power.

In their later research Kofjač et al [7] proposed the use of machine learning methods for forecasting of terminal call rate. They investigated the estimation of cumulative density function with MLM and its impact on the TCR (Terminal call rate) prediction accuracy. The cumulative density functions were modelled with exponential and logistic models and their parameters were estimated with MLM, such as regression trees, neural networks and ensembles of regression trees. Standard error of the estimate (SEE) measure was used to evaluate goodness of fit of cumulative density functions to the actual data was and the best results were achieved by ensembles of regression trees with SEE. Because the stance of the study was focused on fundamental research in prediction of TCR with MLM, authors proposed future studies to improve the prediction accuracy, addressing the impact of other attributes, such as mean time to failure (MTTF) and the optimization of ML methods attributes, for example, number of instances in leaves for regression trees.

Finally the last research on given problem was a student project called PKP (Po kreativni poti do znanja) funded under Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia that addressed the falling prices of processing and storage capabilities and development of new models and techniques, where one could build models as needed using only the most relevant and recent data. A team of students and their mentors were tasked with the development of the prototype application that would provide end users with call rate predictions for the next year to support planning of the expenses. In the core of this prototype was a stacked model which consisted of a linear regression model a simple tree and a random forest regressor. As new data was extracted from the transactional database it was transformed and stored as a matrix of cumulative call rates for each month following a production of an individual series. The model would then be applied on user's demand and it would locate the most appropriate date range and learn all three models on the provided data, followed by a modulation using implementation of techniques from the Forecasting terminal call rate with machine learning methods [7]. The above-mentioned process was successfully implemented in the Guided Machine Learning for Business Users [2]. Models from PKP project would then vote on future predictions for series that are still in the warranty period. Their contribution would be weighted based on how well they performed on test inputs while learning. If prediction was triggered in the following months, when new data was available, the model would incorporate this data and re-learn itself before evaluating predictions for the new time period. Given more data with each subsequent execution the prototype should be less volatile and prone to errors, but this was never empirically tested on unseen data since the project ended before the prototype could be fully implemented and integrated with the transactional database.

This research contributes to the discussion of how successful predictions in the field of warranty claims can be made based on related work.

2 METHODOLOGY

The proposed methodological approach is rooted in Design Science Research. Design Science Research is driven by business needs to ensure relevance and uses theoretical knowledge for rigor [4]. In order to structure data mining process, several models are available, one of the most widespread approaches being the Cross-Industry Standard Process for data mining, CRISP-DM. The process or methodology of CRISP-DM is described in these six major steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment [11].

The research process in this paper is at first discussing previous studies and later using this knowledge or findings to make an evaluation and oppose future research with other possible methods in the process of developing a model for warranty claim predictions. The basic idea of gaining additional insight if another dimension to the data is added [9] and is explored in this paper. Instead of only looking at when the appliance would break, we would also want to look at why it happened. We were particularly interested in the performance of convolutional neural networks, given that we have a 2-dimensional dataset that has the same representation as a single-channel (grayscale) image. Applying the CRISP-DM methodology we visited the factory and spoke with employees inside the QA department to gain a deeper understanding of the business problem at hand. During this stage we managed to reduce the window of prediction from 36 or 60 to 12 months while maintaining to support the business needs of the company. The data was taken from the transactional database of all the service interventions. We received data for 4 specific markets - Nordic region, Germany, Russia and Serbia. There were specifics for each market which we were made aware of during our visits to the factory (for instance: it has been noticed that in the Nordic countries there is a greater time between the appliance failure and the service procedure than in other markets). Data was of varying quality depending on the market and the time period in which it was produced. While newer data was evidently more consistent there was less of a systematic approach to gathering data in the past, thus limiting the usefulness of the data. Still, almost 50 % of all service inquiries did not have the reason for failure listed and that was the important feature in our proposed models. Data preparation consisted of checking for inconsistent and missing data, which was then either cleaned or removed. Next, we transformed the data into the format we wanted to use and split the data based on the product hierarchy to get more samples. We ended up with 11852 useful data points. After that we were able to proceed towards model development.

3 RESULTS

Our goal was to predict call rates 12 months into the future given up to 6 months of initial data for each series. We were familiar with the business problem and the data, which hasn't changed since previous research projects [5], [6], [7], [10]. Dataset contained data from the service procedures, which consisted of product identifier (this is standardised and describes the product down to the model level), production date, failure date and a reason for failure. Reasons for failure were comparable in products close in the product hierarchy but could be more different the further the two products inside the product hierarchy. We did not control for this since our dataset consisted of only data for a very narrow product hierarchy. For the purposes of this new analysis we did have to rewrite the data pre-processing pipelines since a dimension in a matrix was changed from the series (all products produced in a certain month) to a reason for failure. The result of this new pre-processing pipeline was a matrix, presented in Figure 1 where the first dimension represents a few unique failure reasons and the second one is the number of months. This matrix was used as an input for our predictive models.

$$
\begin{array}{c} \\ error_1 \\ error_2 \\ error_3 \\ \vdots \\ error_n \end{array}
\begin{array}{c}
\begin{array}{cccccc} mon_1 & mon_2 & mon_3 & mon_4 & mon_5 & mon_6 \end{array} \\
\begin{bmatrix}
\mathbf{0.01} & \mathbf{0} & \mathbf{0.002} & \mathbf{0.01} & \mathbf{0.003} & \mathbf{0.004} \\
\mathbf{0} & \mathbf{0.02} & \mathbf{0.03} & \mathbf{0.01} & \mathbf{0.01} & \mathbf{0.005} \\
\mathbf{0.1} & \mathbf{0.05} & \mathbf{0.04} & \mathbf{0} & \mathbf{0.004} & \mathbf{0.01} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0.02} & \mathbf{0} & \mathbf{0.01} & \mathbf{0.03} & \mathbf{0.04} & \mathbf{0}
\end{bmatrix}
\end{array}
$$

Figure 1: New matrix as a result of pre-processing pipeline

Upon inspection of this previously unused dimension of the data we learned that there are a lot of missing values. More than 46 % of the service entries had no stated reason for the failure. This was varying from market to market which suggested that we could possibly extract useful data from certain markets while the model would pay less attention to the markets with a lot of missing data, presented in Figure 2.

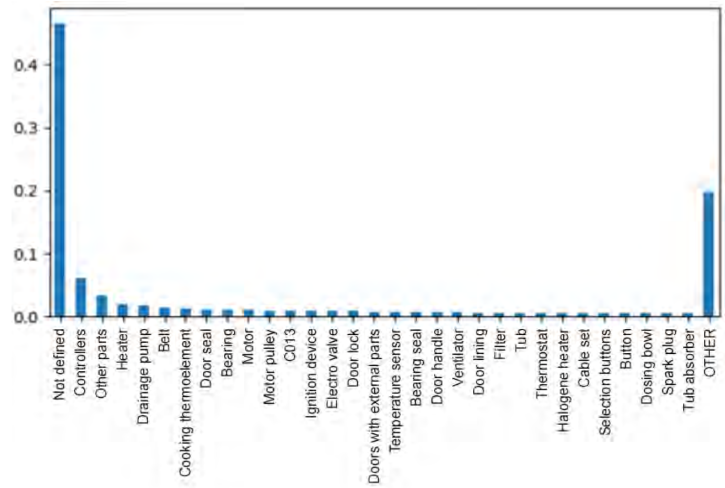


Figure 2: Error frequency in the data (% used on y axis where OTHER is meant as a joined group with less than 0,5% of the data). Adapted from [9]

To provide enough training samples for neural networks, which usually only work with high amounts of data, we prepared separate matrices for each market and we also fragmented products based on a few different levels of product hierarchy provided by the data owner. This gave us 308 fragments on 4 markets for which we had data available for multiple series produced in the past. In total that gave us 11852 data points which is admittedly a small number given that neural networks usually consume datasets that are orders of magnitude larger…

For our model we started with a shallow neural network which was not able to learn well on our data. We proceeded with a series of optimizations which included:

- **Deepening the neural network:** we saw best results with a neural net with 6 hidden layers, after that adding additional layers did not result in better predictions.

- **Testing of a few different activation functions:** ReLU, ELU, softmax, tanh and sigmoid were tested, sigmoid was ultimately selected as the best alternative for our model.
- **Regularization:** we tested dropout with no effect and we also tried L2 regularization which had a detrimental effect on our model.
- **Hyperparameter optimizations:** we ran experiments using different learning rates, keep probabilities for dropout and different number of neurons in each layer. This did help to speed up convergence and reduce overfitting but did not improve prediction accuracy.

Our final model was a simple convolutional neural network with three convolutional layers and a fully connected layer. It also incorporated regularization techniques from previous examples. We generated three different models - a single layer perceptron (which was scrapped immediately since it did not converge), a few variations of deep neural networks and a convolutional net. We used L1 norms+ (absolute difference) to measure accuracy since it is most intuitively interpreted by humans that would look at the results (QA department at Gorenje). There was no significant difference between a deep neural network and a convolutional networks in terms of accuracy but the convolutional network did converge faster, which could mean shorter learning times with less computing resources as you can see on Figure 3.

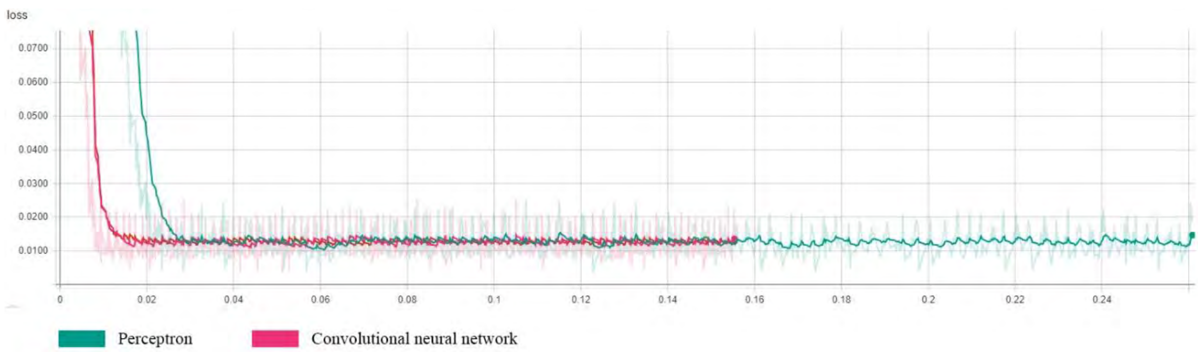


Figure 3: Convolutional neural network with higher level of learning (0,101)

4 CONCLUSIONS

We tested a new approach in predicting failure rates in home appliances. Our research was handicapped by a low amount of data which was also low quality. We employed the CRISP-DM model, learning about business requirements, understanding the data and putting a lot of work into cleaning the data and transforming it to the right format for consumption by neural nets. After that we developed a few models based on different neural network architectures.

For our best two models (deep neural network with 6 layers and a convolutional neural network) the absolute difference between TCR at 12 months and our prediction was 1 percentage point on average. Since it rarely happens that more than 5 % of the items sold will return in a single year that means that our errors in relative terms rise well above 20 % which is not better than the models they are currently using. Based on these results we could not conclude that deep neural networks perform better than conventional machine learning methods for prediction of call rates in home appliances, but there is more research to be done before we could make any conclusions to the contrary. There are many approaches that were not tested yet (for instance: deep belief networks, LSTM networks, recursive networks, etc.)

and our own models would probably gain some predictive power from more and better data which is produced as this paper is being made.

Acknowledgement
This research was supported by the Slovenian Research Agency, ARRS, through research programme P5-0018; and partly funded by Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia, through Creative Path to Practical Knowledge project.

References

[1] Blischke, R. and Murthy, P. (2000) Reliability. Wiley, New York, 18-19. http://dx.doi.org/10.1002/9781118150481

[2] Bourcevet, A., Piller, G., Scholz, M. & Wiesemann, J. (2019). Guided Machine Learning for Business Users. 32nd Bled eConference, Vol(1), pg: 257 - 270.

[3] Chan, J. (2007). A Predictive Analytic Model for Value Chain Management. Journal of International Technology and Information Management, Vol(16), Issue 1, Article 3.

[4] Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, Vol(28), Issue 1, pg: 75 – 105.

[5] Kljajić M., Bernik I., Škraba A. (2000). Simulation Approach to Decision Assessment in Enterprises. Simulation. Simulation Councils Inc., Vol(75), Issue 4, pg: 199 – 210.

[6] Kofjač, D., Škraba, A. & Brglez, A. (2014). Warranty Claims Prediction with a Combined Model of Market Absorption and Failure Process. International journal of computing anticipatory systems, pg: 81-91.

[7] Kofjač, D., Škraba, A., & Mujanović, A. (2016). Forecasting terminal call rate with machine learning methods. 29th Bled eConference, Vol(1).

[8] Murthy, P. & Djamaludin, I. (2002). New product warranty: A literature review. International Journal of Production Economics. 79. 231-260. 10.1016/S0925-5273(02)00153-6.

[9] Sašek, B. (2017). Product failure prediction with deep learning methods. University of Maribor. https://dk.um.si/IzpisGradiva.php?lang=slv&id=67936

[10] Škraba A., Kljajić M., Papler P., Kofjač D., Obed M. (2011). Determination of recruitment and transition strategies. Kybernetes, Vol(40), Issue 9/10, pg: 1503 - 1522, doi: 10.1108/03684921111169512.

[11] Wirth, J. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pg: 29 – 39.

DETERMINING BUSINESS PROCESS MATURITY LEVELS BY USING CLUSTER ANALYSIS: CASE OF CROATIA

Ljubica Milanović Glavan
Faculty of Economics and Business, University of Zagreb, Department of Informatics
Trg J. F. Kennedya 6, 10000, Zagreb
E-mail: ljmilanovic@efzg.hr

Abstract: It has been shown in previous studies that the companies which have reached higher business process maturity level consistently outperform those that have not reached them. Over the past few years different methodologies for analysing maturity state of business process orientation (BPO) have been developed. Based on survey results cluster analysis method was used to determine BPO maturity level of Croatian companies. It has been calculated that companies in Croatia are between the Defined and Linked level of BPO maturity.

Keywords: business process orientation, maturity model, maturity level, cluster analysis, Croatia

1 INTRODUCTION

Competition in today's global economy is now based upon capabilities, or "complex bundles of skills and accumulated knowledge, exercised through organizational processes" [6]. Owing to this new capabilities business approach many organizations are now viewing processes as strategic assets. Under this perspective, organizations are no longer viewed as a collection of functional areas, but as a combination of highly integrated processes [6]. Thus, the concept of BPO is becoming increasingly important. Literature review shows that there are several general definitions of BPO, but the most extended version was delivered by McCormack [5]: BPO is the level at which the company pays attention to its relevant (core) processes. BPO can slim down operational costs, promote customer relations through satisfying customer needs better and increase employee satisfaction. As this is a complex process done over a long period of time, companies can attain various degrees of BPO acceptance through adjustments of their business processes.

The broad adoption of BPO within an organization derives from the understanding that processes have life cycles or developmental stages that can be clearly defined, managed, measured and controlled throughout time. Since the 1980s, a plethora of maturity models have emerged that claim to guide an organization through the process of building levels of maturity that lead to competitive advantage [1;2;8;13]. To date, there has been a lack of quantitative studies documenting these models. The fact is also that the most of the literature on BPO has been in the popular press and lacks research or an empirical focus. There is no explanation on what precisely companies need to do to advance to higher BPO maturity level. So, the aim of this paper was to address this issue. This paper investigates BPO maturity levels in Croatian companies by using cluster analysis method. The paper is organized in the following way: first, background and purpose of the conducted research is explained, second, BPO model and maturity levels are described, third, methodology, data source and results of the survey conducted in Croatian companies are given, and finally, conclusion, limitations and future research are presented.

2 THEORETICAL BACKGROUND ON BUSINESS PROCESS ORIENTATION MATURITY

Higher levels of maturity in any business process result in: better control of results, improved forecasting of performance, greater effectiveness in reaching defined goals and improving managements' ability to propose new and higher targets [5]. As companies increase their process maturity, institutionalization takes place via policies, standards, and organizational structures [4]. As processes mature they move from an internally focused perspective to an externally focused, system perspective. A maturity level represents a threshold, that when reached, will institutionalize a total systems view necessary to achieve a set of process goals [5]. Achieving each level of maturity establishes a higher level of process capability for the company. In the current business environment, there is no scarcity of process maturity models [9]. For the purpose of this research, the BPO maturity model and assessment instruments from McCormack [5] were used as a starting point and adapted as needed for each individual research objective. The McCormack construct describes a four-step pathway for systematically advancing business processes along the maturity continuum ("Ad hoc", "Defined", "Linked" and "Integrated" level). Each step builds on the work of the previous steps to apply improvement strategies that are appropriate to the current maturity level. The following definitions for the levels that the company goes through when becoming BPO are provided [5;6]:

(1) "Ad hoc". The processes are unstructured and ill defined. Process measures are not in place and the jobs and organizational structures are based upon the traditional functions, not horizontal processes.
(2) "Defined". The basic processes are defined, documented and available in flowcharts. Changes to these processes must now go through a formal procedure. Jobs and organizational structures include a process aspect, but remain basically functional. Representatives from functional areas meet regularly to coordinate with each other, but only as representatives of their traditional functions.
(3) "Linked". The breakthrough level. Managers employ process management with strategic intent and results. Broad process jobs, and structures are put in place outside of traditional functions.
(4) "Integrated". The company, its vendors and suppliers, take cooperation to the process level. Organizational structures and jobs are based on processes, and traditional functions begin to be equal or sometimes subordinate to process. Process measures and management systems are deeply imbedded in the organization.

Based on the extensive literature review different viewpoints of BPO have been synthesized into a comprehensive BPO model that takes into account majority of components, frequently mentioned in literature. In order to analyse and improve BPO companies need to take the following domains (components) into account:

A1. "Strategic view"
A2. "Process identification and documentation"
A3. "Process measurement and management"
A4. "Process oriented organizational structure"
A5. "Human resources management"
A6. "Process oriented organizational culture"
A7. "Market orientation"
A8. "Supplier perspective"
A9. "Process oriented information technology"

In each maturity level certain components of BPO become evident and others barely registered. A component of BPO that stabilizes within an organization and leads to the establishment and expansion of other factors that move the organization to the next maturity level is called key turning point [10].

3 THE EMPIRICAL ANALYSIS OF BPO MATURITY IN CROATIAN COMPANIES

3.1 Description on survey, data and method

The main goal of the empirical research was to assess the current state of BPO maturity in Croatian companies. In order to carry out the empirical study a questionnaire was developed. Even though the original instrument from McCormack and Johnson [7] included an overall BPO construct, it was only measured with 3 dimensions. As the goal was to tap deeper into the problem the given construct was enlarged. It contained 60 questions regarding BPO characteristics. The questions were distributed across the nine domains: Strategic view (5 questions); Process identification and documentation (6 questions); Process measurement and management (10 questions); Process oriented organizational structure (7 questions); Human resources management (5 questions); Process oriented organizational culture (6 questions); Market orientation (7 questions); Supplier perspective (3 questions); Process oriented information technology (11 questions). Each question describes a particular BPO characteristic and/or business practice considered important within each domain. The degree of presence of these characteristics in the company is measured on a 7 point Likert scale (1=Strongly disagree, 2=Disagree, 3= Disagree more than agree, 4=Neither agree or disagree, 5=Agree more than disagree, 6=Agree,7 =Strongly Agree).

The main source of data about Croatian companies was the database of The Institute for Business Intelligence and the questionnaire was sent randomly to the 1200 companies. The questionnaire was addressed to the CEOs or the chairpersons of the companies who were instructed to fill out the questionnaire themselves or give it to a competent person within the organization. 127 completed questionnaires were returned, so the final response rate was 10.58%. The selected companies were analysed according to the number of employees. In the resulting data set 40 companies had between 1 and 50 employees, 44 companies had between 50 and 249 employees and 43 companies had 250 or more employees. Companies from all sectors participated in the research, so all business sectors are appropriately captured in the data sample. The most common trade of business in data set was Financial and insurance services (16,53%). It was followed with Manufacturing (15.75%), Trade (11.81%) and Information and communication services (11.09%). 44.82% of the companies in the sample represented other sorts of business. The sample is an adequate representation of the population of big, small and medium sized Croatian companies from all sectors.

Aiming to evaluate maturity levels, cluster analysis was chosen as an approach. Cluster analysis, also called segmentation analysis or taxonomy analysis, seeks to identify homogeneous subgroups of cases in a population. That is, cluster analysis seeks to identify a set of groups that both minimize within-group variation and maximize between-group variation [3]. The first step was to create new construct variables by grouping the collected variables in their specific categories, assuming the maturity model constructs. Next, the two-step cluster analysis was used encompassing all construct variables created in the previous step and the maturity score as the continuous variable. At this point, a fixed number of four clusters was established, corresponding to four maturity levels. The two-step cluster analysis approach grouped cases into pre-clusters which were treated as single cases. Aiming to investigate the relationship between the constructs, hierarchical clustering was performed on the pre-clusters generated by the two-step cluster analysis. This is a method that allows users to select a definition of distance, select a linking method for forming clusters, and then determine how many clusters best suit the data. It requires neither a proximity table like hierarchical classification nor an iterative process like k-means clustering; rather, it is a one-pass-through-the-dataset method. Finally, k-means clustering was used to identify turning points for each component/construct and its respective position regarding maturity classification.

3.2 Results and discussion

Before all, the compound measure of BPO construct was analyzed, which revealed the overall state of BPO. This was done by calculating the average grade for each domain of the questionnaire. Based on the sample the compound measure of the BPO in Croatia is 4.84.

Then, in order to identify the hierarchical relationship between the groupings a set of TwoStep cluster analysis procedures was conducted using Statistical Package for Social Sciences (SPSS). The first step of the two-step procedure is formation of preclusters. The goal of preclustering is to reduce the size of the matrix that contains distances between all possible pairs of cases. Preclusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed precluster or start a new precluster. When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters. In the first step the maturity score was considered as a continuous variable and a fixed number of 4 clusters were defined, each representing one maturity level of BPO maturity model. In this second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. The 127 cases in the sample were classified considering its positions in each of the four clusters, i.e. in each of the four levels of maturity. The research results showed four different centroids. By considering each cluster as a distinguished maturity level, with a centroid determined for each cluster, maturity scores were identified (Figure 1).

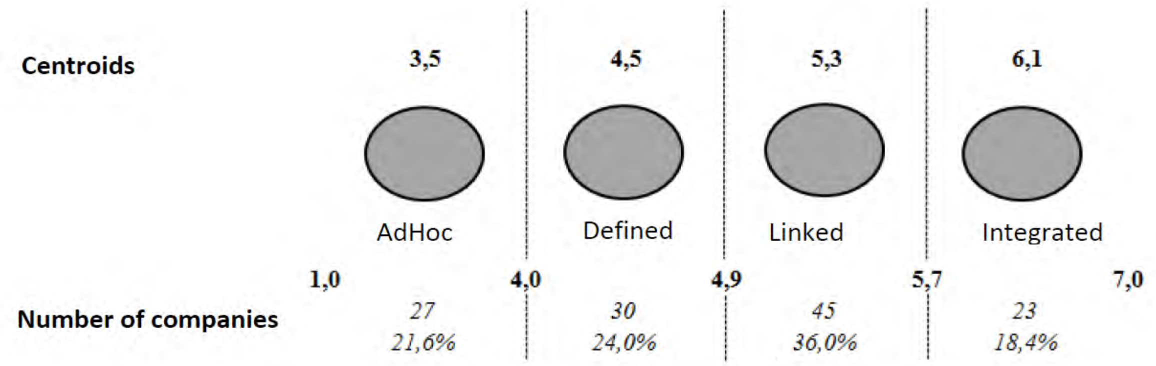


Figure 1. Cluster centroids
Source: Author's calculation

Due to this step, difference in Likert scale of the questionnaire and the number of maturity levels, values were mapped in order to assess the maturity level for each data set record (Table 1).

Table 1. BPO Maturity level mapping

BPO maturity level	Likert scale values
AdHoc	1.0-3.9
Defined	4.0-4.8
Linked	4.9-5.6
Integrated	5.7-7.0

Source: Author's calculation

Companies with BPO value between 1.0 and 3.9 fall into level 1 of BPO maturity (21.6% companies from the sample). Companies with BPO value between 4.0 and 4.8 fall into level 2 of BPO maturity (24.0% companies from the sample). Companies with BPO value between 4.9 and 5.6 fall into level 3 of BPO maturity (36.0% companies from the sample). Companies with BPO value between 5.7 and 7.0 fall into level 4 of BPO maturity (18.4% companies from the sample). Since the compound measure of the BPO is 4.84 it can be concluded that Croatian companies are between the Defined and Linked stage of BPO maturity. Companies with that kind of level of BPO maturity have well defined and documented processes, but don't realize that these business processes are connected.

Additionally, internal turning points in each process grouping, i.e. the points that can be used to define a change in a maturity level for each group, were further identified by means of the cluster analysis with k-means algorithm. The scores in Table 2 represent the percentage score (of the total points available) at the centroid of each level for each maturity component in the maturity model. It is proposed that when the score goes above 50 percent, that component is then established and stable within the set of companies at that maturity level [12]. This represents a key turning point in the maturity continuum.

Table 2: Maturity scores by component and centroid

	Percentage of maturity scores			
	AdHoc	Defined	Linked	Integrated
Strategic view	52%	69%	83%	89%
Process identification and documentation	46%	63%	79%	91%
Process measurement and management	42%	60%	77%	91%
Process oriented organizational structure	51%	70%	80%	90%
Human resources management	42%	62%	72%	89%
Process oriented organizational culture	46%	58%	65%	75%
Market orientation	57%	71%	79%	91%
Supplier perspective	63%	71%	77%	85%
Process oriented information technology	44%	47%	68%	80%

Source: Author's calculation

It can be seen from the Table 2 that the leading component for the Linked level of BPO maturity is strategic view. So, in order to advance to the higher process level, Croatian companies have to improve strategic view. That has to be done by linking process goals to the performance goals and by active involvement of top management in the activities of implementing the principles of BPO into the functioning of the company. A well-developed strategy enables optimal definition, planning and execution of business processes that implement that strategy [11].

4 CONCLUSION

The main goal of this study was to determine the state of BPO adoption in Croatian companies. The data from the empirical study that has been subjected to relevant statistical techniques has shown that the Croatian companies are between the Defined and Linked level of BPO maturity.

The contribution of this paper is multiple, since it offers important implications for research and practice. First, BPO components were systemized. This BPO construct can serve managers as a road map of specific steps that will lead the company to the highest process maturity level.

Second, the empirical results of this research have many practical guidelines for managers of the companies in Croatia. According to the results of the survey, it is of a great importance for the Croatian companies to increase the efforts in stimulating strategic view in order to advance to the higher, Linked level of BPO maturity. That efforts include: the alignment of business processes with organization's strategy possibly achieved by linking process goals to the organization goals and active support of top management in the activities of implementing BPO into the organization. Third, the results of the survey presented in this paper could provide a solid basis for further research in the field it addresses.

Although this research reveals new findings, it is significant to mention that it has few limitations. First, the research was conducted using survey. This means that the conclusions of the research are subject to the general weaknesses of correlational studies. One of the limitations is also a sample size. Success of this method depends highly on the data set used for cluster analysis. A way to improve the reliability of the results would be to increase the sample size of a survey in a future research. Since this survey is limited to respondents from Croatian companies a future research can be done in few other countries to develop a methodology that could be used to compare BPO maturity levels and to detect BPO key turning points in different countries.

References

[1] Bititci, U. S., Turner, T., Begemann, C. (1997), Integrated performance measurement systems: a development guide, International Journal of Operations & Production Management, Vol. 17, No. 5, 522-34.

[2] Bosilj Vukšić V., Milanović Lj, Škrinjar R., Indihar Štemberger M. (2008), Organizational Performance Measures for Business Process Management: a Performance Measurement Guideline, IEEE computer society, ISBN 978-0-7695-3114-4.

[3] Garson, D., Cluster Analysis., CHASS College of Humanities and Social Sciences, available at: https://faculty.chass.ncsu.edu/garson/PA765/index.htm, access in: 15 May 2019.

[4] Kovačič, A., Bosilj Vukšić, V. (2005), Management poslovnih procesov: Prenova in informatizacija poslovanja. Ljubljana: GV Založba d.o.o.

[5] McCormack, K. (2001), Business Process Orientation: do you have it?, Quality Progress, January, 51-58.

[6] McCormack, K. et al. (2009), A global investigation of key turning points in business process maturity, Business Process Management Journal, 15(5), 792-815.

[7] McCormack, K., Johnson, W. C. (2001), Business Process Orientation: Gaining the E-Business Competitive Advantage, New York: St. Lucie Press.

[8] Milanović, Lj. (2011), Understanding Process Performance Measurement Systems, Business Systems Research Journal, 2 (2011), 2; 25-39.

[9] Milanović Glavan, Lj. Konceptualni model sustava za mjerenje procesne uspješnosti poduzeća, Doktorska disertacija, Ekonomski fakultet Zagreb, 2014.

[10] Oliveira, M., Ladeira, M., McCormack, K. (2011), The Supply Chain Process Management Maturity Model – SCPM3, Supply Chain Management - Pathways for Research and Practice, Prof. Dilek Onkal (Ed.), ISBN: 978-953-307-294-4, InTech, Available from: http://www.intechopen.com/books/supply-chain-management-pathways-for-research-and-practice/the-supplychain-process-management-maturity-model-scpm3, access in: 5 May 2019.

[11] Škrinjar R, Bosilj Vukšić V., Indihar Štemberger M. (2010), Adoption of Business Process Orientation Practices: Slovenian and Croatian Survey, Business Systems Research, 1(1-2), 5-20.

[12] Vlahović, N., Milanović Glavan, Lj., Škrinjar, R. (2010), Turning Points in Business Process Orientation Maturity Model: An East European Survey, WSEAS transactions on business and economics, 7 (2010), 1; 22-32.

[13] Zairi M. (1997), Business process management: a boundary less approach to modern competitiveness, Business Process Management Journal, 3(1), 68-80.

CLUSTERING OF ARRIVALS AND ITS IMPACT ON PROCESS SIMULATION

Petra Tomanová
University of Economics, Prague, Department of Econometrics,
W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic
petra.tomanova@vse.cz

Abstract: Precise estimates of times between arrivals is a key aspect of process simulations and subsequently, process optimization and efficient decision making. Times between arrivals typically exhibit strong autocorrelation structure and even after seasonal and diurnal adjustment they still tend to cluster over time. In order to capture the time dependence accurately, we consider Autoregressive Conditional Duration (ACD) models and their generalization to Generalized Autoregressive Score (GAS) models. Once the process of arrivals is estimated, a process assessment can be performed using process simulations. The empirical study of an online bookshop in Prague, Czech Republic shows the importance of the correct treatment of the time dependence of arrivals since the process simulation based on common assumption of exponentially distributed arrivals with a constant rate might lead to suboptimal decisions.

Keywords: Arrival Process, Duration Model, Generalized Autoregressive Score Model, Simulation

1 INTRODUCTION

A common assumption in the queueing theory is that times between arrivals follow the exponential distribution with a constant rate. We relax this assumption by considering a much broader class of distributions. We allow the mean duration (inverse of the rate) to be time-varying in the two following ways. First, seasonal and diurnal patterns in arrivals are investigated by the cubic spline. Second, clustering of arrivals is captured by the generalized autoregressive score model. Specifically, arrivals of Czech online bookshop are analyzed and the following hypotheses are tested: *data sample and simulated sample assuming (i) exponential distribution with constant mean; (ii) generalized gamma distribution with time-varying parameters, are drawn from the same distribution.* The research question is that the utilization of generalized gamma distribution with time-varying parameters leads to a more faithful representation. Further, more accurate predictions of the mean and extreme values of the arrival process are expected.

2 GENERALIZED AUTOREGRESSIVE SCORE MODELS

Generalized autoregressive score (GAS) models [3], also known as dynamic conditional score models [5], capture dynamics of time-varying parameters, $f_i = (f_{i,1}, \dots, f_{i,k})'$, by the autoregressive term and the scaled score of the conditional observation density (or the conditional observation probability mass function in the case of discrete distribution). The time-varying parameters f_i follow the updating scheme

$$f_{i+1} = C + Bf_i + AS(f_i)\nabla(x_i, f_i),$$

where $C = (c_1, \dots, c_k)'$ are the constant parameters, $B = diag(b_1, \dots, b_k)'$ are the autoregressive parameters, $A = diag(a_1, \dots, a_k)'$ are the score parameters and $S(f_i)$ is the scaling function for the score. As the scaling function, we adopt

- unit scaling, $S(f_i) = I$,
- inverse of the Fisher information, $S(f_i) = \mathcal{I}(f_i)^{-1}$,
- square root of inverse of the Fisher information, $S(f_i) = \mathcal{I}(f_i)^{-1/2}$,

where x_i are observed durations, $g = (g_1, \ldots, g_l)$ are static parameters and

$$\mathcal{I}(f_i) = E\left[\nabla(x_i, f_i)\nabla(x_i f_i)|f_i, g\right], \qquad \nabla(x_i, f_i) = \frac{\partial \ln P\left[X_i = x_i|f_i, g\right]}{\partial f_i}.$$

The family of Autoregressive Conditional Duration (ACD) model is $x_i = \mu_i \varepsilon_i$ where different specifications of the conditional mean duration $\mu_i = f_{i,1}$ and the distribution of the error term ε_i result in different models. In this paper, we consider 9 different distributions, three different scaling functions and various combinations of time-varying and static parameters.

3 Data Characteristics

Our data sample is obtained from the same database of an online bookshop as in [6]. The data are collected from June 8, 2018. However, we utilize a much larger data sample. Our raw dataset consists of 5,860 observations collected during seven months, from June 8 to December 28, 2018. First, data cleaning is performed since it is one of the most important aspects of intraday data and duration analysis. The following three steps are utilized.

1. *Extreme values are discarded:* 194 removed observations which duration values are higher than 250 minutes;
2. *Night entries are discarded:* 302 additionally removed observations due to the fact that it is unlikely that the order arrives between midnight and 7:00 AM;
3. *Zero durations are set to a small value:* 81 zero durations are set to 1 minute.

The final dataset consists of 5,364 observations.

Second, durations can be decomposed into deterministic and stochastic components which might result in a better fit. A daily seasonal component in arrivals is investigated by the cubic spline. Seven nodes, which are 3 hours apart (starting from 7 AM, boundaries are included), are used during a day. The estimated daily pattern by cubic spline method is plotted in Figure 1. Since the red line (estimated daily pattern) in Figure 1 is almost flat, we conclude that arrivals do not contain a significant daily seasonal component and thus, a diurnal adjustment does not need to be performed. However, Figure 1 shows that the arrivals tend to cluster over time.

4 Comparison of Different ACD and GAS Models

The same model selection as in [6] is utilized. First, best known specifications of ACD model are estimated with various assumed error term continuous distributions using maximum likelihood estimation procedure. Model performances in terms of Akaike information criterion (AIC) are shown in Table 1. Based on the results, the generalized gamma distribution is the most promising since it fits the data the best.

Second, various GAS models are estimated utilizing continuous distributions, using three scaling functions and various combinations of time-varying and static parameters. Moreover,

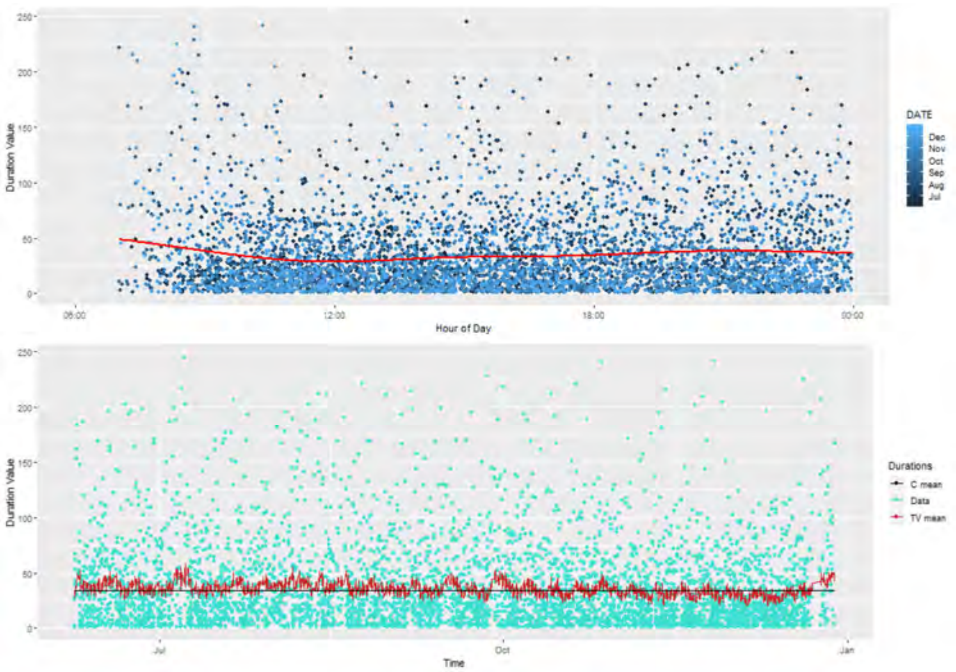


Figure 1: *(upper)* Observed durations (blue dots) and nonparametric estimate of daily pattern for durations (red line). *(lower)* Observed durations (cyan dots), estimated constant (C) mean assuming exponential distribution (black line) and time-varying (TV) mean assuming generalized gamma distribution (red line).

Table 1: AIC values of ACD models (see [1] for details): Standard ACD, Log-ACD (two alternative specifications), Additive and Multiplicative ACD (AMACD), Box-Cox ACD (BACD), Augmented Box-Cox ACD (ABACD), Spline News Impact ACD (SNIACD) and its Log-ACD version (LSNIACD) utilizing various distributions.

Model	Exponential	Weibull	Burr	**Generalized Gamma**	Generalized F	q-Weibull
ACD	48,508.92	48,510.89	48,493.71	48,466.27	48,480.72	48,494.31
LACD1	48,516.92	48,518.92	48,497.26	48,476.94	48,472.58	48,497.26
LACD2	48,510.70	48,512.68	48,491.35	48,465.05	48,516.57	48,491.40
AMACD	48,511.10	48,513.94	48,496.74	48,468.55	48,514.42	48,491.65
BACD	48,525.53	48,511.13	48,516.86	48,486.45	48,504.70	48,497.25
ABACD	48,505.72	48,509.91	48,493.18	48,475.80	48,594.59	48,514.68
SNIACD	48,505.14	48,507.10	48,496.60	48,462.73	48,484.40	48,493.83
LSNIACD	48,504.28	48,506.23	48,504.36	48,461.30	48,592.62	48,507.68

discrete distributions are considered as well since Blasques et. al. [2] showed how discrete distributions might be useful in duration modeling. Main results are reported in Table 3. The selected GAS model with generalized gamma distribution and time-varying scale and shape (λ, κ) parameters outperforms substantially considered ACD models since the difference of AIC values for GAS and best performing ACD model is higher than 22 points. Moreover, it fits the data better than considered discrete distributions. The estimated parameters of the best performing GAS model with generalized gamma distribution are shown in Table 2. The data exhibit a strong positive autocorrelation since the b_1 parameter is close to one and it is highly significant.

Table 2: GAS with generalized gamma distribution.

Parameter	c_1	c_2	c_3	b_1	b_2	a_1	a_2
Estimate	0.0375	0.3847	-0.3631	0.9848	0.3944	0.0255	0.0275
Std. Error	0.0263	0.1149	0.0567	0.0105	0.1560	0.0080	0.0061
P-value	0.0772	0.0004	0.0000	0.0000	0.0057	0.0007	0.0000

5 SIMULATION STUDY

Precise estimates of times between arrivals is a key aspect of process simulations and thus process optimization [6]. We argue that incorrect treatment of the time dependence of arrivals in simulations might lead to suboptimal decisions in process optimization. In this section, we compare outputs from process simulations: first assuming exponentially distributed arrivals with a constant rate; and second assuming generalized gamma distribution with time-varying λ and κ and constant parameters given in Table 2.

In the beginning, process mapping has to be performed. The map of processes of Czech online bookshop under the study is depicted in Figure 2. Orders (books) arrive from two sources: website and e-mail. First, the pre-processing is performed (P1 – P4 nodes). Second, books have to be found and bring from one of seven warehouses (W1 – W7 nodes). Third, the order is finalized and prepared for a customer (F1 – F14 nodes).

It turns out that the critical process is *F13 Packaging* since the corresponding queue is not always empty at the end of the day. This happens on Mondays since not only new-day orders have to be satisfied but also orders which arrived during the weekend (since the resources are available only during working days). As an example, Figure 3 shows a number of books in the packaging queue of one simulated month. On Mondays, the number of books in the queue is highest due to weekends. During the Tuesdays, the queue gets smaller and orders are usually satisfied at the end of Tuesdays. However, it sometimes happens that orders are not satisfied at the end of Wednesday (see the third week in Figure 3), nor at the end of Tuesdays which might cause problems. This occurs more often when the time-varying mean is assumed than when the constant mean is assumed. This is due to the fact that the clustering of arrivals increases the probability that we observe weeks with an extreme number of orders.

Using the Kolmogorov-Smirnov test we reject the null hypothesis that our data sample and simulated sample assuming exponential distribution with constant mean (red density in Figure 3) are drawn from the same distribution using any reasonable significance level (e.g. 1 – 10 %). On the other hand, we cannot reject the null hypothesis that our data sample and simulated sample assuming generalized gamma distribution with time-varying parameters (blue density in Figure 3) are drawn from the same distribution using any reasonable significance level (e.g. 1 – 10 %). Results show that the GAS model with generalized gamma distribution fits the data much better and that the exponential distribution with constant rate is not sufficient. The wrong choice of a distribution might cause problems especially when someone wishes to investigate critical processes.

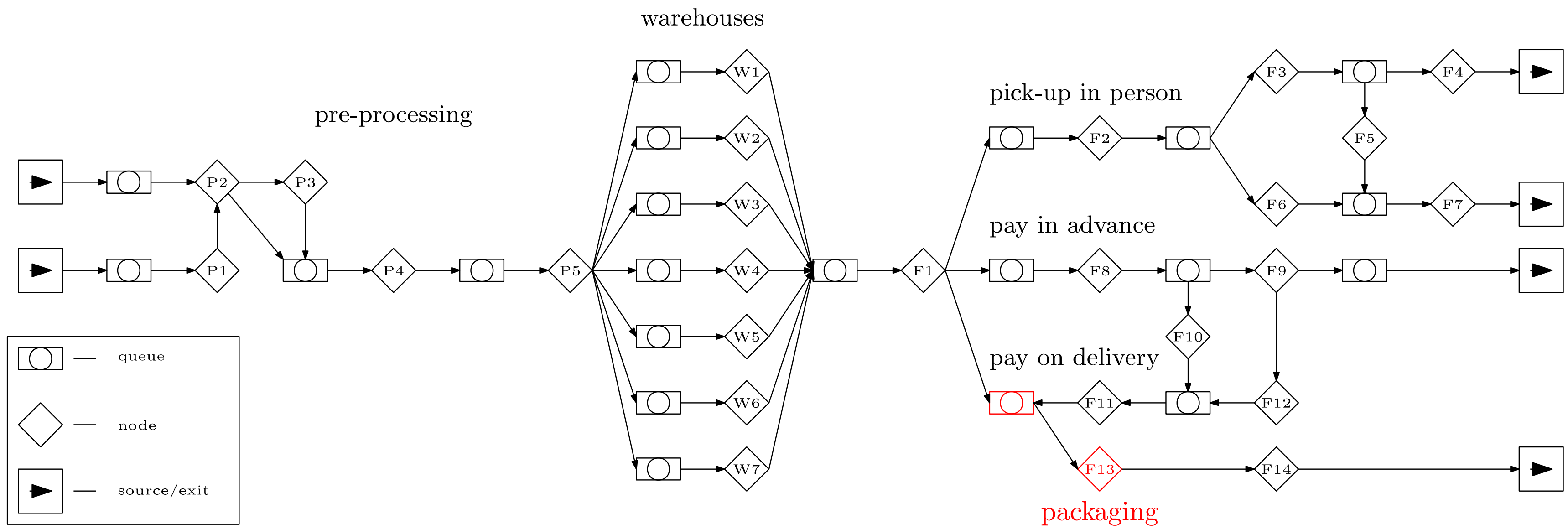


Figure 2: Process mapping.

Distribution	Time-varying parameter	Scaling function	AIC
Poisson	μ	I	186,097.00
Neg. Binomial	μ	I	48,767.83
	μ	$\mathcal{I}(f_i)^{-1/2}$	48,655.35
	μ, α	$\mathcal{I}(f_i)^{-1/2}$	48,652.98
Exponential	μ	I	48,621.76
Gamma	λ	I	48,624.74
	λ	$\mathcal{I}(f_i)^{-1}$	48,624.57
	λ, κ	I	48,625.64
	λ, κ	$\mathcal{I}(f_i)^{-1}$	48,526.52
Generalized Gamma	λ	I	48,459.56
	λ	$\mathcal{I}(f_i)^{-1}$	48,460.53
	λ	$\mathcal{I}(f_i)^{-1/2}$	48,460.51
	λ, κ	I	48,438.81
	λ, κ	$\mathcal{I}(f_i)^{-1}$	48,460.93
	λ, κ	$\mathcal{I}(f_i)^{-1/2}$	48,464.00
	λ, κ, γ	I	48,573.26
	λ, κ, γ	$\mathcal{I}(f_i)^{-1}$	48,456.34
	λ, κ, γ	$\mathcal{I}(f_i)^{-1/2}$	48,544.54

Table 3: Performance of GAS models utilizing different distributions, time-varying parameters and scaling functions.

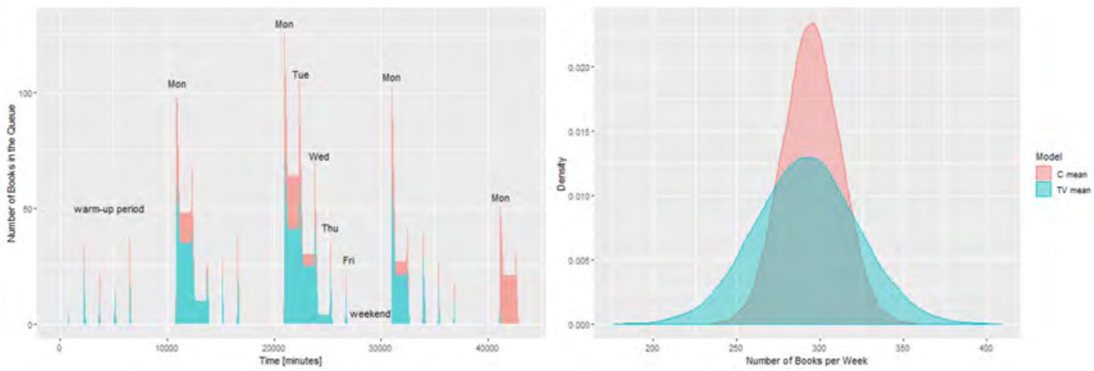


Figure 3: *(left)* A number of books in the packaging queue of one simulated month. *(right)* Number of books per week – distribution based on 10,000,000 simulated observations assuming exponential distribution with constant (C) mean and generalized gamma distribution with time-varying (TV) parameters.

6 CONCLUSION

In this paper, we capture the time dependence in arrivals by the Generalized Autoregressive Score model and show that this approach leads to a more faithful representation of the mean and extreme values of the arrival process. The paper of [6] is extended by analyzing the whole complex system of firm processes. Moreover, a much larger data set was utilized which allowed us to investigate the consequences of the wrong choice of distribution more thoroughly and to show how the common approach of exponentially distributed arrivals with a constant rate might lead to suboptimal decisions since it underestimates the probability of extreme values.

7 Acknowledgements

This work was supported by the IGA of University of Economics, Prague under Grant F4/53/2019.

References

[1] Belfrage, M. (2015). R package ACDm: tools for autoregressive conditional duration model. Available at: `https://cran.r-project.org/web/packages/ACDm/ACDm.pdf`

[2] Blasques, F. and Holý V. and Tomanová, P. (2018). Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros. arXiv preprint arXiv:1812.07318.

[3] Creal, D. and Koopman, S. J. and Lucas, A. (2008). Generalized autoregressive score models with applications. Journal of Applied Econometrics, 28(5): 777–795.

[4] Engle, R. F. and Russell, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica, 66(5): 1127–1162.

[5] Harvey, A. C. (2013). Dynamic models for volatility and heavy tails: with applications to financial and economic time series. Cambridge University Press.

[6] Tomanová, P. (2018). Measuring Intensity of Order Arrivals and Process Quality Assessment of an Online Bookshop: A Case Study from the Czech Republic. Management 2018. Prešov, 768–773.

PREDICTING FUTURE MARKETS FOR PERSONAL SERVICE ROBOTS

Josipa Višić
Faculty of Economics, Business and Tourism/Department of Economics
Cvite Fiskovića 5, 21 000 Split, Croatia
E-mail: josipa.visic@efst.hr

Abstract: Robot market is growing rapidly and sales of service robots are increasing. An important niche of service robots are service robots for personal/domestic use and the main aim of the paper is to estimate which OECD countries are most likely to create the future demand for such sort of products. Cluster analysis has shown that Australia, Austria, Denmark, Germany, Iceland, Ireland, Netherlands, Norway, Switzerland and United States are more likely to become a lucrative markets while remaining 24 countries are grouped in two other clusters depending on variables which served as indicators of probable future consumption of these products.

Keywords: robotization, personal service robots, cluster analysis, market predictions, OECD

1 INTRODUCTION

Strive to produce more, to beat the competition has strongly encouraged development of robots since they are able to produce more, faster, often safer and more precisely than a human worker. In other words, race for efficiency in a form of robotization is changing the way we work and the way we live. Implications of robotization are various. Generally speaking, economic effects can be divided on micro and macroeconomic effects. Regarding microeconomic repercussions of robotization, robots can be observed both as inputs and outputs. When regarded as inputs, issue of human workers being replaced with robots is the most important one and it raises many questions since it deals with the impact of robots on a labour market. Further, robotization affects many organizational aspects of a company and causes many additional problems that managers might face [7]. For instance, if a firm decides to use robots in a production process, it is up to managers to decide on the most appropriate employment policy since workers need to be trained to use these robots. Therefore, it is necessary to choose the most efficient tactic and decide whether to train existing employees, to hire trained ones or to combine tactics. Also, there are other aspects of worker-robot interaction that need to be addressed along with the issues of measuring productivity, especially in terms of developed artificial intelligence. A cost-benefit analysis necessary to decide on adopting this sort of technical progress also becomes more complex and needs to include characteristics of a respective market [5]. When it comes to robots as outputs, as with any other product/service, it is important to predict consumers' needs and preferences but when it comes to robots, significance of economic implications of psychology is increasing. In that manner, producers for example, need to perceive the relevance of consumers' willingness to adopt to a technical solution for their problems as well as the importance of including potential users in the process of designing so that robotic technologies become more socially robust [13]. Regarding macroeconomic implications of robotization, the most analysed aspect is its impact on labour market, i.e. impact on employment and wages. Along with this issue, the question of inequality among countries also captures both scientific and political attention [e.g. 12, 3].

The size of a robot industry is raising and along with a significant growth in absolute vales of revenues in the industry, there is also a raise in the share of non-industrial robots. Non-industrial robots represent 70% of the $39.3 billion robotics market globally in 2017, growing from a 64% share in 2016 [11]. When it comes to non-industrial robots, often classified as service robots, they can be divided in two groups – personal/domestic robots and professional service robots. Market for personal service robots is developing fast and it is projected that sales of all types of robots for domestic tasks (such as vacuum cleaning, lawn mowing, window cleaning etc.) could reach 39.5 million units in the period 2019-2021, with an estimated value of US$ 11.1bn [4]. These numbers do not include projections for all types of entertainment and leisure robots or robots for elderly and handicap assistance. Therefore, when analysing possible future trends in non-industrial robot market it is obvious that this market niche could be very lucrative.

As previously stated, economic implications of robotization are manifold. However, this paper focuses on a microeconomic aspect, or more precisely, robots are perceived as outputs. In that manner, aim of this paper is to set a starting point for future market analysis of demand for personal service robots. The idea is to detect countries that are more likely to have higher demand for such products. Namely, studies has shown that attitudes towards robots are strongly influenced by culture and even religion and these differences between consumers might be seen through the prism of consumers' location i.e. their countries [6, 2]. For instance, Japanese consumers generally tend to have positive attitude towards robots and when compared to US and German consumers have the highest preferences when it comes to necessity of faces for social robots [8]. Although this might seem as irrelevant fact, when it comes to development costs of robots it is obvious that the awareness of differences between consumers' preferences might decrease development costs if we combine information about consumers' preferences and estimates on future markets for service robots. In that sense, this paper seeks to detect which countries are most likely to become lucrative markets for personal service robots. Having in mind scarce data on consumers of service robots, several variables have been used in order to cluster OECD countries in three groups. Respective variables have been selected in a way that these three clusters differentiate counties so that possible investors (or policy makers) can more easily group countries according to their market potential.

The paper is dived into three segments. After the introduction, method and results will be presented in Section 2, while the paper ends with the conclusion containing shortcomings of the analysis and suggestions for future work on the respective theme.

2 METHOD AND RESULTS

2.1 Data description and the method

Aim of the empirical part of the paper is to detect similar groups of OECD countries. Therefore, a cluster analysis (K-Means Cluster) using IBM Statistics SPSS 23 have been performed. Data have been collected from OECD [9] and Turkey has been left out due to lack of data, while Luxembourg has been left out as an outlier regarding data for GDP per capita. By selecting the newest possible data, following six variables have been used to cluster selected countries in three groups:

- Unemployment - rate of unemployment as percentage of Labour Force in 2018
- GERD - GERD as a percentage of GDP in 2017, where GERD is gross domestic expenditure on research and development (R&D)
- GDP - gross domestic product (expenditure approach) per head, current prices, current PPPs in 2018, USD
- GDP productivity - GDP per hour worked as a measure of productivity, current prices, current PPPs, USD in 2017
- Wages - average annual wages, constant prices at 2017 USD PPPs in 2017

- Internet - ICT Access and Usage by Households and Individuals, percentage of households with Internet access at home in 2017.

Rate of unemployment has been selected in order to help predict consumers' behaviour since people that work are more likely to be sufficiently skilled in the use of new technologies to do their jobs [1]. Therefore, it is assumed that they are more likely to use personal service robots as well. If a country invests more money on research and development it is more likely that its residents will be acquainted with new technologies, and are more likely to spend their money on robots. Further, GDP, GDP productivity and wages are used as a wealth indicator for country's residents since personal service robots are luxurious goods and are more likely to be bought when a person is a resident of a rich, productive country with high wages. Percentage of households with Internet access at home has been used due to high correlation between the use of Internet and the use of robots at home. Since data from the European Commission show that the more often a person uses the Internet, the more likely he/she has used a robot [1], it is reasonable to expect that this will reflect on future behaviour regarding purchases of personal service robots.

K-Means Cluster has been selected since the idea was to group OECD countries in three groups, clusters in order to detect countries that are most likely to be lucrative market for personal service robots and countries that should not be in the focus of attention at the moment. Third group, the middle cluster, includes countries that should be further observed in the future since they are potentially interesting market for these products.

2.2 Results

Descriptive statistics on respective data, presented in the Table 1, show that OECD countries are highly heterogeneous group of countries when selected variables are observed. Therefore, data on cluster centers (see Tab. 2) reveal characteristics of an each cluster where countries belonging to the cluster 1 are those with the least desirable economic indicators (high unemployment rates, low expenditure on R&D, low GDP, GDP productivity and wages and low Internet access rate). Third cluster is the most promising one, since it is the one containing countries with strong, healthy economies whose residents are most likely to become buyers of personal service robots due to their favourable financial conditions and familiarity of technology (estimated through GERD levels and Internet access at home).

Table 1: Descriptive statistics

Variable (unit)	***N***	***Minimum***	***Maximum***	***Mean***	***Std. Deviation***
Unemployment (%)	34	2,24	19,29	5,86	3,46
GERD (%)	34	0,37	4,55	1,99	1,08
GDP (USD)	34	20227,34	83945,92	45404,92	13414,87
GDPproductivity (USD)	34	21,56	99,54	54,99	17,93
Internet (%)	34	50,92	99,51	85,15	9,81
Wages (USD)	34	15313,94	62282,57	39133,83	12656,23
Valid N	34				

Table 2: Final cluster centers

Variable	Cluster		
	1	**2**	**3**
Unemployment	6,53	6,47	4,32
GERD	1,01	2,46	2,46
GDP	31741,25	44615,90	61460,67
GDPproductivity	36,85	55,60	74,13
Internet	77,91	86,10	91,88
Wages	24022,90	40877,12	53489,59

When differences between clusters are observed, it is evident that the differences between clusters 1 and 2 on one side, and clusters 2 and 3 on the other side are almost identical (see Tab. 3). These data support the choice of three clusters to sort OECD countries in order to detect which ones are more likely to become a lucrative market for personal service robots. If differences between clusters were less evident, countries could have been divided into two groups. Cluster membership and the Euclidean distance between each case (country) and the cluster center used to classify the case is presented in Tab. 4. Presented distances, as measures of similarity, reveal that clusters are not uniform. Namely, Latvia in cluster 1 is significantly closer to cluster centre (distance is 912,11) than e.g. Mexico (with distance 14436,66). Therefore, more detailed analysis in terms of additional variables should be made in order to more precisely estimate demand for personal service robots. However, since respective market is still developing, data on number and value of purchased personal service robots on a country level are not available.

Table 3: Distance between final cluster centers

Cluster	***1***	***2***	***3***
1		21209,00	41851,30
2	21209,00		21043,32
3	41851,30	21043,32	

Table 4: Cluster membership

Clusters					
1		***2***		***3***	
Country	*Distance*	*Country*	*Distance*	*Country*	*Distance*
Chile	6771,68	Belgium	10783,22	Australia	8550,06
Czech Republic	8328,29	Canada	7426,66	Austria	6064,59
Estonia	4066,60	Finland	4543,70	Denmark	5679,76
Greece	2766,69	France	3113,89	Germany	9238,38
Hungary	1639,05	Israel	6951,27	Iceland	9083,89
Latvia	912,11	Italy	4924,32	Ireland	23230,45
Lithuania	4068,65	Japan	1237,23	Netherlands	4731,77
Mexico	14436,66	Korea	7040,64	Norway	3255,97
Poland	3033,46	New Zealand	3337,08	Switzerland	11564,49
Portugal	2692,79	Slovenia	8458,18	United States	7141,86
Slovak Republic	2612,22	Spain	4452,42		
		Sweden	8688,54		
		United Kingdom	3291,93		
Number of cases in each cluster	**11**		**13**		**10**

When these data are compared with data on the current use of a robot at home from 2017 [1], predictions for Austria and Denmark (belonging to the cluster 3) and Italy, Slovenia and Spain (belonging to cluster 2) are in a line with expectations since these countries already have the highest share of residents that have at least once used a robot at home. However, results for Czech Republic and Slovak Republic are somewhat surprising since their residents use robots at home very often, when compared to other EU countries, but here belong to cluster 1 indicating that these countries should not be considered as a significant market for personal service robots. A possible explanation may be found in the fact that this analysis does not capture current growth of these economies, hence it underestimates their market potential. Since, to the author's best knowledge, there are no existing data on the use of robots at home for all OECD countries, a similar comparison cannot be made for all the other countries.

Determining which countries are going to be a lucrative market for personal service robots is still very complex, since data on past purchases of these types of products are scarce and estimates should be made with caution. For instance, according to the data on the number of installed industrial robots in manufacturing industry [10], countries with the highest number of industrial robots are not, according to here obtained results, those that should be in the focus when we discuss personal service robots. Namely, Portugal and Hungary are among five countries with the highest industrial robot density but at the same time they appear in the cluster 1 indicating that, at the moment, these countries are not a potentially fruitful market for personal service robots. However, information on industrial robot density are not redundant in this context since the United Kingdom (with the highest score), Norway and New Zealand appear as important markets for robots in both analysis.

3 CONCLUSION

Robots are a part of the world we are living in, and eventually their importance in our personal life will increase. In that manner, it is important to estimate future demand for this type of products since previous studies have shown that: a) consumers have different preferences regarding usage of personal robots and their characteristics and b) these preferences are determined by their culture. Therefore, having in mind that development of personal service robots generates high costs, it is opportune to be able to estimate which countries are most likely to become lucrative markets.

Here obtained results group OECD countries in three clusters where cluster 3 contains the most developed countries among 34 selected countries. These countries should be in the focus of personal service robots producers and more detailed characteristics of their residents should be further analysed while developing new products of this sort. Countries in cluster 2 should also be observed with attention, since they are more similar to cluster 3 than to cluster 1. Consumers' features from countries in cluster 1 are not to be ignored but, having in mind analysed data, these markets should not be considered as a priority, at least not at the moment.

Presented results serve as a starting point for further analysis of the respective theme. Namely, the idea was to detect similar group of countries in order to estimate which markets could be more promising so development of personal service robots can be adjusted to their consumers, consequently making costs lower. However, inferences should be made with caution since e.g. Germany appears to be a very promising market while at the same time existing research indicates that people in Germany feel strong resistance to the presence of robots in their households [8]. This could be explained with the fact that consumers' preferences, when it comes to modern technologies, are insufficiently investigated and therefore this information may not reflect the entire German market. Further, Czech Republic and Slovak Republic here belong to cluster 1, while at the same time their residents belong to a group of EU countries whose residents have very high portion of population using robots at home [1]. In that manner, further analysis of the respective theme should be broadened in a way that it should include additional explanatory variables such as countries' growth rates, educational level of the population, data on the frequency of online purchases and usage of online public services. Further, the sample should be broadened to include non-OECD countries as well. Namely, large markets with growing purchasing power of the middle class, such as Brazil, Russia, India and China are becoming lucrative markets for a variety of household products.

Having in mind that the way we organize our private life is changing, personal service robot market is very likely to grow progressively. Therefore, database on consumers and these products will expand and will provide a fruitful path for future, more complex market predictions, useful both to producers of products and policy makers in respective countries.

References

[1] Attitudes towards the impact of digitisation and automation on daily life. Special Eurobarometer 460 – March 2017. European commission. http://ec.europa.eu/commfrontoffice/publicopinion [Accessed 15/3/2019].

[2] Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., Kennsuke, K. 2005. Cultural Differences in Attitudes Towards Robots. Proceedings of the AISB Symposium on Robot Companions: Hard Problems And Open Challenges In Human-Robot Interaction.

[3] DeCanio, S. J. 2016. Robots and humans–complements or substitutes? Journal of Macroeconomics, 49: 280–291.

[4] Executive Summary World Robotics 2018 Service Robots. International Federation of Robotics.https://ifr.org/downloads/press2018/Executive_Summary_WR_Service_Robots_2018.pdf [Accessed 15/4/2019].

[5] Ivanov, S. H., Webster, C. 2017. Adoption of robots, artificial intelligence and service automation by travel, tourism and hospitality companies – A cost-benefit analysis. Conference: International Scientific Conference "Contemporary tourism – traditions and innovations".

[6] Mohammad, Y., Nishida, T. 2015. Cultural Difference in Back-Imitation's Effect on the Perception of Robot's Imitative Performance. In Koh, J.T.K.V. et al. (Eds.). Cultural Robotics. Springer International Publishing.

[7] Moniz, A. B., Krings, B.-A. 2016. Robots Working with Humans or Humans Working with Robots? Searching for Social Dimensions in New Human-Robot Interaction in Industry. Societies, 6(3): 1–23.

[8] Nitto, H., Taniyama D., Inagaki, H. 2017. Social Acceptance and Impact of Robots and Artificial Intelligence – Findings of Survey in Japan, the U.S. and Germany. Nomura Research Institute. NRI Papers, No. 211.

[9] OECD Statistics. 2019. OECD. https://stats.oecd.org/ [Accessed 15/4/2019].

[10] Robot density rises globally. The International Federation of Robotics. https://ifr.org/ifr-press-releases/news/robot-density-rises-globally [Accessed 13/4/2019].

[11] Robotics Market Forecasts. 2019. Tractica. https://www.tractica.com/research/robotics-market-forecasts/ [Accessed 15/5/2019].

[12] Sachs, J., Kotlikoff, L. 2012. Smart Machines and Long-Term Misery. National Bureau of Economic Research. Working Paper No. 18629.

[13] Šabanović, S. 2010. Robots in society, society in robots: Mutual Shaping of Society and Technology as a Framework for Social Robot Design. International Journal of Social Robotics, 2(4): 439–450.

STATISTICAL ANALYSIS OF THE PUBLIC OPINION SURVEY ON FREE SUNDAY

Bože Vuleta
Franciscan Institute for the Culture of Peace,
Poljudsko šetalište 24, 21000 Split, Croatia
Phone: ++385989197 132; Fax: ++38521381257
E-mail: fbvuleta@gmail.com

Elza Jurun
University of Split/Faculty of Economics, Business and Tourism
21000 Split, Cvite Fiskovića 5, Croatia
Phone: ++ 38521430648; Fax: ++38521430701
E-mail: elza@efst.hr

Nada Ratković
University of Split/Faculty of Economics, Business and Tourism
21000 Split, Cvite Fiskovića 5, Croatia
Phone: ++ 38521824171; Fax: ++38521430701
E-mail: nada.ratkovic@efst.hr

Abstract: In the focus of this paper is comprehensive statistical analysis of a public opinion survey on the issue of non-working Sunday. Case study is the survey carried out in Croatia in October 2017. As a member of the European Sunday Alliance, Croatia has been the first EU country to promote free Sunday as one of the measures of active demographic policy. Historical background, European legislation of free Sunday and good practices of its implementation have also been considered as well as its consequences on various spheres of socio-economic life, especially on family well-being.

Keywords: free Sunday, European legislation, public opinion survey, post stratification analysis, hypothesis testing

1 INTRODUCTION

This work has emerged as a practical need to launch further research based on the conclusions of the conference entitled „Free Sunday and Dignified Working Time in Europe: What is the Way Forward?". The conference was organized by the European Sunday Alliance at the European Parliament in Brussels, February 2019. Why the phenomenon of the non-working Sunday as a historical legacy is so important for living in modern Europe?

It is in the nature of man to be free; he is called to be brotherly equal in dignity with all of his brothers and sisters of the human species. These are the leitmotifs of the French Revolution taken from the Gospel [2]. The ideas are very powerful, the foundations of the development of our civilization. However, in a number of European countries, especially in Croatia, life situation of some workers is closer to the state of slavery than the state of freedom. In Croatia people who work in retail stores are forced to work overtime, to work on holidays and Sundays. Formally, everything looks legitimate but in reality they are not appropriately paid for overtime work nor for their work on Sundays and holidays.

Not only does the situation have negative consequences on their families, but also on the society as a whole. The damages are even worse since the most of the respective employees are women. Far too exploited and more absent than present in their families, the women cannot effectively raise up their children, which is one of the generators of violent behaviour among children and young adolescents, and consequently in the entire society. Hence, it concerns all the citizens and the common freedom.

Moreover, what is happening to the employees in the retail trade today could soon happen to everyone else. The investment capital knows no borders. It brings impoverishment, intensifies delinquency, destroys public expenditures, increases expenditures and reduces the quality of public health care, etc. Briefly, the state should first protect its citizens since that is it's primarily task.

Hence, it is necessary to look at the social significance of Sunday. Joining European Sunday Alliance Croatia has enriched the scope of arguments for work-free Sunday by extending its relevance onto demographic state and trends. Croatian Sunday Alliance has been first to propose free Sunday as one of the measures of active demographic policy. By fifty years ago, the main reasons for divorce were mostly behavioural and fairly concrete, such as alcoholism and neglect by a spouse [1]. However, in the last two decades reasons for divorce have become more of affective and abstract nature as feeling unloved or incompatible in the areas of life values and interests. According to the results of some researches, the probability of divorce increases with increased number of working hours, particularly by women [5]. Since the number of divorced marriages in a quarter of the counties of the Republic of Croatia has exceeded the number of new marriages on an annual basis, it is no wonder that free Sunday is for the first time perceived and promoted in Croatia as one of the main measures of active demographic policy.

This paper is organized as follows: after the introduction, historical background and legislation in some European countries are dealt with. The topic of the third part is the case study of Croatia. The statistical analysis of public opinion survey results (October 2017) is carried out. The final section contains conclusion remarks. Consulted literature is listed at the end of the paper.

2 HISTORICAL BACKGROUND AND LEGISLATION IN SOME EUROPEAN COUNTRIES

Throughout the centuries, the justification of a day of celebration, a non-working day, is sociologically infused by Jewish fundamental social attitude that man needs rest. Moreover, Jewish tradition has always emphasized that this vacation belongs to every human being, even to animals. (Exodus 20,8; 23,12; 34,21) [7]. Saturday in Israel was originally a "socio-ethical institution, i.e. exclusive day of the rest and (...) only after the Babylonian exile it become the day of special Divine Worship "[8].

Later on the Roman Empire launched legalization of a free day, and instead of Sabbath, Sunday become the free day, since it is the day of Christ's resurrection, (after the Milan Edict, 313), and afterwards non-working public holiday (321). Modernization, secularization, industrialization and urbanization have brought up the question of free Sunday into the centre again. "Machine" and "profit" imposed working hours, and working conditions become ever worse. On the basis of the scientific analysis of that time, the benefit of workers' vacations was recognized and the free Sunday was reinstated in the legal regulations. Thus, the implementation of free Sunday in the legislations of the western countries at the end of 19th and the beginning of 20th century was not motivated by workers' needs. The benefit of the workers' rest had already been attributed to "human, social and economic reasons" [8]. Thereby non-working Sunday was reintroduced in Switzerland, 1877, in Germany, 1891, in France, 1906 and in Italy, 1907. It is an interesting fact that protestant countries, such as England and the United States, have not abolished free Sunday [6].

Due to the character of this paper, it is possible to mention just a concise description of Sunday work models that are currently valid in certain European countries. In the EU, about 30% of employees regularly work on Sundays. In Austria, this percentage is 16%, tending to grow, and in Germany 23%. According to *Eurostat* data, Sunday is at most work-day in England and Denmark, while it is at lowest in Spain and Italy. Work on Sundays (in retail trade) is prohibited in Belgium, Denmark, France, Greece, Italy, Norway, Germany, Luxembourg and Austria. However, there are exceptions in all of these countries: work in retail shops in Norway is allowed only three weeks before Christmas from 14:00 to 20:00h; in The Netherlands 12 Sundays per year; in Spain 8 Sundays per year; in Finland just from 12:00h to 21:00h; there are no work restrictions in smaller places and tourist zones in Greece. In Germany, since 2006, working time has been allocated to federal states. So, work on Sunday in Bavaria is prohibited altogether, but in most other federal states it is allowed to work only 4-6 Sundays per year. The Constitutional Court of Germany issued an act, in December 2009, according to which only exceptionally can shops be open on Sundays: activities typical for workdays cannot be passed onto Sunday, and *"pure financial interests of the shop owners cannot be sufficient to work on Sundays"*.

Due to various negative effects of work on Sundays throughout the modern Europe in the recent years, numerous initiatives, associations and mass movements for free Sunday have been established, such as European Sunday Alliance and European Citizens' Initiative for a work-free Sunday in Europe, which urge that non-working Sunday be implemented in European legislation and be valid throughout the EU.

3 CASE STUDY OF CROATIA

As an integral part of the EU, Croatia shares its destiny and all that has been said previously refers to Croatia, too. Croatian Sunday Alliance (CSA) was established in 2017. It is an association of trade unions, academic, social and religious institutions as well as NGO's. Free Sunday is particularly opposed to the competition of the burly capital owners. For decades, they have been misrepresenting realities of the work on Sundays, falsely claiming, for example, that it increases economic activities and employment rate in Croatia. However, no economic theory proves that, the available indicators disclaim it and the data in the following table show far the lowest fiscal turnover in retail trade on Sundays and holidays.

Table 1: Average turnover in fiscalization in G47 Retail trade by day in 2015, 2016 and 2017 in Kuna

Average turnover in fiscalization in G47 Retail trade by day in 2015, in Kuna							
Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Sunday*	*Holidays*
217.353.239	215.138.930	219.655.595	232.188.885	242.538.963	256.457.213	135.767.295	105.583.320
Average Annual turnover in 2015							79.348.724
Average Fiscal Daily G47 in 2015							241.263.673
Average turnover in fiscalization in G47 Retail trade by day in 2016 in Kuna							
Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Sunday*	*Holidays*
221.929.611	221.976.522	220.493.272	248.633.722	259.816.483	269.003.246	141.836.200	110.390.856
Average Annual turnover in 2016							82.589.145
Average Fiscal Daily G47 in 2016							226.271.631
Average turnover in fiscalization in G47 Retail trade by day in 2017 in Kuna							
Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Sunday*	*Holidays*
235.914.507	234.048.728	238.983.062	256.322.650	274.511.389	288.579.629	156.616.216	118.345.147
Average Annual turnover in 2017							88.061.240
Average Fiscal Daily G47 in 2017							241.263.673

Source: ISPU – Ministry of finance RH – Tax Department Zagreb, April, 2018.

CSA significantly contributed to the promotion of the values of free Sunday. On behalf of it, Franciscan Institute for the Culture of Peace from Split conducted a survey, in October 2017, in order to find out prevailing public attitudes to the value of non-working Sunday, which authors have used as the case study of Croatia.

Public opinion survey was carried out by a specialized agency *Ipsos Public Affairs*. It was conducted by telephone interviews. Since the views of the Croatian population aged 18 and on were studied, a two-step stratified random sample was used with the following stages:

- By random selection of place of residence within the stratum - the stratums are defined by region (6) and size of the place of residence (4 categories).
- The household was chosen by random selection of the phone number.
- The respondent was chosen by quota.

Post stratification has been created on the basis of gender, age, size of residence, regions and education. The final realized sample consists of 603 respondents.

From the results of this comprehensive research only some basic attitudes and answers to the relevant issues are presented in this paper. The first one is presented on the Figure 1 with the answers to the question: "How often you, if ever, go to the next places on Sundays?"

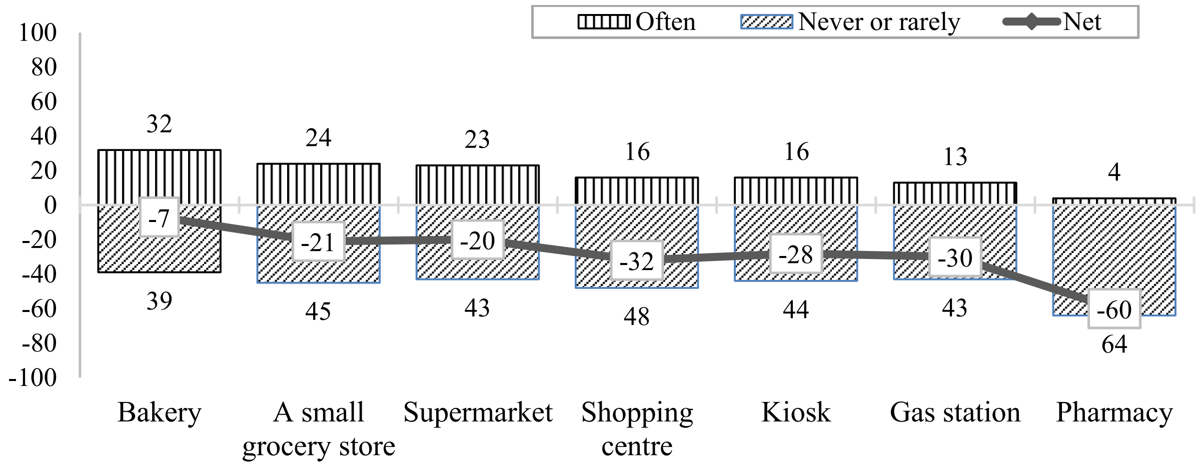


Figure 1: Frequency of visiting to certain places on Sundays
Source: Authors according Ipsos Public Affairs 2017

The random stratified sample enabled authors to statistically analyse the answers to the aforementioned question for some demographic features. The frequency of going to the respective places has been statistically analysed in relation to the answer: "I rarely or never visit them on Sundays". As far as methodology is concerned, testing of hypothesis about the difference in the proportion of the two statistical populations has been applied. The usual level of significance of the 5% test was used. The analysis has been carried out for all of the places visited on Sundays and the results are very similar for each of them. Therefore, more precise results of visiting the shopping centres on Sundays on the basis of demographic categories have been presented below.

Statistical analysis has shown that there is no statistically significant difference between the frequency of going to a shopping centre on Sundays between men and women. The answer "I rarely or never go shopping on Sundays" gave 48% of male and 50% of female respondents. The same answer has been given by 47% of the urban and 52% of the rural population. It shows that neither their responses statistically differ significantly.

Analysis of the profile of the respondents according to the level of education reveals that 45% of them who have elementary school education, 51% middle and 46% high school or college education "rarely or never" go shopping on Sundays. The respondents up to 30 years visit shopping centres on Sundays more often (41% rarely or never) while those between 45 and 60 years old do it the rarest (59% rarely or never). »Rarely or never« go shopping on Sundays 44% of the respondents between 30 and 44 years old and those over 60 do not differ much (45%). The analysis done according to the particular regions of Croatia shows that 56% of the population of Zagreb and its surroundings go shopping on Sundays »rarely or never«; in Slavonija 41%; in Kordun and Lika 57%; in North Croatia 43%; in Primorje and Istra 46% and in Dalmatia 49%.

Even more than two thirds (67.5%) of respondents claim that not-working on Sundays within their regular business is important or exceptionally important for them. What they particularly think about work on Sundays can be seen distinctly from the Figure 2.

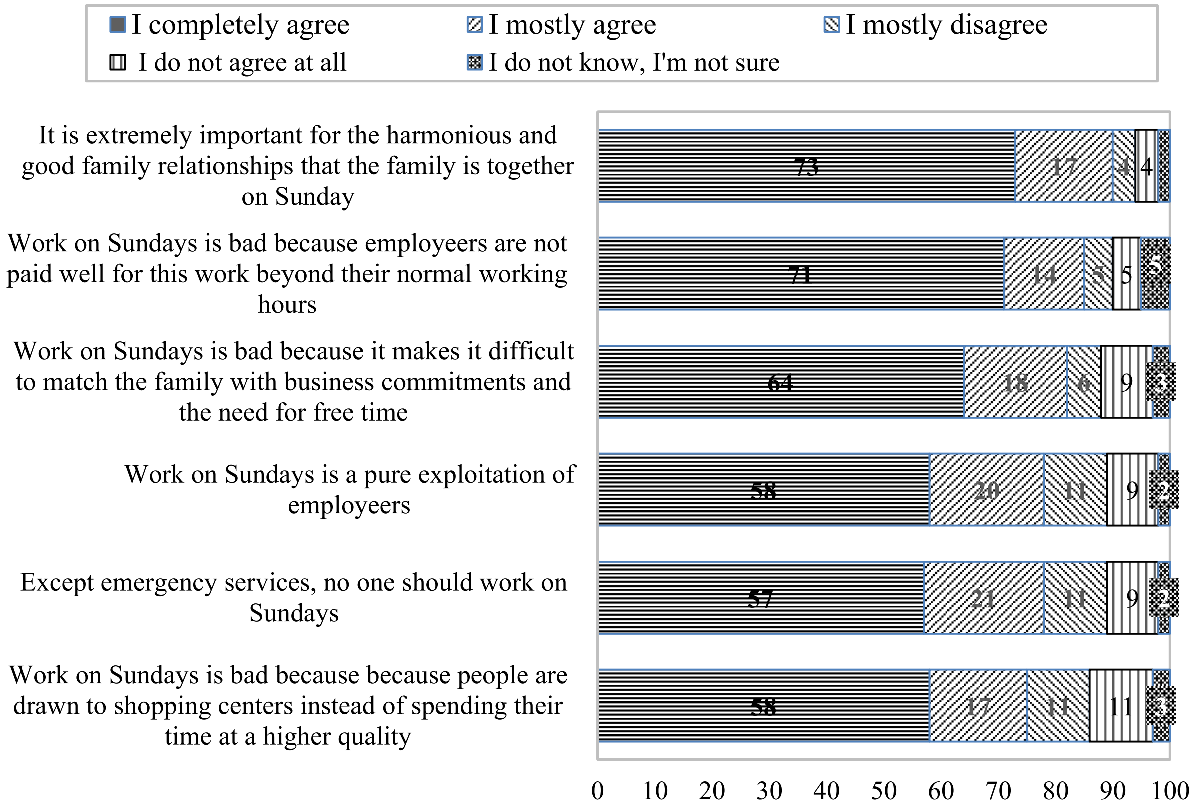


Figure 2: Attitudes about work on Sundays
Source: Authors according Ipsos Public Affairs 2017

Figure 3 shows answers about personal support of maximum limitation of work on Sundays.

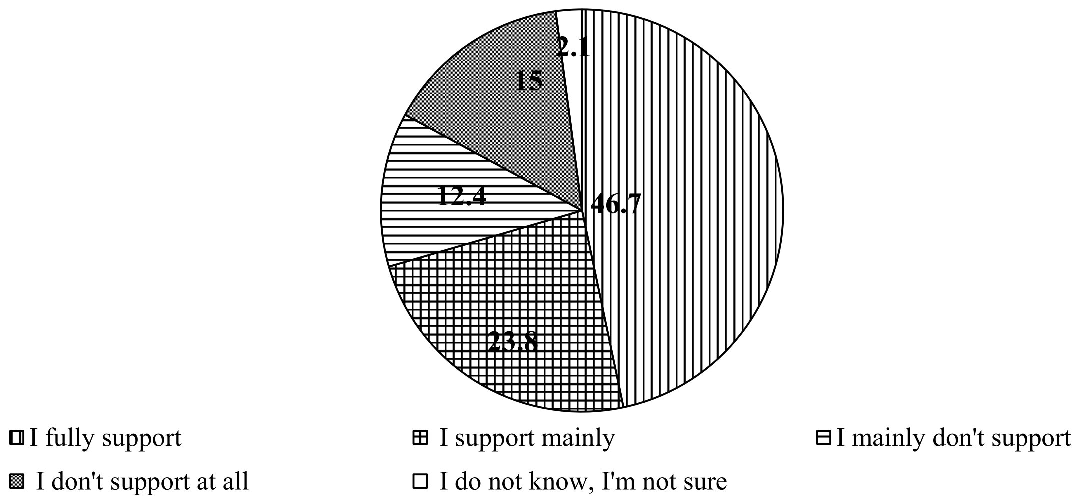


Figure 3: Support of limitation of work on Sundays
Source: Authors according Ipsos Public Affairs 2017

4 CONCLUSION

This paper is meant as a contribution to promoting non-working Sunday as one of the vital basis for human well-being in general. After a brief review of historical background and European legislation, some practices in various European countries are presented. Different influences and consequences of (non-)working Sunday on social, economic, psychological, health, family well-being and demographic trends of social communities in contemporary Europe have been considered. Negative consequences of the fact that Croatia has no legal limitation of the opening time of retail stores is evident in many spheres of socio-economic life and especially intense in family life. To intensify pressure to change the practice, different initiatives joined and founded Croatian Sunday Alliance, 2017, (CSA). It is an association of trade unions, academic, social and religious institutions as well as civic organizations. CSA as a member of European Sunday Alliance has particularly contributed to raising awareness of free Sunday being one of the measures of active demographic policy. On behalf of CSA, Franciscan Institute for the Culture of Peace from Split conducted a survey in order to find out prevailing public attitudes to the issue of (non-)working Sunday in October 2017, which authors have used as their case study. From the results of this survey the authors have provided only some of the most relevant attitudes for the overall goal of this paper. Randomly stratified sample enabled the authors to statistically analyse the outcomes for some demographic features. Methodologically, hypothesis testing of the difference in the proportions of the two statistical populations has been carried out. It has been implemented in the analyses of all the questions in this survey. The fact that even more than two thirds (67.5%) of respondents conceder important or exceptionally important not-working on Sundays within their regular business could be taken out as a general conclusion remark. It should also be pointed out that 70,5% of respondents support maximum limitation of work on Sundays. This paper is representing only a part of the on-going research of the free Sunday phenomenon. The authors have presented just the beginnings of an extensive statistical analysis of public opinion. Therefore, in order to promote non-working Sunday as a vital basis of human well-being, further research by the same authors can be expected soon.

References

[1] Chang, J. (2003). *Self-Reported Reasons for Divorce and Correlates of Psychological Well-Being Among Divorced Korean Immigrant Women.* (Vol.40, No.1-2). (111-128). Journal of Divorce & Remarriage, Retrieved 20.3.2019. from https://www.tandfonline.com/

[2] Črpić, G, Džolan, M. (2014). *Slobodna nedjelja: kultura u nestajanju?* (1). Zagreb: Centar za promicanje socijalnog nauka crkve, Kršćanska sadašnjost, Franjevački institut za Kulturu mira

[3] European Foundation for the Improvement of Living and Working Conditions – Eurofound (2016). Sixth European Working Conditions Survey – Overview report, Publications Office of the European Union, Luxembourg.

[4] Franjevački institut za kulturu mira (2017). *Stavovi i iskustva građana vezana uz rad nedjeljom.* Socijalna istraživanja i korporativna reputacija: Ipsos Public Affairs, Retrieved 22.3.2019.from http://franjevacki-institut.hr/sadrzaj/pdf/2017-11-23-08-34-1312-.pdf

[5] Lyonette, C., Clark, M.(2009). *Unsocial Hours: Unsocial Families? Working Time and Family Wellbeing.* (No.1106460). Cambridge: Relationships Foundation.

[6] Sabotič, I. (2005). O dokolici, slobodnom vremenu i nedjelji u Europi i Hrvatskoj na prijelazu iz XIX. u XX stoljeće. In Baloban, S, Črpić, G. (2014). *Kultura nedjelje i dostojanstvo radnika* (1) (ur.). Zagreb: Centar za promicanje socijalnog nauka crkve, Kršćanska sadašnjost.

[7] Spicq, C., Grilot, P. (1993). *Subota, Riječnik biblijske teologije.* Zagreb: Kršćanska sadašnjost.

[8] Tamarut, A. (1970). Dan Gospodnji i počinak od rada kroz povijest do danas. Bogoslovska smotra (Vol 40) (1) (29.-37.). Zagreb: Kršćanska sadašnjost.

[9] Zakon o trgovini.(2008/9) Narodne novine 87/08,96/08,116/08,76/09-OUSRH br. 642/2009.

E-GOVERNMENT USAGE IN EUROPEAN COUNTRIES: GENDER AND EDUCATIONAL DIFFERENCES

Jovana Zoroja
University of Zagreb, Faculty of Economics and Business
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: jzoroja@efzg.hr

Anton Florijan Barišić
University for applied sciences VERN, Zagreb
Trg Drage Iblera 10, 10000, Zagreb, Croatia
E-mail: afbarisic@chronos.hr

Mirjana Pejic-Bach
University of Zagreb, Faculty of Economics and Business
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: mpejic@efzg.hr

Abstract: E-government refers to various technologies used for enabling better communication and services for citizens and enterprises. E-government enhances public services through openness, efficiency, transparency, democracy, and foster individuals and enterprises to use public services online. The goal of the paper is to investigate possible digital divide among the European countries according to e-government utilization, with the focus to gender and education. In order to shed some light on this issue, data were collected from Eurostat about the e-government usage in 2017 in European countries (female vs. male; lower, medium vs. higher education). Cluster analysis has been conducted, and results revealed that differences between countries are larger than the difference between gender and educational groups.

Keywords: e-government, Internet, education, gender, European countries, cluster analysis

1 INTRODUCTION

Rapid progress in the area of information and communication technologies (ICTs) has a strong impact on the development of digital society while ICTs applications become inevitable in private and business aspect of life [1]. Development and usage of e-government applications increased in the last two decades, especially since it improves communication and collaboration for individuals and enterprises. E-government systems offer the availability of public services 24/7 with no time and space restrictions, lower costs, and higher efficiency [2; 3]. Interaction between public institutions and users (individuals or enterprises) implicates obtaining information and documents, fulfilling, receiving, and sending forms, making payments, voting [4]. However, there are still differences among countries regarding e-government usage, while some prerequisites are necessary for e-government usage. In addition, the latest development of ICTs which implicit high speed Internet access, as well as ICT literacy, present main indicators for e-government usage [5]. Therefore, the digital divide among countries refers to a different level of e-government utilization. Developed countries which encourage and foster ICTs development are leaders in using different web based technology such as e-government while developing countries with a low level of broadband Internet are still lagging behind [6; 7]. At the European Union level, there are many policies and initiatives trying to decrease differences among European Union countries regarding ICTs development level.

The goal of the paper is to investigate are there any differences regarding e-government usage according to selected European countries and gender. Data about e-government usage by individuals in 2017 is collected from Eurostat. Data is analysed using k-means cluster analysis. Research results show that there are more differences among selected European countries than among individuals regarding gender and education level.

Paper consists of six sections. After the introduction, background on e-government usage in 2017 by females and by males with different level of education according to selected European countries has been presented. In the third section, data and k-means cluster analysis have been described. Results are presented in the fourth section. In discussion section countries across clusters and graph of the cluster, means have been defined. The last section concludes the paper.

2 BACKGROUND

Backgrounds on e-government usage in 2017 by individuals in selected European countries are presented below. Data were collected from Eurostat regarding gender and education level [8].

Figure 1 presents e-government usage in 2017 by females with different level of education according to selected European countries. Females with a low level of education use e-government services the least, especially in the following countries: Bulgaria, Croatia, Poland, Romania, and Serbia. Females with a high level of education use e-government services the most, especially individuals in Scandinavian countries: Denmark, Netherland, Finland, Sweden, Iceland, and Norway. Even in developing countries, there is a high percentage of females with a high level of education regarding e-government usage. Only in Bulgaria, Romania, and Croatia, less than 50% of females with a high level of education are using e-government, while in all other selected countries that percentage is much higher.

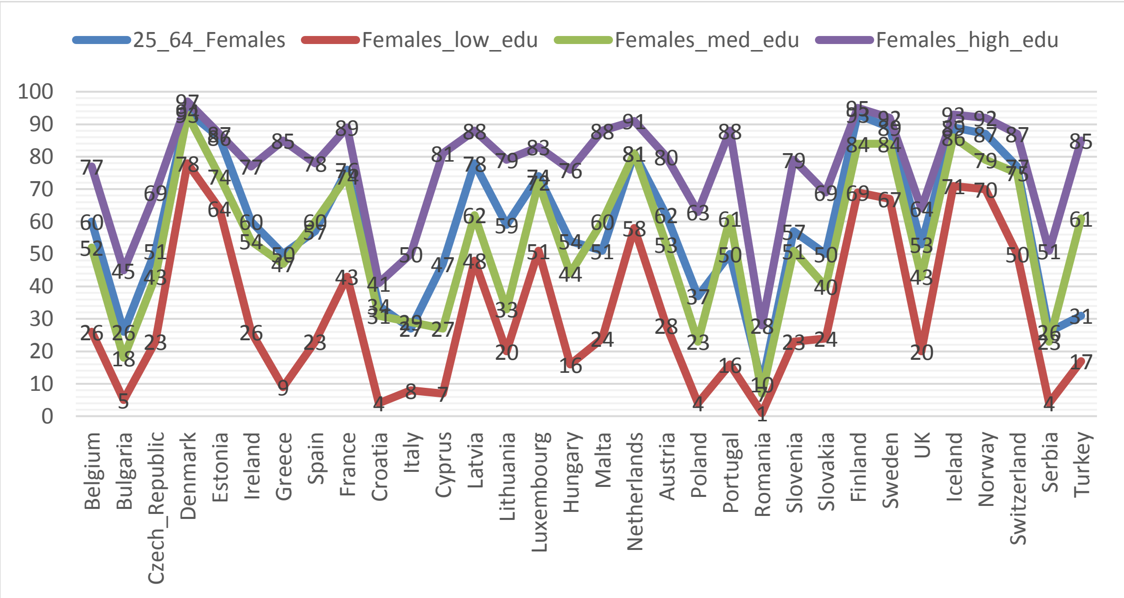


Figure 1: E-government usage in 2017 by females with different level of education according to selected European countries

Source: Authors' work based on Eurostat data (2017)

Figure 2 presents e-government usage in 2017 by males with different level of education according to selected European countries. Males with a low level of education use e-government services the least, especially in the following countries: Bulgaria, Cyprus, Romania, and Serbia. Males with a high level of education use e-government services the most, especially individuals in Scandinavian countries: Denmark, Netherland, Portugal, Finland, Sweden, Iceland, Norway, and Switzerland. Even in developing countries, there is a high percentage of males with a high level of education regarding e-government usage. Only in Bulgaria, Romania, Croatia, and Serbia, less than 50% of males with a high level of education are using e-government, while in all other selected countries that percentage is much higher.

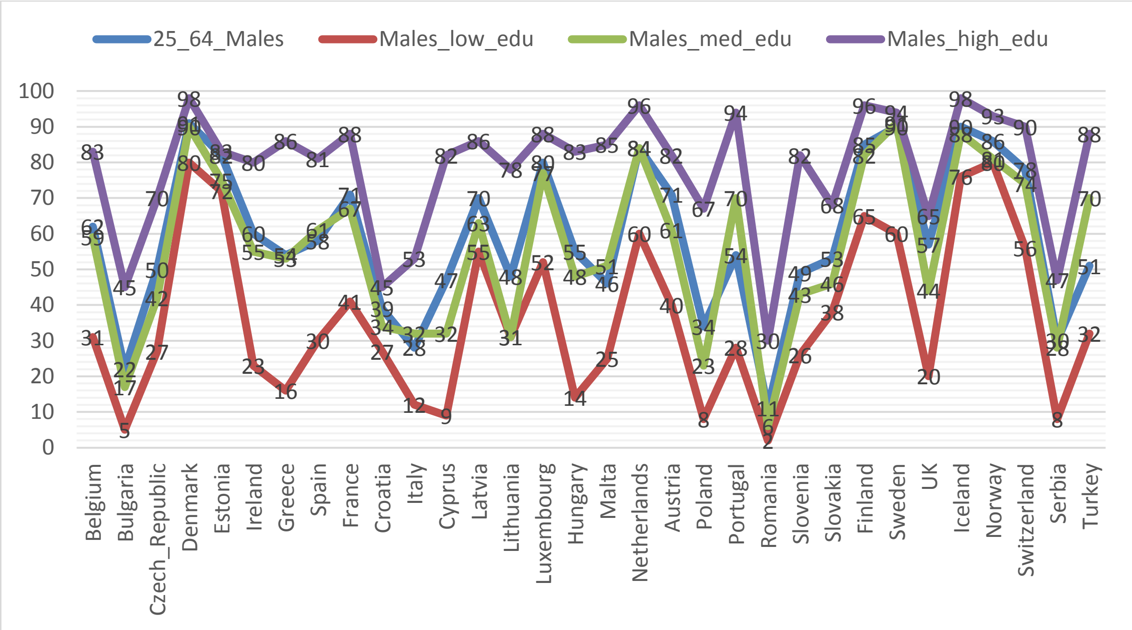


Figure 2: E-government usage in 2017 by females with different level of education according to selected European countries
Source: Authors' work based on Eurostat data (2017)

3 METHODOLOGY

3.1 Data

In order to investigate e-government usage by individuals with different level of education, we collected data from Eurostat for the year 2017, with the focus to 32 European countries (EU countries, Norway, Turkey, Switzerland, and Serbia). We did not include other European countries since the data were missing for the observed period and selected variables. We used eight variables regarding e-government usage:

- 25_64_Males - % of males 25 to 64 years that used at least one element of e-government;
- Males_low_edu - % of males 25 to 64 years, with a low level of education that use at least one element of e-government;
- Males_med_edu - % of males 25 to 64 years, with a medium level of education that use at least one element of e-government;
- Males_high_edu - % of males 25 to 64 years, with a high level of education that use at least one element of e-government;
- 25_64_Females - % of females 25 to 64 years, with a low level of education that use at least one element of e-government;
- Females_low_edu - % of females 25 to 64 years, with a low level of education that use at least one element of e-government;
- Females_med_edu - % of females 25 to 64 years, with a medium level of education that use at least one element of e-government;
- Females_high_edu - % of females 25 to 64 years, with a high level of education that use at least one element of e-government.

3.2 K-means cluster analysis

Cluster analysis was used in order to identify European countries into similar groups for the year 2017. In order to conduct the analysis, we have used eight variables regarding e-government usage among individuals with different levels of education. Research results showed that there are four clusters. We have also conducted Anova analysis and cluster means in order to test differences among defined clusters

4 RESULTS

Results of cluster analysis are presented by (i) Anova analysis (Table 1), (ii) cluster means (Table 2) and (iii) graph of the cost sequence (Figure 3). Figure 3 presents a graph of the cost sequence where it is shown that four clusters are appropriate for this analysis.

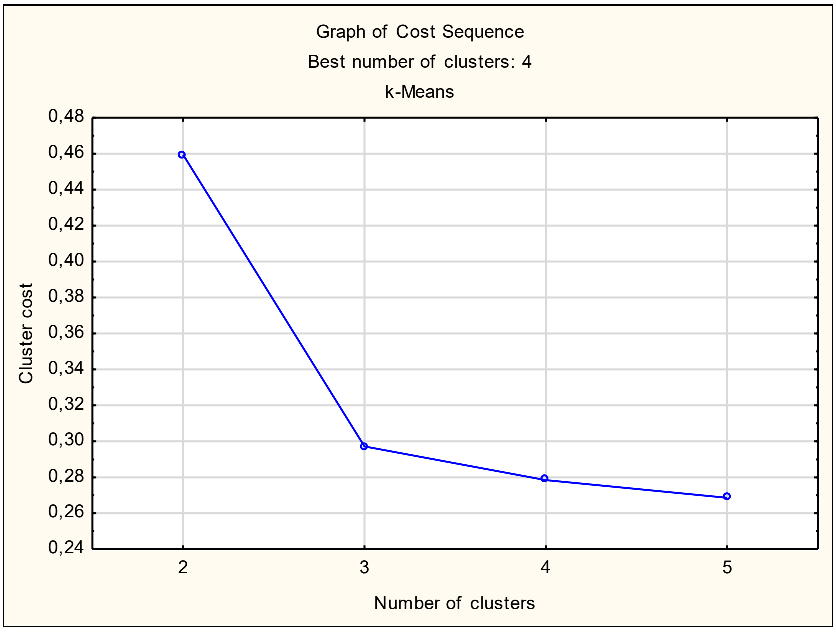


Figure 3: Graph of the cost sequence
Source: Authors' work based on Eurostat data (2017)

Table 1 presents results of Anova analysis and k-means clustering for 8 variables. Results also showed that 32 European countries are grouped into 4 clusters. All of the selected variables are statistically significant at 1%, which implies that the decision of grouping variables into four clusters is valid.

Table 1: Anova analysis, k-means clustering; 8 variables, 4 clusters, n=32 countries

	ANOVA for continuous variables, No of clusters: 4 Total no of training cases: 32					
	Between SS	**df**	**Within SS**	**df**	**F**	**p value**
25_64_Males	13053,97	3	957,905	28	127,1912	0,000000***
Males_low_edu	14762,87	3	1907,848	28	72,2211	0,000000***
Males_med_edu	13268,41	3	2825,462	28	43,8295	0,000000***
Males_high_edu	7715,15	3	1746,848	28	41,2217	0,000000***
25_64_Females	14025,54	3	1437,962	28	91,0351	0,000000***
Females_low_edu	16596,60	3	1073,619	28	144,2799	0,000000***
Females_med_edu	13423,01	3	2426,490	28	51,6307	0,000000***
Females_high_edu	7673,36	3	1468,605	28	48,7661	0,000000***

Source: Authors' work based on Eurostat data (2017)
Note: *** statistically significant at 1%

Table 2 presents cluster means for 12 selected variables. The highest mean value has Cluster 4 where are countries whose young inhabitants with low, medium and high educational level use Internet the most for participation in social networks, for taking part in on-line consultations or voting to define civic or political issues, for posting opinions on civic or political issues via websites and for civic or political participation.

Table 2: Cluster means, k-means clustering; 8 variables, 4 clusters, n=32 countries

Cluster	Centroids for k-means clustering; Number of clusters: 4; Total number of training cases: 32									
	25_64 Males	Males _low_edu	Males _med_edu	Males _high_edu	25_64_ Females	Females _low_edu	Females _med_edu	Females _high_edu	Number of cases	Percentage (%)
1	74,0	48,8	68,4	86,8	73,4	44,0	67,2	85,4	5	15,6
2	27,3	10,3	23,3	47,8	26,7	4,3	21,8	46,3	6	18,8
3	53,1	25,0	50,4	80,4	52,1	19,6	48,3	78,2	14	43,8
4	86,9	70,4	84,4	94,0	88,4	68,1	83,0	92,4	7	21,9

Source: Authors' work based on Eurostat data (2017)

5 DISCUSSION

Table 3 presents countries across clusters. Countries in Cluster 1 are Austria, France, Latvia, Luxembourg, and Switzerland. Countries in Cluster 2 are Bulgaria, Croatia, Italy, Poland, Romania, and Serbia. Countries in Cluster 3 are Czech Republic, Ireland, Greece, Spain, Cyprus, Lithuania, Hungary, Malta, Portugal, Slovenia, Slovakia, UK, and Turkey. Countries in Cluster 4 are Denmark, Estonia, Finland, Iceland, Netherlands, Norway, and Sweden. Individuals from countries which are grouped into Cluster 2 use e-government the least. In most developed north European countries, individuals use e-government the most (Cluster 4). No matter on education level or gender, individuals from developing countries which are not fostering ICTs usage and implementation, use the e-government services the least.
Graph of the cluster means is presented in Figure 4.

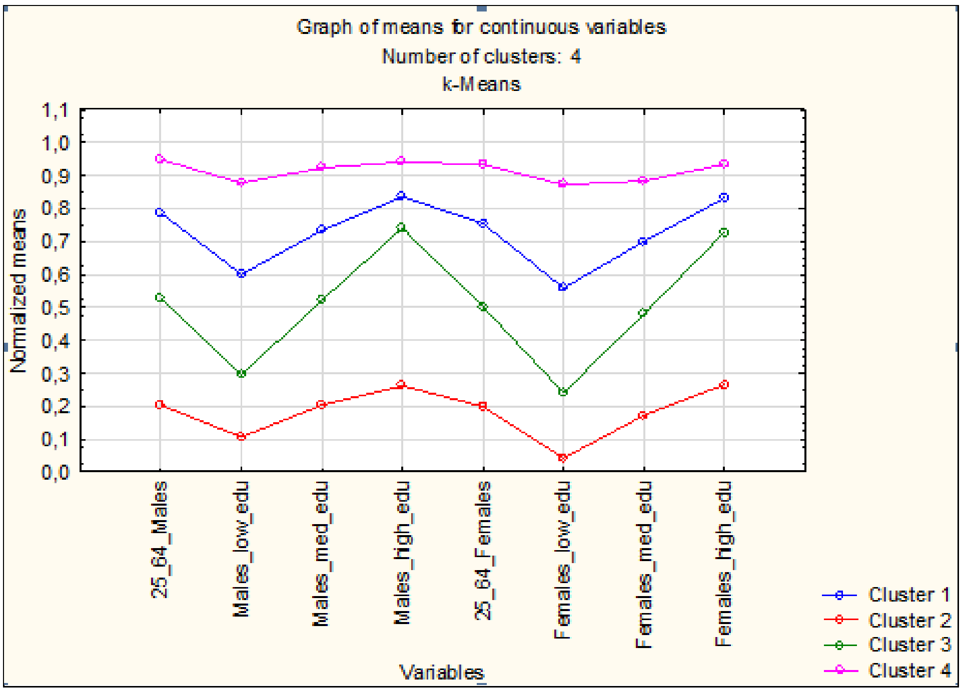


Figure 4: Graph of the cluster means
Source: Authors' work based on Eurostat data (2017)

Table 3: Countries across clusters

Cluster	Countries
Cluster 1	Austria, France, Latvia, Luxembourg, Switzerland
Cluster 2	Bulgaria, Croatia, Italy, Poland, Romania, Serbia
Cluster 3	Czech Republic, Ireland, Greece, Spain, Cyprus, Lithuania, Hungary, Malta, Portugal, Slovenia, Slovakia, UK, Turkey
Cluster 4	Denmark, Estonia, Finland, Iceland, Netherlands, Norway, Sweden

Source: Authors' work based on Eurostat data (2017)

6 CONCLUSION

This paper concentrates on e-government usage by individuals in selected European countries according to their gender and education level. In order to investigate differences among countries, cluster analysis was conducted. Results of cluster analysis revealed that European countries could be grouped into four clusters according to their usage of e-government. Research results showed that differences between countries are larger than the difference between gender and education level of individuals. The digital divide among countries still presents huge differences among developed and developing countries and unable individuals to use e-government services. Another important factor which prevents individuals in using e-government is infrastructure accessibility and IT skills. Developing countries are still lagging behind regarding ICTs development and usage. Therefore, stronger political efforts are needed to foster ICTs progress in developing countries in order to decrease the existing digital divide. The venues on how to accomplish this goal should be the guideline for future researchers.

References

[1] Fang, Y. (2002). E-government in Digital Era: Concept, Practice, and Development. International Journal of the Computer, Internet and Management, 10(2): 1-22.

[2] Guida, J., Crow, M. (2009). E-government and e-governance. ICT4D: Information and Communication Technology for Development, 283–320.

[3] Akman, I., Yazici, A., Mishra, A., Arifoglu, A. (2005). E-Government: A global view and an empirical evaluation of some attributes of citizens. Government Information Quarterly 22(2): 239-257.

[4] Rorissa, A., Demissie, D., Pardo, T. (2011). Benchmarking eGovernment: A Comparison of Frameworks for Computing eGovernment Indeks and Ranking. Government Information Quarterly, 28(3): 354-362.

[5] Sarrayrih, M. A., Sriram, B. (2015). Major Challenges in Developing a Successful eGovernment: A Review on the Sultanate of Oman. Journal of King Saud University-Computer and Information Sciences, 27: 230-235.

[6] Hung, S.Y., Chang, C.M., Yu, T.J. (2006). Determinants of user acceptance of the e-Government services: The case of online tax filing and payment system. Government Information Quarterly, 23(1): 97-122.

[7] Torres, L., Pina, V., Acerete, B. (2005). E-government developments on delivering public services among EU cities. Government Information Quarterly, 22(2): 217-238.

[8] Eurostat (2016). https://ec.europa.eu/eurostat/data/database [Accessed 16/06/2019].

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 2:

Environment and Social Issues

SUSTAINABLE PRACTICES: AN ANALYSIS OF PORTUGUESE COMPANIES

Wellington Alves
Business School, Viana do Castelo Polytechnic Institute, and
ALGORITMI Research Centre, University of Minho,
Escola de Engenharia, Depto Produção e Sistemas, Campus de Azurém, Portugal
E-mail: wellingtonb@esce.ipvc.pt

Ângela Silva
Business School, Viana do Castelo Polytechnic Institute, and
Centro de Investigação em Organizações,
Mercados e Gestão Industrial (COMEGI), Lusíada University, Portugal
E-mail: angela.a@esce.ipvc.pt

Helena Sofia Rodrigues
Business School, Viana do Castelo Polytechnic Institute, and
Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal
E-mail: sofiarodrigues@esce.ipvc.pt

Abstract: The high level of industrialization of companies contributed to the increase of the environmental impacts on the environment and society. Aiming at evaluating the level environmental practices of a set of companies from the North of Portugal conducted the research based on questionnaires. The initial results show that most of the companies are in a progress stage regarding the implementation of measures and practices related to sustainability. The study allowed suggesting some implications for consulted companies, namely the need for effective mechanics to ensure compulsory but expedite environmental procedures along with its control which is a key factor to ensure sustainable and green practices.

Keywords: Green practices, sustainability, eco-growth, companies, Portugal, statistical analysis.

1 INTRODUCTION

In response to the urgent for sustainability in the industrial sector, strategies for environmental and social impacts must be considered. The sustainability strategies are traditionally based on the identification and evaluation of criteria which expose potential impacts on the three dimensions of sustainable development: social, economic and environmental [1]. On account, the globalization, advanced supply chains have become increasingly complex over the years. The sustainability concept has been launched in the green agenda for all industrial sector, aiming to incorporate sustainable strategies with a focus on reduction, or even elimination of the negative impacts generated by products and operational process on the environment [2].

Following the agenda of sustainability, environmental practices have been posted as an important player in companies to the development of sustainable strategies. For instance, a work developed by Rashidi and Cullinane [3] has investigated the sustainability of the operational logistics performance of different countries using the OCDE database as a sample. Also, Dey et al. [4] have investigated the development of initiatives towards sustainability in the field of the supply chain operations, identifying opportunities and providing recommendations for companies regarding the implementation of sustainability in the logistics.

Having in mind the importance of the company's activities as well their significant environmental, social and economic impacts attributed to these activities, the need for sustainable strategies to reduce these impacts emerge as a fundamental research topic. This paper aims to analyze the level of implementation of environmental practices in a set of companies from the North of Portugal.

To achieve the objective of this research, a review of the relevant existing literature related to sustainability and environmental practices was conducted. Then, the case of companies in the North region of Portugal was taken as a sample.

The paper is organized in five main sections, as outlined next. A review of the relevant literature related to sustainability and environmental practices was conducted in Section 2. Then, a method was chosen in order to assess the environmental management practices of the selected companies in the North of Portugal, in Section 3. Section 4 presents and discusses the main results; highlighting aspects such as enterprise category (micro, small, medium or large) environmental policies were analysed. Section 5 presents conclusions and identifies a direction for future work.

2 LITERATURE OVERVIEW

In recent decades, environmental assessment has become commonplace in planning and evaluation at all levels in different organizations. In manufacturing process industries attention has been paid to the environmental impacts of their processes and resulting products. At the forefront, as a pioneer with the scope of sustainability the well-known Bruntland Commission Report [5] defines sustainable development as the capacity of the current generations to meet their needs without compromising the capacity of achieving the same by the future.

Sustainability issues are mostly integrated with different functions of companies which already perceived these concerns as important aspects for their performance [6]. In the last few years, sustainability awareness has been introduced as a forefront subject for companies worldwide; it has been supporting companies towards addressing economic, social and environmental goals for society, additionally adopting common practices for the elaboration of sustainable practices [7].

The relationship between sustainable development and green business growth has gained increasing importance in the literature in the last years. The discussion about environmental strategies in industrial activities is growing in both academia and industries. For instance, the work developed by Aldakhil *et al.* [9] investigates the main determinants of integrated supply chain management for green business growth for BRICS (Brazil, Russia, India, and China) countries, considering some aspects such as economic growth and environmental policies.

Notwithstanding, the development of sustainable practices by companies has not been accomplished similarly by all industrial sector and countries worldwide. In spite of all these concerns, one of the main challenges to sustainable development in the industrial sector remains as how to apply this concept on their activities, contributing positively to environmental, social, and economic aspects.

Under such a background, this research aims to investigate the level of environmental practices implemented by a set of Portuguese companies, which could contribute to understand the actual scenario of sustainable practices in these companies.

3 METHODOLOGY

In order to achieve the objective of this research several stages were considered, namely (1) an analysis of the current literature on sustainable practices; (2) based on the literature review a questionnaire was designed in order to address the issue of sustainable practices; (3) a case study was chosen as a strategy to asses a set of companies; (4) one hundred and two companies were consulted through an online questionnaire (the sample was defined for convenience, due to time and budget constraints); then (5) a statistical was performed to and results and draw results and conclusions.

The steps carried out in this research was inspired in a previous work developed by Jabbour *et al.* [10], where quantitative methods were used to investigate primary data, which support clear benefits to describe and to explore variables as well constructs of interest [11]. The design of the questionnaire is divided into two parts: one related to the companies' characterization, another concerning the measurement of environmental practices (Table 1). For the second part of the questionnaire, a five-point Likert scale was adopted as a tool to assess the company's performance. The scale comprises five levels of agreement, ranging from (1) "Not implemented" to (5) "Completely implemented".

Table 1: Level of implementation in the company of the practices of "Environmental Management"

Question	***Description***
EM1	Clear environmental management policy
EM2	Environmental training for all employees
EM3	3Rs (Reduction, Reuse and Recycling applied in water, electricity and paper)
EM4	Development of products with lower environmental impacts
EM5	Development of productive process with lower environmental impacts
EM6	Selection of suppliers based on environmental criteria
EM7	Environmental management system (ISO 14001 or others)
EM8	Voluntary disclosure of environmental performance information

Regarding the sample, from the invited companies, 102 of them agreed to participate. They were asked to fill out the questionnaire designed through Google Docs forms, and all of them were completed without any irregularity.

4 ANALYSIS OF THE RESULTS

The main findings in this research are discussed below, considering a statistical approach using the software IBM SPSS version 24.

4.1 Sample characterization

The results related to the characterization of the companies are summarized in Table 2. The sample was made up of micro companies (up to10 employees, 32.35%), small companies (between 10 and 50 employees, 25.49%), medium size (between 50 and 250 employees, 16.67%) and large companies (more than 250 employees, 25.49%).

Table 2: Technical record of participating companies

Dimension on the company	***Percent***		***Number of employees associated with logistics***	***Percent***		***Turnover (in euros)***	***Percent***
Micro	32.35		[0;3)	34.31		[0;100k)	13.7
Small	25.49		[3;6)	20.59		[100k; 250k)	10.8
Medium	16.67		[6;9)	5.88		[250k 500k)	9.8
Large	25.49		[9;12)	9.80		[500k; 1M)	10.8
			[12;15)	1.96		[1M; 5M)	18.6
			15 or more	27.45		5M or more	36.3

Regarding the number of works associated, it is possible to observe that a large number of companies have up to three workers. The results also showed that a great number of companies had a turnover (by year), more than five million euros (36.3%).

4.2 Environmental management practices

Environmental impacts are one of the most important issues related to the production process in the industrial sector. In this direction, green practices are considered as key instruments to ensure the minimization of these impacts. The results presented in Table 3 compile a summary of descriptive statistics associated with eight environmental practices proposed in this research. Results showed that all items were answered using the entire scale, meaning that the level of implementation of the environmental practices from the consulted companies are in different stages.

Table 3: Descriptive statistics for environment management practices

Environment managment practices	***Min***	***Max***	***Mean***	***St. Dev.***
EM1	1	5	3.45	1.087
EM2	1	5	3.28	1.146
EM3	1	5	3.40	1.017
EM4	1	5	2.77	1.342
EM5	1	5	2.78	1.302
EM6	1	5	2.51	1.391
EM7	1	5	2.45	1.558
EM8	1	5	2.61	1.415

The results also showed that the lowest averages are related to the environmental management system (EM7) and the selection of suppliers based on environmental criteria (EM6). These values can be explained by the fact of these measures carry a huge financial burden for businesses.

Nonetheless, for the first three items (EM1, EM2, and EM3), the highest means, which means that they are relevant to achieve environmental management practices. The standard deviation does not present great discrepancies between items.

Following this analysis, the results showed in Figure 1 present the intervals of 95% of confidence for the average answers of the companies. It confirms the results from Table 3, showing that the first measures have a higher level of implementation, while the latter is still starting.

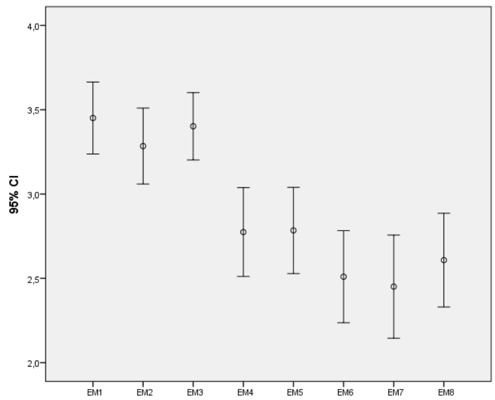


Figure 1 - Confidence interval for environment management practices

Table 4 shows the correlation between environmental practices. For the cases of EM4 and EM5, they achieved the highest correlation coefficient (0.816). These relationships are considered as important measures for companies, because if the company take into account environmental concerns the design of product/service, the production process should takes into account green practices. For the case of EM1 and EM2, the results also show a high correlation (0.778). It can be justified by the fact of these companies have a clear environmental policy in place, also employees are involved in the company’s policies.

Table 4 - Matrix correlation between environment management practices

Item	EM1	EM2	EM3	EM4	EM5	EM6	EM7	EM8
EM1	1.000							
EM2	.778	1.000						
EM3	.578	.674	1.000					
EM4	.403	.467	.553	1.000				
EM5	.468	.565	.619	.816	1.000			
EM6	.436	.548	.497	.635	.662	1.000		
EM7	.621	.570	.390	.385	.380	.372	1.000	
EM8	.425	.478	.558	.511	.620	.615	.350	1,000

For the environmental practices, the results show that it was considered as an important dimension to be addressed by companies, in order to develop strategic/finance policies, including green practices.

4.3 Environmental management practices by companies

Figure 2 presents the level of environmental management practices of the consulted companies, by the dimension. The results showed that the large ones are at the forefront regarding the implementation of environmental practices.

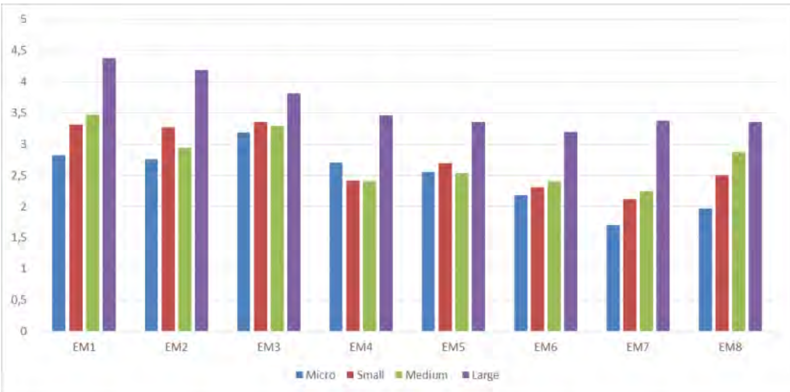


Figure 2- Average level of environment management practices, by companies' dimension

At the same time, it is possible to observe that for all companies, the last three environmental practices present the lowest scores; this fact could be associated with two important aspects, namely the lack of awareness about the benefits of implementing these practices; and also the lack of financial support to allow companies to be connected with green procedures.

To ensure the impact of these measures, a Kruskal Wallis test was conducted to examine the differences in environmental practices according to the types of companies inquired. All the assumptions of the test are assured [12]. The test showed that was a statistically significant difference exists in all the measures except in EM3 and EM5, with $p<0.05$, which means that there are significant differences of the stages of companies related with green issues, taking into account the dimension of them (Table 5).

Table 5: Kruskal Wallis Test (group variable: dimension of the company)

	EM1	*EM2*	*EM3*	*EM4*	*EM5*	*EM6*	*EM7*	*EM8*
Chi-Square	32,201	25,775	6,519	9,157	6,696	8,827	29,325	13,428
df	3	3	3	3	3	3	3	3
Asymp. Sig.	,000	,000	,089	,027	,082	,032	,000	,004

However, the 3Rs policy (EM3) is already a measure very common and the development of productive process with lower environment impacts (EM5) is also a huge concern related to the reduction of waste.

5 CONCLUSIONS

In this work, we addressed the contribution of sustainable practices for companies and sustainability. Recognizing the importance of these initiatives we proposed an analysis of the implementation of these practices taking a set of companies as a sample. Results from the literature confirmed that sustainable practices have led companies to develop environmental strategies, such as Green initiatives which have been contributing to companies save costs, meet compliance requirements, and also to create a sustainable network among customers.

From the companies consulted, the research showed that they have a long path to go toward implementation of sustainable practices, with few exceptions for large companies which have well-defined policies on sustainability field as economic resources to implement it.

Despite being an initial analysis, the results showed that for the consulted companies, environmental issues are not properly addressed and formalized by those. Finally, the results indicate that small companies face several berries to implement green actions, particularly the ones related to certification.

The study allowed suggesting some implications for Portuguese' companies. For instance, the need for effective mechanics to ensure compulsory but expedite environmental procedures along with its control is key factors to ensure sustainable and green practices of companies.

Acknowledgement
This research was supported by the FCT – Fundação para a Ciência e Tecnologia, through the project UID/CEC/00319/2019 (Alves); the project UID/EMS/04005/2019 (Silva); and the project UID/MAT/04106/2019 (Rodrigues).

References

[1] OECD, *Guidance on Sustainability Impact Assessment*. OECD Publishing, 2010.

[2] W. Bahr and E. Sweeney, "Environmental Sustainability in the Follow-Up and Evaluation Stage of Logistics Services Purchasing: Perspectives from UK Shippers and 3PLs," *Sustainability*, vol. 11, no. 9, p. 2460, 2019.

[3] K. Rashidi and K. Cullinane, "Evaluating the sustainability of national logistics performance using Data Envelopment Analysis," *Transp. Policy*, vol. 74, no. March 2018, pp. 35–46, 2019.

[4] A. Dey, P. LaGuardia, and M. Srinivasan, "Building sustainability in logistics operations: a research agenda," *Manag. Res. Rev.*, vol. 34, no. 11, pp. 1237–1259, Oct. 2011.

[5] ComissionBrundtland, "Our common future, report of the World Commission on Environment and Development," 1987.

[6] W. Alves, P. Ferreira, and M. Araújo, "Sustainability awareness in Brazilian mining corporations: the case of Paraíba state," *Environ. Dev. Sustain.*, May 2018.

[7] H. Y. Ching, F. Gerab, and T. H. Toste, "Scoring Sustainability Reports using GRI indicators: A Study based on ISE and FTSE4Good Price Indexes," *J. Manag. Res.*, vol. 6, no. 3, p. 27, 2014.

[8] P. Evangelista, L. Santoro, and A. Thomas, "Environmental sustainability in third-party logistics service providers: A systematic literature review from 2000-2016," *Sustain.*, vol. 10, no. 5, 2018.

[9] A. M. Aldakhil, A. A. Nassani, U. Awan, M. M. Q. Abro, and K. Zaman, "Determinants of green logistics in BRICS countries: An integrated supply chain model for green business," *J. Clean. Prod.*, vol. 195, pp. 861–868, 2018.

[10] A. B. L. de S. Jabbour, C. J. C. Jabbour, W. R. de S. Freitas, and A. A. Teixeira, "Lean and green?: evidências empíricas do setor automotivo brasileiro," *Gestão & Produção*, vol. 20, no. 3, pp. 653–665, 2013.

[11] J. M. Norris, L. Plonsky, S. J. Ross, and R. Schoonen, "Guidelines for Reporting Quantitative Methods and Results in Primary Research," *Lang. Learn.*, vol. 65, no. 2, pp. 470–476, Jun. 2015.

[12] J. C. F. De Winter and D. Dodou, "FivePoint Likert Items: t test versus MannWhitneyWilcoxon," Pract. Assessment, Res. Eval., vol. 15, no. 11, pp. 1–16, 2010.

COST OPTIMAL PROCESS DESIGN WITH RELIABILITY CONSTRAINTS

János Baumgartner
University of Pannonia, Faculty of Information Technology, Department of Computer Science and Systems Technology
Egyetem str. 10., Veszprém, 8200, Hungary
baumgartner@dcs.uni-pannon.hu

Zoltán Süle
University of Pannonia, Faculty of Information Technology, Department of Computer Science and Systems Technology
Egyetem str. 10., Veszprém, 8200, Hungary
sule@dcs.uni-pannon.hu

Abstract: The utilization of renewable resources poses new challenges to process design [1]. The present work shows a methodology based on process graphs (P-graph) [2][3] for generating optimal and near-optimal supply networks where the set of activities and those reliabilities are given.
A branch and bound algorithm was developed for designing supply chains under uncertainty. Such steps comprise superstructure generation, construction of the mathematical model, optimization, calculating the reliability constraints and the solution interpretation.
The developed algorithm is illustrated with an optimization problem for a region that is to support a sustainable environment. This example involving the utilization of renewable feedstock, namely agricultural products and the reliability of the solutions are calculated where lower bound of the expected reliability is input parameter. The optimal design of a supply chain under uncertainties attributable to the availability of the renewable resources as feedstock is carried out via the P-graph-based methodology. The mixed integer linear programming (MILP) problem, generated automatically via the algorithms of the methodology, serves as an input to subsequent optimization to determine the optimal supply chains in terms of multiple criteria (e.g., cost-optimal, performance optimal given a cost limit, etc.). The problem definition contains the materials (raw materials, intermediates and products) and the operating units. A branch and bound algorithm was developed for determining the profit and cost optimal solution using heuristics and cutting steps. The input of the algorithm consists of a set of products (demands to be satisfied), raw materials (available), and potential operating units (manufacturing, transportation, etc.), as well as cost data with capacity constraints and the reliability value of each operating unit. The lower reliability bound defined for the entire process is also given. The output of the algorithm is a ranked list of the *n*-best networks.

Keywords: optimisation, cost, quantity, reliability

References

[1] Čuček L., Lam H. L., Klemeš J., Varbanov P. and Kravanja Z., 2010, Synthesis of Networks for the Production and Supply of Renewable Energy from Biomass, Chemical Engineering Transactions, 21, 1189-1194.

[2] Friedler F., Tarjan K., Huang Y. W. and Fan L. T., 1992, Graph-Theoretic Approach to Process Synthesis: Axioms and Theorems, Chem. Engng Sci., 47, 1973-1988.

[3] Friedler F., 2009, Process Integration, Modelling and Optimisation for Energy Saving and Pollution Reduction, Chemical Engineering Transactions, 18, 1-26.

AGGREGATION OF INDIVIDUAL JUDGMENTS INTO GROUP INTERVAL JUDGMENTS IN AHP

Petra Grošelj, Lidija Zadnik Stirn, Gregor Dolinar
University of Ljubljana, Biotechnical Faculty
Jamnikarjeva 101, 1000 Ljubljana, Slovenia
petra.groselj@bf.uni-lj.si, lidija.zadnik@bf.uni-lj.si, gregor.dolinar@bf.uni-lj.si

Abstract: The aggregation of individual judgments by weighted geometric mean is frequently applied within the analytic hierarchy process. This method can be unsatisfactory, since a single scalar number cannot embrace the diverse views and experiences of all the stakeholders in the decision-making process. We propose a new aggregating approach that is based on the weighted geometric mean and the geometric standard deviation and produces group interval judgments. The parameter that influences the width of the intervals is incorporated. The method is applied to the case study.

Keywords: Multiple criteria evaluation, Analytic hierarchy process, Group decision-making, Geometric standard deviation, Interval judgments, Natural resources.

1 INTRODUCTION

The analytic hierarchy process (AHP) [1] is a widely used method in group multi-criteria decision-making. The decision-maker assigns a number from 1 to 9 on the AHP scale for each pairwise comparison $a_{ij},\ i,j=1,...,n$ of two objects, if we have n objects. There are many approaches that can be used to aggregate individual scalar judgments, including: the weighted geometric mean method [2], models based on data envelopment analysis [3, 4], the group-weighted least-squares method [5], and the group method with aggregation on preferential differences and rankings [6]. The weighted geometric mean method is the one most often employed for aggregating individual scalar judgments into group judgments [7-10]. However, the group may not be satisfied with the scalar-valued judgment because it may not reflect the diverse experiences, views, and perspectives of all the decision-makers. In such a case, a group interval judgment can be used. In the literature, minimum and maximum individual judgments of r decision-makers have been applied to aggregate individual scalar judgments $a_{ij}{}^{(k)},\ k=1,...,\mathrm{r}$ into group judgment $a_{ij}^{group}=\left[l_{ij}^{group},u_{ij}^{group}\right]$ for the lower and upper bounds of the intervals, respectively [11, 12] where

$$l_{ij}^{group}=\min_{k=1,2,...,r}\left\{a_{ij}^{(k)}\right\} \tag{1}$$

$$\text{and } u_{ij}^{group}=\max_{k=1,2,...,r}\left\{a_{ij}^{(k)}\right\}. \tag{2}$$

This approach has some drawbacks as only extreme judgments that could be outliers will affect the interval bounds, whereas the intermediate judgments have no influence on the interval. The aim of this paper is to suggest a new approach for creating group interval-valued judgments based on the geometric mean and the geometric standard deviation. The power factor λ that influences the width of the intervals is also incorporated in the approach.
There are many known methods of deriving an interval priority vector from the interval comparison matrix [13, 14, 12]. In this paper, the uncertain logarithmic least-squares method [15] is employed to derive interval weights from an interval comparison matrix, which generalizes the logarithmic least-squares method [16] for deriving weights in the classical point-valued AHP. For ranking the interval weights the possibility-degree formulae based on the distance between two intervals is used.

The proposed approach for generating group interval judgments is applied to the case study of ranking the strengths, weaknesses, opportunities, and threats (SWOT) groups of Pohorje, a mountain range in Slovenia. One of the objectives of this study is to analyze the width of the interval group judgments and the derived interval group weights depending on the factor λ.

2 AGGREGATING INDIVIDUAL JUDGMENTS INTO GROUP INTERVAL JUDGMENTS

The classical AHP uses a 1 to 9 ratio scale to compare two elements on the same level of the hierarchy with respect to the element on the next highest level. Reciprocal comparisons are used to create the comparison matrix $A=\left(a_{ij}\right)_{n\times n}$, $i,j=1,...,n$, where $a_{ji}=1/a_{ij}$. The priority vector $w=(w_1,...,w_n)$ can be derived from the comparison matrix by many known methods, of which the logarithmic least-squares method [16] is often applied:

$$\min\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\ln a_{ij}-\left(\ln w_i-\ln w_j\right)\right)^2 , \tag{3}$$

the solution of which is the geometric mean of the row elements of matrix A:

$$w_i=\sqrt[n]{\prod_{j=1}^{n}a_{ij}},\ i=1,...,n . \tag{4}$$

The inconsistency of judgments is measured by the consistency ratio [1].

We assume that a group of *r* decision-makers take part in the decision-making process. Let $A^{(k)}=\left(a_{ij}^{(k)}\right)_{n\times n}$, $k = 1,...,r$ be their comparison matrices of scalar-valued judgments. Individual judgments are not necessarily equally important because of variations in the knowledge, experience, or status of the decision-makers. Each decision-maker is assigned a weight ρ_k, $k = 1,...,r$, $\rho_k>0$, $\sum_{k=1}^{r}\rho_k=1$, which determines the importance of their opinion.

Our new approach for aggregating individual scalar-valued judgments into a group interval-valued judgment is based on the acknowledged method used in classical group AHP, which uses the weighted geometric mean of individual judgments [2]. The main statistic for measuring the dispersion of values around the geometric mean is the geometric standard deviation.

Let

$$a_{ij}^{(WGMM)}=\prod_{k=1}^{r}\left(a_{ij}^{(k)}\right)^{\rho_k} \tag{5}$$

be the weighted geometric mean of the set of individual judgments $\left\{a_{ij}^{(1)},...,a_{ij}^{(r)}\right\}$. Weighted geometric means of all judgments are gathered in a matrix $A^{WGMM}=\left(a_{ij}^{(WGMM)}\right)_{n\times n}$.

The geometric standard deviation is defined as:

$$s_{ij}^{(WGMM)}=\exp\sqrt{\frac{V_1}{V_1^2-V_2}\sum_{k=1}^{r}\rho_k\left(\ln\frac{a_{ij}^{(k)}}{a_{ij}^{(WGMM)}}\right)^2} , \tag{6}$$

where $V_1=\sum_{k=1}^{r}\rho_k$ and $V_2=\sum_{k=1}^{r}\rho_k^2$. Because $V_1=1$, equation (6) simplifies to:

$$s_{ij}^{(WGMM)} = \exp\sqrt{\frac{1}{1-V_2}\sum_{k=1}^{r}\rho_k\left(\ln\frac{a_{ij}^{(k)}}{a_{ij}^{(WGMM)}}\right)^2}. \quad (7)$$

The matrix of geometric standard deviations is denoted by $S^{WGMM} = \left(s_{ij}^{(WGMM)}\right)_{n\times n}$ and is symmetric.

Individual judgments are dispersed around the weighted geometric mean, which is a good representative of the group judgment. However, it is unsatisfactory because it may not reflect the differences in the views of individual decision-makers. Our proposed approach intends to present the group judgments as intervals. An interval judgment could satisfy decision-makers better than a point-valued group judgment does. The interval should express a variety of different judgments based on differences in knowledge, views, and experiences. This variety can be expressed by the geometric standard deviation. We define the group interval judgments as:

$$a_{ij}^{group} = \left[l_{ij}^{group}, u_{ij}^{group}\right] = \left[\frac{a_{ij}^{(WGMM)}}{\left(s_{ij}^{(WGMM)}\right)^{\lambda}}, a_{ij}^{(WGMM)}\left(s_{ij}^{(WGMM)}\right)^{\lambda}\right],\ i,j=1,\ldots,n \quad (8)$$

The width of the interval can be regulated by a factor λ, $\lambda \in [0,1]$. If $\lambda = 1$ the group interval judgment can be too wide, therefore one of the goals of this study is to analyze the width of the intervals for different λ.

3 GROUP WEIGHTS FROM THE GROUP INTERVAL MATRIX

The group interval matrix is constructed of the interval judgments from the equation (8):

$$A^{group} = \left(\left[l_{ij}^{group}, u_{ij}^{group}\right]\right)_{n\times n}. \quad (9)$$

and has the reciprocal property $a_{ji}^{group} = \left(a_{ij}^{group}\right)^{-1}$.

To derive interval weights from A^{group}, we apply the uncertain logarithmic least-squares method [15] :

$$w_i = \frac{\left(\prod_{j=1}^{n} a_{ij}^{group}\right)^{1/n}}{\sum_{i=1}^{n}\left(\prod_{j=1}^{n} a_{ij}^{group}\right)^{1/n}} = \left[\frac{\left(\prod_{j=1}^{n} l_{ij}^{group}\right)^{1/n}}{\sum_{i=1}^{n}\left(\prod_{j=1}^{n} u_{ij}^{group}\right)^{1/n}}, \frac{\left(\prod_{j=1}^{n} u_{ij}^{group}\right)^{1/n}}{\sum_{i=1}^{n}\left(\prod_{j=1}^{n} l_{ij}^{group}\right)^{1/n}}\right]. \quad (10)$$

To rank the interval weights, we use the row–column elimination method [17, 12] from the matrix of degrees of preference:

$$P = \begin{bmatrix} - & p_{12} & \cdots & p_{1n} \\ p_{21} & - & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & - \end{bmatrix} \quad (11)$$

Let $A = [A_L, A_U]$ and $B = [B_L, B_U]$ be two intervals. To calculate the probabilities p_{ij} in matrix P we use a method that is based on the distance between two intervals, and assumes that interval weights are uniformly distributed:

$$p_{AB} = P(A \geq B) = \frac{\max\{0, A_U - B_L\} - \max\{0, A_L - B_U\}}{(A_U - A_L) + (B_U - B_L)}. \quad (12)$$

4 THE CASE STUDY

Pohorje is a mountain range in northeastern Slovenia in Europe that covers an area of 840 square kilometers. A large part of Pohorje is included in Natura 2000, a collection of nature protection areas within the European Union. The economic development of Pohorje, which is primarily covered with forests, is dependent on agriculture, tourism, and forestry. Pohorje participated in Project NATREG [18]. One of the results of the project was SWOT analysis of Pohorje development, and the ranking of SWOT factors within each SWOT group [19].

The goal of this case study was to rank SWOT groups. We selected 12 stakeholders that were well acquainted with the problems associated with Pohorje. Their importance was determined also by AHP and their weights were 0.0807 for stakeholders from tourism, 0.0759 for forestry, 0.0814 for agriculture and 0.0953 for stakeholders from nature protection. The selected experts conducted pairwise comparisons of SWOT groups with the optimal development of Pohorje as their goal [20].

4.1 RESULTS

We aggregated individual pairwise comparison matrices of SWOT groups by weighted geometric mean (5) and calculated the geometric standard deviations by equation (7):

$$A^{WGMM} = \begin{bmatrix} 1 & 1.39 & 0.71 & 1.17 \\ 0.72 & 1 & 0.66 & 1.23 \\ 1.41 & 1.51 & 1 & 1.43 \\ 0.85 & 0.81 & 0.70 & 1 \end{bmatrix}, \quad S^{WGMM} = \begin{bmatrix} 1 & 4.24 & 2.42 & 4.06 \\ 4.24 & 1 & 3.22 & 2.66 \\ 2.42 & 3.22 & 1 & 3.26 \\ 4.06 & 2.66 & 3.26 & 1 \end{bmatrix} \tag{13}$$

We calculated the group interval matrices A^{group} for different values of parameter λ by equation (8), λ ranging from 0.1 to 1 with step 0.1. To compare the results we also calculated the interval group judgments with minimum as the lower bound and maximum as the upper bound (eq. (1) and (2)) that we called Min-max. The group judgments from A^{WGMM} and the group intervals for all judgments were calculated. The results of the analysis show that the intervals gained by our new proposed approach are mostly narrower than the Min-max intervals. However the intervals for larger values of parameter λ are wide.

We derived the group weights from A^{WGMM} by equation (4) and the group interval weights from the group interval matrices according to the equation (10). The results are presented in Figures 1-2.

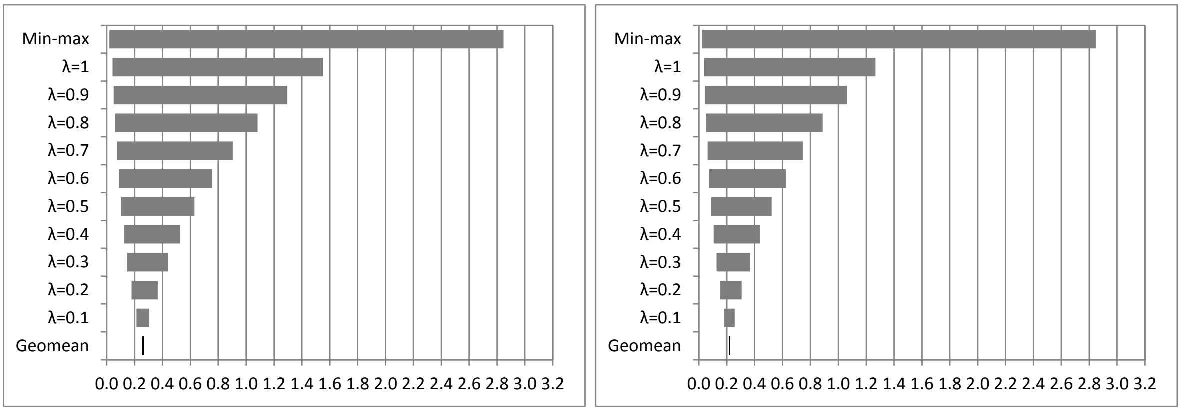


Figure 1: Group weights for the strengths (left graph) and for the weaknesses (right graph)

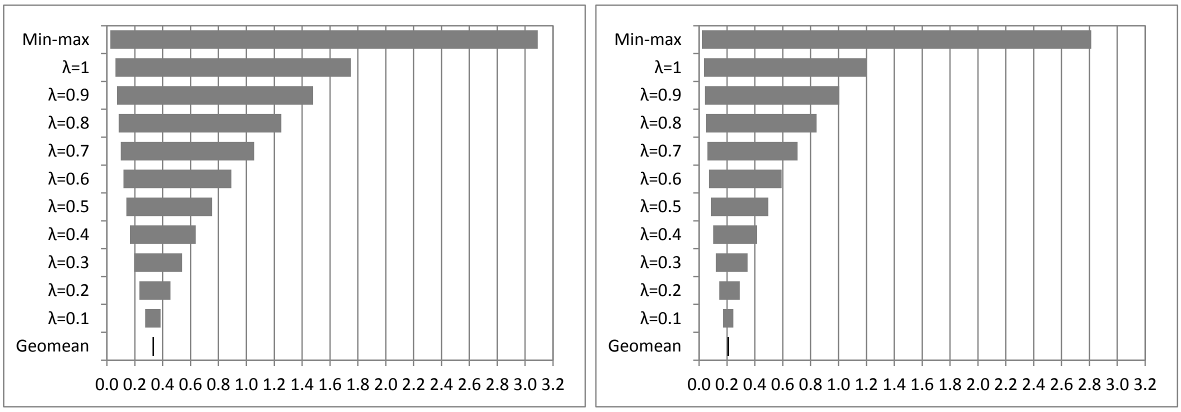


Figure 2: Group weights for the opportunities (left graph) and for the threats (right graph)

The results show that the width of the intervals rapidly increases when the width of the interval judgments increases. The scalar-valued weights gained by geometric mean method represent the percentages of importance. If we would like to preserve this meaning of the weights, then the upper bound of the intervals should be less than 1. Thus, in our case study the values of parameter λ from 0.1 to 0.6 are suitable.

The final ranking of the SWOT groups derived by the equations (11)-(12) show that the ranking of the SWOT groups is identical for all values of parameter λ. Strengths and opportunities are more important than weaknesses and threats with: Opportunities $\overset{53\%}{\succ}$ Strengths $\overset{55\%}{\succ}$ Weaknesses $\overset{51\%}{\succ}$ Threats for $\lambda = 1$ and Opportunities $\overset{85\%}{\succ}$ Strengths $\overset{74\%}{\succ}$ Weaknesses $\overset{57\%}{\succ}$ Threats for $\lambda = 0.1$. When the parameter λ decreases the preferences of one group over the other increase, which is another confirmation that smaller values of λ are more suitable. The ranking is equal to the ranking gained by geometric mean method. The ranking gained by the Min-max method differs which lowers the credibility of Min-max method.

5 CONCLUSIONS

Aggregation of individual judgments into a scalar group judgment using the weighted geometric mean is often unsatisfactory because a single number could hardly embrace the diverse views, perspectives, and ideas of several stakeholders in a decision-making process. This paper proposes a new approach for aggregating individual scalar-valued judgments into a group interval-valued judgment. The new approach is based on the weighted geometric mean and the geometric standard deviation. The group interval takes into account all varieties of individual judgments through the use of the geometric standard deviation and the width of the interval judgments is controlled by parameter λ.

The approach is applied to the case study of ranking SWOT groups, which are important for the development of the Pohorje mountain range in Slovenia. The results show that strengths and opportunities are more important than weaknesses and threats for the development of Pohorje. The analysis of the width of the intervals shows that narrower interval judgments lead to narrower intervals of final weights. In the future work the presented approach should be tested in more applications to confirm our findings.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P4-0059).

References

[1] Saaty, T. L., 1980 The analytic hierarchy process McGraw-Hill, New York.

[2] Saaty, T. L. and Peniwati, K., 2008 Group decision making: Drawing out and reconciling differences RWS Publications, Pittsburgh, PA.

[3] Grošelj, P., Pezdevšek Malovrh, Š. and Zadnik Stirn, L., 2011. Methods based on data envelopment analysis for deriving group priorities in analytic hierarchy process. Central European Journal of Operations Research, 19(3), pp. 267–284.

[4] Hosseinian, S., Navidi, H. and Hajfathaliha, A., 2009. A New Linear Programming Method for Weights Generation and Group Decision Making in the Analytic Hierarchy Process. Group Decision and Negotiation, 21(3), pp. 1–22.

[5] Sun, L. and Greenberg, B., 2006. Multicriteria Group Decision Making: Optimal Priority Synthesis from Pairwise Comparisons. Journal of Optimization Theory and Applications, 130(2), pp. 317–339.

[6] Huang, Y. S., Liao, J. T. and Lin, Z. L., 2009. A Study on Aggregation of Group Decisions. Systems Research and Behavioral Science, 26(4), pp. 445–454.

[7] Ananda, J. and Herath, G., 2008. Multi-attribute preference modelling and regional land-use planning. Ecological Economics, 65(2), pp. 325–335.

[8] De Luca, S., 2014. Public engagement in strategic transportation planning: An analytic hierarchy process based approach. Transport Policy, 33(0), pp. 110–124.

[9] Oblak, L. and Glavonjić, B., 2015. A Model for the Evaluation of Radio Advertisements for the Sale of Timber Products. Drvna industrija, 65(4), pp. 303–308.

[10] Wang, Y.-M. and Chin, K.-S., 2009. A new data envelopment analysis method for priority determination and group decision making in the analytic hierarchy process. European Journal of Operational Research, 195(1), pp. 239–250.

[11] Chandran, B., Golden, B. and Wasil, E., 2005. Linear programming models for estimating weights in the analytic hierarchy process. Computers & Operations Research, 32(9), pp. 2235–2254.

[12] Wang, Y.-M., Yang, J.-B. and Xu, D.-L., 2005. A two-stage logarithmic goal programming method for generating weights from interval comparison matrices. Fuzzy Sets and Systems, 152(3), pp. 475–498.

[13] Arbel, A. and Vargas, L., 2007. Interval judgments and Euclidean centers. Mathematical and Computer Modelling, 46(7-8), pp. 976–984.

[14] Meng, F., Chen, X., Zhu, M. and Lin, J., 2015. Two new methods for deriving the priority vector from interval multiplicative preference relations. Information Fusion, 26(0), pp. 122–135.

[15] Zeshui, X. and Yager, R. R., 2010. Power-Geometric Operators and Their Use in Group Decision Making. Fuzzy Systems, IEEE Transactions on, 18(1), pp. 94–105.

[16] Crawford, G. and Williams, C., 1985. A note on the analysis of subjective judgment matrices. Journal of Mathematical Psychology, 29(4), pp. 387–405.

[17] Facchinetti, G., Ricci, R. G. and Muzzioli, S., 1998. Note on ranking fuzzy triangular numbers. International Journal of Intelligent Systems, 13(7), pp. 613–622.

[18] Urataric, N. (2011) NATREG: Managing Natural Assets and Protected Areas as Sustainable Regional Development Opportunities Regional environmental center, Ljubljana.

[19] Urataric, N. and Marega, M. (2010) Poročilo z zaključne delavnice za izdelavo analize SWOT in oblikovanje elementov vizije [Report of the final workshop of a SWOT analysis and design elements of the vision]; project: NATREG.

[20] Grošelj, P. and Zadnik Stirn, L., 2015. The environmental management problem of Pohorje, Slovenia: A new group approach within ANP – SWOT framework. Journal of Environmental Management, 161, pp. 106-112.

POPULATION DIVERSITY MAINTENANCE USING UNIFORMLY DEPLOYED SET OF P-LOCATION PROBLEM SOLUTIONS

Marek Kvet
University of Žilina, Faculty of Management and Informatics,
Univerzitná 1, 010 26 Žilina, Slovak Republic,
marek.kvet@fri.uniza.sk

Jaroslav Janáček
University of Žilina, Faculty of Management and Informatics,
Univerzitná 1, 010 26 Žilina, Slovak Republic,
jaroslav.janacek@fri.uniza.sk

Abstract: The main research topic refers to and the contribution of the paper is to be seen in the construction of an approximate evolutionary method for the min-sum location problems. To keep the population diversity at a certain level, suggested genetic algorithm uses the uniformly deployed set of associated *p*-location problem solutions. The efficiency of the newly introduced solving technique is studied from the viewpoint of computational time and resulting solution accuracy. Theoretical explanation of the developed heuristic algorithm is here supported by a brief computational study performed on real problem instances, data of which were obtained from the road network of the self-governing regions of Slovakia. The results provided by the suggested approximate approach will be compared to the exact solution of the problems.

Keywords: Min-sum location problems, evolutionary algorithm, uniformly deployed set of solutions, approximate solution accuracy

1 INTRODUCTION

The class of the *p*-location problems comprises a broad spectrum of practical problems from the construction of distribution systems to designing public service systems. The *p*-location problem with a min-sum objective function consists in determination of *p* centre locations in a finite set of possible locations so that the sum of weighted distances from individual user locations to the nearest located centres is minimal. This problem is also known as the weighted *p*-median problem and it is often used to model the public service system design problems [2, 3, 6, 7, 11, 12]. A plethora of exact and approximate algorithms has been developed to solve medium sized instances of these problems [4, 8, 9].

Nevertheless, if the algorithms are applied to larger instances of the problems, the computational time of commonly used branch-and-bound based IP-solvers has proved to be unacceptably long and, in addition, almost unpredictable [1, 10]. Contrary to the exact algorithms, the often quoted evolutionary metaheuristics enable to terminate their computational process in a given time, but a weak side of them is that they can be trapped at a local extreme, what means that the processed population of solutions stays homogenous.

To face the population homogeneity, various approaches have been suggested [5]. Most of them evaluate the diversity of the current population and if the population diversity drops below some threshold, either probability of mutation is increased or randomly generated individuals are inserted into the current population or other procedure of population diversification is employed.

In this paper, we focus on the usage of the uniformly deployed set as a tool of diversity maintenance in a genetic algorithm with an elite set.

The uniformly deployed set of *p*-location solutions can be represented by a set of zero-one vectors of the *m*-dimensional space, where each vector has exactly *p* components equal to one and the remaining components get zero values. The number *m* equals the number of possible centre locations. Each pair of solutions from the uniformly deployed set has Hamming distance greater than or equal to a given even number h and the set is assumed to be maximal, i.e. there is no other solution, which can be added to the set.

Having once the uniformly deployed set for given integer numbers p and m, we can generate a different uniformly deployed set by simple permutation of m components. In this research, we make use of the fact that the uniformly deployed set is a maximally diversified set of solutions and we will study the impact of the uniformly deployed set usage on the performance of a genetic algorithm solving the min-sum p-location problem.

The paper is assembled in the following way. The next section comprises the formulation of the simple min-sum p-location problem and definition of the uniformly deployed set of the p-location problem solutions. The third section deals with the genetic algorithm, which uses the uniformly deployed set for keeping population diversity. The fourth section contains the results of numerical experiments and a comparison of the heuristic solutions to the exact ones, and the last section summarizes the obtained findings.

2 THE P-LOCATION PROBLEM AND UNIFORMLY DEPLOYED SET

The general p-location problem with simple min-sum objective function can be defined as a task of determination of p service center locations from a set I of m possible service centre locations so that the sum of weighted distances from users' locations to the nearest located centre is minimal. Let symbol J denote the set of the users' locations and b_j denote the weight associated with user location $j \in J$. If d_{ij} denotes the distance between locations i and j, then the studied problem known also as the weighted p-median problem can be described by (1).

$$\min\left\{\sum_{j\in J} b_j \min\left\{d_{ij} : i \in P\right\} : P \subseteq I,\ |P| = p\right\} \tag{1}$$

A feasible solution of the problem (1) is presented by a sub-set of p locations from the set I. Any such solution can be also represented by m-dimensional vector of zeroes and ones, where m is the cardinality of I. The vector $\boldsymbol{y}$, which corresponds with a solution P, is defined so that y_i=1 if and only if $i \in P$. This way, the set of all feasible solutions can be studied as a sub-set of vertices of m-dimensional hypercube. A topology on the set of hypercube vertices can be defined by Hamming gauge, which gives the length of zero-one vector $\boldsymbol{y}$ according to (2).

$$length(\mathbf{y}) = \sum_{i=1}^{m} y_i \tag{2}$$

Obviously, each vector corresponding to a feasible solution of the p-location problem has the length equal to the value of p. The Hamming distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by (3).

$$H(\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{m} |y_i - x_i| \tag{3}$$

Hamming distance of two different feasible solutions of the p-location problem is an even integer and varies from 2 to $2p$, where the distance of 2 corresponds with pair of solutions, which differ only in one service centre location, and the distance of $2p$ indicates that the corresponding solutions have no common service centre. The notion of Hamming distance enables to define the maximal uniformly deployed set. The maximal uniformly deployed set of the p-location problem is defined for given distance d as a maximal set of p-location problem solutions, where every two solutions have minimal Hamming distance d. As the maximal uniformly deployed set can be formed based only on the values p, m and d, regardless of the location numbering, each permutation of subscripts of locations gives a different maximal uniformly deployed set. This property together with the fact that the uniformly deployed set is a maximally diversified set of a given number of solutions, can be used in evolutionary metaheuristic discussed in the next section.

3 GENETIC ALGORITHM AND UNIFORMLY DEPLOYED SET

The genetic algorithm belongs to the family of evolutionary metaheuristics, which imitate evolution of some species represented by a population of individuals. Very often, an individual corresponds to a feasible solution to the solved problem. The genetic algorithm comes from an initial population, and then it transforms the current population to a new one imitating a real evolutionary process. A standard schema of the genetic algorithm is given by the following steps.

0. Initialize the current population, compute fitness of each individual in the current population and initialize the best-found solution by the individual with the best fitness. Determine the elite set as a subset of the current population using the evaluated fitness.
1. If the termination rule is met, then terminate, otherwise create a new population according to the following steps.
2. Fill up the set of candidates by repeating these operations: Random selection of a pair of individuals from the current population. Creation of offspring by the operation of crossover. Submit offspring to the random mutation and insert the resulting solutions into the set of candidates.
3. Create a new population so that the elite set of the current population is included in the new one first. Then, based on fitness, the elements of the set of candidates are used to fill up the non-elite part of the new population. The best-found solution and the elite set are updated.
4. Declare the new population as the current one and go to 1.

A common approach to initialization of the starting population consists of a random selection of p locations from m possible centre location, to determine positions of the p service centres. This process is repeated until the demanded number of individuals is obtained.

Contrary to other problems, the special structure of the set of feasible solutions yields the possibility to construct crossover operation so that offspring are feasible solutions of the p-location problem. Similarly, the operation of mutation can be defined as one or more location exchanges and, this way, mutated offspring keeps feasibility. Even if the topology of the set of all p-location solutions enables formulation of smart operations of crossover and mutation, the substantial danger of evolutionary process of being trapped by homogeneous population in a local minimum stays unsolved.

In common genetic algorithms, population diversity is evaluated after a given number of population exchanges and if diversity decreases below a given threshold, then a diversity maintenance attempt is made. Diversity of the population can be increased either by increasing mutation probability or by injecting other very distant solutions in the non-elite part of the population.

We suggest a new version of the above-described genetic algorithm. The version uses the uniformly deployed set of p-location problem solutions both for the constitution of the initial population and for population diversification.

The initial maximally diversified population can be obtained simply by adopting the uniformly deployed set as the initial population.

Diversity maintenance can be performed in step 3, where a set of candidates is used only for updating of the elite set in the new population, but the non-elite part of the new population is filled up from the permutated uniformly deployed set of *p*-location problem solutions.

4 NUMERICAL EXPERIMENTS

This section is devoted to the computational study aimed at studying the suggested genetic algorithm with uniformly deployed set from the viewpoint of computational time demands and the solution accuracy.

The real instances of *p*-location problems were obtained from the road network of the Slovak Republic and the emergency health care systems organized in particular self-governing regions. The mentioned instances are further denoted by the names of capitals of the individual regions followed by triples (XX, *m*, *p*), where XX is commonly used abbreviation of the region denotation, *m* stands for the number of possible centre locations and *p* is the number of service centres, which are to be located in the mentioned region. The list of instances follows: Bratislava (BA, 87, 14), Banská Bystrica (BB, 515, 36), Košice (KE, 460, 32), Nitra (NR, 350, 27), Prešov (PO, 664, 32), Trenčín (TN, 276, 21), Trnava (TT, 249, 18) and Žilina (ZA, 315, 29). All cities and villages with the corresponding number of inhabitants b_j were taken into account. The coefficients b_j were rounded up to hundreds. The set of communities represents both the set of users' locations and the set of possible centre locations as well.

An individual experiment was organized so that the exact solution of the underlying *p*-location problem was computed using the radial approach described in [4, 8, 9], first. To obtain the exact solution of the problem, the optimization software FICO Xpress 7.3 was used and the experiments were run on a PC equipped with the Intel® Core™ i7 5500U processor with the parameters: 2.4 GHz and 16 GB RAM. The obtained results are summarized in the left part of Table 1, which is denoted by "Optimal solution". The column denoted by *minSum* contains the optimal objective function values of the problem (1). The computational times in seconds are given in the column denoted by *CT* [s].

The right part of Table 1, denoted by "Heuristic solution" is dedicated to the results of suggested evolutionary heuristic, which makes use of the uniformly deployed set of solutions. The uniformly deployed sets were obtained by the previously performed process for the individual problems. The process consists of creating an initial uniformly deployed set, and then the process continues with series of optimization problem-solving procedures, when each step either adds a new solution to the set or declares that the set is maximal. The cardinalities of the resulting individual sets are reported in Table 1 in the column denoted by *PopSize*. The experiments with the genetic algorithm were performed for a various maximal time of evolution *CT_Evol*, which varied from 0.5 to 5 seconds. Each experiment for the given self-governing region and given maximal computational time was performed 10 times and the average results are plotted in the associated column. In the columns, the average *gap* is reported. The gap expresses the relative difference in the obtained objective function value from the objective function value of the exact solution.

Table 1: Results of numerical experiments for the self-governing regions of Slovakia

region	Optimal solution		Heuristic solution				
	minSum	*CT* [s]	*CT_Evol* [s]:	5	2	1	0.5
			PopSize	*avg gap* [%]	*avg gap* [%]	*avg gap* [%]	*avg gap* [%]
BA	19325	0.28	23	0.00	0.00	0.00	0.00
BB	29873	2.22	172	2.53	2.88	5.51	8.02
KE	31200	1.44	60	1.10	4.39	6.06	9.00
NR	34041	2.84	83	0.23	1.28	2.33	3.14
PO	39073	35.01	232	1.40	3.08	5.24	11.28
TN	25099	1.38	137	0.01	0.03	0.22	0.67
TT	28206	0.87	212	0.01	0.06	0.36	0.84
ZA	28967	0.71	112	0.14	0.66	1.58	2.67

Analysing the results reported in Table 1 and assuming that the gap of approximately 3% can be considered a satisfactory result, we can claim that the evolutionary algorithm is faster than the exact approach with only one exception (see the benchmark KE). The table also demonstrates the unpredictability of the computational time of the exact branch-and-bound-based method (see the benchmark PO). In this case, the genetic algorithm reached satisfactory results 15 times faster.

5 CONCLUSIONS

The main goal of this paper was to demonstrate the usage of uniformly deployed set in a genetic algorithm for *p*-location min-sum problem. Presented results of performed numerical experiments confirm the usefulness of the suggested heuristic approach, in which we combine the evolutionary process with the usage of predefined uniformly deployed set of solutions. The suggested evolutionary algorithm enables to keep the computational time of the method in acceptable limits contrary to the exact approach. Therefore, we can conclude that we provide the readers with a very useful heuristic method for effective and fast min-sum *p*-location problem-solving.

Future research in this field may be aimed at a more complex form of the *p*-location problem with robustness consideration. The research of evolutionary techniques efficiency could be enlarged by memetic approaches.

Acknowledgement

This work was supported by the research grants VEGA 1/0342/18 “Optimal dimensioning of service systems”, VEGA1/0089/19 “Data analysis methods and decisions support tools for service systems supporting electric vehicles”, and VEGA 1/0689/19 “Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities” and APVV-15-0179 “Reliability of emergency systems on infrastructure with uncertain functionality of critical elements”.

References

[1] Avella, P., Sassano, A. and Vasil'ev, I. (2007). Computational study of large scale p-median problems. Mathematical Programming, 109: pp. 89-114.

[2] Brotcorne, L. and Laporte, G. and Semet, F. (2003). Ambulance location and relocation models. European Journal of Operational Research, 147: pp. 451–463.

[3] Czimmermann, P., Kohâni, M. (2018). Computation of transportation performance in public service systems. In IEEE workshop on complexity in engineering, Firenze, pp. 1-5.

[4] García, S., Labbé, M. and Marín, A. (2011). Solving large p-median problems with a radius formulation. INFORMS Journal on Computing, 23(4): pp. 546-556.

[5] Gendreau, M., Potvin, J. (2010). Handbook of Metaheuristics, Springer Science & Business Media, 648 p.

[6] Ingolfsson, A., Budge, S. and Erkut, E. (2008). Optimal ambulance location with random delays and travel times. Health Care management science, 11(3): pp. 262-274.

[7] Jánošíková, Ľ., Žarnay, M. (2014). Location of emergency stations as the capacitated p-median problem. International scientific conference: Quantitative Methods in Economics-Multiple Criteria Decision Making XVII, Virt, Slovak Republic.

[8] Kvet, M. (2014). Computational study of radial approach to public service system design with generalized utility. Digital Technologies 2014: the 10th International IEEE conference, Zilina, Slovak Republic.

[9] Kvet, M. (2015). Exact and heuristic radial approach to fair public service system design. Information and Digital Technologies 2015: IEEE catalog number CFP15CDT-USB, Zilina: Slovak Republic.

[10] Marianov, V. and Serra, D. (2004). Location problems in the public sector, In Facility location. Applications and theory (by Drezner Z (ed.) et al.), Berlin, Springer: pp. 119-150.

[11] Marsh, M. and Schilling, D. (1994). Equity measurement in facility location analysis. European Journal of Operational Research, 74: 1–17.

[12] Snyder, L. V., Daskin, M. S. (2005). Reliability Models for Facility Location; the Expected Failure Cost Case. Transport Science, 39 (3), pp. 400-416.

THE IMPACT OF HARMONY ON THE PERCEPTION OF MUSIC

Lorena Mihelač
ŠC Novo mesto, IT and Music Department & International Postgraduate School Jožef Stefan
Novo mesto, Ljubljana, Slovenia
E-mail: lorena.mihelac@sc-nm.si

Janez Povh
University of Ljubljana, Faculty of Mechanical engineering and Institute of mathematics, physics and mechanics Ljubljana
Ljubljana, Slovenia
E-mail: janez.povh@fs.uni-lj.si

Abstract: In this paper we present results of continuations of a longitudinal study, started in 2018 and partially reported in [11]. We used a dataset of 160 musical excerpts and measured for each musical piece the complexity (entropy) of the harmonic progression in the piece, and four perceptual variables (difficulty, pleasantness, recognition and repeatability). For this study we used a group of 20 evaluators, 10 with and 10 without formal musical education. After two evaluations in 2018, within a time interval of 1 month, we did a third evaluation in 2019 - one year after the second evaluation, in which the evaluators were involved in two musical courses. The results show that in the third evaluation, the perceptual variables difficulty and recognition significantly increased, while repeatability significantly decreased, regardless musical education. For the fourth variable pleasantness, we can show statistical significance only compared to the second evaluation. We provide plausible explanations of the impact of musical knowledge on the understanding of harmony and directions for a computational in-depth analysis of musical excerpts found to be irregular in structure.

Keywords: Entropy, harmony, harmonic progression, subliminal irregularity, musical knowledge.

1 INTRODUCTION

According to Madsen and Widmer [7], each musical piece can be considered as information. Proceeding from this idea, in the empirical studies [9,10], the information-theoretic measure, entropy (measure of uncertainty) of the harmonic progression (succession of chords) and chords (defined as a combination of at least 3 tones), has been used in explaining the acceptability of music by participants. Data consisting of 160 musical excerpts covering different musical styles, from baroque till the 20th century has been examined. The results we obtained show that the acceptance of a musical piece can be explained by analyzing the entropy, as the entropy of the harmonic progression correlates with the acceptance (pleasantness) of a musical piece. In [9,10], 53 out of 160 musical excerpts have been found to exhibit peculiarities: regardless of the low entropy or absence of entropy, these 53 excerpts have been evaluated by the participants as difficult and less pleasant, the rest of them with frequent chord changes in the harmonic progression as difficult but at the same time as very pleasant.

In 2018, same data has been evaluated twice by twenty new participants, 20 female students from secondary school ŠC Novo mesto, 10 with musical knowledge (at least 5 years of formal musical education) and 10 without musical knowledge [11]. Significance of formal musical knowledge in evaluating all the main four variables (difficulty, pleasantness, repeatability and recognizability) at the $p<.05$ level was found in both evaluations. Increase of average pleasantness, repeatability and recognizability and decrease of average difficulty has been found in the 2nd evaluation compared to the 1st one. Furthermore, significance of impact of musical styles on difficulty at the $p<.05$ level was found in both evaluations, while no significance was found in any of the evaluations of musical styles on pleasantness.

Same 53 musical excerpts, recognized as excerpts with a peculiar harmonic progression in the studies [9,10] have been also recognized in both evaluations in 2018 by new participants [11]. To examine the impact of the harmonic progression in these excerpts on all the four main variables, Two Sample t-tests were conducted in both evaluations, by categorizing the data in two main categories: "regular" (107 musical pieces) and "irregular" (53 musical pieces). Significance of the impact of harmonic progression on all the main four variables (difficulty, pleasantness, repeatability and recognizability) at the $p<.05$ level was found in both evaluations.

In 2019, the same data has been evaluated for the 3rd time with exactly the same participants with the purpose: (i) to explore the impact of musical training on the perception of the aforementioned four perceptual variables and (ii), to check the impact of the same 53 musical excerpts found to be "irregular" on the perception of four main variables.

2 METHOD

In April 2019, a 3rd evaluation of the data used in studies [9,10,11] has been conducted with same 20 participants as in the study [11]. In the period between the 2nd and 3rd evaluation, the participants have been enrolled in two musical courses dealing with music theory and harmony. We hypothesize that the improved musical knowledge has impact on the understanding of the harmonic progression and the complexity of harmony as well, in the sense of lowering the feeling of complexity and difficultness.

2.1 Participants

Twenty female participants (N=20), students from the secondary school ŠC Novo mesto, aged 16-17 years ($M = 15.4$, $SD = 0.49$), of which 10 with musical knowledge (more than five years of formal musical training) and 10 without musical knowledge, have participated in the 3rd evaluation of the data.

2.2 Music stimuli

Data used in the studies [9,10], and evaluated twice in study [11], has been used in the 3rd evaluation in 2019. The data consists of 160 musical examples covering different musical styles, from baroque until the 20th century, shortened to musical excerpts with a length between 14 *s* and 18 *s* in duration, adjusted to an equal loudness level. Each musical excerpt has been evaluated by the same four main variables using a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The four main variables were: a) the difficulty of listening to the musical piece, b) the pleasantness perceived while listening to the musical piece, c) the recognition of the musical piece and d) the repeatability of the musical piece (the readiness of the evaluator to listen to the whole music piece, not only to a part of it).

2.3 Procedure

The open source software Moodle was used for evaluating the musical excerpts also in the 3rd listening session. The evaluation was time limited; however, each participant had two days available to complete the task due to the demanding listening to the complete data. The rating of the musical excerpt could be done only after its conclusion. We have used paired t-test and two sample t-test in analyzing the data.

3 RESULTS

Paired-samples t-tests were conducted to compare the results from the 3rd evaluation with the results from the 1st and the 2nd evaluation obtained on the same 160 musical excerpts. Four main variables (difficultness, pleasantness, repeatability and recognition) were evaluated. The results of the comparison between 1st and 3rd evaluation, and 2nd and 3rd evaluation (all participants) are presented in Table 1.

Table 1: Comparison between 1st and 3rd, and 2nd and 3rd evaluation (all participants).

variable	***difficulty***	***pleasantness***	***variable***	***difficulty***	***pleasantness***
1st ev.	M = 1.75, SD = 0.358	M = 3.58, SD = 0.608	***2nd ev.***	M = 1.61, SD = 0.316	M = 3.77, SD = 0.482
3rd ev.	M = 2.46, SD = 0.549	M = 3.56, SD = 0.609	***3rd ev.***	M = 2.46, SD = 0.549	M = 3.56, SD = 0.609
	t(159) = -25.22 p = 2.2e-16	t(159) = 1.04 p = 0.301		t(159) = -25.60 p = 2.2e-16	t(159) = 8.13 p = 1.111e-13
	repeatability	***recognition***		***repeatability***	***recognition***
1st ev.	M = 3.44, SD = 0.652	M = 2.67, SD = 1.13	***2nd ev.***	M = 3.52, SD = 0.579	M = 3.09, SD = 0.94
3rd ev.	M = 3.29, SD = 0.706	M = 3.26, SD = 1.01	***3rd ev.***	M = 3.29, SD = 0.706	M = 3.26, SD = 1.01
	t(159) = 5.46 p = 1.838e-7	t(159) = -17.81 p = 2.2e-16		t(159) = 7.48 p = 4.739e-12	t(159) = -17.81 p = 7.014e-10

The comparison between 1st and 3rd evaluation show significantly increased difficulty and recognition, and significantly decreased repeatability. No significant difference was found in average pleasantness. The comparison of the 2nd and 3rd evaluation show significantly increased difficulty and recognition, and significantly decreased repeatability and pleasantness.

The results of the comparison of the 1st and 3rd evaluation and 2nd and 3rd evaluation by participants without knowledge are presented in Table 2, by participants with knowledge in Table 3 respectively.

Table 2: Comparison between 1st and 3rd, and 2nd and 3rd evaluation (participants without musical knowledge).

variable	***difficulty***	***pleasantness***	***variable***	***difficulty***	***pleasantness***
1st ev.	M = 1.92, SD = 0.452	M = 3.37, SD = 0.567	***2nd ev.***	M = 1.70, SD = 0.467	M = 3.60, SD = 0.566
3rd ev.	M = 2.66, SD = 0.663	M = 3.43, SD = 0.687	***3rd ev.***	M = 2.66, SD = 0.663	M = 3.43, SD = 0.687
	t(159) = -18.67 p =2.2e-16	t(159) = -1.796 p = 0.074		t(159) = -21.71 p =2.2e-16	t(159) = 4.30 p = 2.916e-05
	repeatability	***recognition***		***repeatability***	***recognition***
1st ev.	M = 3.14, SD = 0.610	M = 2.50, SD = 0.978	***2nd ev.***	M = 3.21, SD = 0.664	M = 2.85, SD = 0.994
3rd ev.	M = 3.08, SD = 0.779	M = 3.09, SD = 1.099	***3rd ev.***	M = 3.08, SD = 0.779	M = 3.09, SD = 1.099
	t(159) =1.81 p = 0.007	t(159) = -14.30 p = 2.2e-16		t(159) = 2.89 p = 0.004	t(159) = -6.19 p = 4.795e-09

The results of the comparison between 1st and 3rd evaluation, and 2nd and 3rd evaluation (Tables 2 and 3) of all the four main variables by participants with and without musical knowledge, show significantly increased average difficulty and recognition, and significantly decreased average repeatability. No significance was found in pleasantness in the comparison between

1^{st} and 3^{rd} evaluation, however significant difference in average pleasantness was found in the comparison between 2^{nd} and 3^{rd} evaluation in participants with and without musical knowledge.

Comparisons between 1^{st} and 3^{rd} evaluation, and 2^{nd} and 3^{rd} evaluation show significantly increased difficulty in both groups at the $p<.05$ level. Plausible explanation could be the involvement of both groups in additional musical subjects (music theory and harmony, musical expression), and the impact of these subjects on the way how the participants have perceived, understood and evaluated the same 160 musical excerpts after one year of rigorous and exhaustive training.

Table 3: Comparison between 1^{st} and 3^{rd}, and 2^{nd} and 3^{rd} evaluation (participants with musical knowledge).

variable	***difficulty***	***pleasantness***	***variable***	***difficulty***	***pleasantness***
1st ev.	M = 1.69, SD = 0.386	M = 3.71, SD = 0.616	***2nd ev.***	M = 1.51, SD = 0.269	M = 3.95, SD = 0.494
3rd ev.	M = 2.26, SD = 0.515	M = 3.69, SD = 0.637	***3rd ev.***	M = 2.26, SD = 0.515	M = 3.69, SD = 0.637
	t(159) = -19.55 p = 2.2e-16	t(159) = 0.82 p = 0.413		t(159) = -22.22 p = 2.2e-16	t(159) = 8.77 p = 2.635e-15
	repeatability	**recognition**		**repeatability**	**recognition**
1st ev.	M = 3.71, SD = 0.672	M = 2.90, SD = 1.13	***2nd ev.***	M = 3.83, SD = 0.615	M = 3.33, SD = 0.970
3rd ev.	M = 3.49, SD = 0.741	M = 3.43, SD = 1.011	***3rd ev.***	M = 3.49, SD = 0.741	M = 3.43, SD = 1.011
	t(159) = 5.60 p = 9.2e-08	t(159) = -12.94 p = 2.2e-16		t(159) = 9.29 p = 2.2e-16	t(159) = -3.08 p = 0.002

According to the findings in [8], where improved pitch discrimination of tones after only 3-weeks of auditory musical training was found, and positive impact of formal musical training on key and harmony perception in 3 to 6 year old children in [1], it was expected that the gained and improved musical knowledge (regardless of the previous musical knowledge) would affect the perception of difficultness in the sense of lowering the feeling of complexity. However, this was not the case in the 3^{rd} evaluation, as the impact of formal musical training on the evaluation of difficultness of the musical excerpts has not manifested itself in the sense of lowering the feeling of difficultness, but only in a renewed approach and understanding of chords and their relations in harmony.

The finding in studies [9,10] have shown that some musical excerpts are perceived as difficult (complex) by participants. Both evaluations in study [11] have confirmed the impact of the same musical excerpts, suggesting that these excerpts are not meeting the listener's expectations, who tends to use a set of basic perceptual principles, applying them to different musical styles, depending on the kind of music the listener is exposed to. If these expectations are not met, the given information in the musical piece is not well understood and/or recognized [2], and the complexity, the feeling of difficultness, sometimes interpreted as a mid-point between order and disorder [4], seem to be perceived more in the sense of a disorder or "irregularity". In the study [11], these "irregularities" have been defined as *subliminal irregularities*, as they appear to be perceived by the listener without being aware of them, affecting the feeling of difficultness and pleasantness. In the 3^{rd} evaluation, these same 53 musical excerpts have been examined with Two Sample t-tests, and the same main four variables (difficultness, pleasantness, repeatability and recognition) have been evaluated.

As in the previous two evaluations, significance of the impact of harmonic progression (irregular/regular) on all the main four variables (difficulty, pleasantness, repeatability and recognizability) at the $p<.05$ level was found in all participants and in both groups (with and without musical knowledge) separately. In a recent study [12], irregularity of the harmonic progression and specific cases of irregularity have been examined in-depth and proposed as a measure for the complexity of harmony. The results of this study are suggesting, that the perception of a musical content as regular/irregular depends on the musical syntax, which is in accordance with the findings in [5].

Unusual musical content (e.g., unusual use of chords in harmonic progression), content perceived as new or content which is expected but is missing (e.g., missing chords in the harmonic progression), are exhibiting a subjective feeling of difficultness in a musical piece, impacting also the pleasantness and listener's acceptability of a musical piece.

4 CONCLUSION AND FUTURE WORK

In this paper, we have presented the 3rd evaluation of the same data used in studies [9-11]. As in the previous two evaluations [11], four main variables have been tested. The comparison between 1st and 3rd evaluation, and 2nd and 3rd evaluation has shown a significant increase of difficultness and recognition, significant decreasement of repeatability and significant decreasement of pleasantness in the 3rd evaluation compared to the 2nd one.

Musical exhaustive training has shown to be important in the 3rd evaluation of musical excerpts. Compared to the 1st and 2nd evaluation, a significantly increase of difficulty has been found in the 3rd evaluation. A plausible explanation could be the improved musical knowledge, regardless of the previous musical knowledge, exhibiting impact on the understanding of the harmonic progression. It seems that the participants have become aware of the chords and their relations in the harmonic progression, trying to understand from a different viewpoint the harmonic progression as it was a year ago in the previous two evaluations.

Harmonic progression and its regularity/irregularity has shown to be important also in the 3rd evaluation. As the regularity of harmonic progression has been proposed as a measure of complexity of the harmony in [12], future work could be extended and focused on: (i) searching of peculiar irregularities in the melody and its impact on the harmonic progression, (ii) examining the interlaced structure of harmony and melody by computational modelling of the listener's perception of the aforementioned 53 irregular musical excerpts. These approaches would certainly highlight the content in the musical pieces, which are perceived in listeners as irregular.

Glossary

Harmony is the structure of music with respect to the composition and progression of chords. *Chord* is a combination of at least 3 tones (sounds) performed simultaneously. *Harmonic progression* is a succession of chords. *Key* is the relationship system based on a scale. The keys are simply named by the scale on which they are based, e.g., (key) C major, C minor, . . . *Pitch* is defined as the highness or lowness of sound. *Musical style* is used in this paper with the meaning as music genre, a conventional category that identifies some pieces of music belonging to a shared tradition or set of conventions. *Subliminal irregularity* is defined in this paper as all the peculiarities in the musical structure, which are impacting the listener's perception of music, causing a higher feeling of complexity and/or difficulty.

Acknowledgement

We would like to thank all the participants for providing their estimates for the acceptability of the musical examples. We also thank the School center Novo Mesto for its partial support of the present study and the Slovenian Research Agency for its support through the research project J1-8155 and program P2-0256.

References

[1] Corrigall, K. A., Trainor, L. J. 2009. Effects of musical training on key and harmony perception. Annals of the New York Academy of Sciences. 1169(7): 164-168.

[2] Edmonds, B. 2018. What is complexity? The philosophy of complexity per se with application to some examples in evolution. http://cogprints.org/357/4/evolcomp.pdf [Accessed 20/06/2019]

[3] Febres, G., Jaffe, K. 2007. Music viewed by its entropy content: A novel window for comparative analysis. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185757 [Accessed 19/06/2019].

[4] Grassberger, P. 1989) Problems in quantifying self-organized complexity. Helvetica Physica Acta, 62: 498-511.

[5] Harrison, P. M. C., Pearce, M. T. 2018. Dissociating sensory and cognitive theories of harmony perception through computational modelling. Proceedings of ICMPC15/ESCOM10. Graz, Austria: Centre for Systematic Musicology, University of Graz.

[6] Krumhansl, C. L. 2004. The cognition of tonality – as we know it today. Journal of New Music research, 33(3): 253-268.

[7] Madsen, S. T. Widmer, G. 2006. Music complexity measures predicting the listening experience. Proceedings of the 9t International Conference on Music Perception & Cognition.

[8] Menning, H., Roberts, L., Pantey, C. 2000. Plastic changes in the auditory cortex induced by intensive frequency discrimination training. Neuroreport, 11: 817-822.

[9] Mihelač, L. 2017. Predicting acceptability of music with entropy of harmony. Master thesis. Novo mesto: Faculty of information studies.

[10] Mihelač, L., Povh, J. 2017. Predicting the acceptability of music with entropy of harmony. SOR'2017, 14(1): 371–375.

[11] Mihelač, L., Wiggins, A. G., Lavrač, N., Povh, J. 2018. Entropy and acceptability: information dynamics and music acceptance. Proceedings of ICMPC15/ESCOM10. Graz, Austria: Centre for Systematic Musicology, University of Graz.

[12] Mihelač, L., Povh, J. 2019. The impact of the complexity of harmony on the acceptability of music. *Submitted*.

WASTE MANAGEMENT CONSEQUENCES - CASE STUDY ON THE ISLAND OF BRAČ

Marija Vuković, Snježana Pivac
University of Split, Faculty of Economics, Business and Tourism, Department of Quantitative Methods
Cvite Fiskovića 5, Split, Croatia
E-mail: marija.vukovic@efst.hr, spivac@efst.hr

Marijana Šemanović
University of Split, Faculty of Economics, Business and Tourism, Postgraduate Specialist Study
Cvite Fiskovića 5, Split, Croatia
E-mail: marijana.semanovic@gmail.com

Abstract: This paper explores waste management on the island of Brač, Croatia through the residents' perception. The impact of waste treatment on life quality and public and local self-government budget were analysed for two waste landfills: Košer and Kupinovica. The research was conducted through a survey on a sample of 395 residents of Brač. Structural equation models and nonparametric tests showed that the place of residence significantly affects the residents' perception of life quality for Košer and the perception of public and local self-government budget for Kupinovica, while age group affects public and local self-government budget perception for Košer.

Keywords: waste management, life quality, nonparametric tests, public and local self-government budget, structural equation modelling.

1 INTRODUCTION

Waste management is one of the major problems today. While human activities have always generated waste, it was not a major issue until urbanisation, human population growth and conurbations. Poor waste management is gradually causing even more problems. It has led to negative environmental consequences, such as contamination of water, soil and atmosphere, which had a major impact on public health. For the last couple of decades, environmental awareness in the area of waste management has reached a certain level. In Europe, the basic legislation on waste management has been adopted in 1975 through The Framework Directive on Waste [4], when waste began to be perceived as a serious threat to health and the environment generally. It is important that all stakeholders, from inspections, manufacturers to end users, find a common way to achieve a complete sustainable waste management system. Waste must be reduced, starting from households, and useful ingredients must be singled out for recycling and production. Consciousness about waste management should reach the general level [5, 9, 14].

When waste management is considered on small islands, as isolated and limited territories, it does not have a simple solution. Since many small islands are popular tourist destinations, they have to deal with high waste generation as a consequence. Landfill sites may cause the loss of environmental quality or the reduction of tourism, while incineration plants may not be economically efficient and waste reduction strategies may not be very effective. Shipping waste to the mainland, as another alternative, usually causes large costs [3]. Therefore, options for alternative waste management systems on islands are limited. There are also difficulties like limited space availability, restricted recycling and resale opportunities and impacts on the local environment, especially increased when the island is small, densely populated and tourist dependent [2, 12]. Landfilling, as one of the waste management options, is still highly practiced in many small island developing states and islands in general. The main reason for high landfilling is the absence of other waste management techniques and the fact that landfilling is the cheapest way of waste disposal. However, it still remains the most undesirable option, due to the loss of potentially useful waste materials, such as recyclables, which can be used again [11]. Therefore, there is no simple solution to this problem. Skordilis [13] presented a system's engineering model for the strategic planning of an integrated solid waste management at local level and demonstrated that the combination of material recovery at the source with the utilization of the organic fraction is the optimum solution for small local communities.

According to European Parliament statistics, Croatia is among the worst countries in waste management, using the practice of landfilling 78% of their municipal waste in 2016, while the target of landfilling by 2035 is below 10% [16]. The island of Brač is especially interesting, since it is a tourist oriented island near Split, the second largest city of Croatia [14]. Currently waste management on Brač takes place at two official landfills: Košer in the area of Pučišća and Kupinovica near Supetar [14]. The citizens of Supetar have expressed their disagreement with the solution of the Ministry of Environmental Protection and Energy regarding the waste disposal of the entire island just in Supetar (Kupinovica) and a referendum was organized for June 9th, 2019. Although only 47% of voters approached the referendum, 99% of them were against that decision. However, due to the low voter response, the referendum was not valid [8]. Considering those recent events, the aim of this paper is to explore the residents' perception about waste management on the island of Brač. The research hypothesis is that demographic characteristics of residents of the island of Brač affect life quality (LQ) and public and local self-government budget (PLSG) in the context of waste management consequences for Košer and Kupinovica waste landfills.

2 DATA AND METHODOLOGY

For the purpose of the research, a questionnaire was designed according to Aleksić [1], and it was focused on the 13 956 residents of the island of Brač. The population was divided into 8 locations (public and local self-government units): Bol, Milna, Nerežišća, Postira, Pučišća, Selca, Supetar and Sutivan. A stratified sample was then taken, since the population structure for each of the 8 defined locations corresponds to the total population structure for those locations [14]. The final sample included 395 residents of Brač. The survey included questions about the residents' opinions about waste management consequences on life quality (LQ) and public and local self-government budget (PLSG) for Košer and Kupinovica waste landfills. The answers to those questions were formulated as a 1-5 Likert scale. Structural equation modelling (SEM) was used to find the impact of demographic characteristics of residents on the perception of life quality, as well as the perception of public and local self-government budget. SEM is a multivariate technique, which combines aspects of factor analysis and multiple regression. It enables the simultaneous examination of a series of interrelated dependence relationships between the measured variables and latent constructs, as well as between several latent constructs [6]. Latent constructs (factors) represent the operationalization of a construct in SEM, which is not directly observable, but can be measured by one or more variables (indicators). The indicators are measured (observed, manifest) variables, for which scores are collected and entered in a data file [6, 10]. Data was analysed with statistical packages SPSS 23.0 and Mplus 7. In addition to SEM, reliability analysis was conducted using Cronbach's alpha coefficients to determine the extent to which the measured variables (indicators) of the model are consistent in their values [6] and nonparametric Kruskal Wallis test was used to explore the significant influences in detail.

3 EMPIRICAL RESULTS

The model used in this paper included four influential demographic variables: level of education (EDU), gender, age group and place of residence (RES). The independent latent variables in the model are the opinions of waste management consequences on LQ and PLSG. Both of the latent variables are assumed to be correlated and they are measured through several measured variables, represented by the corresponding survey questions. Since the answers to those questions are coded as ordinal variables (1-5 Likert scale), the mean and variance adjusted weighted least squares (WLSMV) method of estimation is used [10]. The internal quality of latent variables is shown in table 1. It can be seen that all of the factor loadings are higher than 0.5 and are statistically significant, which means that the indicator variables almost perfectly reflect the latent variable they measure [6]. Cronbach's alpha coefficients were used to test the reliability. Coefficients for both factors are very good, above 0.70, which shows an internal consistency of the items [15].

Table 1: The internal quality of latent variables.

	Location Košer			***Location Kupinovica***		
Latent variable	*Item*	*Standardized factor loading*	*Cronbach's alpha*	*Item*	*Standardized factor loading*	*Cronbach's alpha*
Perception of life quality (LQ)	LQ1	0.743**	0.828	LQ_1	0.803**	0.850
	LQ2	0.753**		LQ_2	0.785**	
	LQ3	0.778**		LQ_3	0.840**	
	LQ4	0.799**		LQ_4	0.785**	
	LQ5	0.713**		LQ_5	0.736**	
Perception of public and local self-government budget (PLSG)	PLSG1	0.711**	0.767	PLSG_1	0.711**	0.727
	PLSG2	0.936**		PLSG_2	0.885**	
Items: LQ1 (LQ_1) – Do you consider that the waste treatment on the island of Brač would have a negative impact on the life conditions of the local population?; LQ2 (LQ_2) – Do you consider that the waste treatment on the island of Brač would have a negative impact on the visual experience?; LQ3 (LQ_3) – Do you consider that the waste treatment on the island of Brač would have a negative impact on the island tourism?; LQ4 (LQ_4) – Do you consider that the waste treatment on the island of Brač would have a negative impact on the groundwater?; LQ5 (LQ_5) – Do you consider that the waste treatment on the island of Brač would have a negative impact on flora and fauna?; PLSG1 (PLSG_1) – Do you consider that the waste treatment on the island of Brač would have a positive impact on the island's economy?; PLSG2 (PLSG_2) – Do you consider that the waste treatment on the island of Brač would significantly contribute to the public and local self-government budget?						

*** Significant at 0.01 level*

Table 2 shows the goodness of fit measures for structural models and path diagrams with standardized estimates for both models are shown in figure 1. It can be concluded that, for location Košer, only RES significantly affects the perception of consequences on LQ, while age group significantly affects the perception of consequences on PLSG. The model Chi-square is significant, but other goodness of fit measures are acceptable. The model is retained, since the Chi-square increases with the sample size [6, 7, 10]. On the other hand, for location Kupinovica, only the impact of RES on PLSG is significant. The goodness of fit measures are somewhat better than the first model, although the Chi-square value is also significant.

Table 2: Goodness of fit measures for structural models.

	Location Košer	***Location Kupinovica***
Model fit	*χ^2 = 138.558, p-value ≤0.001, RMSEA (90% C.I.) = 0.088 (0.073-0.104), CFI=0.955, TLI=0.926*	*χ^2 = 116.418, p-value ≤0.001, RMSEA (90% C.I.) = 0.078 (0.063-0.094), CFI=0.972, TLI=0.955*

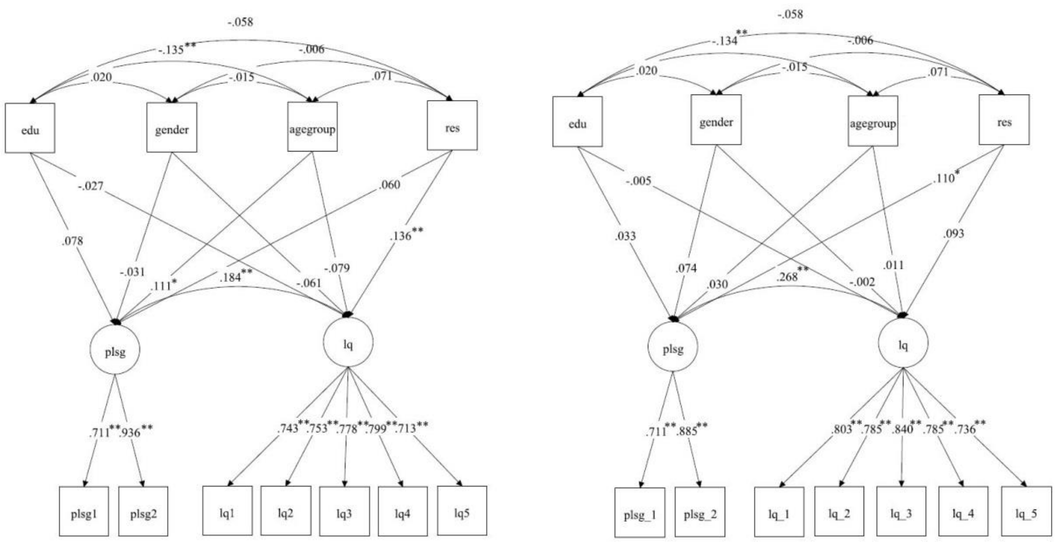


Figure 1: Path diagrams with standardized estimates for locations Košer and Kupinovica.

Considering the results of SEM, nonparametric Kruskal Wallis test was used to determine the way these significant paths influence the latent variables. Table 3 shows the results for RES for Košer. The higher mean ranks represent the higher level of agreement with the statements of negative impact of waste treatment. Low p-values of LQ1, LQ4 and LQ5 indicate that there are significant differences in residents' opinions on waste management consequences on LQ. Specifically, they consider that it would have a negative impact on the life conditions of the local population, the groundwater and flora and fauna. The highest mean ranks are those of residents of Sutivan and Selca, which indicates that those residents have the most negative opinion of waste management consequences on LQ regarding the location Košer.

Table 3: Results of Kruskal Wallis Test for place of residence (location Košer)

Place of residence (RES)	*N*	*Mean Rank LQ1*	*Mean Rank LQ2*	*Mean Rank LQ3*	*Mean Rank LQ4*	*Mean Rank LQ5*
Bol	47	177.76	176.35	195.20	204.69	183.73
Milna	33	169.06	187.35	147.80	126.59	146.58
Nerežišća	23	200.02	209.98	210.24	214.20	208.35
Postira	46	182.66	222.42	208.41	206.36	200.98
Pučišća	60	174.33	181.23	193.37	167.65	168.70
Selca	51	224.33	217.51	219.77	237.01	229.75
Supetar	111	215.17	202.35	200.42	205.28	215.26
Sutivan	24	228.69	177.13	194.94	210.83	206.96
Total	395					
Chi-Square		14.757	8.487	9.623	25.809	19.082
p-value		0.039	0.292	0.211	0.001	0.008

As for age group, for location Košer, it can be seen that PLSG2 shows significant differences in residents' opinions. The higher mean ranks represent the higher level of agreement with the statements of positive impact on PLSG. Therefore, it can be concluded that the residents, especially those aged 31-35 and above 60 years old, consider that the waste treatment would significantly contribute to the public and local self-government budget (Table 4).

Table 4: Results of Kruskal Wallis Test for age group (location Košer)

Age group	*N*	*Mean Rank PLSG1*	*Mean Rank PLSG2*
18-25	65	185.47	159.77
26-30	45	189.04	170.39
31-35	64	217.91	222.84
36-40	42	202.14	209.54
41-50	63	204.30	204.07
51-60	64	201.59	197.92
>60	52	181.50	222.53
Total	395		
Chi-Square		4.712	16.816
p-value		0.581	0.010

For location Kupinovica, there are significant differences in residents' opinions on waste management consequences on PLSG for RES (Table 5). It can be concluded that the residents consider that the waste treatment would significantly contribute to the public and local self-government budget, especially residents of Nerežišća, followed by those from Supetar. This is an interesting finding, since Kupinovica is near Supetar, and recent events have shown that the residents of Supetar strongly disagree with Kupinovica being the only waste disposal on the island of Brač.

Table 5: Results of Kruskal Wallis Test for place of residence (location Kupinovica)

RES	*N*	*Mean Rank PLSG1*	*Mean Rank PLSG2*
Bol	47	179.66	166.85
Milna	33	216.80	186.48
Nerežišća	23	214.59	243.28
Postira	46	167.74	185.23
Pučišća	60	187.60	202.18
Selca	51	212.23	182.38
Supetar	111	201.06	215.74
Sutivan	24	231.79	196.58
Total	395		
Chi-Square		9.920	12.492
p-value		0.193	0.085

4 CONCLUSION

This paper explores waste management consequences on the island of Brač, Croatia, through the influence of the residents' demographic characteristics on life quality and public and local self-government budget for Košer and Kupinovica waste landfills.

There is not much research of the social component, such as demographic characteristics, as key influential factors in waste management. It is assumed that demographic characteristics of residents affect waste management consequences on the island of Brač. For the purpose of the research, a survey questionnaire was designed and applied on a sample of 395 residents of Brač. The research hypothesis is confirmed for some of demographic characteristics. Structural equation models indicated the significant influence of the place of residence on the perception of consequences on life quality for Košer and the perception of public and local self-government budget for Kupinovica. Age group of the residents significantly affects the public and local self-government budget perception for Košer. This research gives a new perspective on waste management through the residents' personal perception about waste management consequences. It is interesting, especially for decision makers, to find out how the residents perceive certain waste disposal consequences and alternatives in order to find the best alternative, not only from economic point of view, but also from the social point of view. Future research could include extended structural models, determining the residents' opinions about recycling and home waste disposal, as well as economic and technological aspect of waste management.

References

[1] Aleksić, A., 2011. Upravljački pristup izboru optimalne metode obrade komunalnog otpada na primjeru Splitsko-dalmatinske županije/Managerial approach to choosing the optimal method of municipal waste treatment on the example of the Split-Dalmatia County. Master thesis, Faculty of Econimics, Business and Tourism Split.

[2] Camilleri-Fenecha, M., Oliver-Solà, J., Farreny, R., Gabarrell, X. 2018. Where do islands put their waste? – A material flow and carbon footprint analysis of municipal waste management in the Maltese Islands. Journal of Cleaner Production, 195: 1609-1619.

[3] Chen, M. C., Ruijs, A., Wesseler, J. 2005. Solid waste management on small islands: the case of Green Island, Taiwan. Resources, Conservation and Recycling, 45(1): 31–47.

[4] Council Directive 75/442/EEC of 15 July 1975 on waste. 1975. European Economic Community (EEC).https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31975L0442&from=EN [Accessed 10/06/2019].

[5] Giusti, L. 2009. A review of waste management practices and their impact on human health. Waste Management, 29(8): 2227-2239.

[6] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. 2010. Multivariate Data Analysis. New York: Pearson Prentice Hall.

[7] Hooper, D., Coughlan, J., Mullen, M. R. 2008. Structural Equation Modelling: Guidelines for Determining Model Fit. The Electronic Journal of Business Research Methods, 6(1): 53–60.

[8] Izvješće o utvrđivanju rezultata glasovanja lokalnog referenduma/The report on determining the results of a local referendum. 2019. Local referendum committee of the city of Supetar. https://gradsupetar.hr/attachments/article/1125/izvje%C5%A1%C4%87e%20referendum.pdf [Accessed 11/06/2019].

[9] Kemeter, D. 2013. Održivo gospodrenje otpadom u Republici Hrvatskoj/Sustainable waste management in Croatia. Zbornik radova Međimurskog veleučilišta u Čakovcu, 4(2): 75-82.

[10] Kline, R. B. 2011. Principles and Practice of Structural Equation Modeling. New York: The Guilford Press.

[11] Mohee, R., Mauthoor, S., Bundhoo, Z. M. A., Somaroo, G., Soobhany, N., Gunasee, S. 2015. Current status of solid waste management in small island developing states: A review. Waste Management, 43: 539–549.

[12] Santamarta, J. C., Rodríguez-Martín, J., Paz Arraiza, M., López, J. V. 2014. Waste Problem and Management in Insular and Isolated Systems. Case Study in the Canary Islands (Spain). IERI Procedia, 9: 162-167.

[13] Skordilis, A. 2004. Modelling of integrated solid waste management systems in an island. Resources, Conservation and Recycling, 41(3): 243–254.

[14] Šemanović, M. 2019. Analiza lokacijskih čimbenika u kontekstu gospodarenja otpadom na otoku Braču/Analysis of location factors in the context of waste management on the island of Brač. Master thesis, Faculty of Economics, Business and Tourism Split.

[15] Taber, K. S. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education, 48(6): 1273–1296.

[16] Waste management in the EU: infographic with facts and figures. 2018. European Parliament. http://www.europarl.europa.eu/news/en/headlines/society/20180328STO00751/eu-waste-management-infographic-with-facts-and-figures [Accessed 10/06/2019].

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 3:

Finance and Investments

OPTIMIZATION OF TRANSITION RULES BASED ON CLAIM AMOUNTS IN A BONUS-MALUS SYSTEM

Kolos Csaba Ágoston
Department of Operations Research and Actuarial Sciences,
Corvinus University of Budapest, H-1093, Fővám tér 8., Budapest, Hungary,
Institute of Economics, Research Centre for Economic and Regional Studies,
Hungarian Academy of Sciences, H-1097, Tóth Kálmán u. 4., Budapest, Hungary,
Email: kolos.agoston@uni-corvinus.hu
Márton Gyetvai
Institute of Economics, Research Centre for Economic and Regional Studies,
Hungarian Academy of Sciences, H-1097, Tóth Kálmán u. 4., Budapest, Hungary,
Department of Operations Research and Actuarial Sciences,
Corvinus University of Budapest, H-1093, Fővám tér 8., Budapest, Hungary,
Email: gyetvai.marton@krtk.mta.hu
László Kovács
Department of Computer Science,
Corvinus University of Budapest, H-1093, Fővám tér 8., Budapest, Hungary,
Email: laszlo.kovacs2@uni-corvinus.hu

Abstract: Bonus-malus systems are well-known actuarial techniques to reduce the welfare loss caused by Adverse Selection. In a bonus-malus system, the insurer classifies its policyholders into a finite number of classes based on their claim history. The policyholders' payment depends on the class, which they are currently classified into. In this paper, we investigate bonus-malus systems where classification is based only on the claim amount. To determine the optimal premiums, we use simulated annealing. Results show the claim amount based bonus-malus system can reduce the welfare loss better if the difference between the risks of the policyholders is high enough.

Keywords: Adverse Selection, Bonus-malus system, Simulated annealing

1 Introduction

Bonus-malus (BM) systems are well-known actuarial techniques, to incite the insured people to be more cautious and distinguish riskier ones from less risky, at the same time. A BM system has finite number of classes, where the insurer classifies policyholders based on the claims reported in the current period and classification of the previous period. Therefore, policyholders with great losses pay more premiums and policyholders without losses (or fewer losses) pay fewer premiums.

At the end of each period (usually years), the insurance companies reclassify their policyholders into new classes. The transition rules determine this reclassification.

In practice — most often — the transition rules are based on the number of claims. In the Hungarian Third-party vehicle insurances, for instance, the policyholders without claims improve their classification by one, while they deteriorate their class by two per claim. Policyholders with four or more claims get straight into the worst BM class. As we see, the size of the claims does not affect the transition rules.

This is reasonable since empirical studies suggest that 'good' and 'bad' policyholder differs more in the probability of the number of claims than in amount (assuming that there is at least one claim). Despite the fact that differences in the number of claims are more significant than differences in the claim amounts, we can observe deviations in the (conditional) amount of claims as well.

We think that it is worth an investigation to see what happens if we consider a BM system, where the transition rules depend only on the size of claims. Undoubtedly, the best choice would be if the number, as well the size of claims would affect the transition rules, but this investigation would go beyond the limits of this article.

2 Preliminaries

Among the policyholders, there are I disjoint groups. The (aggregate) claim amount is described with random variable L^i for group i, which differs in each group. We denote ψ^i as the ratio of the group i among all of the policyholders.

The insurance company knows the size and distributions of each group, but cannot distinguish these. Hence, the insurer does not know which group, a specific policyholder, belongs to. It is the common framework of the models focusing on adverse selection (see for instance [8]).

We have a BM system with K classes. Class 0 is the worst one (i.e. it has the highest premium) and class K is the best one (i.e. where the premium is the lowest). We denote the premium of class k with π_k.

We define a discrete stochastic process X_t^i as the classification of a policyholder from group i at the time period t. The X_{t+1}^i depends on the classification and on the claim amount of the current period t. So, in other words, the process X_t^i satisfies the Markov-property, hence the classification of policyholders is a Markov chain.

The transition rules define how the classification of policyholders change from time period t to $t+1$. The insurance company is not able to distinguish the groups, so the transition rules are the same for every policyholder.

Let the number of classes in the BM system be K. In order to consider the transition rules based on the claim amount, we define K breakpoints, for every classes k: $\ell_1^k > \ell_2^k > \cdots > \ell_K^k$. The transition rules are based on these breakpoints: if a policyholder is assigned to class k and its claim amount between ℓ_m^k and ℓ_{m+1}^k, the policyholder will be transitioned to the class m in the next period. If the claim amount of the policyholder is higher than ℓ_1^k, she/he gets into class 0; if it is less than ℓ_K^k, then the policyholder gets into class K.

In this case, for the transition rules, we have to find K^2+K optimal breakpoints, that can be quite large (in Hungarian BM system, for instance, there are 15 classes). We can reduce the number of breakpoints if these do not differ in the classes. Accordingly, we define $2K+1$ breakpoints $\ell_{-K} > \ell_{-(K-1)} > \cdots > \ell^{-1} > \ell_0 > \ell_1 > \cdots > \ell_K$.

In this case, if the claim amount is between ℓ_m and ℓ_{m+1}, the policyholder moves m classes upward (if $m < 0$ actually it will be a downward move). Surely, the policyholders cannot move higher than class K or lower than class 0, hence in these cases, the policyholder will get to (or remain in) class K or class 0 respectively. For instance, suppose that a policyholder is currently in class 2 and it should move 3 class downward, which would mean, that the policyholder should go to the (non-existing) class -1. In this case, the policyholder will be classified to class 0, so the total number of downward steps will only be two.

In the second approach, we can reduce further the number of breakpoints. We can consider breakpoints $\ell_{-D} > \ell_{-(D-1)} > \cdots > \ell^{-1} > \ell_0 > \ell_1 > \cdots > \ell_U$ with $U, D < K$. Thus the policyholder can move downward at most D classes, and upward at most U classes.

In both of the previously shown approaches, the policyholder can move upward and downward one class, with positive probability which ensures the irreducibility of this Markov-chain. Besides, because of the finite number of classes with these transition rules, the Markov-chain is aperiodic as well.

Let $c^i_{k,t}$ the probability of a policyholder from group i is classified into the class k at the period t. Irreducible and aperiodic Markov-chains converge to a unique $c^i_0, \ldots, c^i_K$ probabilities (regardless of where the policyholder starts), which are called as stationary probabilities.

BM systems are usually analysed with these stationary probabilities (see [4], [5], [7],[2]).

3 Optimization process

The goal of our optimization is to determine the optimal breakpoints and premiums. The purpose of the BM system is to decrease the welfare loss caused by adverse selection.

The perfect outcome would be if each policyholder would pay their expected claim amount value ($\mathbb{E}(L^i)$). However, in the investigated BM system this situation is unreachable. Therefore, our goal is to get as close to the perfect situation as possible. In order to achieve this, we calculate the absolute deviation of premium and expected claim amount in each group and class. Then, we minimize the weighted sum by the size of groups of these deviations. We denote g^i_k the absolute deviation for the group i in the class k.

We must also take into account that the BM system should not cause loss to the insurance company, thus the expected premiums have to be at least as large as the expected claim amounts.

With fixed breakpoints, and with an assumption of the policyholders' claim amount distribution, we can calculate the stationary probabilities c^i_k. After we determined the stationary probabilities, then we can calculate the optimal premiums (the idea appears in [3]):

$$
\begin{aligned}
& \sum_{i=1}^{I} \psi^i \sum_{k=0}^{K} g^i_k \to \min \\
s.t. \quad & \\
& c^i_k \pi_k + g^i_k \geq c^i_k \mathbb{E}(L^i)\,, && \forall i,k \\
& c^i_k \pi_k - g^i_k \leq c^i_k \mathbb{E}(L^i)\,, && \forall i,k \\
& \pi_{k+1} \leq \pi_k\,, && k = 0,\ldots,K-1 \\
& \sum_{i=1}^{I} \psi^i \sum_{k=0}^{K} c^i_k \pi_k \geq \sum_{i=1}^{I} \psi^i \mathbb{E}(L^i) \\
& g^i_k, \pi_k \geq 0\,, && \forall i,k
\end{aligned}
\tag{1}
$$

To solve this linear programming (LP) model, we have to calculate the stationary probabilities, that depends on the breakpoints. We generate (randomly) these breakpoints, then we optimize the LP with the calculated stationary probabilities.

Let's assume a BM system with 2 classes. We use the second approach for the transition rules, thus we have to determine only two breakpoints. We assume that there are two equal sized groups among the policyholders.

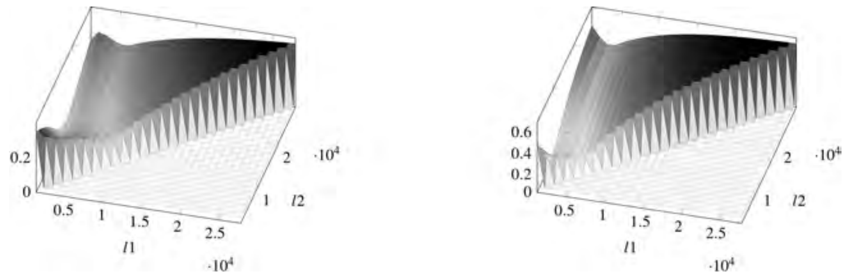


Figure 1: The expected value of absolute deviation in the case of two groups, depending on breakpoints

Figure 1 shows the values of the objective function of the (1) LP model with changing breakpoints. At the left one, we assumed 6200- and 9300-exponential distributions and at the right side, these results come from a 3500- and 7400-exponential distributions. In both cases, the value of the objective function gives a non-convex, smooth surface with local minima.

If we have more classes, then the dimension of the optimization problem increases quadratically. If we use a Monte Carlo type method, we will need a very large number of random points, which is not fortunate.

Using a classical optimization tool, such as the Newton-Raphson method, is also not a suitable solution, because we expect multiple local optima and the surface is not convex. Therefore, we use simulated annealing which is more suitable for problems such as these.

4 Simulated annealing

Simulated annealing is a very simple heuristical search algorithm (see for instance [1]). It first appeared in [6]. It starts from a random solution as the actual solution: s_{act}. In each iteration the algorithm examines a neighbour of s_{act}. If the chosen neighbour is more optimal regarding the aim function, we consider that the new s_{act}. In the following iterations, we examine the neighbours of the new actual solutions.

However, in order to avoid local extremes of the aim function, we can also change s_{act}, if the Boltzmann condition can be accepted: $e^{(-\frac{\delta}{T})} > random(0,1)$. Where δ is the difference between s_{act} and the neighbour in the aim function and T is a parameter that controls the maximum number of iterations of the algorithm. The Boltzmann condition basically means that, if δ is not really high (aim function of the neighbour is not much worse than s_{act}), then we can accept the neighbour as the new s_{act} with a high probability, even if it is less optimal than s_{act}. With this extra condition in the search process, we can escape local extremes that are present in our non-convex surfaces as seen in Figure 1.

The parameter T is called the temperature parameter and it controls the annealing schedule. The starting value for T is given as a parameter and it continuously decreases in each iteration of the algorithm until it reaches a minimum temperature given also as a parameter. If the minimum temperature is reached, the algorithm terminates. We can also specify a time parameter which gives the number of iterations to use the same T. This way the decrease of T is not continuous but monotonous. Together with the Boltzmann condition, the annealing schedule controlled by T can result that the probability for accepting a worse solution as s_{act} is high at the first few iterations and relatively low at the last iterations when we need the formation of a stable convergence.

This algorithm is very similar to annealing in metallurgy, a technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. That is why the algorithm is named simulated annealing. In our applications, we chose the starting value of T as 0.001 and the minimum temperature as 0.0001. Time spent on each T is 10 iterations and T decreases by 0.0001 in every 10 iterations. This way, if we have constant δ of 0.001, the probability of accepting a worse neighbour as s_{act}) decreases from 0.3679 to 0.00005 during the iterations.

5 Numerical results

For calculating the numerical results, we used an Intel Core i5-7300HQ CPU 2,50 GHz computer, with 8 GB DDR3 RAM. We run the program in Python 3.7.3. and we used the Cbc 2.10. solver for the optimization of the (1) model.

In our numerical examination, we assumed two risk groups with different expected claim amounts with exponential distribution. We calculated a BM system with four classes, and we used the second type approach, therefore, in this case, we only need to determine three breakpoints.

$\mathbb{E}(L_1)$	$\mathbb{E}(L_2)$	Best Obj.	SD of Obj.	π_0	π_1	π_2	π_3	OP_1	OP_2
500	550	22.15	0.0091	550	550	500	500	4.88%	-3.62%
500	600	23.13	1.8348	600	500	500	500	9.29%	-2.97%
500	700	38.11	0.0363	700	500	500	500	5.61%	-3.99%
500	900	45.83	1.2438	900	500	500	500	10.91%	-4.12%
500	1200	39.05	3.1273	1200	1200	500	500	7.83%	-3.25%
500	1600	21.78	0.4441	1600	1600	500	500	4.36%	-1.36%
500	2100	10.83	0.0671	2100	2100	500	500	2.17%	-0.52%
500	3100	3.90	0.0005	3100	3100	500	500	0.78%	-0.13%
500	5100	1.14	0.0004	5100	5100	500	500	0.23%	-0.02%
500	9100	0.30	0.0024	9100	9100	500	500	0.06%	0.00%

Table 1: Results of exponential distributions.

The first two columns show the parameters of the two groups, the two adjacent the best objective value and the standard deviations of the objective values, then the premiums of each class, and lastly the overpayment (OP) of each group

The results of ten different models are shown in Table 1. In each case, the optimal premium scale is staircase type with only two steps. The lower premium always equals the expected claim amount of the less risky group while the higher premium is consistently matched with the same parameter of the riskier one.

The simulated annealing always stopped after 470 seconds. The standard deviations of the objective values are relatively small, as seen in the fourth column.

The columns OP shows the overpayment of the risk groups, compared to the expected claim amount. The less risky group is always paying more than their real risk, while the riskier almost

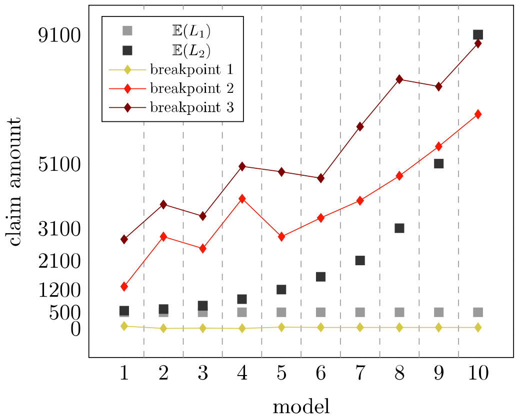


Figure 2: The changes of the breakpoints in the models.

always pay less. However, the rate of the overpayment becomes smaller if the difference between the groups' parameters is higher. Thus, the BM system can handle the Adverse Selection in a better way if the difference between the risks of the policyholders is high.

The breakpoints of the models can be seen in figure 2. The lowest breakpoint always stays close to zero, however, the other two higher breakpoints grow together with the difference in parameters. The two higher breakpoints in most cases are over the higher-risk policyholder's parameter, except the last case, where the difference between the policyholders the highest.

References

[1] Datta, S., Roy, S. and Davim. J. P. (2019). Optimization Techniques: An Overview. In Datta, S. and Davim. J. P. (Eds.). Optimization in Industry Present Practices and Future Scopes (pp. 1–11). Cham, Switzerland : Springer.

[2] Gyetvai, M. and Ágoston, K. Cs. (2018). Optimization of transition rules in a Bonus-Malus system. Electronic Notes in Discrete Mathematics, Vol(69): 5-12.

[3] Heras, A. T., Gil, J. A, García-Pineda, P. and Vilar, J. L. (2004). An Application of Linear Programming to Bonus Malus System Design. ASTIN Bulletin: The Journal of the IAA, Vol(34(2)):435-456.

[4] Kaas, R., Goovaerts, M., Dhaene, J. and Denuit, M. (2001). Modern Actuarial Risk Theory Using R. Berlin Heidelberg: Springer.

[5] Kafkova, S. (2015). Comparison of Several Bonus Malus Systems. Procedia Economics and Finance, Vol(26): 188–193.

[6] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. Science, Vol(220(4598)): 671-680.

[7] Loimaranta, K. (1972). Some asymptotic properties of bonus systems. ASTIN Bulletin: The Journal of the IAA, Vol(6(3)): 233-245.

[8] Rothschild, M. and Stiglitz, J. (1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. The Quarterly Journal of Economics, Vol(90(4)): 629-649.

CO-MOVEMENTS OF EXCHANGE RATE RETURNS: MULTIVARIATE GARCH APPROACH

Michaela Chocholatá
University of Economics in Bratislava, Department of Operations Research and Econometrics
Dolnozemská cesta 1, 852 35 Bratislava, Slovakia
E-mail: michaela.chocholata@euba.sk

Abstract: This paper deals with analysis of co-movements of exchange rate returns of the selected Central European currencies (the Czech koruna, Hungarian forint and Polish zloty) to the European euro based on the VAR(6)-BEKK-GARCH(1,1) model. The empirical analysis is based on daily data from May 2004 to April 2019 and enables to study the dynamic character of relationships both in the first and the second moment. The results confirmed the return spillover effects and volatility co-movements among analysed currency markets. Considerably dynamic character of conditional correlations during the analysed period indicated the strongest linkages during the post-entry period to the European Union 2005 – 2006, during the global financial crisis period 2008 – 2009 and during worsening of the euro area sovereign debt crisis in 2011.

Keywords: exchange rate returns, volatility, spillover, conditional correlation, MGARCH model

1 INTRODUCTION

The approaches to modelling of exchange rate returns are relatively broadly developed in the literature. Taking into account not only the level of exchange rate returns, but also their volatility, it is suitable to use the popular univariate ARCH (autoregressive conditional heteroscedasticity) class methodology. However, modelling of returns' volatility based on univariate ARCH class models [4] is not sufficient to assess the co-movement of multiple exchange rate returns. In addition to features such as volatility clustering or fat tails taken into account in univariate modelling, it is also necessary to consider the multivariate analysis of time-varying covariance/correlation movement. Therefore, the generalization of univariate ARCH models to their multivariate version of MGARCH (Multivariate Generalised ARCH) plays an important role. In recent years, various versions of MGARCH models were published (for a detailed survey see e.g., [2]). Bollerslev, Engle and Wooldridge [5] extended the univariate GARCH model to its multivariate version denoted as VECH (Vectorized GARCH), very popular are also the Baba-Engle-Kraft-Kroner (BEKK) model defined in Engle and Kroner [8], CCC (Constant Conditional Correlation) model of Bollerslev [3] and DCC (Dynamic Conditional Correlation) model of Engle [7].

The most common application of MGARCH models is the analysis of stock market volatility and verifying the impact of different crises on their development as well as assessing the contagion effect[1]. However, the issue of common volatility movements and the volatility spillover between exchange rate returns is considerably less addressed in the literature. Studies analysing the volatility spillovers between exchange rates are beginning to emerge in the 1990s. Bollerslev [3] applied the aforementioned CCC model to the European currency exchange rates of German mark, French franc, Italian lira, Swiss franc and British pound against the US dollar and revealed a common movement of volatility of these five exchange rates. Kearney and Patton [10] used the BEKK model to analyse the volatility co-movement

[1] In the literature there are various definitions of contagion, e.g., Forbes and Rigobon [9] distinguish the stock market co-movement during the stability periods and during the post-crisis (post-shock) periods. They use the term contagion in case of „a significant increase in cross-market linkages after a shock to one country (or group of countries)“ (p. 2223). On the other hand, they speak about interdependence if the co-movement does not increase significantly after a shock or crisis.

between the selected European currencies and the US dollar. Antonakakis [1] based on the BEKK and DCC models confirmed the common volatility movement as well as volatility spillovers among the exchange rates of European euro, British pound, Japanese yen and Swiss franc against the US dollar. Výrost [12] applied the DCC model to exchange rates of Visegrad 4 countries (Slovakia, Czech Republic, Hungary and Poland) against the euro and confirmed the highest correlations between the Polish zloty and Hungarian forint vis-à-vis the European euro.

The aim of this paper is to analyse the co-movements of exchange rate returns of the selected Central Eastern European (CEE) currencies, namely the Czech koruna (CZK), Hungarian forint (HUF) and Polish zloty (PLN) to the European euro (EUR) based on the VAR(6)-BEKK-GARCH(1,1) model. The analysis is based on daily data covering the 15 years' period of membership of these countries in the EU (May 2004 – April 2019).

The paper is organised as follows: section 2 deals with the methodology of BEKK model, section 3 describes the data and estimation results and section 4 concludes.

2 METHODOLOGY – BEKK MODEL

Modelling of currency markets co-movements consists of two parts – specification of the conditional mean equation and specification of the appropriate MGARCH model. Regarding the empirical part of the paper, from the wide variety of MGARCH models we will briefly present the trivariate BEKK model[2].

The conditional mean equation capturing the dynamic relationship in returns can be characterised in general as a vector autoregressive (VAR) model with *k* lags:

$$\mathbf{r_t} = \boldsymbol{\omega} + \sum_{i=1}^{k} \mathbf{\Gamma_i r_{t-i}} + \boldsymbol{\varepsilon}_\mathbf{t} \tag{1}$$

where $\mathbf{r_t}$ is a three-dimensional vector of daily exchange rate returns, $\boldsymbol{\omega}$ is a three-dimensional vector of constants, $\mathbf{\Gamma_i}$ ($i=1,2,\dots k$) are matrices of parameters of dimension (3×3) and $\boldsymbol{\varepsilon}_\mathbf{t}|\Omega_{t-1} \sim N(0,\mathbf{H_t})$ is a vector of disturbances conditional on information at time $t-1$ (Ω_{t-1} represents the information set at time $t-1$).

The specification of the conditional variance-covariance matrix $\mathbf{H_t}$ (3×3) in the BEKK-GARCH(1,1) model (matrix $\mathbf{H_t}$ can be received through the generalization of the univariate GARCH model) is defined as follows:

$$\mathbf{H_t} = \mathbf{C'C} + \mathbf{A'}\boldsymbol{\varepsilon}_\mathbf{t-1}\boldsymbol{\varepsilon}'_\mathbf{t-1}\mathbf{A} + \mathbf{B'H_{t-1}B} \tag{2}$$

where **C** denotes a positive definite (3×3) dimensional upper triangular matrix of parameters, **A** and **B** are (3×3) matrices of parameters[3]. In the diagonal version of the BEKK model the

[2] Nowadays, the most often used models for modelling of the conditional covariances and correlations are the BEKK model and the DCC model. However, some studies (see e.g., Caporin and McAleer [6]) proved that it is not possible to judge in general about which one of these models to prefer. The diagonal version of the trivariate BEKK model is used in the empirical part of the paper since the conditional variance-covariance matrix of this model is guaranteed to be positive definite. Diagonal BEKK model can be considered as a restricted version of the diagonal VECH model and it furthermore requires the estimation of fewer parameters than the diagonal VECH model.

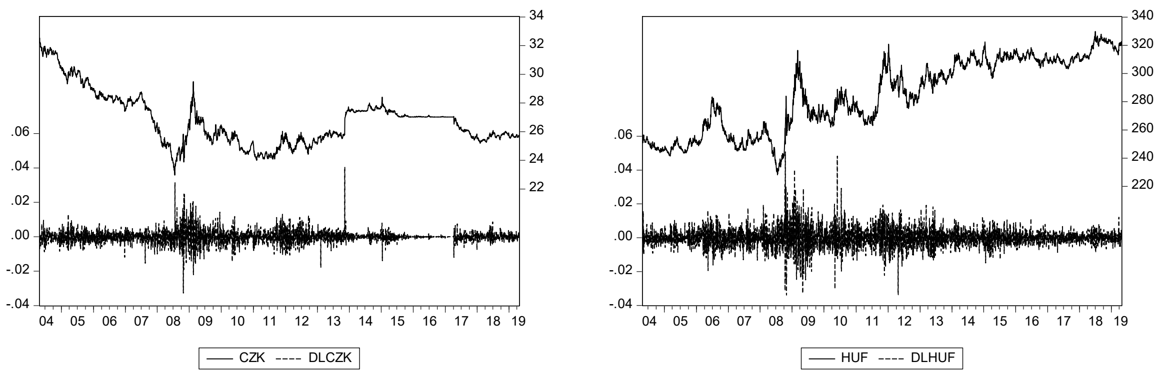
[3] Since the second and the third term of the right-hand-side in equation (2) are expressed in quadratic forms, the positive definiteness of the matrix $\mathbf{H_t}$ is ensured, so there are no additional constraints for parameter matrices **A** and **B**.

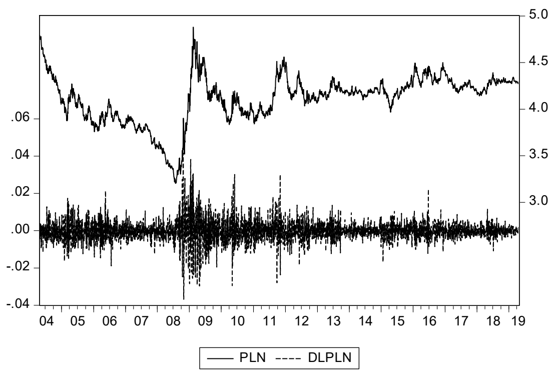
matrices **A** and **B** are diagonal[4]. The coefficients of the matrices **A** and **B** reflect the impact of news in each individual market and the persistence of news in each individual market, respectively [1]. Parameters of the presented model can be estimated by the maximum likelihood method [2].

3 DATA AND EMPIRICAL RESULTS

The analysis in this paper is based on the daily data of exchange rates CZK/EUR, HUF/EUR and PLN/EUR for the period May 3, 2004 – April 30, 2019 (totally 3841 observations) retrieved from the web page of the European Central Bank [14]. The analysed exchange rates together with the corresponding logarithmic returns (calculated as the difference between the natural logarithms of the exchange rate in time t and t-1)[5] are graphically depicted in Figure 1 which also contains the descriptive statistics of analysed return series together with the Jarque-Bera test statistics testing the normality.

Since the time series of daily exchange rates were clearly non-stationary, the logarithmic returns were already stationary[6]. Concerning the logarithmic return series there is a clear evidence of volatility clustering, i.e. that large/small returns tend to be followed by another large/small returns. Based on descriptive statistics, the mean values of all the logarithmic returns oscillated around the 0, the most volatile were the HUF/EUR returns (0.559 %) followed by the PLN/EUR returns (0.536 %) and CZK/EUR returns (0.346 %). All the analysed return series were positively skewed with higher kurtosis than the normal distribution. The non-normality of the distribution was also confirmed by the values of the Jarque-Bera test statistics.





	DLCZK	DLHUF	DLPLN
Mean	-0.00006	0.000065	-0.000029
Std. Dev.	0.003463	0.005591	0.005362
Skewness	0.597853	0.420559	0.337355
Kurtosis	16.94734	10.51772	10.41492
Jarque-Bera	31353.28	9155.765	8869.801
p-value	0.000000	0.000000	0.000000
Observations	3840	3840	3840

Figure 1: Development of exchange rates and logarithmic return series, descriptive statistics of return series
Source: author's calculations in EViews

[4] The more restricted version of the diagonal BEKK model, the scalar BEKK model, was not considered – detailed information can be found e.g., in [2].
[5] Logarithmic return series are denoted with prefix „DL“.
[6] The results are available from the author upon request.

One of the approaches to analyse the co-movements of the return series is to calculate the pair-wise unconditional Pearson's correlations. The unconditional correlation coefficients for the whole analysed period were as follows: DLCZK – DLHUF 0.4257, DLCZK – DLPLN 0.4614 and DLPLN – DLHUF 0.6334, which indicates the strongest positive linear relationship between the Polish and Hungarian exchange rate returns. Forbes and Rigobon [9] pointed out the limitations connected with these unconditional correlation coefficients which are biased and inaccurate due to heteroscedasticity in return series. To analyse the development of conditional correlations against time, the VAR(k)-BEKK-GARCH(1,1) model was estimated based on non-linear maximum likelihood method assuming the multivariate normal distribution of disturbances. The number of lags k used in the VAR model (1) was specified by the information criteria to be 6 in order to ensure the uncorrelatedness. The estimated mean equations of the trivariate VAR(6)-BEKK-GARCH(1,1) model are as follows:

DLCZK = -0.0024*DLCZK(-1) - 0.0007*DLCZK(-2) **- 0.0404***DLCZK(-3) + 0.0100*DLCZK(-4) - 0.0116*DLCZK(-5) + 0.0019*DLCZK(-6) - 0.0061*DLHUF(-1) + 0.0006*DLHUF(-2) - 0.0171*DLHUF(-3) + 0.0069*DLHUF(-4) **- 0.0194***DLHUF(-5) **- 0.0235***DLHUF(-6) + 0.0114*DLPLN(-1) - 0.0102*DLPLN(-2) + 0.0186*DLPLN(-3) - 0.0066*DLPLN(-4) + 0.0119*DLPLN(-5) **+ 0.0066***DLPLN(-6) **- 0.0001**

DLHUF = 0.0362*DLCZK(-1) + 0.0425*DLCZK(-2) - 0.0048*DLCZK(-3) + 0.0311*DLCZK(-4) - 0.0122*DLCZK(-5) + 0.0162*DLCZK(-6) - 0.0060*DLHUF(-1) **- 0.0436***DLHUF(-2) + 0.0090*DLHUF(-3) + 0.0076*DLHUF(-4) **- 0.0582***DLHUF(-5) - 0.0060*DLHUF(-6) - 0.0219*DLPLN(-1) - 0.0232*DLPLN(-2) **- 0.0481***DLPLN(-3) - 0.0111*DLPLN(-4) + 0.0190*DLPLN(-5) + 0.0111*DLPLN(-6) + $3*10^{-5}$

DLPLN = **0.0540***DLCZK(-1) + **0.0552***DLCZK(-2) - 0.0356*DLCZK(-3) + 0.0198*DLCZK(-4) + 0.0127*DLCZK(-5) + 0.0286*DLCZK(-6) - 0.0087*DLHUF(-1) - 0.0102*DLHUF(-2) **+ 0.0462***DLHUF(-3) + 0.0035*DLHUF(-4) - 0.0137*DLHUF(-5) - 0.0005*DLHUF(-6) + 0.0022*DLPLN(-1) + 0.0011*DLPLN(-2) - **0.0379***DLPLN(-3) - 0.0012*DLPLN(-4) - 0.0131*DLPLN(-5) - 0.0109*DLPLN(-6) – $1*10^{-4}$

Based on the statistically significant parameters (significance level of 5 %, statistically significant parameters denoted in bold) from the above specified mean equations we can conclude, that there exist some return spillovers from the Hungarian and Czech currency markets to Polish market, from both the Hungarian and Polish markets to the Czech market and from the Polish market to Hungarian market.

The parameter estimates of the variance and covariance equations (2) are gathered in Table 1. The estimates indicate that the BEKK-GARCH(1,1) model's parameters were all (with exception of two constant terms from the conditional covariance equations) statistically significant which strongly confirms the adequate use of this model. The highest impact of shocks (given by elements of matrix **A**) was proved for the Czech currency market, followed by Polish and Hungarian currency markets, respectively. The highest persistence of shocks (given by elements of matrix **B**) was identified in the Hungarian currency market, followed by the Polish and the Czech market. The Czech koruna is also the most stable currency. The diagnostic checking of the standardized residuals proved no residual autocorrelation (Cholesky of covariance) – values for 12 and 24 lags, respectively were as follows: 65.3774 and 150.6606, respectively. The Doornik-Hansen multivariate test with the Jarque-Bera test statistic of 23370.77 clearly rejects the hypothesis that the residuals are multivariate normal (for more information about normality conditions see e.g. [6]).

The pair-wise conditional correlations from the estimated VAR(6)-BEKK-GARCH(1,1) model smoothed by the Hodrick-Prescott (HP) Filter[7] together with the corresponding descriptive statistics for the pair-wise conditional correlations (before application of the HP filter) are in Figure 2.

[7] Smoothing parameter of 6812100, for more information about HP filter see e.g., [11].

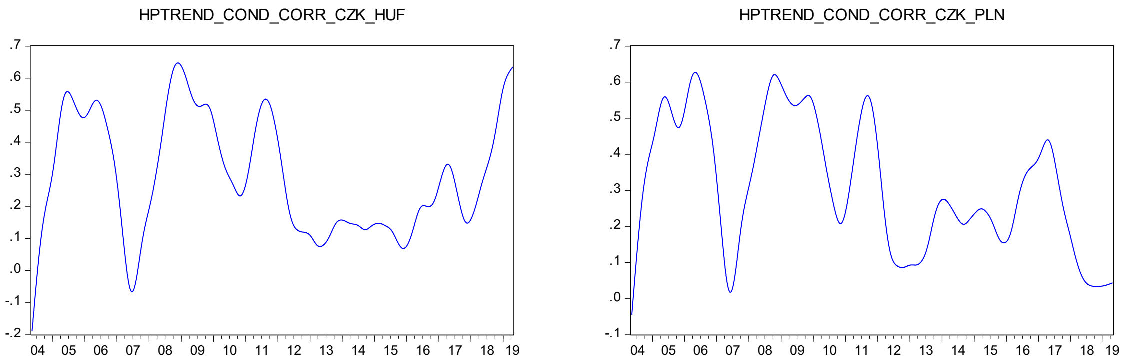
Table 1: Estimates of variance and covariance equations – diagonal BEKK-GARCH(1,1) model

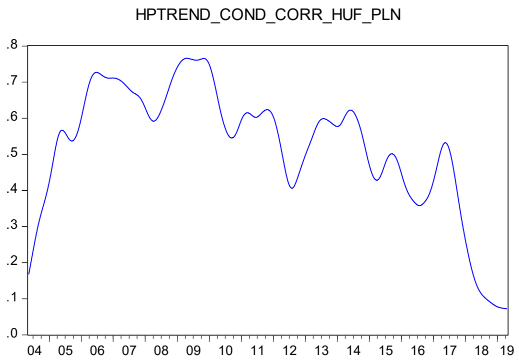
	C(1,1)	C(1,2)	C(1,3)	C(2,2)	C(2,3)	C(3,3)	A(1,1)	A(2,2)	A(3,3)	B(1,1)	B(2,2)	B(3,3)
Value	3.10^{-8}	6.10^{-9}	9.10^{-9}	6.10^{-8}	3.10^{-8}	7.10^{-8}	0.2805	0.2030	0.2229	0.9643	0.9791	0.9740
Prob.	0.0000	0.3023	0.1116	0.0000	0.0042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: CZK/EUR=1, HUF/EUR=2 and PLN/EUR=3

Source: author's calculations in EViews

Taking into account the conditional correlations in Figure 2, it seems to be clear, that among the analysed currency markets, characterized by the CZK/EUR, HUF/EUR and PLN/EUR exchange rate returns, exist significant dynamically changing conditional correlations. Similarly, as in case of unconditional correlations, the conditional correlations were in average the highest for the pair HUF/EUR and PLN/EUR exchange rate returns, followed by the pair CZK/EUR – PLN/EUR and CZK/EUR – HUF/EUR, indicating the strongest relationships between the Hungarian and Polish currency market. These results are in accordance with those of [12] and [13] analysing the integration of the CEE stock markets with the Eurozone market.





	CZK_HUF	CZK_PLN	HUF_PLN
Mean	0.290997	0.313612	0.523145
Median	0.291880	0.314000	0.560808
Max	0.879932	0.898826	0.901239
Min	-0.522826	-0.409715	0.052938
Std. Dev.	0.242647	0.237309	0.203240
Skewness	-0.113335	-0.022705	-0.603838
Kurtosis	2.840065	2.583701	2.471254

Figure 2: Pair-wise conditional correlations smoothed by the HP filter and descriptive statistics for conditional correlations

Source: author's calculations in EViews

Based on the descriptive statistics of conditional correlations it is clear that the differences between the minimum and maximum values were huge in all analysed cases, standard deviations are considerably high as well. Conditional correlations for the pairs CZK/EUR – HUF/EUR and CZK/EUR – PLN/EUR, respectively, had the similar trend till the beginning of 2018, indicating the highest peaks in the EU' post-entry period 2005 – 2006 as well as during the global financial crisis period of 2008 – 2009 and in 2011 (worsening of the euro area sovereign debt crisis). Since the conditional correlations for CZK/EUR – HUF/EUR have had the rising trend since 2018, the values were declining in case of CZK/EUR – PLN/EUR returns. The highest conditional correlations reaching almost the values of 0.9 during the crisis period 2008 – 2009 for the return pair HUF/EUR – PLN/EUR was followed by slightly declining trend during the next analysed period indicating the declining linkages between Hungarian and Polish currency markets.

5 CONCLUSION

This paper was focused on the co-movements of the CZK/EUR, HUF/EUR and PLN/EUR exchange rate returns for the period May 3, 2004 – April 30, 2019. The estimation of the VAR(6)-BEKK-GARCH(1,1) model revealed the existence of the return spillover effects from the Hungarian and Czech currency market to Polish market, from both the Hungarian and Polish market to the Czech market and from the Polish market to Hungarian market. With regard to development of the conditional variance and covariance it was proved that the volatility generated on one currency market was transmitted to the remaining analysed currency markets.

The values of conditional correlations enabled to detect quite strong linkages between the analysed currency markets especially during the 2008 – 2009 crisis period and provide a useful information of the dynamic evolution of the co-movement of the analysed currency markets in time.

Acknowledgement
This work was supported by the Grant Agency of Slovak Republic – VEGA grant no. 1/0248/17 „Analysis of regional disparities in the EU based on spatial econometric approaches “.

References

[1] Antonakakis, N. 2008. Exchange Rate Volatility Comovements and Spillovers before and after the Introduction of Euro: A Multivariate GARCH Approach. http://www.econ.jku.at/members/Department/files/LunchTimeSeminar/NikolasAntonakakis.pdf [Accessed 20/08/2013].
[2] Bauwens, L., Laurent, S., Rombouts, J.V.K. 2006. Multivariate GARCH Models: A Survey. Journal of Applied Econometrics, 21: 79–109.
[3] Bollerslev, T. 1990. Modeling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model. Review of Economics and Statistics, 72: 498–505.
[4] Bollerslev, T. 2009. Glossary to ARCH (GARCH). http://public.econ.duke.edu/~boller/Papers/glossary_arch.pdf [Accessed 20/08/2013].
[5] Bollerslev, T., Engle, R. F., Wooldridge, J. M. 1988. A Capital-Asset Pricing Model with Time-Varying Covariances. Journal of Political Economy, 96(1): 116–131.
[6] Caporin, M., McAleer, M. 2009. Do We Really Need Both BEKK and DCC? A Tale of Two Covariance Models. http://eprints.ucm.es/8590/1/0904.pdf [Accessed 20/02/2013].
[7] Engle, R. F. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business and Economic Statistics, 20: 339–350.
[8] Engle, R., Kroner, F.K. 1995. Multivariate simultaneous generalized ARCH. Econometric Theory, 11: 122–150.
[9] Forbes, K. J., Rigobon, R. 2002. No Contagion, Only Interdependence: Measuring Stock Market Comovements. Journal of Finance, 57(5): 2223–2262.
[10] Kearney, C., Patton, A. J. 2000. Multivariate GARCH Modeling of Exchange Rate Volatility Transmission in the European Monetary System. The Financial Review, 41: 29–48.
[11] Lütkepohl, H. 2004. Univariate Time Series Analysis. In Lütkepohl, H., Krätzig, M. (Eds.). Applied Time Series Econometrics (pp. 8-85). New York: Cambridge University Press.
[12] Výrost, T. 2008. Modelovanie dynamickej podmienenej korelácie menových kurzov V4. http://www3.ekf.tuke.sk/konfera2008/zbornik/files/cena_dekana.pdf [Accessed 17/05/2014].
[13] Wang, P., Moore, T. 2008. Stock Market Integration for the Transition Economies: Time-Varying Conditional Correlation Approach. The Manchester School, 76(1): 116–133.
[14] http://www.ecb.int/stats/eurofxref/eurofxref-hist.zip [Accessed 05/05/2019].

BARRIERS TO INTERNATIONAL TRADE AND EXPORT COMPETITIVENESS OF THE EU NEW MEMBER STATES

Nataša Erjavec
Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: nerjavec@efzg.hr

Boris Cota
Faculty of Economics and Business, University of Zagreb, Croatia, Department of Macroeconomics and Economic Development
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: bcota@net.efzg.hr

Saša Jakšić
Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: sjaksic@efzg.hr

Abstract: The aim of this paper is to analyse the relationship between trade reform and export competitiveness for a set of selected EU new member states in the period from 2000 to 2016. In particular, the dataset included countries at different levels of development and that had their national currencies in most of the analysed period. Furthermore, the analysed period was very turbulent for the analysed countries considering that it included a period prior to the EU accession, as well as the economic and financial crisis. For that reason, the performed empirical analysis should contribute to a better understanding of the importance of fostering free trade and removing barriers to trade for promoting export competitiveness.

Keywords: trade reform, exports, EU, new member states, freedom to trade internationally, export competitiveness.

1 INTRODUCTION

As free trade is one of the EU's founding principles, all countries that are EU members or on the way of becoming a member, committed themselves to reduce barriers to trade. There are numerous possible barriers to trade ranging from tariffs, quotas, regulatory trade barriers (non-tariff) to various controls of movement of capital and people.

Trade barriers affect export competitiveness in various ways. When a country imposes trade barriers this reduces exports of other countries. However, this also indirectly leads to a decrease in export competitiveness of a country that imposed trade barriers through several channels [5]. First of all, barriers to trade reduce competition, which can have a negative impact on innovation and productivity. Secondly, restrictions on imports also limit firms' access to foreign technology and inputs that can increase productivity and provide a competitive advantage to exporters. And, last but not least, imposing import restrictions restrains firms from taking advantage of participating in the global supply chain.

Generally, the literature stresses the importance of structural reforms for macroeconomic performance [3], [7] and export performance in particular [5], [6]. Among the structural reforms, this article looks at the impact of trade reform (reducing barriers to international trade) on export competitiveness.

As a suitable indicator of the countries' incentives in reducing the trade barriers, this paper focuses on the freedom to trade internationally, the essential component of economic freedom index [4] calculated by Fraser Institute.

The article analyses the relationship between trade reform and exports of the selected eastern European countries that are now a part of the EU. These countries also known as the New Member States (NMS). To be precise, the paper considers six countries from the 2004 enlargement wave (Czech Republic, Poland, Hungary, Estonia, Latvia and Lithuania), and countries that joined the EU afterward: Romania and Bulgaria (2007) and Croatia (2013). Two countries from the 2004 enlargement wave (Slovenia and Slovakia) were not considered because, in contrast to the analysed countries, these countries adopted euro in a substantial part of the analysed period. Namely, Slovenia and Slovakia joined the euro zone, in 2007 and 2009, respectively. Since then, they have followed a common monetary policy and are unable to stimulate their exports through exchange rate changes for almost half of the analysed period.

The main contribution of the article is in analysing the impact of trade reforms for a set of eastern European countries that joined the EU in different accession periods and are therefore in different stages of transition and levels of development.

The rest of the paper is organized in the following manner. Section 2 provides a description of the analysed data sets. The model employed in the empirical analysis and the results of the estimated model is presented and discussed in section 3. Finally, section 4 concludes.

2 DATA SET

The empirical analysis is performed for a cross-section of selected EU NMS ($\boldsymbol{N = 9}$). The selection of the analysed period depended on the data availability. Data on Freedom to Trade Internationally are reported regularly on a yearly basis starting from 2000. Prior to 2000, data were reported on a five year basis. Furthermore, as the latest data used for calculating the Economic Freedom index in the 2018 edition of Economic Freedom of the World is from 2016, the dataset applied in the empirical analysis spans from 2000 to 2016 ($\boldsymbol{T = 17}$).

Variables selected for the empirical analysis are the share of country's export in total world exports (EX) as a dependent variable while independent variables are Freedom to trade internationally (FREE), a real effective exchange rate (REER) and unit labor cost (ULC).

Variable EX is a measure of a country's competitiveness calculated as the country's share in world exports. The data source is the Direction of Trade Statistics database (DOTS) of the International Monetary Fund (IMF).

Variable FREE stands for Freedom to Trade Internationally, a component of the Economic Freedom index which measures the degree of economic freedom in several areas and is calculated by the Fraser Institute. The components of variable FREE that affect international trade are tariffs, quotas, administrative restraints, controls on the exchange rate and the movement of capital. The variable FREE is measured on a scale of 0 to 10, with 0 being the lowest, while 10 is the highest score. The higher rating of this variable means that country has low tariffs, fewer controls on the movement of physical and human capital, efficient administration of customs or freely convertible currency.

Variable REER is the real effective exchange rate. It is included in order to assess the country's price (or cost) competitiveness relative to its principal competitors in international markets. The data source is AMECO (the annual macro-economic database of the European Commission's Directorate General for Economic and Financial Affairs). The indicator is deflated by the price index (total economy) against a panel of 42 countries (EU28 and 14 other industrial countries: Australia, Canada, United States, Japan, Norway, New Zealand, Mexico, Switzerland, Turkey, Russia, China, Brazil, South Korea, and Hong Kong). A rise in the index means real appreciation (a loss of competitiveness).

Variable ULC (unit labor cost) measures the average cost of labor per unit of output. It is calculated as the ratio of labor costs to labor productivity. A decrease in the relative unit labor cost index is regarded as an improvement of a country's competitive position relative to their trading partners. The data source is AMECO.

All analysed variables are expressed in logs. The impact of the 2008 world economic and financial crisis was also considered. However, the crisis dummy variable turned out to be statistically insignificant which is not surprising as the crisis affected both the numerator (countries' exports) and the denominator (world exports) of the dependent variable EX.

Figures 1 and 2 depict the values of the analysed variables for the beginning (2000) and the end (2016) of the analysed period. As unit labor cost and real effective exchange rate primarily act as control variables, the descriptive statistics focus on Freedom to trade internationally as a proxy for trade reform incentives and share of countries' exports in world exports as a proxy for countries competitiveness.

Although the reported figures are just simple snapshots as they neglect the dynamics in between, both figures enable fundamental insights into the analysed countries. Namely, although Figure 1 indicates that Poland and Czech Republic have a substantially larger share of exports in world exports, their share is just slightly above 1%. Basically, this means that all analysed countries are small open economies (SOE).

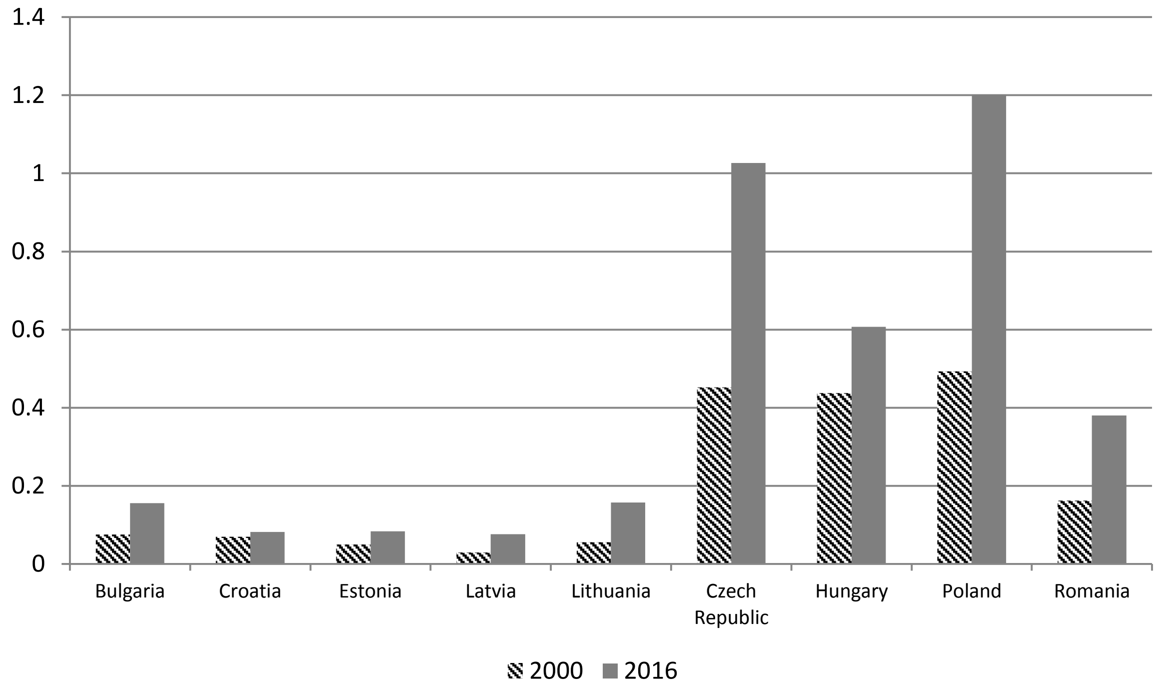


Figure 1: Share of exports in world exports in percent

In 2000 (prior to the EU accession), all countries had a smaller share of exports in the world exports compared to 2016. Therefore, all countries due to trade liberalization with EU, reaped the benefits of EU membership. However, Figure 1 indicates that all countries were not equally successful. For instance, the Czech Republic and Hungary had a similar share in 2000 while Poland's share was slightly higher. In 2016 the Czech Republic and Poland doubled their share, while Hungary did not manage to achieve a similar increase. Bulgaria, Latvia and Lithuania doubled their share. However, their share was quite low at the outset, and is therefore still lower than 0.2%.

Additionally, to account for differences in sizes of the analysed economies, figure 2 illustrates the share of exports in countries' GDP. Figure 2 shows that share of exports in GDP increased following the EU accession. While Lithuania and Poland almost doubled the share in 2016 compared to 2000, the increase was modest in Romania, Croatia and Estonia. It is interesting to note that these are the countries from three different accession waves.

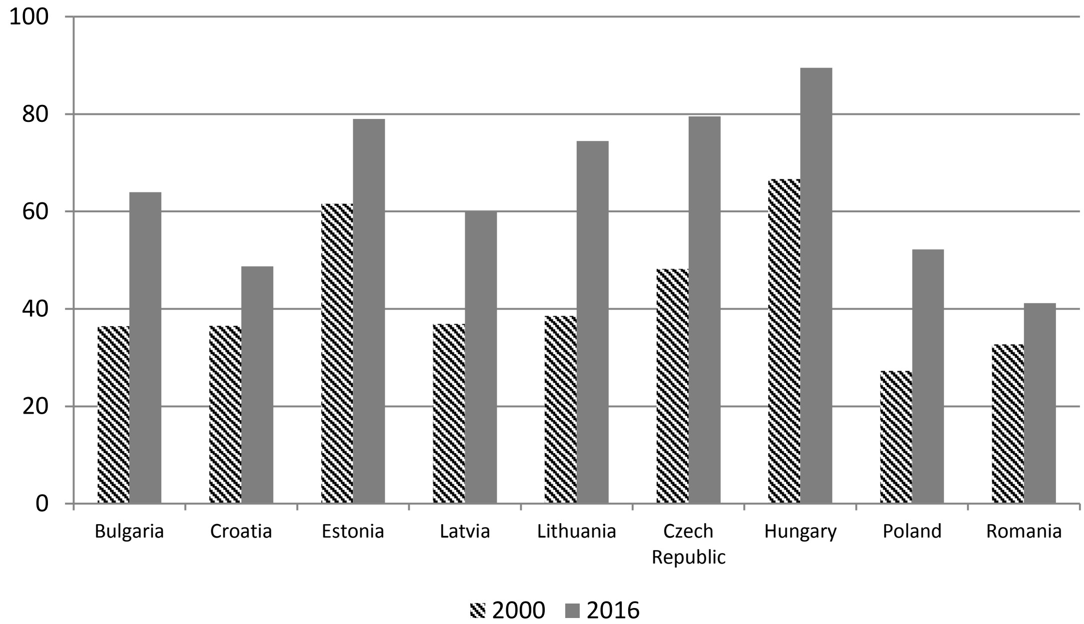


Figure 2: Share of exports in countries' GDP in percent

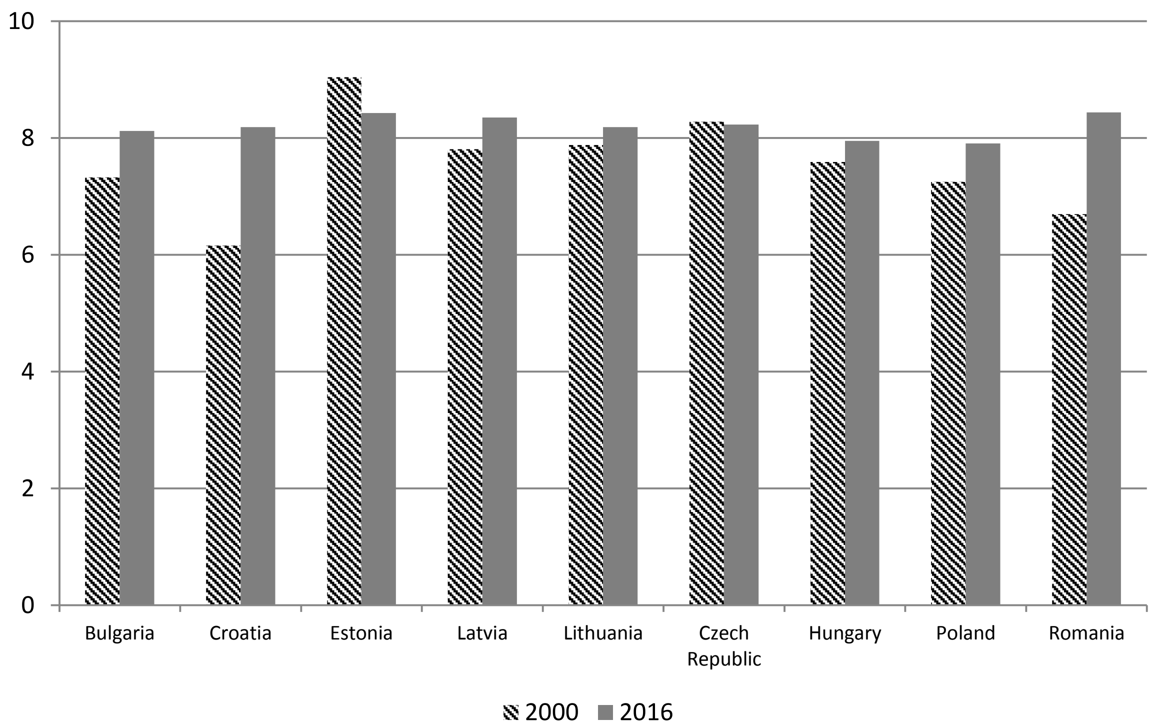


Figure 3: Freedom to trade internationally

Figure 3 depicts the Freedom to trade internationally for the selected EU NMS. Countries from the 2004 enlargement wave recorded a decrease or a slight increase in score. Estonia, which had the largest score in 2000, recorded the largest decline (6.7%). On the contrary, countries that joined the EU after 2004 accession wave had a substantial (double digit percentage) increase in score. Croatia, which had the worst score in 2000, had the largest increase (33%).

Another interesting finding from the Figure 3 is that the range of the FREE variable was substantially higher prior to the EU accession in 2000 (2.88) compared to 2016 (0.53), i.e. after the EU accession. Obviously, all countries had to reduce the trade barriers and foster free trade. However, as the score is still below 10, there is still room for improvement above the trade barriers removal that was obligatory in the pre-accession negotiations.

For comparison purposes, it is interesting to note that Hong Kong and Singapore are the highest ranked world countries with scores 9.32 and 9.29, respectively. Highest among the analysed countries, Romania (score 8.44), is ranked 9th, and the lowest, Poland (score 7.91) is ranked 44th.

3 METHODOLOGY AND EMPIRICAL RESULTS

In order to capture unobserved heterogeneity across analysed countries, panel data model was applied. In particular, a dynamic panel data model [1], [2] with one lagged dependent variable $\boldsymbol{y_{it-1}}$ was estimated:

$$y_{it} = X_{it}'\beta + \gamma \cdot y_{it-1} + \alpha_i + \varepsilon_{it}, \qquad i = 1, \dots, N, t = 1, \dots, T. \qquad (1)$$

y_{it} is a dependent variable, X_{it} is the matrix of independent variables, α_i is the (unobserved) individual effect, and ε_{it} is the error (idiosyncratic) term with $E(\varepsilon_{it}) = 0$, and $E(\varepsilon_{it}\varepsilon_{js}) = \sigma_\varepsilon^2$ if $j = i$ and $t = s$ and $E(\varepsilon_{it}\varepsilon_{js}) = 0$ otherwise.

The results of the estimated model are presented in table 1.

Table 1: Dynamic panel model estimation results (dependent variable: a share of exports in world exports).

Variable	*Coefficient*	*Robust standard errors*	*t-statistic*	*p-value*
Lagged dependent variable	0.8382322	0.0421324	19.90	0.000
lnREER	-0.072397	0.0732109	-0.99	0.349
lnULC	0.0237524	0.1160298	0.20	0.842
lnFREE	0.2570367	0.09366	2.74	0.023

Additionally, to check for model adequacy, the results of diagnostic tests [8] are reported in Table 2.

Table 2: Model diagnostics.

Arellano-Bond test for AR(1) in first differences	z = -2.59 (p-value = 0.010)
Arellano-Bond test for AR(2) in first differences	z = -0.73 (p-value = 0.465)
F-test	F = 136.16 (p-value = 0.000)
Hansen test of overid. restrictions	chi2(14) = 7.46 (p-value = 0.916)
Difference-in-Hansen tests of exogeneity of instrument subsets:	
Hansen test excluding group:	chi2(11) = 7.46 (p-value = 0.761)
Difference (null H = exogenous):	chi2(3) = 0.00 (p-value = 1.000)
Number of observations	133
Number of instruments	18
Number of groups	9

The test results indicate the appropriateness of the dynamic panel model specification (F = 136.16, p-value = 0.000). The dynamic panel model estimators are consistent as there is no second-order serial correlation for the idiosyncratic errors of the first-differenced equation (p-

value = 0.465). The first order serial correlation is expected due to the lagged dependent term (p-value = 0.010). Hansen test (p-value = 0.916) indicates that the crucial assumption for the validity of the dynamic panel model (that the instruments are exogenous) is satisfied. Furthermore, Hansen test excluding group (p-value = 0.761) does not reject the null that the model with additional instruments is correctly specified. Thus, all specification tests support the model adequacy.

Turning to the results of the estimated model (Table 1), FREE variable is statistically significant at the 5% level and has the expected positive sign, which means that the reduction of trade barriers leads to improvement in export competitiveness in selected EU NMS.

Furthermore, as expected, the results of the estimated model indicate that the lagged dependent variable is statistically significant because the current level of export competitiveness is heavily determined by its past values.

4 CONCLUSION

The empirical literature [5] provides evidence that trade restrictions affect export competitiveness through various channels. For that reason, this paper analyses the relationship between trade reform and export competitiveness for a set of selected eastern European countries, known as an EU New Member States in the period from 2000 to 2016. The sample included countries that experience quite a few turbulences ranging from global economic and financial crisis to accession into a political and economic union. The main finding of the paper indicates that reducing trade barriers indeed leads to improvement in export competitiveness in selected EU NMS. While priorities differ across NMS (such as exchange rate regimes) more effective institutions and better governance are *condicio sine qua non* for their export competitiveness. Although some of these NMS have committed to joining the euro area, it does not mean they will improve their export competitiveness and perform trade reforms. Besides, the legal framework harmonizing process with euro area standards is under sovereign control.

Future research should take into regard more specific and detailed indicators of various possibilities of restricting trade. Obvious candidates are components of the Free to trade internationally. This could lead to a more precise pinpointing of the trade barriers that affect export competitiveness the most.

References

[1] Alvarez, J., Arellano, M. 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica, 71(4): 1121–1159.

[2] Baltagi, B. H. 2013. Econometric Analysis of Panel Data. 5th ed. Chichester, UK: Wiley.

[3] Fatas, A. 2015. The Agenda for Structural Reform in Europe. No 10723, CEPR Discussion Papers.

[4] Gwartney, J., Lawson, R., Hall, J., Murphy, R. 2018. Economic Freedom of the World: 2018 Annual Report. Fraser Institute.

[5] Hallaert, J. J. 2012. Structural Reforms and Export Performance. IMF Country Report No. 13/3.

[6] Hallaert, J.J., Cavazos Cepeda, R., Kang, G. 2011. Estimating the Constraints to Trade of Developing Countries, Paris: OECD Trade Policy Working Papers, No. 116.

[7] International Monetary Fund 2015. Structural Reforms and Macroeconomic Performance - Initial Considerations for the Fund. Policy Papers.

[8] Roodman, D. 2009. A note on the theme of too many instruments. Oxford Bulletin of Economics and Statistics, 71(1): 135-158.

ARE INVESTMENT CONSTRAINTS OF MANDATORY PENSION FUNDS RESTRICTING THEIR PERFORMANCE: CASE OF CROATIA

Margareta Gardijan Kedžo
University of Zagreb Faculty of Economics and Business, Department of Mathematics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: mgardijan@efzg.hr

Ana Škrlec
J&T Banka d.d.
Aleja kralja Zvonimira 1, Varaždin
E-mail: ana.skrlec@jtbanka.hr

Abstract: The pension system and the performance of mandatory pension funds are in the public interest due to their social and economic function. Therefore, their investments are under tight restrictions defined by legal acts. Some of the restriction imposed by law are sometimes questioned since they might limit the funds' ability to improve their performance. In this paper, we investigate the asset allocation of mandatory pension funds in the Republic of Croatia and simulate their portfolios using the current investment constraints. We introduce possible changes of these constraints and analyse their effect on portfolios' performance. The empirical analysis focuses on the period of 2015-2017. The results suggest that better performance of our portfolios could be achieved within the regulatory investment constraints, but a further lessening of some constraints could improve their performance in terms of risk and the return.

Keywords: mandatory pension funds, portfolio optimization, investment constraints.

1 INTRODUCTION

Pension funds are institutional investors whose main goal is to accumulate financial resources and invest them in a certain asset (most commonly financial asset) in order to make a profit and maintain a pension system. Specialised pension companies managing pension funds are responsible for investing the resources of fund members (future pensioners) in a responsible and profitable way. This means that those who manage the funds must ensure the yield and take risks in accordance with the degree of risk of the fund [11]. Their functions, responsibilities and obligations are defined by legal acts. Pension beneficiaries, together with pension companies managing pension funds, constitute a pension system that is a fundamental part of a contemporary legal state. Similar as in other countries, the pension system in the Republic of Croatia has social, economic and political significance. Its basic function is to allocate individuals' consumption throughout their lifetime, especially after a period of employment, in the event of disability or loss of the caregivers [14]. The pension system in Croatia, as we know it today, is defined by a series of legal changes that had been gradually introduced until 2002. Today, the pension system is based on three pillars [12]. The first and second pillars are mandatory pension insurance, while the third pillar represents voluntary pension insurance. This scheme was introduced in 2001. for several reasons: unsustainability of the previous pension system, caused by the structural features and unfavourable demographic situation in Croatia during the 1990s, the improvement of the pension system by creating a direct link between labour and pension, and the development of the domestic capital market. The latter role was obviously attained: the activities of the pension funds positively influenced the domestic capital market in terms of the increased turnover, successful IPOs of national companies (INA and HT), and they financed the private sector with 9.5 billion HRK, which led to economic growth, decrease of interest rate on Croatian debt and consequentially increased standard of living of the future pensioners [18] [13]. However, the continuation of unfavourable economic and social trends in Croatia prevented reforms from achieving any success related to the first two reasons. Nowadays, it is obvious that the pension system is not sustainable in this form and many papers dealt whit this topic ([2], [3]). In 2014, Mandatory Pension Act [7] introduced three categories of funds: A, B and C. Differentiation between each fund is based on the remaining period the members have until they retire, and subsequently each category has different investment restrictions and investment strategies (category A is the riskiest, and C the least risky). Given that the pensions depend on the returns made by an individual fund, the entire pension system rests on the success of fund management. Therefore, their investments are strictly regulated in order to avoid situations of major losses that could cause socio-economic and political problems. Croatian Financial Services Supervisory Agency (HANFA) has a regulatory role in Croatia. The assets of the A, B and C funds may be invested in securities and money market instruments issued by the Republic of Croatia, another member of the European Union (EU) or member states of the Organization for Economic Co-operation and Development (OECD). Moreover, A and B funds can use derivatives, but only for the purpose of effective management and protection of fund assets. The most important legal restrictions on the asset allocation for funds per category are given in Table 1.

Table 1: Asset allocation restrictions in each category of mandatory pension fund [7]

Asset	Category A	Category B	Category C
Bonds issued by the Republic of Croatia, other EU or OECD member states	at least 30% of the fund's net assets	at least 50% of the fund's net assets	at least 70% of the net assets of the fund
Shares from issuers from Croatia, other EU or OECD member states	no more than 55% of the fund's net assets	a maximum of 35% of the fund's net assets	Investment in shares is not allowed
Corporate bonds and commercial bills of issuers from Croatia, other EU or OECD member states	no more than 50% of the net assets	no more than 30% of the funds' net assets	no more than 10% of the net assets
UCITS shares	no more than 30% of the net assets	no more than 30% of the net assets	no more than 10% of the net assets
Property traded or settled in HRK		at least 60% of the net assets of the fund	at least 90% of the net assets

By observing the investment restrictions for certain fund category, one could raise a question of these restrictions limiting the funds' performance at the expense of the pension beneficiaries and in the favour of developing the domestic market. This question is tackled in several papers before the introduction of the 'ABC system', such as in [1] who states that the legal constraints are not an obstacle for pension funds' construction of the optimal portfolio, but that the limitation of Croatian capital market could be. On the other hand, [17] investigates the introduction of alternative investments and reaches the opposite conclusions. Even though funds are restricted in their asset allocation, the law defines that if a fund underperforms given the past 3-years average yield of all funds in the same category, its pension company is obligated by law to pay the difference defined by the guaranteed reference yield [4]. While there are several expert studies which suggest how to reform and improve the pension system in Croatia, the main concern of this paper are the investment restrictions and an idea that they might be limiting the funds' performance and market competition. [9] benchmarks the performance of mandatory pension funds' to a portfolio consisting of CROBIS and CROBEX indices, and concludes that pension funds' performance is satisfactory compared to the benchmark. On the other hand, [10] conclude that pension funds' portfolio in Croatia has a high sensitivity to risk due to its highly conservative asset allocation. They state that investment constraints prevent the competitiveness and the development of the pension funds, which leads to inefficient asset allocation and the degradation of economic development and employment.

Even though Croatia was not analysed in [15], this study observed the pension funds' excess real return over GDP growth in seven Central European countries and concluded that their investment limits are also too conservative and that performance is not satisfactory. In other studies, such as [6] and [16], the position of Croatia is relatively worse in several indicators than in other countries included in [15]. These findings motivate us to analyse the influence of the investment restrictions on the performance of simulated portfolios whose structure is inspired by the actual asset allocation of mandatory pension funds in Croatia. Currently, there are 4 pension companies in Croatia which operate since 2002.: Allianz ZB d.o.o., Erste d.o.o., PBZ Croatia Osiguranje d.d. and Raiffeisen d.d.. Each of them runs one mandatory investment fund from each category. The average structure of the funds' consolidated balance sheet during the period of 2015.-2017. for each category is shown in Figure 1.

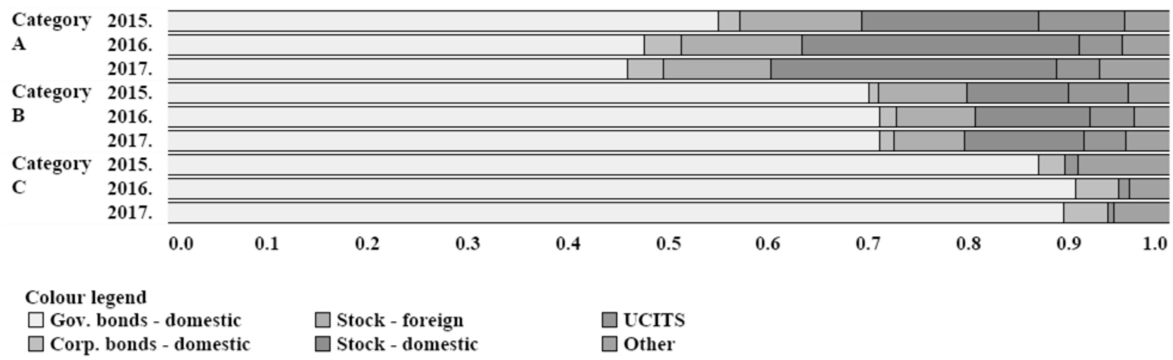


Figure 1: Average asset allocation of mandatory pension funds in the period of 2015-2017 [8]

By analysing the funds' performance since their inception, it is interesting to find that category C funds, which are more conservative, were overall more successful than B categorized funds managed by the same pension company [5]. The reason was that C funds invested largely in bonds that have been profitable over the last couple of years, given the historically low interest rates on the financial markets. Although funds do not have an upper limit for investing in bonds, B funds did not increase their bond positions accordingly. This observation indicated that investment limitations (which are tighter for C funds), could not be always used as excuses for poor funds' performance. Therefore, we had another reason for investigating the possible effects of investment restrictions on funds' performance.

2 METHODOLOGY

In the empirical part, we simulate the portfolios whose asset allocation follows approximately the actual asset allocation of pension funds in Croatia (Figure 1). Therefore, our portfolios consist of: stocks issued by issuers from Republic of Croatia (represented by the stock market index CROBEX), Croatian government bonds (represented by the bond market index CROBIS), cash, treasury bills (represented by the return on government bonds with 3 months to maturity), investments in domestic UCITS (The Undertakings for the Collective Investment in Transferable Securities) (represented by the average monthly change of UCITS's net assets), investments in foreign UCITS (represented by Barclays UCITS funds index) and foreign shares (represented by the S&P100). We use data of daily returns of market indices in the period of 2015.-2017. and interest rates are calculated proportionally. We calculated the average asset allocation for each fund category in each month during the period of the analysis and constructed portfolios that hold the same structure throughout the period of 2015.-2017. (36 portfolios in total). The lower and upper values of the asset allocation for each category are shown in Table 2.

Table 2: Lower and upper bounds of the asset allocation of simulated portfolios in the period of 2015-2017

	Simulation A		Simulation B		Simulation C	
	min	max	min	max	min	max
Stocks - domestic	10%	32%	10%	13%	0%	0%
Stocks - foreign	9%	13%	6%	9%	0%	0%
Government bonds	42%	64%	68%	73%	78%	94%
Cash	0%	6%	1%	4%	0%	6%
UCITS domestic	0%	6%	1%	3%	0%	5%
UCITS foreign	0%	9%	2%	5%	0%	0%
Treasury notes	0%	0%	0%	0%	0%	8%

The second step of our analysis includes questioning whether a better performance of funds (in terms of risk and return) could be achieved within the given restrictions. For that purpose, we solve a portfolio optimization problem for each fund category, formulated as:

$$\begin{array}{llll}
 & \text{Category A:} & \text{Category B:} & \text{Category C:} \\
 & \max \sum_t r_t & \max \sum_t r_t & \max \sum_t r_t \\
s.t. & VaR_{0.95} \le c_A & VaR_{0.95} \le c_B & VaR_{0.95} \le c_C \\
 & \sigma^2 \le \sigma^2_{A\min} & \sigma^2 \le \sigma^2_{B\min} & \sigma^2 \le \sigma^2_{C\min} \\
 & \pi^e \le 0.55 & \pi^e \le 0.35 & \pi^e = 0 \\
 & \pi^b \ge 0.3 & \pi^b \ge 0.5 & \pi^b \ge 0.7 \\
 & \pi^u \le 0.3 & \pi^u \le 0.3 & \pi^u \le 0.1 \\
 & \pi^e + \pi^b + \pi^u = 1 & \pi^e + \pi^b + \pi^u = 1 & \pi^e + \pi^b + \pi^u = 1 \\
 & \pi^e, \pi^b, \pi^u \ge 0 & \pi^e, \pi^b, \pi^u \ge 0 & \pi^e, \pi^b, \pi^u \ge 0
\end{array} \quad \text{(M1)}$$

where r_t is a daily return of a portfolio, $VaR_{0.95}$ is *Value at Risk* of a portfolio at 95% (calculated as the 5th percentile of the portfolio return distribution), σ^2 is the portfolio variance, c is the minimal $VaR_{0.95}$ and σ^2_{min} is the minimal variance of all the portfolios from Table 2 within each category, π^e is the share of equity, π^b is a share of bonds and π^u is the share of investments in UCITS funds, t=1,…,N, and N is a number of observations.

The last step in our analysis includes solving the modification of the model (M1) without the restriction on asset allocation. We will refer to this modified model as the model (M2).

3 RESULTS

The average performance of 36 simulated portfolios from Table 2 is shown in the first 3 columns of Table 3. Here, the performance of the portfolios is consistent with their investment strategy – portfolios A have the highest risk and return, portfolios B are in the middle, and C portfolios have the smallest total return and risk. We use the values of the minimal variance and maximal $VaR_{0.95}$ of these portfolios as right-hand side constraints in the model (M1). By solving the model (M1) for each fund category, we obtain portfolios presented in Table 4. Their structure shows that the share in UCITS is always maximal, that is, the constraint for UCITS portfolio share (π^u) is active in all solutions of model (M1). The constraints for π^e and π^b are not active. However, if we observe upper bounds in the actual average asset allocation (Table 2), the investment in UCITS is far below the defined investment constraint despite its favourable performance throughout the years. The performance of the obtained portfolios is shown in Table 3 (A-M1, B-M1, C-M1). All portfolios achieve an incomparably better return for a given level of variance (standard deviation) and $VaR_{0.95}$. These results indicate that better performance could be achieved by reducing the investments in shares and increasing the investments in UCITS without increasing the riskiness measured by both VaR and the variance.

Table 3: Performance measures of simulated portfolios (in %)

	A	B	C	A-M1	B-M1	C-M1	A-M2	B-M2	C-M2
(Avg) Total return	9.953	8.750	5.138	21.196	23.877	10.145	54.582	54.965	49.633
Max. total return	12.292	9.722	5.463						
Min. total return	7.578	7.871	4.859						
(Avg.) St. dev.	0.199	0.145	0.126	0.140	0.138	0.111	0.130	0.131	0.111
Max st. dev.	0.232	0.149	0.133						
min st. dev.	0.140	0.138	0.138						
(Avg.) $VaR_{0.95}$	-0.279	-0.202	-0.192	-0.187	-0.173	-0.163	-0.187	-0.195	-0.162
Max $VaR_{0.95}$	-0.19	-0.20	-0.16						
Min $VaR_{0.95}$	-0.34	-0.21	-0.20						

Table 4: Asset allocation of portfolios obtained by solving the model (M1) and its modification (M2)

	A-M1	B-M1	C-M1	A-M2	B-M2	C-M2
Stocks - domestic	0%	1%	0%	0%	0%	0%
Stocks - foreign	15%	13%	0%	2%	1%	0%
Government bonds	43%	56%	78%	0%	0%	0%
Cash	12%	0%	0%	0%	0%	0%
UCITS foreign	4%	0%	0%	0%	0%	14%
UCITS domestic	26%	30%	10%	98%	99%	86%
Tresury notes	0%	0%	12%	0%	0%	0%

By solving the model (M2), without the asset allocation restrictions, we obtain portfolios A-M2, B-M2 and C-M2 whose structure is shown in Table 4. Their performance is shown in Table 3. The solution even more strongly suggests that increasing the investments in domestic UCITS should be beneficial for the portfolios' overall performance. Lessening the limitations in favour of investments in UCITS leads to an increased return and reduced risk.

4 CONCLUSION

The pension fund management is extremely important to individuals as well as to the state that is actually responsible for pensions. That is the reason why pension funds' asset allocation is defined by legal acts and under the strict supervision of the regulatory agencies. This paper analyses these restrictions and simulates the portfolios imitating the asset allocation of the mandatory pension funds in Croatia in the period of 2015.-2017. The portfolio optimization models are solved for finding optimal portfolios in compliance with the legal restrictions for each fund category as well as the portfolios outside the legal limitations. Given the data and assumptions used, the obtained results show that portfolio performance could be significantly improved within the legal restrictions by asset reallocation. Accordingly, the analysis shows that legal restrictions do leave additional space for increasing the funds' performance. Therefore, more attention should be put towards the market situation and investments should be made accordingly. The results also indicate that pension funds' asset allocation could be brought closer to the limits in order to increase the return without increasing risk. This holds especially for investments in UCITS. Moreover, it is shown that further lessening of the investment constraints for investment in UCITS could additionally increase the overall funds' performance. The posed restrictions foster activities of pension funds which positively influence the domestic financial system and country's economic growth, but their main concern should be the benefit of the future pensioners. These findings do not offer the solution for pension reform or the financial stabilization of the pension system, but they do show that investment restrictions are just another indicator that things are not set as they should be. Limitations of the research are the data series that were used as the proxies for the returns of certain assets and the dependence of conclusions to the period from which data were taken. Further research should consider longer time series and a greater number of asset categories including foreign and alternative asset categories to reach conclusions that can be more directly related to the pension funds' performance.

References

[1] Bakić, N. 2002. Investments of Pension Funds in the Republic of Croatia. Financial Theory and Practice, 26(2): 435-445.

[2] Bejaković, P. 2011. Mirovinski sustav u RH - problemi i perspektiva. Analiza mirovinskih sustava. Banka magazine & The Institute of Public Finance. Available at https://www.ijf.hr/upload/files/file/AMS/zbornik.pdf. [Accessed 26/05/2019]

[3] Bejaković, P. 2019. The causes of problems in the public pension system and reasons why funded pension insurance should be preserved in Croatia. Rev. soc. polit., 26(1): 37-53.

[4] Decision on the guaranteed return of mandatory pension funds. Official Gazette, 11/16, 8/17, 8/18.

[5] HANFA Monthly reports. (2015-2017). HANFA. https://www.hanfa.hr/publikacije/mjesecna-izvjesca/ [Accessed 26/052019]

[6] Krpan, M., Pavković, A., Galetić, F. 2019. Comparison of Sustainability Indicators of Pension Systems in the New EU Member States. In: Tipurić, D., Hruška, D. (Eds.). 7th International OFEL Conference on Governance, Management and Entrepreneurship: Embracing Diversity in Organisations. Zagreb: CIRU & University of Dubrovnik.

[7] Mandatory Pension Funds Act. Official Gazette, 19/2014.

[8] Mandatory pension funds investment structure. (2015-2017). HANFA.

[9] Novaković, D. 2015. Evaluation of the financial performance of pension funds in Croatia. Econviews, 28(1): 199-212.

[10] Olgić Draženović, B., Pečarić, M., Suljić, S. 2017. Institutional investors in the selected new members EU countries. Journal of Economy and Business, 23: 104-120.

[11] Pennacchi, G. M., Rastad, M. 2011. Portfolio Allocation for Public Pension Funds. Journal of Pension Economics and Finance, 10(2): 221-245.

[12] Pension Insurance Act. Official Gazette, 157/2013.

[13] Pension Reform in the Republic of Croatia: Past Effects, Current Situation and Proposals for the Future. 2011. UMFO. https://mirovinskifondovi.hr/wp-content/uploads/2015/02/Mirovinska-reforma-1.pdf. [Accessed 26/05/2019].

[14] Puljiz, V. 2007. The Croatian pension system: origins, evolution and perspectives. Rev. soc. polit., 14(2): 163-192.

[15] Stanko, D. 2017. Analysis of the functioning of funded parts of the pension system in selected CEE countries in the context of the implemented changes. In: Bielawska, K., Chlon-Dominczak, A., Stanko, D. (Eds.). Retreat from mandatory pension funds in countries of the Eastern and Central Europe in result of crisis. Warsaw: Munich Personal RePEc Archive, 21-46.

[16] Škuflić, L., Pavković, A., Novinc, F. 2018. Challenges Facing Pension Systems of Ex-Yugoslavian Countries. In: Bevanda, V. (Ed.). LIMEN 2018 Conference Proceedings: Leadership & Management: Integrated Politics of Research and Innovations. Belgrade.

[17] Štimac, D. 2002. Portfolio analysis of justification of pension funds' investment limits (Doctoral dissertation). Zagreb: Faculty of Economics and Business Zagreb.

[18] The past, present and future of the pension system in the Republic of Croatia. 2014. Raiffeisen Research. https://www.rmf.hr/UserDocsImages/dokumenti/Proslost,%20sadasnjost%20i%20buducnost%20mirovinskoga%20sustava%20u%20RH.pdf. [Accessed 26/05/2019].

SCORE-DRIVEN COUNT TIME SERIES

Vladimír Holý
University of Economics, Prague
Winston Churchill Square 1938/4, 130 67 Prague 3, Czech Republic
vladimir.holy@vse.cz

Abstract: We review the use of the generalized autoregressive score (GAS) framework in the analysis of count time series. We formulate GAS models based on the Poisson distribution for equidispersed data, the negative binomial distribution for overdispersed data, the generalized Poisson distribution for moderately underdispersed data and the double Poisson distribution for significantly underdispersed data. The presented models are count data analogues to the ARMA and GARCH models.

Keywords: Count Time Series, Score-Driven Model, Poisson Distribution, Negative Binomial Distribution, Generalized Poisson Distribution, Double Poisson Distribution.

1 INTRODUCTION

Count time series are series of observations with non-negative integer values ordered in time. For a review of count time series models, see [10]. A sensible approach for modeling count time series is the generalized autoregressive score (GAS) model of [6], also known as the dynamic conditional score (DCS) model of [12]. It is an observation-driven model providing a general framework for modeling of time-varying parameters for any underlying probability distribution. It captures dynamics of time-varying parameters by the autoregressive term and the scaled score of the conditional density function (or the conditional probability mass function). The GAS class includes many commonly used econometric models such as the GARCH model based on the normal distribution and the ACD model based on the exponential distribution. The GAS model can be estimated in a straightforward manner by the maximum likelihood method.

In the literature, the GAS framework is successfully utilized for count time series. The paper [2] uses the Poisson count panel model for the number of patent applications. The paper [14] utilizes the bivariate Poisson distribution for the number of goals in football matches and the Skellam distribution for the score difference. The paper [11] consideres the Poisson distribution as well as the negative binomial distribution for offensive conduct reports. The paper [1] uses the zero-inflated negative binomial distribution for trade durations with frequent split transactions. The Poisson count model discussed in [7] belongs to the GAS class as well.

In the paper, we present specifications of the GAS model based on four distributions suitable for count data – the Poisson distribution, the negative binomial distribution, the generalized Poisson distribution and the double Poisson distribution. The presented models can capture both time-varying mean and time-varying volatility.

2 GENERALIZED AUTOREGRESSIVE SCORE MODEL

2.1 Theoretical Framework

Our goal is to model non-negative random variables $Y_t \in \mathbb{N}_0$, $t = 1, \dots, n$. We denote the observed values as $y_t \in \mathbb{N}_0$, $t = 1, \dots, n$. In general, we assume that Y_t follow a distribution with k time-varying parameters $f_t = (f_{t,1} \dots, f_{t,k})'$, $t = 1, \dots, n$ and l static parameters $g = (g_1, \dots, g_l)'$. We denote the conditional probability mass function as $P[Y_t = y_t | f_t, g]$. In the dynamic model for the time-varying parameters, we utilize the score and the Fisher information. The score for the time-varying parameters f_t is given by

$$\nabla(y_t; f_t, g) = \frac{\partial \log P[Y_t = y_t | f_t, g]}{\partial f_t}. \tag{1}$$

The Fisher information for the time-varying parameters f_t is given by

$$\mathcal{I}(f_t, g) = \mathrm{E}\Big[\nabla(y_t; f_t, g)\nabla(y_t; f_t, g)'\Big| f_t, g\Big] = -\mathrm{E}\left[\frac{\partial^2 \log P[Y_t = y_t | f_t, g]}{\partial f_t \partial f_t'}\middle| f_t, g\right]. \tag{2}$$

Note, that the latter equality requires some regularity conditions. The expected value of the score is zero and its variance is equal to the Fisher information under some regularity conditions.

2.2 Model Specification

The generalized autoregressive score (GAS) model of [6] specifies the dynamics of the time-varying parameters f_t. In the GAS(p, q) model, the parameters f_t follow the recursion

$$f_{t+1} = c + \sum_{j=1}^{q} B_j f_{t-j+1} + \sum_{i=1}^{p} A_i S(f_{t-i+1}, g)\nabla(y_{t-i+1}; f_{t-i+1}, g), \tag{3}$$

where $c = (c_1, \ldots, c_k)'$ are the constant parameters, $B_j = \mathrm{diag}(b_{j,1}, \ldots, b_{j,k})$ are the autoregressive parameters, $A_i = \mathrm{diag}(a_{i,1}, \ldots, a_{i,k})$ are the score parameters, $S(f_{t-i+1}, g)$ is the scaling function for the score and $\nabla(y_{t-i+1}, f_{t-i+1}, g)$ is the score. The score for the time-varying vector f_t is the gradient of the log-likelihood with respect to f_t. It indicates how sensitive the log-likelihood is to parameter f_t. In the dynamic model, it drives the time variation in the parameter f_t and links the shape of the conditional probability mass function directly to the dynamics of f_t. Usually, the scaling function is choosen to be the unit matrix I, the square root of the inverse of the Fisher information $\mathcal{I}(f_t, g)^{-\frac{1}{2}}$ or the inverse of the Fisher information $\mathcal{I}(f_t, g)^{-1}$. Note that each scaling function results in a different model. In the case of $\mathcal{I}(f_t, g)^{-\frac{1}{2}}$ scaling, the scaled score has unit variance.

In Section 3, we specify possible distributions for the count variable Y_t together with their score and Fisher information used for the dynamics of time-varying parameters. All presented distributions have a location parameter and (except the Poisson distribution) a dispersion parameter. The time-varying GAS dynamics can be utilized for the location parameter (an analogue to the ARMA model), time-varying dispersion parameter (an analogue to the GARCH model) or both.

2.3 Estimation

Let us denote $\theta = (c, B_1, \ldots, B_q, A_1, \ldots, A_p, g)'$ the vector of $(1+q+p)k+l$ unknown parameters. Using a sequence of n observations $y_1, \ldots, y_n$, we can estimate the vector θ by the maximum likelihood method as

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \log \mathrm{P}\left[Y_t = y_t | \hat{f}_t(\theta), \theta\right], \tag{4}$$

where $\hat{f}_t(\theta)$ are the filtered time-varying parameters. Note that we need to specify the first $\max\{p, q\}$ values of $\hat{f}_t(\theta)$.

3 PROBABILITY DISTRIBUTIONS FOR COUNT VARIABLES

3.1 Poisson Distribution

The most basic distribution for count data is the Poisson distribution. It has only one parameter $\mu > 0$ determining its mean and variance. The variance equal to the mean is known as equidispersion and is very limiting in applications. The probability mass function is given by

$$\mathrm{P}[Y = y | \mu] = \frac{\mu^y e^{-\mu}}{y!}. \tag{5}$$

The expected value and variance are given by

$$\begin{aligned} \mathrm{E}[Y] &= \mu, \\ \mathrm{var}[Y] &= \mu. \end{aligned} \tag{6}$$

The score is given by

$$\nabla(y;\mu) = \frac{y-\mu}{\mu}. \tag{7}$$

The Fisher information is given by

$$\mathcal{I}(\mu) = \frac{1}{\mu}. \tag{8}$$

3.2 Negative Binomial Distribution

When data exhibit overdispersion (i.e. the variance greater than the mean), the standard approach is to consider the negative binomial distribution. It is derived as the Poisson-gamma mixture. We present the NB2 parameterization of [3]. It has a location parameter $\mu > 0$ and a dispersion parameter $\alpha \geq 0$. For $\alpha = 0$, it reduces to the Poisson distribution. In the following text, let $\Gamma(\cdot)$ denote the gamma function, $\psi_0(\cdot)$ the digamma function and $\psi_1(\cdot)$ the trigamma function. The probability mass function is given by

$$\mathrm{P}[Y = y|\mu,\alpha] = \frac{\Gamma(y+\alpha^{-1})}{\Gamma_i(y+1)\Gamma_i(\alpha^{-1})} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}. \tag{9}$$

The expected value and variance are given by

$$\begin{aligned} \mathrm{E}[Y] &= \mu, \\ \mathrm{var}[Y] &= \mu(1+\alpha\mu). \end{aligned} \tag{10}$$

The score is a vector with 2 elements given by

$$\begin{aligned} \nabla_1(y;\mu,\alpha) &= \frac{y-\mu}{\mu(1+\alpha\mu)}, \\ \nabla_2(y;\mu,\alpha) &= \frac{1}{\alpha^2}\ln(1+\alpha\mu) + \frac{y-\mu}{\alpha(1+\alpha\mu)} + \frac{1}{\alpha^2}\psi_0(\alpha^{-1}) - \frac{1}{\alpha^2}\psi_0(y+\alpha^{-1}). \end{aligned} \tag{11}$$

The Fisher information is a 2×2 matrix with elemets given by

$$\begin{aligned} \mathcal{I}_{11}(\mu,\alpha) &= \frac{1}{\mu(1+\alpha\mu)}, \\ \mathcal{I}_{12}(\mu,\alpha) &= \mathcal{I}_{21}(\mu,\alpha) = 0, \\ \mathcal{I}_{22}(\mu,\alpha) &= \frac{2}{\alpha^3}\ln(1+\alpha\mu) - \frac{\mu}{\alpha^2(1+\alpha\mu)} + \frac{2}{\alpha^3}\psi_0(\alpha^{-1}) + \frac{1}{\alpha^4}\psi_1(\alpha^{-1}) \\ &\quad - \mathrm{E}\left[\frac{2}{\alpha^3}\psi_0(y+\alpha^{-1}) + \frac{1}{\alpha^4}\psi_1(y+\alpha^{-1})\right]. \end{aligned} \tag{12}$$

3.3 Generalized Poisson Distribution

Underdispersion (i.e. the variance lower than the mean) occurs less often in data and is more complex to model. There is no universally accepted approach. One option is to consider the generalized Poisson distribution proposed in [5]. It is derived as the Poisson-lognormal mixture. We present the mean parametrization of [9]. It has a location parameter $\mu > 0$ and a dispersion parameter α. For $\alpha = 0$, it reduces to the Poisson distribution. Values $\alpha < 0$ result in underdispersion while values $\alpha > 0$ result in overdispersion. However, the parameter space is more complex with additional restrictions which can accommodate only limited underdispersion. Furthermore, the support is limited from above in the case of underdispersion. See e.g. [4] or [13] for more details. The probability mass function is given by

$$\mathrm{P}[Y = y|\mu, \alpha] = \frac{1}{y!(1+\alpha y)} \left(\frac{\mu(1+\alpha y)}{1+\alpha\mu} \right)^{y} e^{-\frac{\mu(1+\alpha y)}{1+\alpha\mu}}. \tag{13}$$

The expected value and variance are given by

$$\begin{aligned} \mathrm{E}[Y] &= \mu, \\ \mathrm{var}[Y] &= \mu(1+\alpha\mu)^2. \end{aligned} \tag{14}$$

The score is a vector with 2 elements given by

$$\begin{aligned} \nabla_1(y;\mu,\alpha) &= \frac{y-\mu}{\mu(1+\alpha\mu)^2}, \\ \nabla_2(y;\mu,\alpha) &= \frac{(y-\mu)^2}{(1+\alpha y)(1+\alpha\mu)^2} - \frac{y}{1+\alpha y}. \end{aligned} \tag{15}$$

The Fisher information is a 2×2 matrix with elemets given by

$$\begin{aligned} &\mathcal{I}_{11}(\mu,\alpha) = \frac{1}{\mu(1+\alpha\mu)^2}, \\ &\mathcal{I}_{12}(\mu,\alpha) = \mathcal{I}_{21}(\mu,\alpha) = 0, \\ &\mathcal{I}_{22}(\mu,\alpha) = \frac{1}{\alpha^3(1+\alpha\mu)^2} - \frac{3}{\alpha^3(1+\alpha\mu)} - \frac{1}{\alpha^2} - \mathrm{E}\left[\frac{1+\alpha}{\alpha^3(1+\alpha y)^2} - \frac{3+2\alpha}{\alpha^3(1+\alpha y)} \right]. \end{aligned} \tag{16}$$

3.4 Double Poisson Distribution

Another option for underdispersion is the double Poisson distribution proposed in [8]. It has a location parameter $\mu > 0$ and a dispersion parameter $\alpha > 0$. For $\alpha = 1$, it reduces to the Poisson distribution. Values $\alpha > 1$ result in underdispersion while values $0 < \alpha < 1$ result in overdispersion. The probability mass function is given by

$$\mathrm{P}[Y = y|\mu, \alpha] = \frac{1}{C(\mu,\alpha)} \sqrt{\alpha} \frac{y^y}{y!} \left(\frac{\mu}{y} \right)^{\alpha y} e^{\alpha y - \alpha\mu - y}, \tag{17}$$

where $C(\mu, \alpha)$ is a normalizing constant given by

$$C(\mu,\alpha) = \sum_{y=0}^{\infty} \sqrt{\alpha} \frac{y^y}{y!} \left(\frac{\mu}{y} \right)^{\alpha y} e^{\alpha y - \alpha\mu - y}. \tag{18}$$

The normalizing constant is not available in a closed form but can be approximated by

$$C(\mu,\alpha) \simeq 1 + \frac{1-\alpha}{12\alpha\mu} \left(1 + \frac{1}{\alpha\mu} \right) \simeq 1. \tag{19}$$

Alternatively, [15] suggest to approximate the normalizing constant by the sum of the first m terms in the infinite sum, where m should be at least twice as large as the sample mean. The expected value and variance can be approximated by

$$\begin{aligned} \mathrm{E}[Y] &\simeq \mu, \\ \mathrm{var}[Y] &\simeq \frac{\mu}{\alpha}. \end{aligned} \tag{20}$$

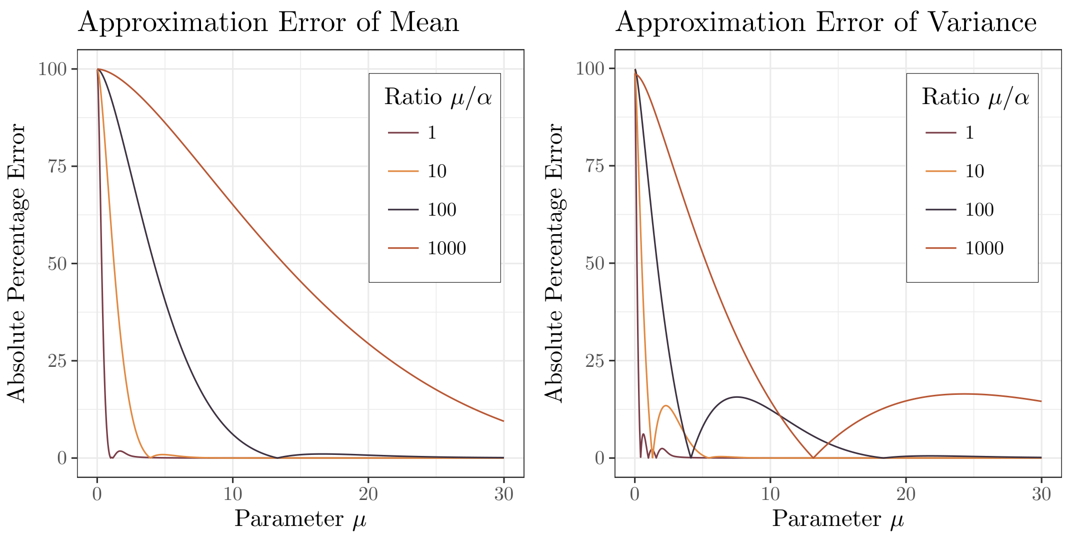


Figure 1: Absolute percentage errors of the approximations of the first and second moments for the double Poisson distribution.

The score is a vector with 2 elements approximated by

$$\begin{aligned} \nabla_1(y;\mu,\alpha) &\simeq \frac{\alpha}{\mu}(y-\mu), \\ \nabla_2(y;\mu,\alpha) &\simeq \frac{1}{2\alpha} + y\log(\mu) - \mu - y\log(y) + y. \end{aligned} \tag{21}$$

The Fisher information is a 2×2 matrix with elemets approximated by

$$\begin{aligned} \mathcal{I}_{11}(\mu,\alpha) &\simeq \frac{\alpha}{\mu}, \\ \mathcal{I}_{12}(\mu,\alpha) &= \mathcal{I}_{21}(\mu,\alpha) \simeq 0, \\ \mathcal{I}_{22}(\mu,\alpha) &\simeq \frac{1}{2\alpha^2}. \end{aligned} \tag{22}$$

In Figure 1, we illustrate the approximation error of the mean and variance using the absolute percentage error measure. We see that the approximations are quite precise for high values of mean and low values of variance. For more details about the approximations, see [8].

4 CONCLUSION

We review the score-driven approach for count time series. We present models based on four distributions. The Poisson distribution exhibits equidispersion and is commonly considered the benchmark distribution. The negative binomial distribution is the most popular choice for overdispersion. The generalized Poisson distribution can be utilized for moderate underdispersion while the double Poisson distribution is suitable for significant underdispersion with high mean and low variance. To our knowledge the latter two distributions have not yet been considered in the context of score-driven models.

Acknowledgements

The work on this paper was supported by the grant No. F4/21/2018 of the Internal Grant Agency of University of Economics, Prague.

References

[1] Blasques, F., Holý, V. and Tomanová, P. (2018). Zero-Inflated Autoregressive Conditional Duration Model for Discrete Trade Durations with Excessive Zeros. Working Paper. https://arxiv.org/abs/1812.07318.

[2] Blazsek, S. and Escribano, A. (2016). Score-Driven Dynamic Patent Count Panel Data Models. Economics Letters, 149: 116–119. https://doi.org/10.1016/j.econlet.2016.10.026.

[3] Cameron, A. C. and Trivedi, P. K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. Journal of Applied Econometrics, 1(1): 29–53. https://doi.org/10.2307/2096536.

[4] Consul, P. C. and Famoye, F. (1992). Generalized Poisson Regression Model. Communications in Statistics - Theory and Methods, 21(1): 89–109. https://doi.org/10.1080/03610929208830766.

[5] Consul, P. C. and Jain, G. C. (1973). A Generalization of the Poisson Distribution. Technometrics, 15(4): 791–799. https://doi.org/10.1080/00401706.1973.10489112.

[6] Creal, D., Koopman, S. J. and Lucas, A. (2013). Generalized Autoregressive Score Models with Applications. Journal of Applied Econometrics, 28(5): 777–795. https://doi.org/10.1002/jae.1279.

[7] Davis, R. A., Dunsmuir, W. T. M. and Street, S. B. (2003). Observation-Driven Models for Poisson Counts. Biometrika, 90(4): 777–790. https://doi.org/10.1093/biomet/90.4.777.

[8] Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. Journal of the American Statistical Association, 81(395): 709–721. https://doi.org/10.1080/01621459.1986.10478327.

[9] Famoye, F. and Singh, K. P. (2006). Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. Journal of Data Science, 4(1): 117–130. http://www.jds-online.com/file_download/96/JDS-257.pdf.

[10] Fokianos, K. (2011). Some Recent Progress in Count Time Series. Statistics, 45(1): 49–58. https://doi.org/10.1080/02331888.2010.541250.

[11] Gorgi, P. (2018). Integer-Valued Autoregressive Models with Survival Probability Driven By a Stochastic Recurrence Equation. Journal of Time Series Analysis, 39(2): 150–171. https://doi.org/10.1111/jtsa.12272.

[12] Harvey, A. C. (2013). Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series. New York: Cambridge University Press. https://doi.org/0.1017/cbo9781139540933.

[13] Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). Univariate Discrete Distributions (Third Edition). Hoboken: Wiley. https://doi.org/10.1002/0471715816.

[14] Koopman, S. J. and Lit, R. (2017). Forecasting Football Match Results in National League Competitions Using Score-Driven Time Series Models. Working Paper. https://www.tinbergen.nl/discussion-paper/608.

[15] Zou, Y., Geedipally, S. R. and Lord, D. (2013). Evaluating the Double Poisson Generalized Linear Model. Accident Analysis and Prevention, 59: 497–505. https://doi.org/10.1016/j.aap.2013.07.017.

BENEFITS OF INVENTORY INFORMATION SHARING IN A HYBRID MTS/MTO SYSTEM

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

Abstract: We model an inventory problem faced by a retailer sourcing a product from the supplier working under hybrid make-to-stock/make-to-order production plan. The retailer can source a product by: (1) immediate supply from the currently available on-hand inventory at the supplier, or/and (2) placing a regular order with the supplier, which is fully replenished with a delay. On-hand stock availability at the supplier is variable and possibly unknown to the retailer. The retailer also faces uncertain market demand for the product and thus uncertain requirements in future periods. We derive the optimal dynamic programming formulation that minimizes the total expected inventory holding and backorder costs over a finite planning horizon. We show the optimality of the myopic optimal policy in the case where on-hand inventory availability information is shared with the retailer. We evaluate the policy numerically to determine the benefits of inventory information sharing, and show relevant implications for the retailers sourcing decisions.

Keywords: inventory, information sharing, dual sourcing, stochastic models, myopic policy.

1. INTRODUCTION

In this paper we address the problem of a retailer procuring a product from a supplier to satisfy stochastic end-customer demand. The supplier's production policy is to produce the component both to stock as well as according to regular orders placed by, possibly many, retailers. Therefore, the supplier's production system can be characterized as a hybrid make-to-stock make-to-order system (MTS/MTO system). On-hand inventory availability at the supplier allows for an immediate replenishment to a retailer, however the size of the replenishment may be limited by the current availability. In the case of the regular order placed with the supplier, the longer delivery lead time allows the supplier to fully process the order and guarantee full delivery.

To reduce retailer's uncertainty about the on-hand inventory availability, the supplier may share the information about the current inventory on-hand. Supplier may even provide information about the future inventory availability, which can be considered as a form of inventory or capacity reservation policy. In this case, an amount of inventory would be allocated to a particular retailer for a number of future periods. This will likely increase the share of sourcing directly from stock, which can eventually lead to a pure MTS system where the supplier would capture the majority of the risk in a supply chain.

We focus on the retailer's perspective, where in each period the retailer needs to decide how to allocate the replenishment between a reliable regular order with positive lead time, and an immediate replenishment from limited on-hand inventory at the supplier. Therefore, if on-hand inventory information is not shared, the retailer faces both uncertain end-customer demand as well as uncertain immediate replenishment. Even in the case where the supplier shares information on current inventory availability and potential future inventory reservations with the retailer, the availability might vary considerably from period to period.

The goal of this study is to analyze how proposed on-hand inventory information scenarios affect the retailers sourcing policy. More specifically, we want to characterize the retailer's optimal policy when replenishment is possible from two sourcing options with different lead times, where sourcing directly from supplier's on-hand inventory is potentially constrained. Further, we want to explore how the optimal policy structure is affected by the inventory information and estimate the benefits of information sharing.

We proceed with a review of the relevant literature. The model we propose is related to two fields of inventory research: capacitated single stage inventory models with information sharing and dual sourcing inventory models. The limited immediate supply availability could be modelled as a an output of the random production capacity at the supplier. This assumption was first discussed in the model by [3]. They show that the optimal policy can be characterized as the base-stock policy, where the optimal base-stock level represents a target inventory position to which order needs to be placed. The base-stock level should account for possible capacity shortfalls in future periods, which is done through the accumulation of inventory. Similar setting is also analysed in papers by [8, 5]. In addition, the work by [6, 1, 2] is closely related to the way we model inventory information sharing. They assume that the retailer is able to obtain the information about the supply availability for a limited number of future periods, usually denoted as advance supply/capacity information, while the supply availability in the following periods remains uncertain. They show that the optimal policy remains of a base-stock type, however the base-stock level is now state-dependent on the available advance supply information. Our paper differs from the above due to considering a dual sourcing setting in which advance information is shared.

As the retailer can also opt to place a regular order with the supplier with a delivery lead time of one period, the problem can also be seen as a dual sourcing inventory problem. The dual sourcing inventory literature mostly explores the trade-off between the procurement cost and the responsiveness of the two supply channels. The early papers by [4, 9] analyze the setting in which the lead times of the two supply channels differ by a fixed number of periods. For the case of so-called consecutive lead times (lead times differ by one period) they show the optimality of the two-level base-stock policy. However, all these papers assume that the inventory availability for immediate replenishment is not limited.

Thus, it is the dual sourcing papers that assume either one or both supply channels are capacitated (have limited inventory availability), which are of particular interest to us. While most of the research is focusing on identical lead time supply channels, the work by [12, 10, 11] assumes different lead times and fixed capacity limit at the faster supply channel. When faster supplier has stochastic capacity, [7] show that the optimal policy has a complex structure even in the simplest case of lead time of zero and one, however it can be approximated well by the myopic policy. Our work extends these results by assuming the inventory availability varies from period to period, and the information is shared with the retailer about the current (and future) on-hand inventory availability in every period.

The remainder of the paper is organized as follows. We present the model formulation and derive the optimal cost formulation in Section 2. In Section 3, the myopic policies for the inventory information sharing scenarios are derived. The benefits of information sharing are assessed by means of numerical analysis in Section 4. We summarize our findings in Section 5.

2. MODEL FORMULATION

We consider a one-product inventory system with the retailer sourcing from a single supplier working under hybrid MTS/MTO production policy. Retailer is facing stochastic end-customer demand D_t in period t, assumed to be stationary and independent across periods. In each period t retailer places orders with the supplier, where quick order q_t is satisfied immediately from the on-hand stock s_t at the supplier (with zero lead time), depending on the current availability. The on-hand stock availability at the supplier is assumed exogenous to the retailer, but with known probability distribution (we assume i.i.d. on-hand stock availability over periods). In addition to the quick order, the retailer places a regular order with the supplier, which is always replenished in full with a fixed one-period lead time.

The inventory information is given by the vector $(s_t, \vec{s}_t) = (s_t, s_{t+1}, \ldots, s_{t+n})$, where we study the problem under different lengths of the information horizon n. For $n = 0$ the information is only available for the current on-hand inventory availability s_t, while in the case of $n > 0$ future on-hand inventory availability in the form of inventory or capacity reservations is represented by $\vec{s}_t$. We assume that in each period t the retailer receives the information about the inventory reservation s_{t+n} in period $t+n$. Observe that in the *no-information case* the retailer is faced with unknown on-hand stock availability and the actual realization of a quick order is only revealed upon replenishment.

The goal is to find the optimal policy that minimizes the relevant retailer's costs for different inventory visibility scenarios. Unsatisfied demands at the retailer are backordered, and inventory holding and backorder costs are charged in each period. We reasonably assume there is no price difference whether a product is sourced directly from stock or through the regular order. While immediate fulfillment from on-hand inventory could be considered as a premium service, unreliable (and potentially unknown) availability does not justify additional charges. We summarize the notation in Table 1.

Table 1: Summary of the notation.

T	:	number of periods in the finite planning horizon
h	:	inventory holding cost per unit per period
b	:	backorder cost per unit per period
α	:	discount factor ($0 \leq \alpha \leq 1$)
q_t	:	quick order placed in period t
r_t	:	regular order placed in period t
x_t	:	inventory position before ordering in period t
y_t	:	inventory position after placing a quick order in period t
w_t	:	inventory position after placing a regular order in period t
d_t, D_t	:	realization and random variable of demand in period t
s_t, S_t	:	realization and random variable of on-hand inventory at the supplier in period t
n	:	length of on-hand inventory information horizon, $n \geq 0$
$s_t, \vec{s}_t$	:	on-hand inventory information in period t

We assume the following sequence of events. (1) At the start of the period, the retailer receives the regular order r_{t-1} placed in the previous period and information about the on-hand stock availability s_{t+n} for period $t+n$ (if available), and reviews the inventory position before ordering x_t. (2) Quick order q_t and regular order r_t are placed with the supplier, where q_t is placed depending on the current on-hand inventory availability at the supplier s_t, $q_t \leq s_t$. (3) Quick order q_t is replenished. Observe that in the no information case s_t is revealed upon replenishment and correspondingly the inventory position is corrected by $(q_t - s_t)^+$ if q_t is not fully replenished, where $(q_t - q_t)^+ = \max(q_t - s_t, 0)$. (4) At the end of the period, demand d_t is observed and satisfied, where the relevant costs are incurred based on the end-of-period net inventory.

It is convenient to introduce two inventory positions after placement of orders. First, after placing order q_t the inventory position is raised to y_t, $y_t = x_t + q_t$, and subsequently, after order r_t is placed, the inventory position is raised to w_t, $w_t = x_t + q_t + r_t$. Observe that the sequence in which orders are placed makes no difference, as long as both are placed after s_t is obtained in the case of $n \geq 0$ (or before s_t is revealed in the *no-information case*) and before demand d_t is revealed.

The expected single-period holding and backorder costs in period t are $C_t(y_t) = E_{D_t}(h(y_t - D_t)^+ + b(D_t - y_t)^+)$. The minimal discounted expected cost function that optimizes the cost over finite planning horizon T from time t onward and starting in the initial state $(x_t, s_t, \vec{s}_t)$

can be written as:

$$f_t(x_t, s_t, \vec{s}_t) = \min_{x_t \le y_t \le x_t + s_t, y_t \le w_t} \{C_t(y_t)$$
$$+\alpha \begin{cases} E_{D_t} f_{t+1}(w_t - D_t, s_{t+1}, \vec{s}_{t+1})\}\}, & \text{if } T-n \le t \le T, \\ E_{D_t, S_{t+n+1}} f_{t+1}(w_t - D_t, s_{t+1}, \vec{s}_{t+1})\}\}, & \text{if } 1 \le t \le T-n-1, \end{cases} \tag{1}$$

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

3. MYOPIC OPTIMAL POLICIES UNDER INVENTORY INFORMATION SHARING

In this section, we construct the myopic policy with two base-stock levels that optimizes the extended single period problem. We show that the two base-stock levels correspond to the optimal solutions of (1) for the $n = 0$ scenario, which means that myopic policy is also optimal in the multiperiod setting. Observe first that (1) is additively separable in the two decision variables y_t and w_t. We first analyze the scenario where only current on-hand inventory availability is shared with the retailer($n = 0$). We assume that $\hat{y}$ minimizes $C(y)$, $\hat{w}$ minimizes $C(\hat{y}) + \alpha E_{S_{t+1}} C(\max(w - D, min(w - D + S_{t+1}, \hat{y})))$. The intuition behind these formulations is that $\hat{y}$ minimizes the single period cost, and $\hat{w}$ minimizes the possible overshoot or undershoot of $\hat{y}$ in the next period. The overshoot and the additional holding costs occur when $w - D > \hat{y}$, while the undershoot occurs when $w - D + S_{t+1} < \hat{y}$ and incurs additional backorder costs.

We rewrite (1) into an equivalent multiperiod cost formulation where $\tilde{f}_t(x_t, s_t) = f_t(x_t, s_t) - C(\max(x_t, \min(x_t + s_t, \hat{y})))$:

$$\tilde{f}^n_t(x, s, \vec{s}) = \min_{x \le y \le x+s, y \le w} \{C^n(y) + \alpha E_{D, S_{n+1}} C^n(w - D \vee \hat{y} \wedge w - D + s_{+1})$$
$$+\alpha \left\{ \begin{array}{l} E_D \tilde{f}^n_{t+1}(w - D, s_{+1}) \\ E_{D, S_{n+1}} \tilde{f}^n_{t+1}(w - D, s_{+1}, \vec{s}_{+1}) \end{array} \right\}\} - C(x \vee \hat{y} \wedge x + s), \quad \begin{array}{l} \text{if } T-n \le t \le T, \\ \text{if } 1 \le t \le T-n-1, \end{array} \tag{2}$$

and the resulting myopic formulation of (2) is:

$$\min_{x_t \le y_t \le x_t + s_t, y_t \le w_t} \{C_t(y_t) + \alpha E_{S_{t+1}} C(\max(w_t - D_t, \min(w_t - D_t + S_{t+1}, \hat{y})))\} \tag{3}$$

The two inventory positions after ordering y_t and w_t under the optimal myopic sourcing policy are therefore as follows:

$$y_t(x_t) = \max(x_t, \min(\hat{y}, x_t + s_t)) \tag{4}$$

$$w_t(y_t) = \max(y_t, \hat{w}) \tag{5}$$

To show that the optimal myopic base-stock levels $\hat{y}$ and $\hat{w}$ also satisfy (2) and therefore (1), we first see that myopic policy does not depend on t. Further, we need to check whether the myopic policy guarantees that the set of consistent states are visited in the next period from which the optimal decisions can be taken. We omit the full proof, but one can quickly see that when $x_t \le \hat{w}$ holds, only steady states will be visited in the future periods.

Due to the space constraint we limit the analysis for the $n = 1$ scenario to the following observation. The construction of a myopic policy follows the same steps as above. The only difference is that s_{t+1} is known when orders are placed in period t, thus s_{t+1} is a part of the state description in period t and the optimal base stock level $\hat{w}(s_{t+1})$ now becomes a function of the future on-hand inventory availability. However, due to the dependence of $\hat{w}_t$ on s_{t+1} it can happen that we overshoot the optimal base-stock level, thus the inventory position w_t is not consistent. Therefore, the myopic policy is not optimal in this case, however numerical results show that it provides a near-perfect approximation to the optimal one. We provide additional intuition behind the myopic policy and the effect of the information sharing in the following section.

4. BENEFITS OF INVENTORY INFORMATION SHARING

In this section, we evaluate the benefits of inventory information sharing by means of the numerical analysis. Due to the myopic optimality of the sourcing policy, we already know that the full benefits are gained by sharing information on current on-hand inventory s_t and future availability s_{t+1} in the next period. Observe that the main driver of cost reduction is decrease in uncertainty of on-hand inventory availability through information sharing. Therefore we are primarily interested in how underlying uncertainty of inventory availability (and subsequently variability in s_t, when uncertainty is resolved) affects the value of information sharing.

In Figure 1 we present the cost comparison between the base scenario without information sharing and the scenario in which on-hand inventory availability is known for the current and future period. We define the relative on-hand inventory availability as the ratio of the average on-hand inventory and the average demand, $\zeta = E(S)/E(D)$. CV_D and CV_S denote the coefficients of variation of demand and on-hand inventory variability.

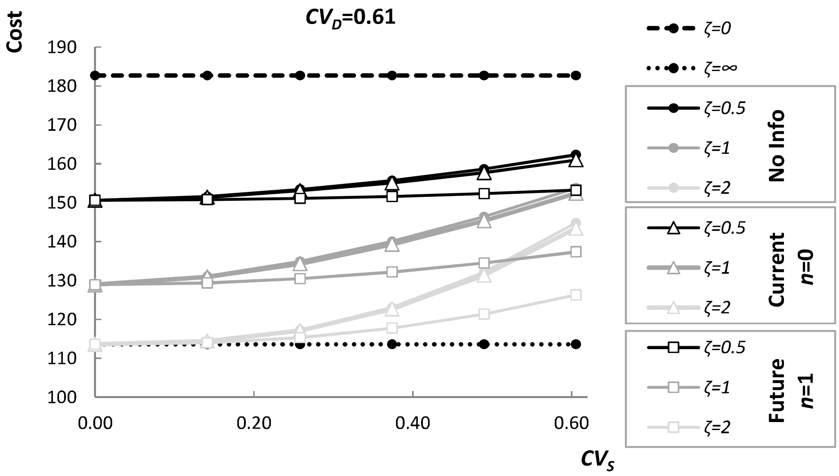


Figure 1: System costs under inventory information sharing scenarios

The cost curve depicted as $\zeta = \infty$ represents the case with the lowest costs, where sourcing is done solely from abundant on-hand inventory (effectively pure MTS system). In the worst case, depicted by $\zeta = 0$, supplier only fulfills regular orders (pure MTO system) and on-hand inventory availability is zero. Our interest lies in studying the intermediate cases.

While one might expect that knowing current on-hand inventory availability is critical for a retailer, we show that this is not the case. For a single sourcing system with stochastic limited capacity [6] even show that knowing immediate capacity availability (a synonym for the on-hand inventory availability) does not improve the performance. In a dual sourcing system under $n = 0$ we observe the cost reduction of up to 1.5%. Intuitively, knowing s_t enables a retailer to react to the lack of on-hand availability in the current period by increasing the regular order r_t. With this, the inventory position before ordering $x_t = \hat{w} - D_t$ in period $t + 1$ is closer to the optimal one, and the likelihood of inventory shortage in period $t + 1$ is decreased.

For the case of future inventory visibility ($n = 1$) the cost savings become considerable, approaching 20% for the cases presented. In a myopic sense, all inventory availability uncertainty is resolved. Knowing s_{t+1} enables the retailer to adapt the optimal base stock level $\hat{w}(s_{t+1})$ accordingly. When s_{t+1} is low, $\hat{w}(s_{t+1})$ is increased to reduce the likelihood of a stockout in $t + 1$. And vice versa, when s_{t+1} is high, $\hat{w}(s_{t+1})$ is reduced to avoid the probability of an overstock. Observe that in both information sharing scenarios a quick order q_t is always placed up to $\hat{y}$, if s_t allows it. Thus, the benefit of inventory information sharing lies in choosing the regular order r_t so that we will end up in the best possible starting position in the next period.

5. CONCLUSIONS

In this paper, we analyze the effect of sharing supplier's on-hand inventory information on the retailer's optimal sourcing policy. When sourcing from a hybrid MTS/MTO system, the retailer can take advantage of immediate replenishment from the limited on-hand inventory availability as an alternative to placing regular orders with longer lead time. We show that myopic policy is optimal if the current inventory availability is known to the retailer, and is characterized by the two base-stock levels to which the immediate and regular order are placed in every period. For the case of sharing future availability information, the myopic policy provides a near perfect approximation of the optimal policy. By means of numerical analysis, we show the benefits of inventory information sharing. The cost reduction is the highest in a setting with high variability of on-hand inventory availability, especially when average inventory availability relative to end-customer demand is high and demand uncertainty is low. This work can be extended to analysis of a combined supplier-retailer system, where one could examine how the information sharing affects the supplier's production decisions.

References

[1] Atasoy, B., Güllü, R. and Tan, T. (2012). Optimal inventory policies with non-stationary supply disruptions and advance supply information. Decision Support Systems, Vol(53(2)): 269–281.

[2] Çinar, E., and Güllü, R. (2012). An inventory model with capacity flexibility in the existence of advance capacity information. Decision Support Systems, Vol(53(2)): 320–330.

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

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

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

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

[7] Jakšič, M., and Fransoo, J. (2018). Dual sourcing in the age of near-shoring: trading off stochastic capacity limitations and long lead times. European Journal of Operational Research, Vol(267): 150–161.

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

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

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

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

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

RATIONAL OR IRRATIONAL? - Pension Expectations in Hungary

Professor Dr. Erzsébet Kovács
Head, Department of Operational Research and Actuary Sciences
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: erzsebet.kovacs@uni-corvinus.hu

Ágnes Vaskövi
Assistant Professor, Department of Finance
Corvinus University of Budapest
Fővám tér 8, Budapest 1093, Hungary
E-mail: agnes.vaskovi@uni-corvinus.hu

Abstract: This study demonstrates expectations on social security benefits and pension awareness of Hungarian university students. Data were collected from a survey. Information was composed by principal component analysis to test two hypotheses using two factors. The authors attempt to find explanations for the differences in expectations according to gender and university majors.

Keywords: retirement, social security expectations, financial literacy

1 INTRODUCTION

In our study we examined survey data on university students' expectations as to their future retirement income and social security benefits as a major source of this income in Hungary. We defined the group of respondents to be highly educated and fairly far from retirement in order to have a sample with relatively good financial literacy but also long uncertain period before retirement. We conceived four groups of questions that explore (i) expectations as to future retirement income, (ii) generosity of benefits, (iii) retirement age to be eligible for social security benefits, and (iv) demographic data. Similar survey was conducted before us in the USA, Canada and Ireland (Turner et al., 2019), and parallel to our research the same survey is conducted in the University of Gdansk, Poland.

We stated two hypotheses:

1. respondents with finance major have higher pension literacy than other students, and
2. pension expectations do not depend on gender.

As per the first hypothesis, we found that students in finance major are aware of the interdependencies of longer working period and higher retirement benefits, however we did not identify significant difference between the respondents' expectations, i.e. most of survey participants expect low level state pension. Concerning the second null-hypothesis, we explored male respondents expect later retirement than female participants.

2 HUNGARIAN PENSION SYSTEM

In this section we shortly present the characteristics of Hungarian pension system in order to form context for the expectations analysis.

The Hungarian pension system is currently a statutory, one-pillar pay-as-you-go scheme where all employees are covered and the pension is calculated on the basis of earnings and working years. The mandatory second pillar was abolished in 2010 when most of savings in mandatory private pension funds were redirected back to state pension system. Third and fourth pillars of the pension system are voluntary with individual and employer-financed saving forms. Pillars of Hungarian pension system are shown in Table 1.

Table 1: Pillars of Hungarian pension scheme

	Mandatory	Finance	Attributes
Pillar I	yes	pay-as-you-go (DB)	statutory
Pillar II	-	-	abolished in 2010 and funds transferred to state funds, appr. 50 000 individual accounts remained
Pillar III	no	defined contribution	voluntary pension funds with individual accounts
Pillar IV	no	individual contributions	other retirement saving schemes with tax subsidies

The retirement age was increased gradually from 62 years up to 65 by 2022. Thus, Hungarian pension system is unisex with one exception, i.e. women with 40 years eligibility period can retire before this increased statutory retirement age (this program is called "Women40").

Basically in the last 10 years some fundamental changes were applied on the Hungarian pension system, however sufficient financing of DB system is still an open question after 2035 (Bajkó et al., 2015). This may have effect on expectations and employees should have higher pension literacy to maintain their living standard after retirement.

3 SURVEY DATA AND METHODOLOGY

Our survey consists of 14 questions: five questions on demographic data and other nine questions on social security benefits. These nine questions are grouped as follows:

i. expectations as to future retirement benefits (4 questions)
   we asked if students expect to get social security benefits, what level of trust they have that the government would provide these benefits, and what is their attitude to Hungarian pension system,
ii. generosity of benefits (2 questions)
   these two questions are examining what proportion of total future retirement income would come from the state pension system, and the estimated salary replacement rate (it is the proportion of social security benefits related to the worker's salary near the end of his or her working career),
iii. receipt of social security benefits and retirement age (3 questions)
   we also inquired if students plan to leave Hungary for their working years and which are the destination countries, and we formed two questions about retirement age and earliest possible age at which retirement benefits might be received.

We gathered 320 responses out of which we excluded 70 because of any invalid data, thus in our final dataset there are 250 records with the demographic data distributions as Table 2 shows:

Table 2: Demographic data distributions of survey data

Gender	Male Female	58.0% 42.0%
Age	Between 19 and 24 years Between 25 and 42 years	85.6% 14.4%
Present major	University – finances University – economics (not fin) Other university	35.0% 64.0% 1.0%
Region of origin	Budapest and Pest county Other counties	40.0% 60.0%
Residence	Budapest and Pest county Other counties	87.6% 12.4%

We used factor analysis which is a multidimensional statistical method to compress information of the original variables, thus to reduce dimension of the analysis. From the numerous methods of factor analysis, we applied Principal Component Analysis (PCA) for factor extraction, where the uncorrelated linear combination of variables is calculated based on eigenvalue-eigenvector decomposition of correlation matrix of the original variables included in the model. Factors driven from PCA were used to test the two hypotheses.

4 RESULTS

Based on data of Hungarian Central Statistical Office and EUROSTAT, slightly above 20% of total population received retirement benefits, which amount was 65% of the net average wage in 2017. This salary replacement rate decreased from 74% to 65% within 6 years from 2011 to 2017. The increase of population receiving social security benefits and the fall of salary replacement rate are warning signals of the Hungarian pension system. Most of the population might not be aware of the exact pension figures however the survey outcome shows that people's expectations are overly pessimistic.

4.1. Expectations as to future retirement benefits

4.0% of the respondents expect to not receive any future retirement benefits, and on the other hand, also 4.0% are 100% confident that the state-pension will be there when they retire. 50% of the respondents considered that the probability of receiving state-provided social security retirement benefit is less than 50%, however 72.8% expressed their negative expectations as to their state pension. The aspects why the survey participants' attitude might be negative are described in Table 3.

Table 1: Different reasons of negative attitude to social security benefits

Why do you have certain views on social security benefits?	Frequency
The government has frequently changed social security benefits.	5.6%
Hungary's future is uncertain.	8.8%
I don't trust the government.	10.0%
I plan to leave Hungary thus I might not get Hungarian social security benefits.	4.4%
The increasing percentage of the population that is retired will cause a reduction in the generosity of social security benefits.	40.4%
Others	3.6%

The level of trust that the government will provide future promised social security retirement benefits is overly low. 20% of respondents say that they do not trust in government at all, and 66% have less than 50% trust. There was only 1.6% who fully trust in the social security payments of the government.

4.2. Generosity of benefits

Currently the salary replacement rate in Hungary depends on the years worked and the average net salary. If a person has 40 years of working period he receives 80% of last average net wages, if he has worked 50 years the rate increases up to 90%. Despite, 58.4% of respondents said they would receive less than 50% of the salary as social security benefits and only 3,6% estimated the salary replacement rate at 80% or higher.

The state pension proportion of total retirement income was underestimated, 61.6% of students said it would be lower than 50% and only 1.2% expects to get more than 90% of total income from the social security funds.

4.3. Receipt of social security benefits (age related questions)

We asked students where they want to spend their working years and only 46.8% responded they want to stay in Hungary. We also asked the expected retirement age, where 54.4% expect to retire later than age of 66 years and only 16.8% expect early retirement (before 65 years).

Respondents also expressed their expectations as to the earliest age at which they would receive social security retirement benefits. More than ¾ (77.6%) said this earliest age would be at least 65 years, and only 10% expected this age somewhat before 62 years.

4.4. Factor Analysis

To test our null-hypothesis, a linear factor model was created with Principal Component Analysis. With the development of uncorrelated components, we sought to find out the relationship between the three different groups of social security questions. The best fitting model includes six variables equally distributed among the 3 types of pension expectations variable groups: 2 of expectations, 2 of generosity and other 2 of receipt (age). In this factor model, two orthogonal, uncorrelated components with eigenvalue higher than 1 were extracted from the six original variables, with an extraction of 72.247% of information as a ratio of the original variances. The suitability of our data for PCA is shown in Table 4.

Table 2: Measures of data suitability for PCA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.762
Bartlett's Test of Sphericity	Approx. Chi-Square	592.789
	df	15
	Sig.	0.000

KMO value is 0.762 meaning the data is suitable for principal component analysis. The correlations between the original variables and the two components are shown in Table 5. The empty cells represent very weak correlations.

Table 3: Rotated component matrix of two-factor solution

		Component	
		Expectations &generosity	Age
Expectations	Expectations to get social security benefits	0.838	
	Level of trust that the government would provide social security benefits	0.829	
Generosity	Proportion of total future retirement income coming from the state pension system	0.902	
	Salary replacement rate	0.832	
Receipt (age)	The earliest age at which the pension system allows to collect social security retirement benefits		0.829
	Expected retirement age		0.842

4.5. Hypothesis testing

We formed two hypotheses on the students' pension expectations:

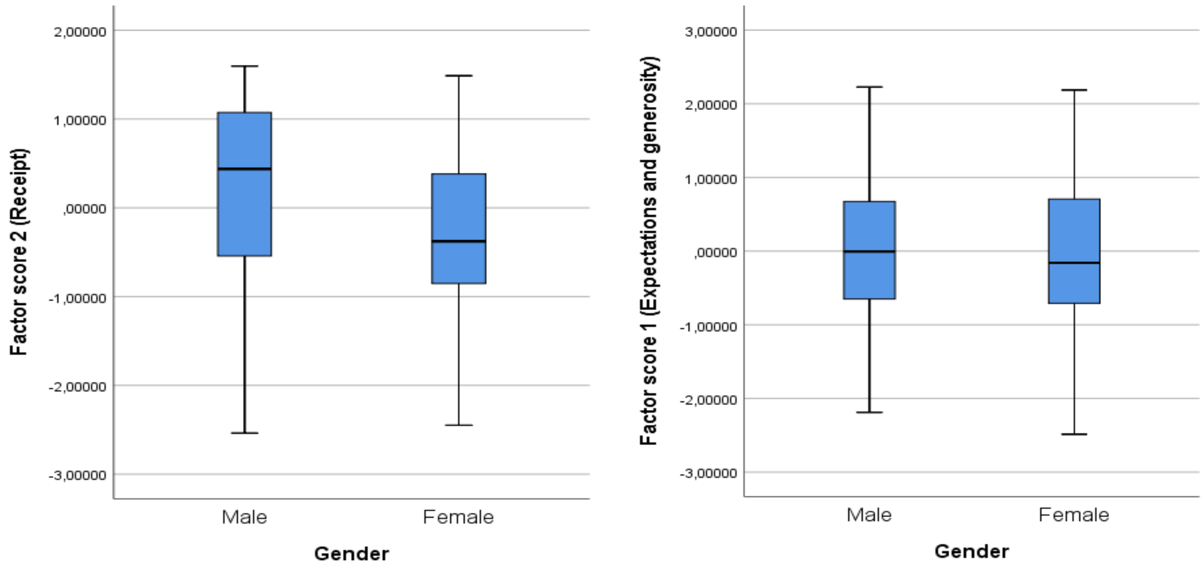
H1: respondents with finance major have higher pension literacy than other students, and

H2: pension expectations do not depend on the gender, which is in line with the unisex pension scheme.

Considering H1 we found that students in finance major are aware of the interdependencies of longer working period and higher retirement benefits (mean of receipt factor score is 0.124 by finance major and -0.058 by respondents in other major), however we did not find significant expectations difference between the groups of majors (independent t-test value for Factor 1 (Expectations & Generosity) is -0.605 with p=0.546, and t-test value for Factor 2 (Age) is 1.368 with 0.173 p-value, which is not significant on any of significance levels).

Considering H2 we found statistical evidence that male respondents count on higher retirement age (mean of age factor score is 0.176 by male and -0.248 by female respondents). Despite of the unisex pension scheme the reason of later retirement expectations for men might be the awerness of Women40 program or the former gender-defined system. Figure 1 shows gender differencies with regards on two factors. Factor 2 (Age) is significantly different on male and female respondents on any significance levels (t-test value is 3.373 with p-value of 0.001). Factor 1 is not significantly different (t-test value=0.517 and p=0.605).

Figure 1: Differences according to gender expectations



4.6. Results comparing to surveys in the USA, Canada, and Ireland

In spite of the similarities between our research and the research conducted by Turner et al. (Turner et al., 2019) the comparison of the results should be very cautious. The survey countries are fairly different with regards on their pension schemes and their future and past reforms, all countries' respondents are overly pessimistic regarding the future social retirement benefits. All countries' survey data indicate that underestimating the future social retirement income does depend nor on the gender, neither on financial literacy of the respondents. In Hungary, students are as much pessimistic as workers in Ireland, in the USA or in Canada. Turner et al. found that these perceptions might change slightly when respondents get older. In our survey we could not explore this phenomenon since our survey participants were university students, but this question could lead us to further research area.

Other important statement is that the pessimistic attitude of respondents does not depend on the actual state of pension systems, whether their financing is sufficient on long term. In Canada and Ireland, where the pension system is considered rather stable than in the USA or in Hungary, expectations are also pessimistic.

5 CONCLUSIONS

In our study we explored survey data collected among Hungarian university students about their future pension expectations. We found that most of the respondents are characterized with a general pessimism, and despite of unisex pension scheme male participants expect later retirement age. We also wanted to find contingency between the financial literacy and pension expectations, but we could only identify students with finance major might have more accurate information about proportionality of higher retirement age and higher potential social security benefits. Further research is necessary to make conclusions about the rational or irrational expectations, and the reasons of the pessimistic approach on pension among university students.

Acknowledgement
This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project EFOP-3.6.2-16-2017-00017, titled "Sustainable, intelligent, and inclusive regional and city models".

References

[1] Bajkó, A., Maknics, A., Tóth, K., Vékás, P. 2015. On the Sustainability of the Hungarian Pension System – The Long-term Effects of Demographic Trends. (in Hungarian) Közgazdasági Szemle. Vol. 62. December 2015.1229-1257.

[2] Turner, J., Zhang, S., Hughes, G., Rajnes, D. 2019. Irrational Expectations, Future Social Security Benefits, and Life Cycle Planning. The Journal of Retirement. Winter 2019, 6 (3) 60-68.

[3] Dr. Barta, J. 2018. A nyugdíjbiztosítások szerepe és jellemzői a kiegészítő nyugdíjpillérben. Biztosítás és Kockázat, Vol (5) No 1: 42-49. (in Hungarian)

[4] Borlói, R. 2016. Gondolatok a magyar nyugdíjrendszerről. Budapest: Gondolat Kiadó. (in Hungarian)

[5] Thompson, L.:Older &Wiser, 1998. The Economics of Public Pensions, Univ. Press of America

[6] Mészáros, J.(Ed): Pension Adequacy and Sustainability, 2013. Pension Seminar, Budapest

EFFICIENCY TEST AS THE BENCHMARK FOR MINIMUM-RISK PORTFOLIO OPTIMIZATION STRATEGIES

Aleš Kresta and Anlan Wang
VŠB – Technical University of Ostrava, Department of Finance
Sokolská tř. 33, 702 00 Ostrava, Czech Republic
E-mail: ales.kresta@vsb.cz, anlan.wang.st@vsb.cz

Abstract: In the paper we propose the efficiency test, which can be considered as the benchmark for portfolio optimization strategies. The test is based on the comparison of out-of-sample performance of considered strategies with randomly created portfolios. We present the practical application of the test in the period of global financial crisis and we perform one-year rolling window tests in the period 2009-2019. According to the empirical results, considered minimum-risk portfolios worked efficiently during the global financial crisis, however, the results are not robust as confirmed by rolling-window approach.

Keywords: financial crisis, hypothesis test, minimum-risk portfolio, portfolio optimization.

1 INTRODUCTION

Portfolio optimization involves the efficient investment strategy considering optimal allocation of limited funds. Investing in a portfolio of different assets is a good way to spread the risk with respect to risk aversion. Under the framework of mean-variance model by Markowitz [3], it states that there exists the inter-relationship between the risk and return of a portfolio. A higher risk level means a higher potential return. Along with the considerations of real-life conditions and enhancements of algorithms, additional constraints have been developed to the early classical Markowitz model in the later studies, especially with the development of the measurements of financial risk, the variance in the mean-variance model is replaced by various risk measures such as the mean absolute deviation, value-at-risk, conditional value-at-risk, etc.

Different models have been proposed to address the various needs with respect to investors. In the pioneers' researches, although the applied performance measurements are made to test the proposed models with experimental data, it's still significant to involve a benchmark in the analysis to verify the efficiency of models. Below we list some commonly used benchmarks.

One of the mostly applied benchmarks is naive strategy also called as the 1/N diversification strategy. DeMiguel et al. [1] evaluated and compared the performance of several optimization methods with respect to the performance of the 1/N strategy and they found that the effect of estimation error on return probability distribution is so large in those optimization models, but this type of error can be avoided by using the 1/N weights. Owing to these good points of 1/N strategy, Xidonas et al. [8] applied the uniform 1/N assets allocation method as a benchmark in order to gauge the robust allocation strategy performance in their research.

Another commonly applied benchmark is the classical mean-variance model. For instance, Fulga [2] presents an approach which incorporates loss aversion preferences in the mean-risk framework and the efficiency of the new approach is tested against the mean-variance model. The mean-VaR optimization model is also a popular benchmark, which appears in the portfolio optimization researches. For instance, to illustrate the effectiveness of the proposed univariate-GARCH-VaR model Ranković et al. [4] compared the results with two benchmarks, the first one is the mean-historical-VaR optimization model, and the second one is the mean-multivariate-GARCH-VaR optimization model.

The indices are also commonly applied as benchmark to compare the performance of proposed methods. For instance, Solares et al. [7] used the dataset of 13 years' historical monthly prices of stocks in the Dow Jones Industrial Average index (DJIA), and they also made an extensive evaluation comparing the performance of the proposed approach with respect to the DJIA index. Interestingly, in their research, they pointed out that the main contraindication of using market indices as the benchmark is that the profitability of portfolios is often compared to popular indices, so most investors expect to reach or exceed the yields of these indices over time, but the problem with this expectation is that they are at a disadvantage because there is no guarantee that the characteristics of the stocks in their portfolio coincide with the characteristics of the stocks contained in the index. So, to avoid this trap, it is suggested to incorporate into the portfolio only the stocks of the index being considered as a benchmark.

However, we also found that in some researches the application of benchmark is missing. This is not rigorous enough from the basis of scientific evaluation, because the verification of a new findings is a significant step to test the efficiency of the approaches when it's applied in the real world.

In this sense, to verify the efficiency of considered portfolio optimization strategies, in this paper we also generate the random-weights portfolios to make the hypothesis tests by comparing their performances with those of the strategy portfolios. The goal of the paper is thus twofold. Firstly, we propose the efficiency test for out-of-sample performance of portfolio optimization strategies. Secondly, we apply the test to find out which portfolio optimization strategy minimizes the risk in the out-of-sample period.

The paper is divided into four sections. In following section, the theoretical basis of this paper is introduced. In order to verify the efficiency of the portfolio optimization strategies, the empirical analysis is made in section 3. In section 4, we conclude the paper.

2 PORTFOLIO OPTIMIZATION METHODS

Portfolio optimization aims at obtaining a portfolio with maximum return, minimum risk or their combination. In this paper we consider only the minimum-risk portfolio, however, we apply three measures of the risk, namely variance (alternatively standard deviation, henceforth STD), Mean Absolute Deviation (henceforth MAD) and Conditional Value at Risk (henceforth CVaR). In this section, we describe the algorithm of these optimization models as well as the basic rules of hypothesis tests applied in the out-of-sample period.

2.1 Mean-Risk optimization method

Mean-Risk model is based on the framework of analyzing the inter-relationships between expected return and risk of a portfolio. We denote x_i as the weight of asset i in the portfolio. In our case, we exclude short sales, so the values of x_i satisfies $x_i \geq 0$ for all assets. We suppose that the expected stock return is identical to the average of the historical stock returns within the chosen in-sample period according to [9]. If we denote $E(R_i)$ as the expected return of asset i then the expected return of a portfolio $E(R_p)$ is the weighted average of $E(R_i)$:

$$E\left(R_p\right) = \sum_{i=1}^{N} x_i \cdot E(R_i) = x^T \cdot E(R), \tag{1}$$

where N is the totoal number of assests in the portfolio, $x = [x_1, x_2, \dots, x_N]^T$ and $E(R) = [\,E(R_1), E(R_2), \dots, E(R_N)]^T$. The sum of x_i in a portfolio equals to 1.

The mean-variance model regards portfolio's variance or standard deviation as the risk measure. They are calculated by the covariances $\sigma_{i,j}$ of the asset returns for all asset pairs

(i,j). We denote a $N \times N$ covariance matrix as $\mathbf{Q}$, $\mathbf{Q} = [\sigma_{i,j}, i = 1,2,\dots,N, j = 1,2,\dots,N]$, we show the calculations of the variance ${\sigma_p}^2$ and the standard deviation σ_p of a portfolio separately in equation (2) and equation (3), where the standard deviation is the square root of the variance.

$$ {\sigma_p}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} x_i \cdot \sigma_{i,j} \cdot x_j = x^T \cdot \mathbf{Q} \cdot x, \tag{2} $$

$$ \sigma_p = \sqrt{{\sigma_p}^2}. \tag{3} $$

The minimum-variance portfolio can be found by solving the following optimization problem,

$$ \begin{aligned} &\text{minimize } {\sigma_p}^2 \\ &\text{subject to} \\ &\sum_{i=1}^{N} x_i = 1 \\ &x_i \geq 0, i = 1,\dots,N. \end{aligned} \tag{4} $$

The mean-MAD optimization method was proposed as an alternative to the Markowitz model, the only difference in mean-MAD method is the risk measure, which changes from the variance to the mean absolute deviation of the portfolio's returns. The minimum-MAD portfolio can be found in the similar way as in (4), the only difference is to change the objective function to $minimize\ MAD$, where the calculation of MAD is shown in equation (5), where T is number of observations, $R_{i,t}$ is the return of asset i for each time t,

$$ MAD = \frac{\sum_{t=1}^{T} \left| \sum_{i=1}^{N} [x_i \cdot R_{i,t} - E(x_i \cdot R_{i,t})] \right|}{T}. \tag{5} $$

With the development of the measurements of financial risk, Vale at Risk (henceforth VaR) has been widely used as the risk measure. VaR is defined as the worst-case loss associated with a given probability and time horizon. However, rather than the application of VaR, the CVaR which indicates the expected loss under the condition of exceeding VaR is applied as the risk measure in this paper. The CVaR for a portfolio is also known as the expected shortfall and its definition can be found in [5, 6].

2.2 Hypothesis tests

To evaluate the performances of the strategy portfolios under the applied portfolio optimization methods, we calculate the portfolio's out-of-sample maximum drawdown (henceforth MDD), STD and MAD of the portfolio returns and CVaR.

We verify the efficiency of strategies under the optimization models by generating random-weights portfolios and making hypothesis tests. The risk measures applied in the hypothesis tests are the ones listed above. We will demonstrate the applied test on MDD, which indicates the maximum loss from a peak to a trough of an investment's wealth, so the smaller the MDD, the better performance of the portfolio.

As it literally means, the weights of assets in each random-weights portfolio are generated randomly, in our case we set up 50,000 random-weights portfolios, and in each portfolio the sum of weights equals to 1. We know that a hypothesis test relies on the method of indirect proof. That is, to prove the hypothesis that we would like to demonstrate as correct, we show that an opposing hypothesis is incorrect. In our case, the strategy portfolios under the optimization methods are more likely to be demonstrated as efficient, so according to the rule of hypothesis tests, we can make the null hypothesis and alternative hypothesis as follows:

null hypothesis—H_0: $MDD_s = MDD_r$,
alternative hypothesis—H_A: $MDD_s < MDD_r$.

MDD_s is the MDD of wealth evolutions of strategy portfolio, and MDD_r is the MDD of wealth evolutions of random-weights portfolio. In our hypothesis test, the p-value is the proportion of the random-weights portfolios which meet the condition$MDD_r < MDD_s$. We set the significance level to 10%. If p-value < 10% then we reject H_0, which means the performance of the strategy portfolio is better than that of the random-weights portfolio, so the strategy is efficient; if p-value ≥ 10%, we fail to reject H_0, which means the performance of the strategy portfolio makes no difference from that of the random-weights portfolio, so the strategy is inefficient in this case.

3 EMPIRICAL ANALYSIS

The empirical analysis is made in this section, we obtain the minimum risk portfolios (in-sample) and measure their performances (out-of-sample). The random-weights portfolios are also generated to make the hypothesis tests to verify the efficiency of the obtained portfolios. In the first subsection we demonstrate the approach in 2007-2009 period. In the second subsection we present the results on the one-year rolling window basis in the period 2009-2019.

3.1 Global financial crisis period

The dataset is the daily closing prices of the components of Dow Jones Industrial Average index (DJIA). There are 29 stocks included in our analysis and the missing one is the stock of Visa Inc. due to the incomplete data in the chosen period. The sample time duration is 3 years, which is from March 7, 2006 to March 2, 2009, and we divide the whole sample evenly into the in-sample period (March, 7, 2006 – August 31, 2007) and the out-of-sample period (September 1, 2007 – March 2, 2009). In Figure 1, we can see that in the in-sample period the DJIA shows an increasing trend, however, in the out-of-sample period, it keeps decreasing due to the cover of the 2007-2008 financial crisis. We obtain the efficient minimum risk portfolios based on the in-sample data, then we make the back-tests of the obtained portfolios by applying the out-of-sample data. We assume the initial wealth to be 1 dollar in all portfolio investments.

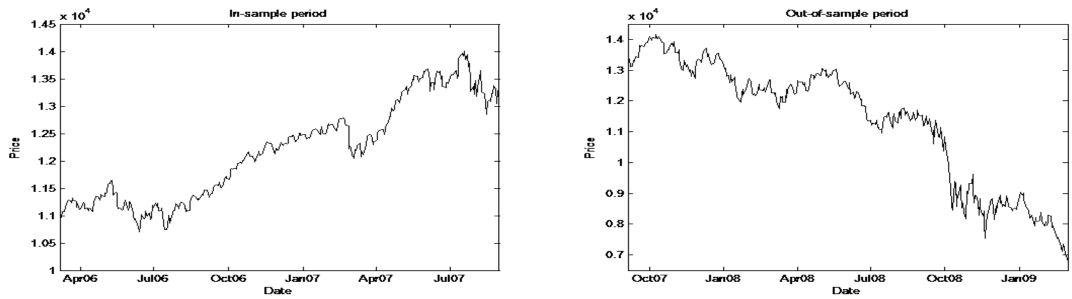


Figure 1: Historical price evolutions of DJIA

We make the back-tests of the obtained minimum risk portfolios based on the out-of-sample data. The obtained performance measures are shown in Table 1. From the table, we can see that the values of mean return are negative due to the financial crisis during the out-of-sample period. Concerning the risk in the out-of-sample period, we can conclude that the risk is the lowest for minimum variance portfolio measured by all four risk measures (except MDD), which is surprising in the case of MAD and CVaR measures, because the other portfolio strategies minimize specifically these measures in-sample. In Table 1 we also show the p-values of the hypothesis tests as described in section 2.2. We can see that all the p-values are very low. Thus, we can reject the null hypotheses and accept the alternative hypotheses, i.e. these strategies minimize the risk of the portfolio (measured by all four measures). Moreover, in Figure 2 we demonstrate that MDD statistics is almost normally distributed. Also other statistics are almost normally distributed, however, due to the limited space, we do not show their distributions.

Table 1: Efficient minimum-risk portfolios' performances in the out-of-sample period

	Minimum Variance	**Minimum MAD**	**Minimum CVaR**
mean daily return	-0.079%	-0.073%	-0.077%
MDD (p-value)	35.42% (0.05%)	35.00% (0.03%)	36.06% (0.08%)
STD (p-value)	1.66% (0.00%)	1.72% (0.00%)	1.71% (0.00%)
MAD (p-value)	1.06% (0.00%)	1.08% (0.00%)	1.12% (0.00%)
CVaR (p-value)	0.30% (0.00%)	0.30% (0.00%)	0.31% (0.00%)

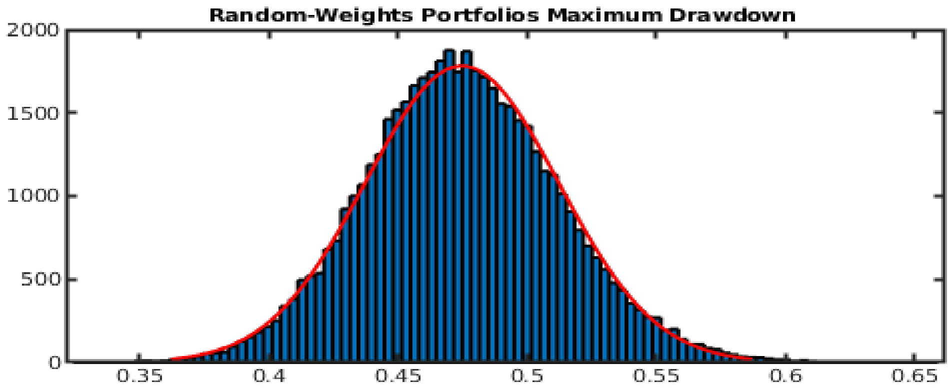


Figure 2: Histogram of MDD of 50,000 random portfolios (MDD statistics) with fitted normal distribution (mean value 47.46% and standard deviation 3.75%)

3.2 Rolling window approach

In order to prove that the results are robust to the change of the period, we perform the tests on one-year rolling window basis. The dataset is the daily closing prices of the DJIA components in the period from March 7, 2006 to June 10, 2019. There are 27 stocks included in our analysis. Components Dow Inc., NIKE Inc and Visa Inc. are missing due to the incomplete data in the chosen period. We always take three years (750 days) as the in-sample period and one year (250 days) as the out-of-sample period, i.e. we move the out-of-sample period day by day from February 27, 2009 to June 11, 2018 (the start of the out-of-sample one-year period). The results are shown in Figure 3. From the figure we can see mixed results. Firstly, the most strict risk measures are MDD and CVaR – for these measures there are long periods in which the strategies did not minimize the risk in the out-of-sample period efficiently. Secondly, it seems that the best strategies are to minimize either the variance or MAD and these two strategies have similar results. On the other hand, minimizing the CVaR does not seem to provide good results in the out-of-sample period. Lastly, minimizing the chosen risk measure in the in-sample period does not guarantee its lowest value in the out-of-sample period (e.g. in 2012-2013 we get the best out-of-sample standard deviation by minimizing in-sample MAD).

4 CONCLUSION

The goal of the paper is twofold. Firstly, we propose the efficiency test for out-of-sample performance of portfolio optimization strategies. Secondly, we apply the test to find out which portfolio optimization strategy minimizes the risk in the out-of-sample period. We found out that all three minimum risk portfolios worked efficiently during global financial crisis (2007-2009). However, the results are not robust as confirmed by rolling-window approach in the period 2008-2018.

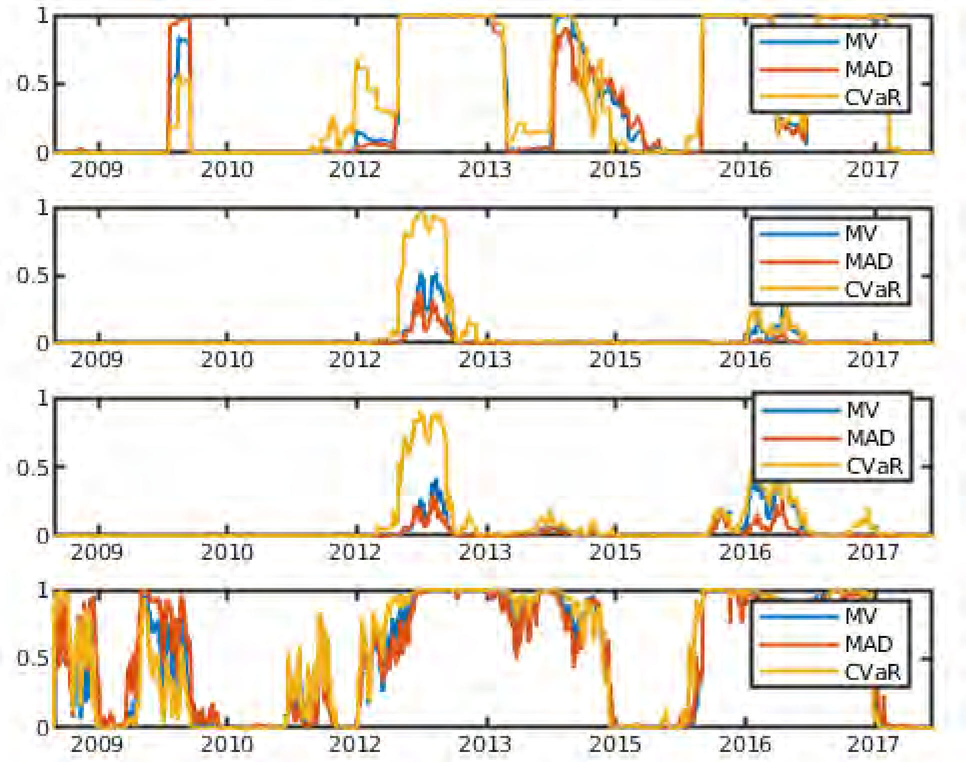


Figure 3: Rolling window p-values (statistics from top to down: MDD, STD, MAD, CVaR)

Acknowledgement

The authors were supported through the Czech Science Foundation (GACR) under the project no. 18-13951S and the SGS research project of VSB-TU Ostrava under the project no. SP2019/5. The support is greatly appreciated.

References

[1] Demiguel, V., Garlappi, L., Uppal, R. 2009. Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? Review of Financial Studies, 22(5): 1915–1953.

[2] Fulga, C. 2016. Portfolio Optimization under Loss Aversion. European Journal of Operational Research, 251(1): 310–322.

[3] Markowitz, H. 1952. Portfolio Selection. The Journal of Finance, 7(1): 77–91.

[4] Ranković, V., Drenovak, M., Urosevic, B., Jelic, R. 2016. Mean-Univariate GARCH VaR Portfolio Optimization: Actual Portfolio Approach. Computers & Operations Research, 72(1): 83–92.

[5] Rockafellar, R. T., Uryasev, S. 2000. Optimization of conditional value-at-risk. Journal of Risk, 2(3): 21–42.

[6] Rockafellar, R. T., Uryasev, S. 2002. Conditional value-at-risk for general loss distributions. Journal of Banking & Finance, 26(7): 1443–1471.

[7] Solares, E., Coello, C., Fernandez, E., Navarro, J. 2019. Handling Uncertainty through Confidence Intervals in Portfolio Optimization. Swarm and Evolutionary Computation, 44(1): 774–787.

[8] Xidonas, P., Hassapis, C., Soulis, J., Samitas, A. 2017. Robust Minimum Variance Portfolio Optimization Modelling under Scenario Uncertainty. Economic Modelling, 64(1): 60–71.

[9] Zmeškal, Z., Dluhošová, D., Tichý, T. 2004. Financial Models. Ostrava: VŠB-TU Ostrava.

INVESTOR ATTENTION AND RISK PREDICTABILITY: A SPILLOVER INDEX APPROACH

Tihana Škrinjarić and Mirjana Čižmešija
University of Zagreb, Faculty of Economics and Business, Department of Mathematics and Department of Statistics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: tskrinjar@net.efzg.hr, mcizmesija@efzg.hr

Abstract: This paper observes shock spillovers between realized volatility and Google search volume regarding the DAX stock index in the period from January 2004 to April 2019. In that way, a dynamic relationship is estimated between DAX risk and search volume. The search volume variable is interpreted as the investor's attention towards the stock market index. Results indicate that a bidirectional time varying relationship is found. This means that potential users of search volume in forecasting the DAX risk should take into account the bidirectional causality.

Keywords: volatility prediction, spillover index, stock market, Google search volume

1 INTRODUCTION

Financial modelling applications and forecasting face many challenges today. One group of research focuses on the possibilities of return and risk forecasting. [10] introduced the online search volume of keywords regarding specific stocks and stock indices as a term called investor's attention; also called the revealed attention measure within financial modelling applications. Although its relevance was already observed theoretically in [20], this measure has become popular in the prediction of stock return and risk only in the last decade. The reasoning lies upon mostly in data availability of online search volume. The most used online search engine today is Google [23], thus the product called Google Trends is now commonly used in financial modelling applications [3, 9, 10, 26, 6, 25,21].

Newer research uses the data available from Google trends in forecasting returns series: [26] analyzes GARCH models for several currencies; [7] on Portuguese stock market; [24] on developed stock markets (US, Canada, Australia, and UK) and simulate trading strategies based upon forecasting results. Other developed markets have been researched as well (Shangai in [29]; France in [2]; Norway in [17]). American indices were mostly in focus [28]; [13]; [15] as some examples. Less developed markets are still not much researched (Croatian market was observed in [25]). Although the research on this topic is increasing in size rapidly[1] over the last few years, there still exist unanswered questions. Thus, this paper contributes to the existing literature as follows. Existing research uses the online search volume only as an explanatory variable in ARMA-GARCH[2] models. This means that no feedback is assumed in research. Moreover, the majority of research utilizes static methodology and models; meaning that the parameters of models stay unchanged over the whole observed period. In this research, we allow for a feedback relationship from one variable to another within a VAR (vector autoregression) model and upgrading it with Diebold and Yilmaz [11, 12] spillover index which allows for shock spillovers from one variable to another in the system. Moreover, the spillover index will be estimated based upon rolling windows, which allows for dynamic analysis. The reasoning lies upon previous empirical research finding that bidirectional relationship exists between stock risk and Google search volume ([15] or [28]).

Main findings of this research indicate that the feedback relationship exists between DAX volatility and Google search volume regarding word DAX. Relationship changes over time by

[1] Other economic and financial variables are being researched as well: initial public offerings [10]; unemployment forecasting [22]; [5]; exchange rate [8]; etc.
[2] AutoRegressive Moving Average – Generalized AutoRegressive Conditional Heteroskedasticity.

one variable becoming a net emitter of shocks after being a net receiver. This is important for potential investors' in future modelling and forecasting risk on the Frankfurt stock exchange. The rest of the paper is structured as follows. The second section describes the methodology used in the empirical part of the paper (third section). The final section concludes the research.

2 METHODOLOGY

Brief methodology description is based on [27] and [18, 19]. Let us observe a stabile[3] Vector AutoRegression (VAR) model of order p can be written in a compact form as follows[4]: $\boldsymbol{Y}_t = \mathbf{v} + \boldsymbol{A}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$. The impulse response functions and variance decompositions are based upon the MA(∞) representation of the VAR model: $\boldsymbol{Y}_t = \boldsymbol{\mu} + \sum_{i=1}^{\infty} \boldsymbol{A}^i \boldsymbol{\varepsilon}_{t-i}$, $\boldsymbol{\mu} \equiv \left(\boldsymbol{I}_{Np} - \boldsymbol{A}\right)^{-1} \mathbf{v}$, i.e. the polynomial form $\boldsymbol{Y}_t = \Phi(\mathrm{L})\boldsymbol{\varepsilon}_t$. $\Phi(\mathrm{L})$ is the polynomial of the lag operator L, in which values $\phi_{jk,i}$ represent the impulse response coefficients of the VAR model. Since the error terms in $\boldsymbol{\varepsilon}_t$ are correlated, the variance-covariance matrix Σ_ε needs to be ortogonalized. Another approach is to observe the generalized forecast error variance decomposition (GEFVD), where non-linear impulse responses are constructed. We follow the GEFVD approach, due to it not depending upon variable ordering as Choleski decomposition does. Firstly, the difference between the true values and forecasted values of variables in the model are estimated for h steps ahead as the difference $\boldsymbol{Y}_{t+h} - E(\boldsymbol{Y}_{t+h})$. Secondly, the mean squared errors are then calculated for every element of the difference as $E(y_{j,t+h} - E(y_{j,t+h}))^2$, which represents the h step ahead variance of variable j. This variance is then decomposed into shocks due to variable j and other variables in the system: $\omega_{jk,h} = \sigma_j^{-1} \sum_{i=0}^{h-1} \left(e_j^{'} \Phi_i \Sigma_\varepsilon e_k\right)^2 / e_j^{'} \Phi_i \Sigma_\varepsilon \Phi_i{'} e_j$, where $\omega_{jk,h}$ is the variance portion of variable j in the h-th forecast step which is due to shocks in variable k; e_j and e_k are the j-th and k-th vectors of matrix I_{Np}. The Diebold and Yilmaz [11, 12] spillover index is now defined as $S = \sum_{\substack{j,k=1 \\ j\neq k}}^{N} \omega_{jk,h} / \sum_{i=0}^{h-1} \sum_{j,k=1}^{N} \omega_{jk,h} 100\%$. The numerator is the sum of all variance portions defined in $\omega_{jk,h}$, whereas the denominator is the total forecast variance. Directional spillover indices are calculated in order to obtain information on which variable is in which period the spillover shocks emitter or receiver. The "to" and "from" spillover indices are calculated as follows: $S_{j\cdot,h} = 1/N \sum_{\substack{k=1 \\ j\neq k}}^{N} \omega_{jk,h} 100\%$ and $S_{\cdot j,h} = 1/N \sum_{\substack{k=1 \\ j\neq k}}^{N} \omega_{kj,h} 100\%$. The net spillover index is the difference between the two directional indices. Finally, rolling indices can be calculated by determining the length of the rolling window. In that way, a dynamic analysis can be performed.

3 EMPIRICAL RESULTS AND DISCUSSION

For the purpose of empirical research, daily data on DAX index value for the period from the 2nd January 2004 until 18th April 2019 was downloaded from [16]. Since Google trends [14]

---

[3] A VAR(p) model is stabile if det($\boldsymbol{I}_{Np}$-Aw)≠0 for |w|≤1.

[4] Where $Y_t = \left[y_t \; y_{t-1} \ldots y_{t-p}\right]'$, $\mathbf{v} = \left[v \; 0 \ldots 0\right]'$, $A = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ I_N & 0 & \cdots & 0 & 0 \\ 0 & I_N & & \vdots & \vdots \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_N & 0 \end{bmatrix}$ and $\varepsilon_t = \left[\varepsilon_t \; 0 \ldots 0\right]'$. It is assumed that $E\left(\varepsilon_t\right) = 0$, $E\left(\varepsilon_t \varepsilon_s^{'}\right) = \Sigma_\varepsilon < \infty$ and for $t \neq s$ $E\left(\varepsilon_t \varepsilon_s^{'}\right) = 0$.

enables to download monthly data on search volume, based upon daily index values we calculated daily returns r_t and estimated monthly realized volatilities (DAX RV henceforward) by using formula $RV_\tau = \sqrt{\sum_{t=1}^{\tau} r_t^2}$ ([1]; [4]). The chosen start point in the research period (from January 2004) is based on Google data availability. We downloaded the monthly Google search volume regarding the word DAX, as the name of the official index is usually used in literature. Moreover, the rationale on using Google search engine is found in the fact that this is the most used search engine online (more than 75% of total market share, [23]). Since stationary data is a necessary condition to use in VAR models, we test with unit root tests the stationarity of both variables in the analysis. Results in Table 1 show that the DAX RV is stationary. However, the Google search volume was found trend-stationary. Thus, this variable was detrended for further analysis. Both stationary variables are shown in Figure 1 (where the Google search volume is in a detrended form). It can be seen that in the financial crisis of 2008 both variables have a great spike. Moreover, the Eurozone crisis in 2012 regarding Greece was reflected in the stock market as well. The departure of J. Ackermann from Deutsche Bank in the spring of 2012 could have contributed to greater volatility as well. Finally, regarding 2012, many small Chinese firms had initial public offerings on the Frankfurt stock exchange in that year, with sharp price declines afterward. Increase of activity on the stock market resulted in greater volatility in 2015. The Brexit uncertainty in 2016 increased the volatility of DAX, as on other European stock markets. Similar patterns can be found in the Google search volume movement. Uncertainty around the Chinese economic slowdown in 2016 has reflected on world markets in 2016, as well as in DAX index via greater volatility.

Table 1: Unit root test results for variables in the study

Deterministic regressors / Test	***DAX RV***		***Google search volume***	
	Level	*-*	*Level*	*Detrended*
None	-5.0938***	-	-0.5073	-5.5917***
Drift	-6.6408***	-	-2.5549	-5.5723***
Drift and trend	-6.6634***	-	-5.5393***	-

Note: *** denotes statistical significance on 1%.

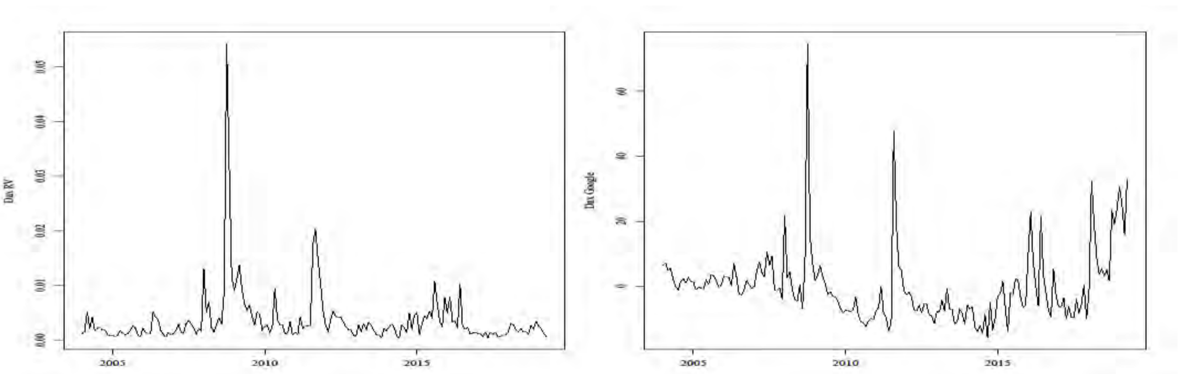


Figure 1: Comparison of DAX RV (left panel) and Google search volume (right panel)

Next, we estimated a VAR(2) model, based upon information criteria and non rejection of null hypotheses of the heteroskedasticity and autocorrelation tests. Both tests were performed up until lag 12, with test values of 39.57 and 47.66 with *p*-values 1 and 0.487 respectively. Moreover, the VAR model was found to be stabile[5]. The spillover table for the whole sample was estimated and is depicted in Table 2. The total spillover index is equal to 33.47%. 63.86%

[5] The roots of the characteristic polynomial are equal to 0.849; 0.4999; 0.1505 and 0.1505.

of the RV variable variance is explained via shocks in RV variable and other 36.14% via shocks in Google search volume variable. Similar interpretations can be made for the variance of Google search volume. As can be seen, almost the third of each variable variance is explained by shocks in another variable. We interpret this as these two variables being very connected over time.

Table 2: Spillover table for total sample

	DAX RV	***GOOGLE***	***FROM***
DAX RV	63.86	36.14	18.07
GOOGLE	30.81	69.19	15.40
TO	15.40	18.07	**33.47**

Dynamic analysis was performed in order to see how the shocks in each variable spilled over to the other over time. We chose different lengths of *h* in order to check for robustness of the results[6]. Rolling spillover indices for different lengths of rolling windows are shown in Figure 2. Several conclusions arise by observing Figure 2 (left panel). Firstly, the value of the total spillover indices increased in the period regarding the financial crisis of 2008 and stayed at a higher level until 2014. The Brexit shocks lead to increase of the indices in 2016. Moreover, the results are more robust until the end of the period, in which greater deviation of each index is found. The reasoning could be in the overall economic uncertainty increase in Europe. Another robustness checking was made by varying the length of *h* for the 24 months rolling window forecast. Results are shown in Figure 2 (right panel). It is obvious that the results are very robust due to very small differences between the rolling indices.

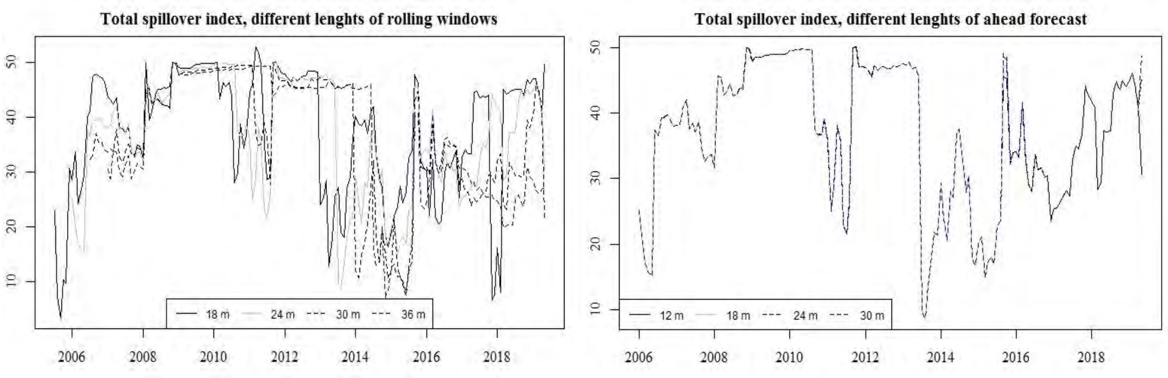
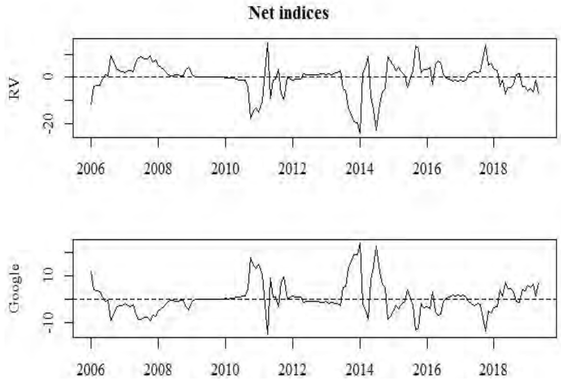


Figure 2: Comparison of total spillover indices, different length of rolling window (left panel) and Comparison of total spillover indices, different *h* length (right panel)

Finally, we observe the rolling net spillover indices in order to have greater insights into which variable is the net emitter/receiver of shocks. These indices were estimated for the *h* = 12- and the 24-month rolling windows. Results are shown in Figure 3. The RV variable is the net emitter of shocks when the values of realized volatilities were higher in the observed sample (crisis of 2008, 2012, period 2015-2016). We interpret this as investors searching more online due to greater volatility on the market. The opposite is true for the Google search volume: this variable emits spillover shocks in less turbulent times on the market. Thus, it would be better to use the online search volume as a predictor of volatility on the market in times of smaller volatility.

[6] The robustness of results is checked by following Diebold and Yilmaz [11, 12] where authors recommend to change the value of the *h*, as well as change the length of the rolling window. The idea is that all of the indices should follow similar patterns over time if the results are robust.



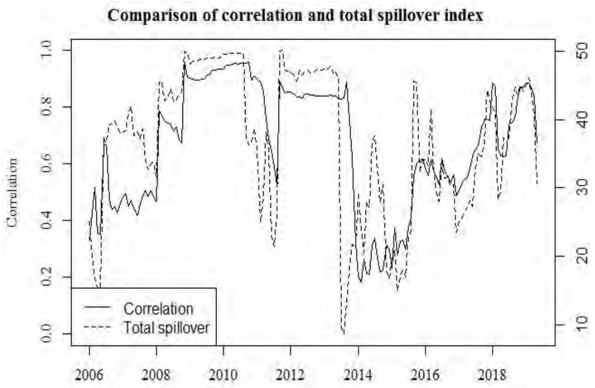


Figure 3: Net spillover indices, 24 month rolling windows *Figure 4:* Correlation coefficient (left axis) and the total spillover index (right axis) between two variables

Finally, we compared the rolling correlation between the two variables and the spillover index in Figure 4. This reassures us of the robustness of the results, due to the spillover being highly correlated with the correlation coefficient. Namely, when the spillover of shocks is great from one variable to another (and vice versa), the correlation among them is greater. Moreover, it can be seen that the spillover index has a leading effect on the correlation coefficient. This means that this methodology can be useful to predict the future correlation between the two variables as well. Several conclusions and recommendations can be provided based on empirical analysis. Since we found a bidirectional relationship between two variables (via spillover indices), future research in this field should take this into account. This is especially important for forecasting models. Next, the relationship is changing over time. Thus, models with constant parameters could be questionable. We found in times of smaller volatility that the Google search index is a greater emitter of shocks. This is useful for prediction of the volatility in times when the market is less volatile. However, in times of greater market volatility, the realized variance is a net emitter of shocks. When interested parties use volatility forecasting models, this should be also taken into consideration.

4 CONCLUSION

Financial risk modelling and forecasting impose many challenges for investors in stock markets today. The online search volume nowadays presents the investor's attention towards specific stocks (or indices). Much research exists which shows the usefulness of the search volume in forecasting risk and return. However, a lot of papers assume only a one-directional relationship with constant parameters over time. This paper overcomes these shortcomings for the DAX realized volatility and Google search volume regarding the mentioned index.

Main results show that there exist bidirectional spillover shock effects between the volatility and search volume regarding the DAX index. Thus, dynamic modelling should incorporate these findings in future work. However, some of the shortfalls of the study were a relatively small number of observations due to data unavailability. Moreover, we used monthly data due to weekly or daily data unavailability of Google search volume. Future work will focus on sector diversification possibilities and investigate the same issues as this work. Thus, simulation of possible investment strategies could be obtained in future work as well. Finally, future work should focus on some kind of new index construction which would be based upon online search volume. This would be in service of better risk or return prediction on stock markets, for investment purposes.

References

[1] Andersen, T. G., Bollerslev, T. 1998. Answering the sceptics: yes standard volatility models do provide accurate forecasts, International Economic Review, 39(4): 885–905. DOI: 10.2307/2527343

[2] Aouadi, A., Aroudi, M., Teulon, F. 2013. Investor attention and stock market activity: evidence from France, Economic Modeling, 35: 674-681. DOI: 10.1016/j.econmod.2013.08.034

[3] Bank, M., Larch, M., Peter, G. 2011. Google search volume and its influence on liquidity and returns of German stocks, Financial Markets and Portfolio Management, 25: 239–264.

[4] Barndorff-Nielsen, O. E., Shephard, N. 2002. Econometric analysis of realised volatility and its use in estimating stochastic volatility models, Journal of the Royal Statistical Society, Series B., 64(2): 253–280.

[5] Barreira, N., Godinho, P., Melo, P. 2013. Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends, Netnomics, 14(3): 129-165. DOI: 10.1007/s11066-013-9082-8

[6] Bijl, L., Kringhaug, G., Molnar, P., Sandvik, E. 2016. Google searches and stock returns. International Review of Financial Analysis, Vol. 45: 150-156. DOI: 10.1016/j.irfa.2016.03.015

[7] Brochado, A. 2016. Investor attention and Portuguese stock market volatility: We'll Google it for you!, EcoMod2016 9345, EcoMod.

[8] Bulut, L. 2015. Google Trends and Forecasting Performance of Exchange Rate Models, No 1505, IPEK Working Papers from Ipek University, Department of Economics.

[9] Da, Z., Engelberg, J., Gao, P. 2011. In search of attention, The Journal of Finance, 66(5): 1461–1499. DOI: 10.1111/j.1540-6261.2011.01679.x

[10] Da, Z., Engelberg, J., Gao, P. 2015. The Sum of All FEARS Investor Sentiment and Asset Prices, The Review of Financial Studies, 28(1): 1–32. DOI: 10.1093/rfs/hhu072

[11] Diebold, F. X., Yilmaz, K. 2009. Measuring Financial Asset Return and Volatility Spillovers with Application to Global Equity Markets. The Economic Journal, 119(534): 158-171. DOI: 10.1111/j.1468-0297.2008.02208.x

[12] Diebold, F. X., Yilmaz, K. 2012. Better to Give than to Receive: Predictive Directional Measurement of Volatility Spillovers, International Journal of Forecasting, 28(1): 57-66. DOI: 10.1016/j.ijforecast.2011.02.006

[13] Dimpfl, T., Jank, S. 2016. Can Internet Search Queries Help to Predict Stock Market Volatility?. European Financial Management, 22(2): 171–192. DOI: 10.1111/eufm.12058

[14] Google trends 2019. Available at https://trends.google.com/trends/ [Accessed 2/5/2019].

[15] Hamid, A., Heiden, M. 2015. Forecasting volatility with empirical similarity and Google Trends, Journal of Economic Behavior & Organization, 117: 62-81. DOI: 10.1016/j.jebo.2015.06.005

[16] Investing 2019. Available at https://www.investing.com [Accessed 2/5/2019].

[17] Kim, N., Lučvijanska, K., Molnar, P., Villa, R. 2019. Google searches and stock market activity: Evidence from Norway. Finance Research Letters, 28: 208-220. DOI: 10.1016/j.frl.2018.05.003

[18] Lütkepohl, H. 2006. New Introduction to Multiple Time Series Analysis, Berlin: Springer.

[19] Lütkepohl, H. 2010. Vector Autoregressive Models, European University Institute. Economics Working Paper No. ECO 2011/30.

[20] Merton, R.C. 1987. A simple model of capital market equilibrium with incomplete information, The Journal of Finance, 42: 483–510. DOI: 10.1111/j.1540-6261.1987.tb04565.x

[21] Mondria, H., Wu, T., Zhang, Y. 2010. The determinants of international investment and attention allocation: using internet search query data, Journal of International Economics, 82: 85–95. DOI: 10.1016/j.jinteco.2010.04.007

[22] Naccarato, A., Falorsi, S., Loriga, S., Pierini, A. 2018. Combining official and Google Trends data to forecast the Italian youth unemployment rate, Technological Forecasting and Social Change, 130: 114-122. DOI: 10.1016/j.techfore.2017.11.022

[23] Netmarketshare 2019. Available at https://trends.google.com/trends/ [Accessed 2/5/2019].

[24] Perlin, M. S., Caldeira, J. F., Santos, A. A. P. 2017. Can We Predict the Financial Markets Based on Google's Search Queries?, Journal of Forecasting, 36(4): 454-467. DOI: 10.1002/for.2446

[25] Škrinjarić, T. 2018. Mogu li Google trend podaci poboljšati prognoziranje prinosa na Zagrebačkoj burzi?. Zbornik radova Ekonomskog fakulteta Sveučilišta u Mostaru, 24: 58-76.

[26] Smith, G. P. 2012. Google Internet Search Activity and Volatility Prediction in the Market for Foreign Currency, Finance Research Letters, 9(2): 103-110. DOI: 10.1016/j.frl.2012.03.003

[27] Urbina, J. 2013. Financial Spillovers Across Countries: Measuring shock transmissions. MPRA Working paper.

[28] Vozlyublennaia, N. 2014. Investor attention, index performance, and return predictability. Journal of Banking & Finance, 41(C): 17–35. DOI: 10.1016/j.jbankfin.2013.12.010

[29] Zhang, W., Shen, D., Zhang, Y., Xiong, X. 2013. Open source information, investor attention and asset pricing, Economic Modelling, 33: 613-619. DOI: 10.1016/j.econmod.2013.03.018

GREY SYSTEMS MODELING AS A TOOL FOR STOCK PRICE PREDICTION

Tihana Škrinjarić and Boško Šego
University of Zagreb, Faculty of Economics and Business, Department of Mathematics
Trg J. F. Kennedyja 6, 10000 Zagreb, Croatia
E-mail: tskrinjar@net.efzg.hr, bsego@efzg.hr

Abstract: This paper observes Grey models (GM) of modelling and forecasting stock prices on the Zagreb Stock Exchange. Grey models belong to the Grey Systems Theory, which focuses on uncertain sets, events and decision-making. Since the Grey methodology is relatively unknown in applications in Europe, the purpose of the paper is to give a concise overview of Grey models as one part of it. The other purpose was to evaluate empirically if Grey models could be a useful tool in stock price prediction. Based upon daily data for the period 2 January – 12 June 2019, the GM (1,1) and (2,1) are estimated and compared to usual approaches of forecasting: average, moving average and ARMA models. Out of sample results show that the GM (1,1) model has the best forecasting ability. Thus, the potential exists in further implementing this methodology in portfolio selection.

Keywords: Grey model (1,1), stock price prediction, Grey Systems, stock exchange, forecasting

1 INTRODUCTION

Quantitative finance represents a very useful tool in portfolio management. However, it is a complex discipline because it includes having great knowledge of quantitative methods and finance theory. The last couple of decades have experienced rapid development of different quantitative methods for answering many specific questions regarding portfolio selection. These include financial econometrics (as maybe one of mostly used), operational research, non-parametric models and methods, etc. Some categorizations can be found in [4], [6], [19] or [21]. Forecasting within the portfolio management represents one of its crucial parts, due to many investing decisions being based on forecasting time series data. A methodology that is still relatively unknown today in the Western countries and Europe is the Grey System Theory (GST), which was developed in the Far East in the 1980s [11]. Although, some improvements in spreading this methodology are observed in last the several years (see [9]).

GST refers to the set of models and methods applied to uncertain systems, scarce or uncertain data, incomplete or partial information when modelling any phenomena and decision-making process. Some of the applications of GST in the area of finance are bankruptcy prediction [2, 3]; financial performance of banks and firms ([22], [17]); forecasting financial ratios of companies [10], etc. Grey methodology is very wide because it includes many different approaches in modelling. One approach is the Grey Relational Analysis (GRA). For an overview of the literature on GRA and empirical applications within finance, please see [18].

This research belongs to the group of papers that focuses on forecasting prices of financial assets [24, 23, 7]. The research on stock price prediction with GST is growing today. This is not surprising as this methodology is useful to implement on financial and especially stock markets. The reason lies upon the fact that many different, conflicting and sometimes distorted information can be found on stock markets today. Majority of existing research compares the predictive capabilities of several GST models and those from other methodologies. However, the results are not used in order to simulate trading strategies, which can be based upon the forecast (see [15, 8]). Thus, the purpose of this paper is to fill the gap in the literature by simulating several trading strategies and compare their performances in terms of risk and return. This will be done on a sample on stock market index CROBEX on the Zagreb Stock Exchange. The research problem of this paper is to test if implementing the Grey methodology leads to better forecasting of stock prices. Another problem which follows from the first one is if using this methodology can enhance portfolio performance compared to approach which does not utilize the Grey methodology. First research hypothesis is: Grey Models have better goodness of fit than benchmark ARMA models; and the second one is: portfolio strategies based on Grey models have greater portfolio value compared to benchmark ARMA ones. The rest of the paper is structured as follows. The second section describes the methodology of the empirical, third section. The last section concludes the paper.

2 METHODOLGY

We follow [12, 13, 14] in briefly describing the methodology, where proofs, theorems and propositions which must hold can be found for interested readers. The basic idea comes from the first order differential equation:

$$\frac{dx}{dt} + ax = b, \tag{1}$$

where dx/dt is the derivative of function $x(t)$, $x(t)$ is the unknown function (called the background value in the literature on Grey systems), a, b∈□ . If $x(t)>0$ holds, the information density of $x(t)$ is infinite[1], background values are a grey number[2] and the derivative dx/dt and $x(t)$ satisfy the horizontal mapping[3], then the Grey differential equation can be formed.

Assume that $X^{(0)}=\{x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n)\}$ is a sequence of non-negative numbers, where $n \geq 4$. The first order accumulated generating operation (AGO) of $X^{(0)}$ is defined as $X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)\}$, where

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), \tag{2}$$

where $k \in \{1,2,..., n\}$. The AGO smoothes out the original sequence $X^{(0)}$ in order to reduce randomness. It is obvious from (2) that the AGO sequence is a monotonically increasing sequence (this ensures that $x^{(1)}(k)>0$, see previously mentioned conditions for a Grey differential equation).

Next, the mean sequence of $X^{(1)}$, $Z^{(1)}$ is defined as $Z^{(1)} = \{z^{(1)}(2), z^{(1)}(3), \ldots, z^{(1)}(n)\}$, where

$$z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(0)}(k-1)}{2}. \tag{3}$$

Construction of (3) is necessary so that the horizontal mapping condition holds. Now, the Grey Model, GM(1,1), is defined as follows:

$$x^{(0)}(k) + az^{(0)}(k) = b, \tag{4}$$

In which the first unit value refers to the order and the second to number of variables observed. (4) is a Grey differential equation, which has the following solution:

$$x^{(0)}(k) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-a(k-1)} (1 - e^{a}). \tag{5}$$

(5) is used to forecast values $x^{(0)}(l)$, where $l \in \{n+1, n+2, ...\}$.

Wavelike sequences can be modelled by GM (2,1) model as follows. Additionally, we need to define the IAGO sequence $\alpha^{(1)}X^{(0)} = \{\alpha^{(1)}x^{(0)}(1), \alpha^{(1)}x^{(0)}(2), \ldots, \alpha^{(1)}x^{(0)}(n)\}$, in which

$$\alpha^{(1)}x^{(0)}(k) = x^{(0)}(k) - x^{(0)}(k-1). \tag{6}$$

Then the equation

$$\alpha^{(1)}X^{(0)} + a_1 X^{(0)} + a_2 Z^{(1)} = b \tag{7}$$

---

[1] Information density of $x(t)$ is infinite if when $\Delta t \to 0$, it is always true that $x(t+\Delta t) \neq x(t)$.

[2] Grey number is that of which true value is unknown but its range is known.

[3] Assume that X and Y are sets, and R is the operation between the two sets. If for $x_1, x_2 \in X$, $y \in Y$ holds that $x_1Ry = x_2Ry$ then y is called a horizontal mapping of x_1 and x_2.

is called the GM (2,1) grey differential equation. The solution of (7) depends upon the characteristic equation $r^2 + a_1 r + a_2 = 0$. Solutions from the models (4) and (7) will be used to forecast future prices on the stock market in the empirical part of the paper, where the following accuracy measures of forecasting will be applied: *RMSE*, *MAE*, *MAPE* and *TIC*[4].

3 EMPIRICAL RESULTS

For the purpose of empirical analysis, daily data on stock market index CROBEX on Zagreb Stock Exchange [25] was collected for the period 2 January 2019 – 12 June 2019. The total sample consists of 113 observations. We divide the sample into 2 subsamples, in sample estimation until 31 May 2019 and the remaining days in June for out of sample forecasts.

As benchmark models, we estimate[5] in sample the average value of the index value, the moving average and ARIMA(1,1,1)[6] model. Sample average is a naïve approach often used as a benchmark for the comparison, the moving average is an extension in order to fit data better to real movements and ARIMA is used due to it being one of the basic approaches when modelling stock price, i.e. return movements. Grey models (4) and (7) have been estimated as well. Table 1 provides information on the estimation results with the goodness of fit comparisons. It can be seen that all of the estimated parameters in all models are significant. Parameters in Grey models are not interpreted as usual models in price and stock modelling. Parameters *a* and *b* in (4) are called development coefficient and grey action quantity respectively (for details please see [14]). End of Table 1 shows in sample diagnostics in terms of maximum likelihood (Log L) and information criteria (Akaike, Schwartz and Hannan-Quinn). The best model in sample is GM (2,1), followed by ARIMA. However, since we test for predictive accuracy, out of sample forecasts were made from every model and the error forecasts were calculated. They are shown in Table 2.

Table 1: Comparison of estimation results, in sample

Parameters / diagnostics	***Benchmark models***			***Grey models***	
	Sample average	*Moving average*	*ARIMA(1,1,1)*	*GM(1,1)*	*GM(2,1)*
$\hat{\mu}$	1790.95 *** (8.21)	1790.92 *** (4.24)	1.23 *** (0.234)	-	-
$\hat{\phi}_1$	-	0.926 *** (0.05)	0.738 *** (0.133)	-	-
$\hat{\varphi}_1$	-	-	-0.920 *** (0.079)	-	-
$\hat{a}$	-	-	-	-0.001 *** (0.00004)	-
$\hat{a}_1$	-	-	-	-	-316.732 *** (99.082)
$\hat{a}_2$	-	-	-	-	0.184 *** (0.057)
$\hat{b}$	-	-	-	1727.641 *** (4.071)	-0.0001 *** (0.00004)
Log L	-532.838	-472.946	-347.534	-403.303	**-345.957**
AIC	10.16833	9.065631	6.760270	7.794296	**6.710713**
SIC	10.19361	9.141458	6.861977	7.845150	**6.786993**
HQC	10.17858	9.096358	6.801474	7.814898	**6.741616**

Note: *** denotes statistical significance on 1%. Standard errors are given in brackets. Bolded values indicate best values when comparing by rows.

Table 2 compares 4 different forecast accuracy measures for out of sample forecast. It can be seen that now, the best model is GM (1,1) according to all of the measures. It is followed by

[4] Root mean squared error, mean absolute error, mean absolute percentage error and Theil inequality index.

[5] Estimations were done via GMM method. Where it was needed, estimations were done with [16] correction of standard errors.

[6] ARIMA stands for autoregressive integrated moving average. Order (1,1,1) was chosen via Box-Jenkins procedure as being the best fitting one among 225 different model specifications.

GM (2,1) in the second place. Based on the results in Tables 1 and 2, it can be said that potential exists in using the GM models for forecasting for portfolio composition purposes. Moreover, the out of sample time span is 8 days long. Out of those 8 days, the GM (2,1) model forecasts were closest to the actual values of the index in the first 2 days, and rest of the 6 days this was true for GM (1,1) model. Based on results in Tables 1 and 2, the first research hypothesis is confirmed.

Table 2: Comparison of out of sample diagnostics

Forecast accuracy measures	***Benchmark models***			***Grey models***	
	Sample average	*Moving average*	*ARIMA(1,1,1)*	*GM(1,1)*	*GM(2,1)*
RMSE	88.09	86.57	33.42	**26.21**	32.45
MAE	84.61	81.57	25.50	**19.73**	25.34
MAPE	4.49	4.33	1.34	**1.04**	1.34
TIC	0.0240	0.0236	0.009	**0.007**	0.009

Note: bolded numbers denote best values in each row

Finally, we simulate trading strategies for all of the models with the exception of the first one (average) due to it being the worst. It is assumed that investor forecasted the values for the mentioned 8 days and he is comparing the values of CROBEX every day in order to decide if he wants to buy or sell the index. If the forecasted value in day t+1 is greater than on day t, the investor decides to buy on day t and sell on day t+1. The opposite is true if the value on day t is greater compared to day t+1. Every strategy started with a unit portfolio value. The values of every portfolio are shown in Figure 1. The returns used in order to construct Figure 1 are the true returns on CROBEX. Thus, it is assumed that investor decided on buying and selling based upon forecasted values, but true returns have to be used in order for the strategies to be comparable and true. The best performing strategy is the GM (1,1) which is not surprising due to results in Table 2. It is followed by the moving average strategy.

Finally, investors are not interested only in returns, but in risks and other performance measures as well. That is why we calculate several measures in Table 3: average return, portfolio standard deviation, Certainty Equivalent[7] (CE) for 3 levels of risk aversion, risk-reward ratio and the total return. Since bolded values denote the best performance for a given measure, it can be seen that the GM (1,1) portfolio is best in terms of achieving best average and total return, which leads to smallest risk-reward ratio. Since investors look at return and risk simultaneously, besides the risk-reward ratio, we look at CEs. Three levels of risk aversion are observed, in order to include very risk-averse investor (γ=10), an average investor (γ=5) and more aggressive investor (γ=1). Only the risk-averse investor should consider the ARIMA model for forecasting and possible construction of trading strategies. However, the rest of the performance measures point to GM (1,1) being the best one.

The results of estimation and out of sample simulations provide evidence in favour of using the GM methodology in forecasting future stock market movements. Not only does the GM (1,1) model have the best forecast accuracy, but it provided the investor with the best performances of his portfolio as seen in Table 3. The methodology of Grey models is relatively simple and straightforward compared to some other complex approaches. Thus, future work should rethink to include this methodology in some parts of the portfolio management. Based on these results, the second research hypothesis is confirmed as well.

[7] Certainty Equivalent is the value which gives the investor utility equal to the expected utility of an uncertain gamble. It is calculated as $CE \approx E(\mu)-0.5\gamma\sigma^2$, where $E(\mu)$ is the average return of the portfolio, γ is the coefficient of absolute aversion of risk and σ^2 is the portfolio risk. For details, see [1] or [20]. We chose 3 different levels of γ as the usual literature does (see [5]).

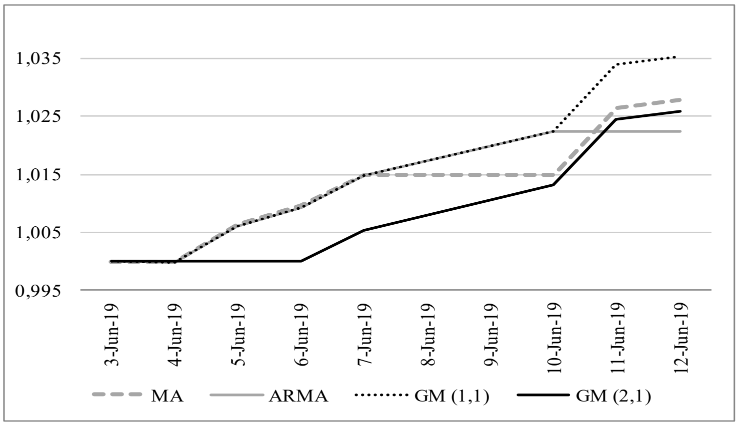


Figure 1: Out of sample simulated portfolio values

Table 3: Comparison of portfolio values

Characteristics	*MA*	*ARMA*	*GM (1,1)*	*GM (2,1)*
Average return	0.0039	0.0032	**0.0050**	0.0036
Standard deviation	0.0041	**0.0034**	0.0039	0.0045
CE 1 (γ=1)	0.0019	0.0015	**0.0030**	0.0014
CE 2 (γ=5)	-0.0042	-0.0035	**-0.0029**	-0.0054
CE 3 (γ=10)	-0.0165	**-0.0136**	-0.0148	-0.0189
Risk-reward ratio	1.05	1.06	**0.78**	1.25
Total return	0.0273	0.0222	**0.0347**	0.0254

Note: bolded numbers denote best values in each row. CE denotes Certainty Equivalent.

4 CONCLUSION

It is not surprising that today there exist many different models, methods and approaches as tools in portfolio management. This paper focused on the Grey Models methodology in order to get initial insights into their usefulness in forecasting future stock market movements. When doing so, it is important to obtain the in and out of sample comparisons due to this methodology being developed for forecasting purposes. Previous research has shown the usefulness of GMs, as well as this one. Since existing literature does not focus on using the forecasts for portfolio simulation purposes, this paper tried to fill that gap. Results here show that portfolios constructed based upon GM results can provide the best performance in terms of portfolio value and the overall utility obtained from portfolio risk and return. The scientific contribution of this research includes incorporating the results of estimation in simulating portfolio strategies in order to show how the GM results can be utilized. This is rarely found in the literature.

However, there are some limitations of this study as well. We observed two rather simple GM models in order to obtain initial information on their performance. Other models have been developed over the years as well, in order to tailor different characteristics of time series. Recent literature has extended the GARCH[8] methodology with GM models (see [11]). Thus, future work should focus on those models and extensions of existing popular ones in order to see the full potential of GM methodology in financial series forecasting. Moreover, we focused only on the total stock market index. In the future, we will extend the analysis on individual stock level and try to construct portfolios based upon individual GM models for many stocks. This will be an exercise of the complexity and time needed to construct portfolios of good characteristics in terms of their risk and return performance.

[8] Generalized Autoregressive Conditional Heteroskedasticity.

References

[1] Cvitanić, J., Zapatero, F. 2004. Introduction to the Economics and Mathematics of Financial Markets, London: MIT Press.

[2] Delcea, C., Scarlat, E. and Maracine, V. 2012. Grey relational analysis between firm's current situation and its possible causes: a bankruptcy syndrome approach. Grey Systems: Theory and Application, 2(2): 229-239. DOI: 10.1108/20439371211260199

[3] Delcea, C., Scarlat, E., Cotfas, L-A. 2013. Grey Relational Analysis of the Financial Sector in Europe. The Journal of Grey System, 25(4): 19-30

[4] Granger, C. J. 1989. Combining forecasts – twenty years later. Journal of Forecasting, 8(1): 167–173.

[5] Guidolin, M., Timmermann, A. 2008. International Asset allocation under Regime Switching, Skew, and Kurtosis Preferences. The Review of Financial Studies, 21(2): 889-935. DOI: 10.1093/rfs/hhn006

[6] Ho, S., Xie, M., Goh, T. 2002. A comparative study of neural network and box-jenkins ARIMA modeling in the time series prediction. Computers and Industrial Engineering, 42(2): 371-375. DOI: 10.1016/S0360-8352(02)00036-0

[7] Huang, K.Y., Jane, C.J. 2009. A hybrid model for stock market forecasting and portfolio selection based on ARX, grey system and RS theories. Expert Systems with Applications 36(3): 5387–5392. DOI: 10.1016/j.eswa.2008.06.103

[8] Huang, K.Y., Jane, C.J. and Chang, T.C. 2010. A hybrid model for portfolio selection based on grey relational analysis and RS theories. 2010 International Computer Symposium: 1014-1019. DOI: 10.1109/COMPSYM.2010.5685542

[9] Kayacan, E., Ulutas, B., Kaynak, O. 2010. Grey system theory-based models in time series prediction. Expert Systems with Applications, 37(2): 1784-1789. DOI: 10.1016/j.eswa.2009.07.064

[10] Khajavi, S., Najafi, Z., Zinedine Zadeh, S. 2012. Applying the grey models and the development of a simple way to forecast free cash flow of listed companies in Tehran stock exchange. Journal of Financial and Management Engineering Exchange, 13(2012): 21-34.

[11] Lin, C-T., Hung, J-C., Wang, Y-H. 2005. Forecasting the One-period-ahead Volatility of the International Stock Indices: GARCH Model vs. GM(1,1)-GARCH model. Journal of Grey System, 8(1): 1-12.

[12] Liu, S., Lin, Y. 2006. Grey Information Theory and Practical Applications. New York: Springer.

[13] Liu, S., Lin, Y. 2010. Grey systems, Theory and Applications. Berlin/Heidelberg: Springer.

[14] Liu, S., Yang Y., Forrest, J. 2016. Grey Data Analysis: Methods, Models and Applications. Singapore: Springer Science + Business Media.

[15] Mohammadi, A., Zeinodin Zade, S. 2011. Appling grey forecasting method to forecast the portfolio's rate of return in stock market of Iran. Australian Journal of Business and Management Research, 1(7): 1-16.

[16] Newey, W. K., West, K. D. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica, 55(3): 703-708. DOI: 10.2307/1913610

[17] Pashaki, M. K., Ahadzadeh, M., Shahverdiani, S. 2018. Ranking Active Companies in Cement Industry with the Combined Approach of Grey Relational Analysis and Data Envelopment Analysis. The third international conference on Intelligent Decision Science (IDS-2018), Iran.

[18] Škrinjarić, T., Šego, B. 2019. Using Grey Incidence Analysis Approach in Portfolio Selection. International Journal of Financial Studies, 7(1): 1-16. DOI: 10.3390/ijfs7010001

[19] Taylor, M., Allen, H. 1992. The use of technical analysis in the foreign exchange market. Journal of International Money and Finance, 11(3): 304-314. DOI: 10.1016/0261-5606(92)90048-3

[20] Varian, H. R. 1992. Microeconomic Analysis, Third edition, New York: W. W. Norton & Co.

[21] Wallis. K. F. 2011. Combining forecasts – forty years later. Applied Financial Economics, 21(1-2): 33-41, DOI: 10.1080/09603107.2011.523179

[22] Wu, C-R., Lin, C-T., Tsai, P-H. 2010. Evaluating business performance of wealth management banks. European Journal of Operational Research, 207(2): 971-979. DOI: 10.1016/j.ejor.2010.04.034

[23] Xu, Y.D., Wu, Z.Y. 2003. Grey pricing of stocks based on limited rationality and inefficiency of stock markets. Journal of Management Engineering 17(2): 115–117

[24] Yongzhong, C., Hongjuan, L. 2005. A Comparison of Two Grey Models in Predicting Stock Index. The Journal of Grey System, 1(2005): 73-76.

[25] Zagreb Stock Exchange. 2019. Data retrieved from https://www.zse.hr [Accessed 12/6/2019].

OPTIMIZATION OF COSTS OF PREVENTIVE MAINTENANCE

Petr Volf
Institute of Information Theory and Automation, AS CR,
Pod vodarenskou vezi 4, Prague 8, Czech Republic
volf@utia.cas.cz

Abstract: In various fields of real life, many interesting optimization problems appear. The present contribution deals with optimization of maintenance of a technical device. Namely, both the period of maintenance and its level are controlled, the costs are compared with the cost caused by the device failure and necessary repair after it. We consider a variant the Kijima model assuming that the consequence of such a repair is the decrease of 'virtual' age of the object. The main objective is to formulate a proper stochastic objective function evaluating the costs of given maintenance strategy and then to present an optimization method for selected characteristics of the costs distribution.

Keywords: reliability, degradation model, maintenance, stochastic optimization.

1 INTRODUCTION

In reliability analysis, the models of imperfect repairs are mostly based on the reduction of the hazard rate, either directly or indirectly (by shifting the virtual age of the system). If the state of the system is characterized by its degradation, the repair degree can be connected with the reduction of the degradation level. We shall concentrate mainly to selection of certain repair schemes, their consequences, and possibilities of an 'optimal' repair policy leading to the hazard rate stabilization and costs minimization. The organization of the paper is the following: First, general models of repairs will be recalled. Then, a variant of the Kijima II type model for preventive repair [6] will be considered and its scheme applied to the case of a model with degradation process. Repair then will be connected with the reduction of the level of degradation. Finally, a solution searching for optimal repair parameters will be demonstrated on an artificial example.

2 BASIC REPAIR MODELS

Let us first recall briefly the most common schemes of repair of a repairable component and the relationship with the distribution of the time to failure (cf. [1]). The renewal means that the component is repaired completely, fully (e.g. exchanged for a new one) and that, consequently, the successive random variables – times to failure – are distributed identically and independently. The resulting intensity of the stream of failures is defined as

$$h(t) = lim_{d\to 0^+} \frac{P(failure\, occurs\, in\, [t, t+d))}{d}.$$

Its integral (i.e. cumulated intensity) is then $H(t) = E[N(t)] = \sum_{k=0}^{\infty} k \cdot P(N(t) = k)$, where $N(t)$ is the number of failures in (0,t].

2.1 Models of partial repairs

There are several natural ways how the models of complete repairs can be widen to repairs incomplete. One of basic contribution is in the paper [6]. Let F be the distribution function of the time to failure of a new system. Assume that at each time the system fails, after a lifetime T_n from the preceding failure, a maintenance reduces the *virtual age* to some value $V_n = y, y \in [0, T_n + V_{n-1}]$ immediately after the n-th repair ($V_0 = 0$). The distribution of the n-th failure-time T_n is then

$$P[T_n \leq x | V_{n-1} = y] = \frac{F(x+y) - F(y)}{1 - F(y)}.$$

M. Kijima then specified several sub-models of imperfect repairs. Denote by A_n the degree of the n-th repair (a random variable taking values between 0 and 1). Then in Model I the n-th repair cannot remove the damages incurred before the (n-1)th repair, $V_n = V_{n-1} + A_n \cdot T_n$. On the contrary, the Model II allows for such a reduction of the virtual age, namely $V_n = A_n \cdot (V_{n-1} + T_n)$. Special cases contain the perfect repair model with $A_n = 0$, minimal repair model, $A_n = 1$, and frequently used variant with constant degree $A_n = A$. Naturally, there are many other approaches, e.g. considering a randomized degree of repair, the regressed degree (based on the system history), accelerated virtual ageing, change of hazard rate etc., see for instance [3], [1], [5]), or [7].

2.2 A variant of Kijima model of preventive maintenance

Let us recall the following simple case of the Kijima II model with constant degree δ of virtual age reduction, and assume that it is used for the description of consequence of preventive repairs. Further, let us assume that after the failure the system is repaired just minimally, or that the number of failures is much less than the number of preventive repairs. Let Δ be the (constant) time between these repairs, V_n, V_n^* the virtual ages before and after $n-$th repair. Hence:

$$V_n = V_{n-1}^* + \Delta \text{ and } V_n^* = \delta \cdot V_n.$$

If we start from time 0, then $V_1 = \Delta,\ V_1^* = \delta\Delta,\ V_2 = \delta\Delta + \Delta = \Delta(\delta+1), V_2^* = \Delta(\delta^2+\delta),\ V_3 = \Delta(\delta^2 + \delta + 1)$ etc. Consequently, $V_n \rightarrow \frac{\Delta}{1-\delta}$, i.e. it 'stabilizes'.

Now, let us consider a variant, in which the reduction of "virtual age" means just reduction of the failure rate to the level corresponding to virtual age. I.e., for each δ and Δ there is a limit meaning that the actual intensity of failures $h(t)$ 'oscillates' between $h_0(\frac{\delta\Delta}{1-\delta})$ and $h_0(\frac{\Delta}{1-\delta})$, where $h_0(t)$ is the hazard rate of the time-to-failure distribution of the non-repaired system.

Simultaneously, the cumulated intensity increases regularly through intervals of length Δ by $dH = H(\frac{\Delta}{1-\delta}) - H(\frac{\delta\Delta}{1-\delta})$, i.e. 'essentially' with the constant slope $a = dH/\Delta$. Figure 1 shows graphical illustration of such a stabilization in the case that the hazard rate $h_0(t)$ increases exponentially.

Example: Let us consider the Weibull model, with $H_0(t) = \alpha \cdot \exp^{\beta}$, ($\beta > 1$, say). In that case

$$dH = \alpha\Delta^{\beta}\frac{1-\delta^{\beta}}{(1-\delta)^{\beta}} \text{ and } A = \alpha\Delta^{\beta-1}\frac{1-\delta^{\beta}}{(1-\delta)^{\beta}}.$$

As special cases, again the perfect repairs with $\delta = 0$, minimal repairs with $\delta \sim 1$, and the exponential distribution case with $\beta = 1$ can be considered.

Remark 1. If the model holds (with constant times between repairs Δ) it is always possible to stabilize the intensity by selecting the upper value of H^* and repair always when $H(t)$ reaches this value. Then $V_n = V = H^{-1}(H^*)$, $V_n^* = \delta V_n$ again, and the interval between repairs should be $\Delta = V(1-\delta)$.

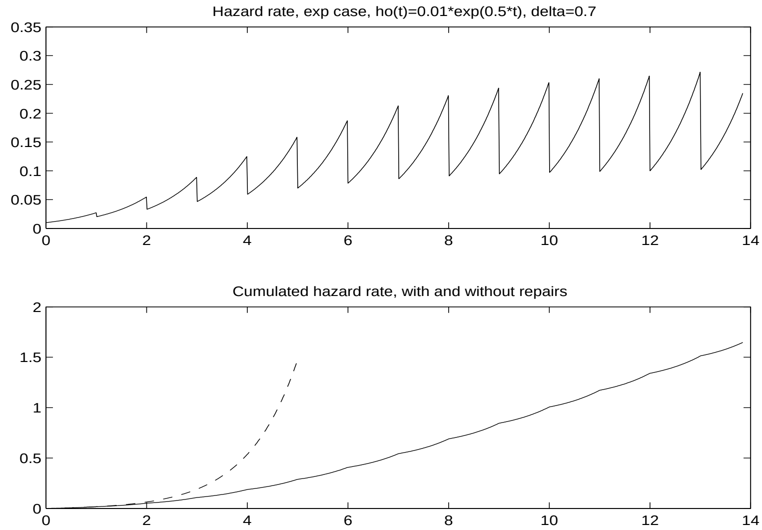


Figure 1: Case of exponentially increasing $h_0(t) = 0.01 \cdot \exp(0.5 \cdot t), \delta = 0.7, \Delta = 1$. Above: Intensity after repairs. Below: Cumulated intensity with repairs (full) and without (dashed curve).

3 TOWARDS OPTIMAL MAINTENANCE STRATEGY

Let us consider the stabilized case, as in Figure 1, and assume that the failures are much less frequent than preventive repairs, then there quite naturally arises the problem of selection of δ to given repair interval Δ (or optimal selection of both). By optimization we mean here the search for values yielding the minimal costs of repairs, which has a sense especially in the case when the repairs after failures are too expensive.

Let C_0 be the cost of failure (and its repair), $C_1(\delta, \Delta)$ the cost of the preventive repair. Then the mean costs to a time t can be written as

$$C \approx C_0 \cdot E(N(t)) + \frac{t}{\Delta} \cdot C_1(\delta, \Delta),$$

where $E(N(t))$ is the mean number of failures up to t, which actually equals $H(t)$, the cumulated intensity of failures under our repairs sequence. The proportion $\frac{t}{\Delta}$ is the number of preventive actions till t. The problem is the selection of function C_1, it should reflect the extent of repair. It leads to the idea to evaluate the level of system degradation and to connect the repair with its reduction.

3.1 Maintenance as a reduction of system degradation

Let us therefore consider a function $S(t)$ (or a latent random process) evaluating the level of degradation after a time t of system usage. In certain cases we can imagine $S(t) = \int_0^t s(u)du$ with $s(u) \geq 0$ is a stress at time u. We further assume that the failure occurs when $S(t)$ crosses a random level X. Recall also that the cumulated hazard rate $H(t)$ of random variable T, the time to failure, has a similar meaning, namely the failure occurs when $H(t)$ crosses a random level given by $Exp(1)$ random variable.

As $T > t <=> X > S(t)$, i.e. $\bar{F}_0(t) = \bar{F}_X(S(t))$, where $\bar{F}$ denote survival functions, then

$$H_0(t) = -log\bar{F}_X(S(t)).$$

We can again consider some special cases, for instance:

– $X \sim Exp(1)$, then $H_0(t) = S(t)$,

– $S(t) = c \cdot t^d$, $d \geq 0$, and X is Weibull (a, b), then T is also Weibull $(\alpha = ac^b, \beta = b + d)$, i.e. $H_0(t) = \alpha \cdot t^\beta$.

Let us now imagine that the repair reduces $S(t)$ as in the Kijima II model, to $S^*(t) = \delta \cdot S(t)$. In the Weibull case considered above we are able to connect such a change with the reduction of virtual time from t to some t^*: $S(t^*) = S^*(t) => t^* = \delta^{\frac{1}{d}} \cdot t$, so that the virtual time reduction follows the Kijima II model, too, with $\delta_t = \delta^{\frac{1}{d}}$. As it has been shown, each selection of δ, Δ leads (converges) to a stable ('constant' intensity) case.

For other forms of function $S(t)$, e.g. if it is of exponential form, $S(t) \sim e^{ct} - 1$, such a tendency to a constant intensity does not hold. Nevertheless, it is possible to select convenient δ and Δ, as noted in Remark 1.

3.2 Degradation as a random process

In the case we cannot observe the function $S(t)$ directly, and it is actually just a latent factor influencing the lifetime of the system, it can be modelled as a random process. What is the convenient type of such a process? There are several possibilities, for instance:
1. $S(t) = Y \cdot S_0(t)$, $Y > 0$ is a random variable, $S_0(t)$ a function.
2. Diffusion with trend function $S_0(t)$ and $B(t)$-the Brown motion process, $S(t) = S_0(t) + B(t)$.
3. $S(t)$ cumulating a random walk $s(t) \geq 0$.
4. Compound Poisson process and its generalizations, see for instance [4].

Though the last choice, sometimes connected also with the "random shock model", differs from the others, because its trajectories are not continuous, we shall add several remarks namely to this case. The compound point process is the following random sum

$$S(t) = \sum_{T_j < t} Y(T_j) = \int_0^t Y(u) dN_s(u)$$

with the counting (mostly Poisson) process $N_s(t)$ yielding the random times T_j and random variables $Y(t) > 0$ giving the increments. Let λ be the intensity of Poisson process, μ, σ^2 the mean and variance of increments, then it holds

$$ES(t) = \int_0^t \lambda(u) \cdot \mu(u) du,$$

$$var(S(t)) = \int_0^t \lambda(u) \cdot (\mu^2(u) + \sigma^2(u)) du.$$

Again, let us assume that the failure occurs when the process $S(t)$ crosses a level x. Then $S(t) < x <=> t < T$, therefore $\bar{F}_0(t) = F_{S(t)}(x)$, where $\bar{F}_0(t)$ denotes again the survival function of the time to failure and $F_{S(t)}(x)$ is the compound distribution function at t. If X is a random level, then the right side has the form $\int_0^\infty F_{S(t)}(x) dF_X(x)$. The evaluation of the compound distribution is not an easy task, nor in the simplest version of compound Poisson process. There exist approximations (derived often in the framework of the financial and insurance mathematics, see again [4]). Another way consists in random generation.

4 PARTIAL MAINTENANCE OPTIMIZATION

What occurs when, as in the preceding cases, the repairs reducing degradation, with degree δ, are applied in regular time intervals Δ? It is assumed that when we decide to repair, then we are able to observe actual state of $S(t)$. Random generation shows that the system then has the tendency to stabilize the intensity of failures, as in Figure 1.

We can now return to the 'cost optimization' problem. Function $C_1(\delta, \Delta)$ can now be specified for instance as $C1 \cdot (dS(t))^\gamma + C_2$, where $dS(t) = S(t)(1 - \delta) = S(t_{end}) - S(t_{init})$. Here

C_1 and C_2 are constants, the later evaluating a fixed cost of each repair. Of course, a proper selection of costs and function $C1$ in real case is a matter of system knowledge and experience. We performed several randomly generated examples, with different variants of the objective function (which was stochastic), with the goal to find optimal maintenance parameters, in the sense of minimization of costs (i.e. their mean, or median, or other quantile).

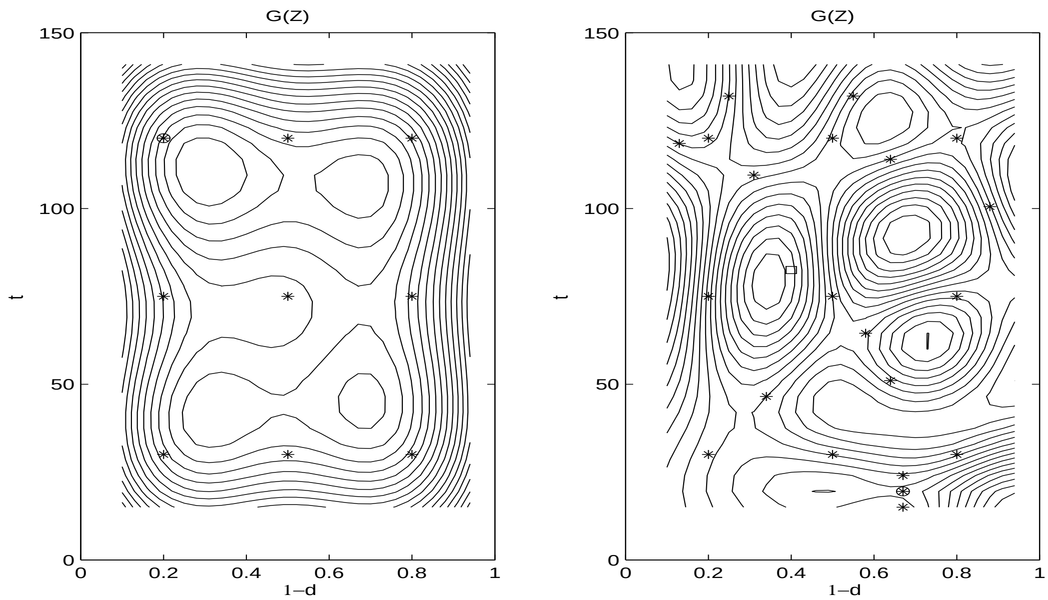


Figure 2: Example of optimal maintenance: Initial phase of search (left), state of search after 12 iterations (right); $1-\delta$ on horizontal, τ on vertical axis.

4.1 Example of optimal maintenance

Let us again assume, in a Kijima II model of non-complete repair, that the device is repeatedly repaired in its virtual age τ with a degree $1-\delta$, which means that after repair the virtual age of the device is $\delta \cdot \tau$. Then the parameter Δ of inter-maintenance times equals $(1-\delta)\cdot\tau$.

In the example it is assumed that the Kijima model concerns to preventive repairs, meanwhile after the failure the device has to be renewed completely. We are given the costs of renewal, C_0, and of preventive repair, $C_1(\delta,\tau)$. It is due the problem assumptions that the objective can be formulated as to maximize, over τ and δ, selected characteristics of random objective function $\varphi(T,\delta,\tau)$ equal to proportion of the time to renewal to the costs to renewal. Here T is the random time to failure of the device. This proportion equals

$$\varphi(T,\delta,\tau) = \frac{T}{C_0} \quad \text{with probability} \quad P(T \le \tau),$$

$$\varphi(T,\delta,\tau) = \frac{\tau + \tau\cdot(1-\delta)\cdot(k-1) + T_k}{C_0 + k\cdot C_1} \text{ with } P(T>\tau)\cdot P(T_1>\tau)^{k-1}\cdot P(T_1 \le \tau),$$

where $T_1 = \{T | T > \tau \cdot \delta\}$ and k is the number of preventive repairs before the failure. It is due the fact revealed in sect. 2.2 and shown in Figure 1, that the hazard rate stabilizes and after each preventive action the conditional distribution above is (approximately) the same.

The direct evaluation of objective function is not easy, moreover, it is strongly non-concave. Therefore, the distribution of variable $Y(\delta,\tau) = \varphi(T,\delta,\tau)$, for different δ,τ, is obtained ‚empirically' by random generation, its characteristics then as sample characteristics. The choice could be the mean, median, or certain quantiles.

For numerical illustration we selected $T \sim \text{Weibull}(a = 100, b = 2)$, with survival function $\overline{F}(t) = \exp\Big(-(\frac{t}{a})^b\Big)$, $\text{E}T \sim 89$, $\text{std}(T) \sim 46$. Further, the costs $C_0 = 40$, $C_1 = 2 + ((1-\delta)\cdot\tau)^\gamma$, $\gamma = 0.2$. Such a selection of C_1 corresponds to case when the degradation $S(t) \sim t$, in the sense of previous discussion, value 2 stands for fixed costs. We decided to maximize the $\alpha = 0.1$ quantile of distribution of $\varphi(T, \delta, \tau)$. Optimal parameters were found with the aid of the Bayes optimization method (cf. [2]) using 2-dimensional Gauss process as an approximation of the 10% quantile of the objective function. Such a choice says that (roughly) with 90% probability the value of $\varphi(T, \delta, \tau)$ will be larger than found maximal value.

Figure 2 shows the results. The procedure started from its Monte Carlo generation in 9 points showed in the left plot. Maximum is denoted by a circle, its value was 0.876. The plot contains also contours of resulting Gauss process surface. The right plot shows the situation after 12 iterations. It is seen how the space was inspected, maximal value was stabilized around 1.124, the corresponding point $(1 - \delta \sim 0.7, \tau \sim 20)$ is again marked by a circle.

5 CONCLUSION

In the present paper, first, several variants of the Kijima II model were presented, relating the maintenance degree to the reduction of the followed technical object degradation. The main objective then was to show how such models can be connected with maintenance costs evaluation, and, finally, with stochastic optimization problem. One example of such a task was formulated in detail and solved, with the aid of Bayes optimization approach, though other procedures of randomized search are applicable as well.

Acknowledgements

The research has been supported by the grant No. 18-02739S of the Grant Agency of the Czech Republic.

References

[1] Bagdonavicius, V.B. and Nikulin, M.S. (2002). Accelerated Life Models: Modeling and Statistical Analysis. Boca Raton: Chapman and Hall/CRC.

[2] Brochu, E., Cora, V.M. and de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599v1.

[3] Dorado, C., Hollander, M. and Sethuraman, J. (1997). Nonparametric estimation for a general repair model. The Annals of Statistics, Vol(25): 1140–1160.

[4] Embrechts, P., Klüppelberg, K. and Mikosch, T. (1997). Modeling Extremal Events. Berlin: Springer.

[5] Finkelstein, M. (2007). Imperfect repair and lifesaving in heterogeneous populations. Reliability Engineering and System Safety, Vol(92): 1671–1676.

[6] Kijima, M. (1989). Some results for repairable systems with general repair. Journal of Applied Probability, Vol(26): 89–102.

[7] Percy, D.F. and Alkali, B.M. (2005). Generalized proportional intensities models for repairable systems. Management Mathematics, Vol(17): 171-185.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 4:

Location and Transport, Graphs and their Applications

EVALUATION OF NET PRESENT VALUE IN SUPPLY CHAINS USING NETWORK SIMULATION METHOD*

Francisco Campuzano-Bolarín[1], Fulgencio Marín-García[2], José Andrés Moreno-Nicolás[3],
[1,2,3]Universidad Politécnica de Cartagena, Campus Muralla del Mar 30202 Cartagena (Murcia) España
E-mails: [1]Francisco.cmpuzano@upct.es; [2]pentxo.marin@upct.es ;[3]josea.moreno@upct.es;

Marija Bogataj[4]
[4,5]University of Ljubljana, SEB & INRISK, Slovenia,
E-mail: [4]marija.bogataj@ef.uni-lj.si

David Bogataj[5]
[5]Università di Padova, Department of Management and Engineering, Str. S. Nicola, 3 - 36100 Vicenza, Italy
E-mail: [5]david.bogataj@unipd.it

Abstract: The purpose of this paper is to present the Network Simulation Method (NSM) to evaluate Supply Chains and investments to their capacities. Time delays are essential when providing food and drugs. Investments in the expansion of capacities are justified if the increase in NPV is greater than the investment, but NPV is overestimated if it is valued on an infinite horizon, which is characteristic in the case of analyses of economic consequences in the frequency domain, using Laplace transforms in MRP models. Therefore, we suggest parallel use of NSM to reduce these overestimations. The first NSM applications concern different science and engineering fields and has not been used for studying financial flows. In this article, the authors point out the possibility of using the NSM so that the Net Present Value (NPV) of annuity streams would be evaluated on a final horizon even in case of stochastic behaviour of some parameters. An assessment of long-term profitability can be defined easily at any perturbations. To check the tool, the circuit simulator NGSPICE, which was never used for the evaluation of the financial consequences of a SC management or any other economic consequences, has been applied. This approach holds good for both stochastic and deterministic processes in SC finances.

Keywords: Industrial Engineering, Supply Chain, Net Present Value, Network Simulation Method

1 INTRODUCTION

The practitioners who measure the long-term profitability of investment very often face problems that involve calculating the discounted or cumulated value of cash flow or the internal rate of return of an investment project, where parameters of the model could change [22]. In the providing drug and food in the Supply Chains (SC), or an appropriate timing at other medical interventions, also time delays should be better coordinated and controlled. Planning such systems are often based on the evaluation of investments for higher efficiency, which reduce the time delays in total chain [9]. We can use The Material Requirements Planning (MRP) Theory to evaluate the impact of timing in the production system [11, 12]. When using Laplace or z-transforms, the economic evaluation of perturbation of timing on the infinite horizon was easier, but for the solutions on the finite horizon where the interest rate and other parameters are changing stochastically, the method gives only approximate solutions. Therefore, we wish to add an advanced support with the Network Simulation Method (NSM), which is highly developed and has found a wide variety of applications-solutions of differential equations, as in the fields of fluid flow, mechanical vibrations, tribology, dry friction, elasticity, electrochemical reactions - ion-exchange membrane systems [17,18], other transports across membranes [13] in general magneto-hydrodynamics [4]. Recently, it has been employed in the mechanics of deformable solids [15], and also in dry friction in atomic terms [14, 24]. Several programs use the NSM as a tool for numerical calculations: PRODASIM for designing simple fins [10], PROCCA-09 for the design and optimization of thermal problems [1], FATSIM-A for simulating flow fluids with solute transport problems [3], FAHET for the simulation of flow fluids with heat transfer problems [2], EPSNET_10 for simulation of elasticity [16] and OXIPSIS_12 for simulating corrosion problems [23]. Remer and Nieto [22] present 25 different methods and techniques used to evaluate the economic consequences of projects. They categorized 25 methods into five types of evaluation criteria: net present value (NPV), the rate of return, ratio, payback, and accounting methods. Already a quarter of a century ago, surveys of the project evaluation techniques used by some of the largest "Fortune 500" companies, such as DuPont, Kodak, Ford, IBM and Westinghouse, were analysed by Remer et al. [21). According to these surveys, a shift from the internal rate of return method to the net present value criterion methods has taken place as more challenged. Beullens and Janssens [5, 6] introduced the Anchor Point (AP) in NPV models and show that its position in the SC can affect the valuation of capital costs in different system stages. The paper considers infinite horizon models with constant demand. The NPV economic consequences of multistage production systems can be evaluated using MRP Theory [11, 12] and in its extension to EMRP, which enable the NPV of activities and their delays to be evaluated in a SC [9]. The overview of development the MRP Theory is given in Bogataj and Bogataj [8]. The study of linear approximations derived from the annuity stream remains useful for many reasons. First, these models often offer analytical solutions and insight. Second, they are typically accurate if we account for the fact that managers commonly do not wish to implement solutions with very long cycle times which enable an approximation on infinite horizon when calculating NPV. Therefore we have short cycle time and long horizon, but in case of high frequency of the innovative solutions and technical obsolescence the linear approximation is not always acceptable. Therefore we have to be aware that the MRP Theory as derived is indeed a theory about the linear approximation, but. the NPV Equivalence Analysis (NPVEA) shows that a linear approximation is needed to prove which first-order effects are responsible for the difference. Therefore, the NPVEA opens up the route to improve the theory [5,6].

In this paper, we wish to present how useful is a reliable efficient network model for NPV evaluations that can be simulated by NGSPICE [19]. NGSPICE is known as an open source mixed-level/mixed-signal circuit simulator which code is based on three packages: Cider1b1, Spice3f5, and Xspice. It is a stable and reliable simulator, incorporated into many projects, whose efficiency has been proven in many science and engineering problems [24]. Once the equivalence between electrical and mechanical variables has been chosen, linear terms are easily implemented by linear electrical devices, such as resistors, capacitors and coils, while non-linear and coupled terms are implemented using auxiliary circuits, or with controlled current and voltage sources. The last type is a special kind of source whose output can be defined by NGSPICE as dependent or independent variables, determined in any node or any electrical component of the model. Boundary and initial conditions, which can be either linear or not, are also easy implemented by suitable electric components. Once the network model has been designed, it runs without having to resort to other mathematical manipulations as the simulation code does this. NGSPICE requires relatively short computing times, thanks to the continuous adjustment of the internal time step requiring convergence. Besides, the solution simultaneously provides all the variables of interest: revenue flow, demand, production intensity, inventory and satisfied demand. It is also possible to implement new expressions for components with small changes in codes. As a result, the NPV can be plotted simultaneously, which is shortage when Laplace transform is used. Therefore for very complex SC a parallel procedures are advised. A summary of the network method and the economic evaluation factors is provided in Section 2. In Section 3, cash flow is defined by a time function and the NPV is expressed by integration with an exponential expression for the discount rate, following the classic format used by Grubbström in his several papers, recently listed in [8]. Sections 4 and 5 present two examples. The differences of results between them show the power of the proposed method. In Section 6, the designs of the network models are indicated, and the results are depicted in Section 7, in which the evolution of the cash flow and NPV components is presented, based on the NGSPICE. Therefore, the general conclusion is, that the NSM, based on NGSPICE can be used not only for simulation of the biological, chemical and technological processes but also for evaluation of many economic and business processes, giving the exact values of the annuity stream and the NPV evaluation.

2 FUNDAMENTALS AND THE GOVERNING EQUATIONS

The formal approach to the NSM governing equations, which is the basis of the development of NPV evaluation, is the 'Network Theory' of Peusner [20] in which his principles of network thermodynamics were develped and biophysical applications have been described. The variables that characterize the problem must satisfy Kirchhoff's laws, and their relationships must determine the corresponding circuit elements. In each process, once the conjugate variables have been chosen, the information about the circuit elements involved in the network model should be determined, and how they connect each other should be formalized in the mathematical model.

Let's consider the case of an investment project. The net present value of a set of cash transactions with the amounts of c_1, c_2, c_3,... at future points of time t_1, t_2, t_3,…is considered. We shall write c_{ij} as the amount of money to be received ($c_{ij}>0$) or as that paid out ($c_{ij}<0$) for the i-th alternative at the j-th time point. The net present value of this alternative is expressed as:

$$NPV_i = \sum_{j=1}^{N_i} c_{ij} \cdot (1 + r_{1i} \cdot T)^{-n_{ij}} = \sum_{j=1}^{N_i} c_{ij} \cdot (1 + r_{1i} \cdot T)^{-\frac{t_{ij}}{T}} \qquad (1)$$

where r_{1i} is the discounted interest rate and T is the unit time period to which r_{1i} is related, n_{ij} is the number of unit time periods until the j-th transaction for the i-th alternative takes place (t_{ij}/T), and N_i is the total number of transactions for the i-th alternative. In Eq. (1) we chose an interest rate related to a unit time period, T, e.g. a year. Let's now define a new quantity r_{2i}, so that Eq. (1) becomes:

$$NPV_i = \sum_{j=1}^{N_i} c_{ij} \cdot e^{-r_{2i} \cdot t_{ij}} \qquad (2)$$

Let's now consider the more general case where receipts and disbursements can exist continuously in time together with transactions at discrete time points which is easy to simulate with NGSPICE. Cash flow can now be suitably described by the time function:

$$v(t) = v'(t) + \sum_{k=1}^{K} NPV_k \cdot \delta(t - t_k) \qquad (3)$$

where v'(t) is the difference between the continuous in- and out-payments per time unit, NPV_k is the cash receipt or disbursement at the discrete time point t_k (monetary unit/time unit), and $\delta(t-t_k)$ is the Dirac pulse that exists at time point t_k. Therefore

$$dNPV = e^{-r \cdot t} \cdot v(t) \cdot dt = \left(v'^{(t)} \cdot dt + \sum_k' NPV_k\right) \cdot e^{-r \cdot t} \qquad (4)$$

where $\sum'$ indicates that the summation is performed over the set of transactions $\{NPV_k\}$ which exist within the domain *N*, and r is the discounted rate as in (2). The *NPV* becomes:

$$NPV = \int_N v(t) \cdot e^{-r \cdot t} \cdot dt = \int_N v'^{(t)} \cdot e^{-r \cdot t} \cdot dt + \sum_{k=1}^{K} NPV_k \cdot e^{-r \cdot t_k} \qquad (5)$$

3 ECONOMIC PROCESS

To check the method, a simple stochastic economic process was chosen. This case consists of analysing a seasonal variation of the revenue flow with a demand described by

$$d(X, t) = X \cdot cos(\omega \cdot t) + A \qquad (6)$$

where ω is related to the period of complete seasonal fluctuation, T, is represented by $2\pi/\omega$, *X* is the stochastic amplitude of fluctuation, and *A* is a constant. Given the market price of each unit, a_d, the value of revenue flow is

$$f(X, \mathrm{t}) = \mathrm{a_d} \cdot [X \cdot \cos(\omega \cdot \mathrm{t}) + \mathrm{A}] \qquad (7)$$

From Eq. (5), the following equivalence is applied:

$$NPV = \lim_{T\to\infty} \int_0^T f(X,\mathrm{t}) \cdot e^{-r\cdot t} dt \tag{8}$$

At given intensity of production at each activity cell p(t), demand d(t) and inventory i(t),

$$d(t) = D \cdot \big(1 - cos(\omega \cdot t)\big);\; p(t) = \begin{cases} p_0 & if\ i(t) < i_0 \\ 0 & if\ i(t) \geq i_0 \end{cases} \tag{9}$$

The following function represents the inventory level:

$$i(t) = \begin{cases} \int_0^t [p(t) - d(t)] \cdot dt & if\ d(t) - p(t) - i(t) < 0 \\ 0 & if\ d(t) - p(t) - i(t) \geq 0 \end{cases} \tag{10}$$

Not necessarily all demand can be satisfied. The expression for satisfied demand is:

$$d_S(t) = \begin{cases} d(t) & if\ p(t) + i(t) - d(t) > 0 \\ p(t) + i(t) & if\ p(t) + i(t) - d(t) \leq 0 \end{cases} \tag{11}$$

Then the following equivalence is applied

$$NPV = \lim_{T\to\infty} \int_0^T \big(a_d \cdot d_S(t) - a_p \cdot p(t) - a_i \cdot i(t) - c \cdot t\big) \cdot e^{-r\cdot t} dt \tag{12}$$

where a_d is the market price of each unit, a_p is the production and sales cost per unit, a_i is the inventory cost per unit and c is the fixed cost per time unit.

4 RESULTS

The initial conditions are inserted into the specifications of the capacitors' initial conditions (initial voltage value) and the coils (initial current value), respectively. The whole network model now runs by employing NGSPICE.

4.1: The parameters in the stochastic model are the following: T = 1, A = 1, r =0.1 and a_d is 1, the stochastic amplitude of fluctuation, X, is considered Gaussian distribution with a mean of 1 and standard deviation of 0.01.

$$[\mathcal{L}\{a_d \cdot (m_X \cdot \cos \omega t + A)\}]_{s=r} = a_d \cdot \left(\frac{r \cdot m_X}{r^2+\omega^2} + \frac{A}{r}\right) \tag{13}$$

After 70 T the NPV goes to 10.08. Fig. 1 shows the NPV T=1 and 10. Fig.2 show how the variation of r influence the approximation of the value got by MRP Theory.

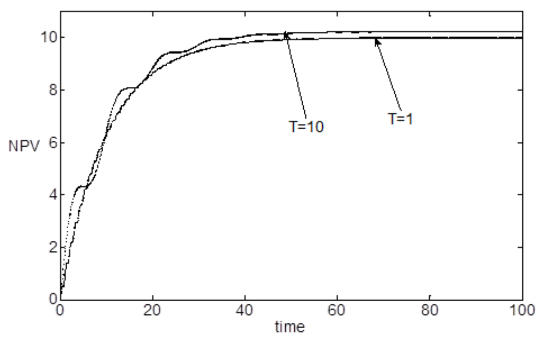


Figure 1: Evolution of the net present value for parameter T= 1 and 10

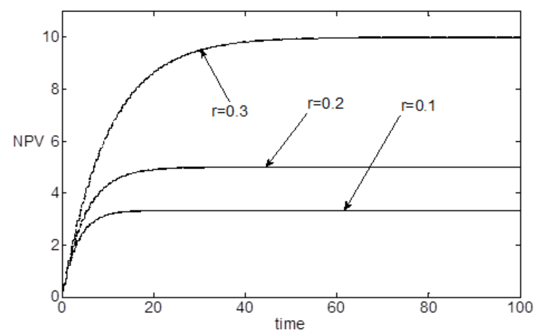


Figure 2: Evolution of the net present value for r: 0.1, 0.2 and 0.3; where D =1; p_0= 1, i_0=1 ; T=1, a_d=1, a_p= 0.01; a_i=0.1 c=0.01

In such a system the evolution of the demand, production and inventory is presented in fig. 3.

4.2: Let us consider the numerical example described in [8]. Activity cell D assembles 2 units of E and 1 unit of F; activity cell B demands 3 units of D for the production of 1 unit of B, A demands 1 unit of B, and 2 units of C for the production of 1 unit of A. The BOM of this example is presented in fig. 4. The average production lead times $\bar{\tau}_i$ are described in the cells and the average transportation lead times are described by $\bar{\tau}_{ij}$.

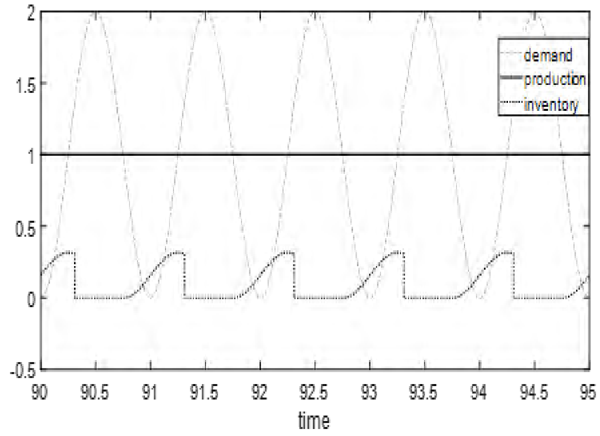


Figure 3: Evolution of the demand, production and inventory functions for the system 4.1

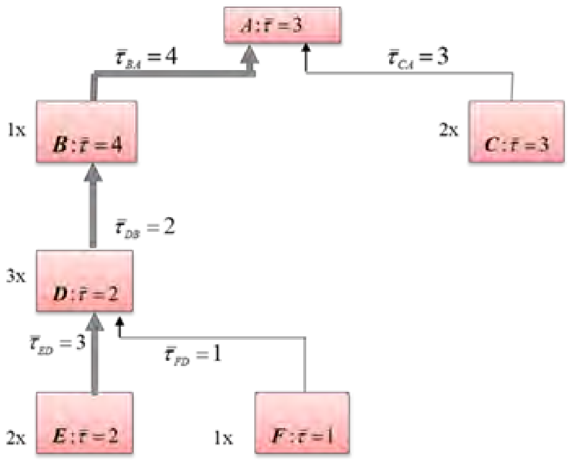


Figure 4: The SC where the risk at simultaneous robust perturbations was analysed in Bogataj et al., [8]

For such a system the impact of number of cycles to the end of activities and the impact of interest rate on the overestimation of NPV evaluated by MRP approach using Laplace transforms (assumption of the infinite horizon) is given in table 1.

Table 1: The overestimation of NPV evaluated by MRP approach using Laplace transforms (assumption of the infinite horizon)

Interest rate	**Horizon in the number of cycles**					
	100	**200**	**300**	**400**	**500**	**1000**
0.01	58.7%	16.3%	5.3%	2.0%	1.0%	0.0%
0.03	5.3%	1.0%	0.0%	0.0%	0.0%	0.0%
0.04	2.0%	1.0%	0.0%	0.0%	0.0%	0.0%
0.06	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

5 CONCLUSIONS

In this paper, we present how useful is the NSM for NPV evaluations in case of th stochastic processes and a finite value of the horizon to study perturbations, which have been analysed for an infinite horizon by Bogataj et al. [7]. The SC can be simulated by NGSPICE, for several stochastics (like latency because of ageing workforce) and deterministic cases, where the time horizon can also be finite. The results were compared with those obtained by Laplace transform. The software is flexible enough to implement different values of the parameters involved in a model with no additional effort, and allows several types of graphs to be produced in short time. The model can be further developed for sensitivity analysis of the SCs in cases where proper timing is often of high importancy. The proposed examples involve all the details of actual process and the method is ready to be check with data in future works. The previous study of SC in the frequency domain, where the time horizon of the reverse Laplace is infinite, was not very convinient to study the impact of the technical obsolescence of the final product. Therefore, for further research we propose to analyse the perturbations in a SC (i.e. time delays) in the skeleton of the MRP theory, using the NSM to study the finacial impact of the product lifespan on the NPV of the total SC.

***Acknowledgement:** The paper was partly financed by ARRS under the contracts J6—939 and P5—0364.*

References

[1] Alhama, F., Del Cerro, F., 2010. PROCCA: Programa de conducción de calor. NAR 08/2010/18 [software]. Universidad Politécnica de Cartagena.

[2] Alhama, I., Cánovas, M., Alhama, F. 2015. Simulation of fluid flow and heat transport coupled processes using FAHET software. *Journal of Porous Media,* 18(5): 537-546.

[3] Alhama I, Soto A, et al. , 2010. FATSIM-A: Fluid flow and solute transport simulator. N.D. Mu-1093-2010 [software]. Cartagena: Universidad Politécnica de Cartagena.

[4] Beg, O.A., Zueco, J., Takhar, H.S., Beg, T.A., Sajid, A., 2009. Transient nonlinear optically-thick radiative-convective double-diffusive boundary layers in a Darcian porous medium adjacent to an impulsively started surface: network simulation solutions. *Communications in Nonlinear Science and Numerical Simulation,* 14(11): 3856-3866.

[5] Beullens, P., Janssens, G.K., 2014. Adapting inventory models for handling various payment structures using net present value equivalence analysis. *Int. J. Prod. Econ.,* 157: 190–200.

[6] Beullens,P., Janssens, G.K., 2011. Holding costs under push or pull condi- tions: the impact of the anchor point. *European Journal of Operational Research* 215:115–125.

[7] Bogataj, D., Aver, B., Bogataj, M. 2014. Supply chain risk at simultaneous robust perturbations *Int. J. Prod. Econ*. 181(A): 68-78.

[8] Bogataj, D., Bogataj, M. , 2019. NPV approach to material requirements planning theory: a 50-year review of these research achievements. *Int. J. Prod. Research,* 57(15-16): 5137-5153.

[9] Bogataj, M., Grubbström, R.W., Bogataj, L., 2011. Efficient location of industrial activity cells in a global SC. *Int. J. Prod. Econ.,* 133(1): 243-250.

[10] Gómez-Lopera S, Alhama F, et al.,2005. PRODASIM: Programa de diseño de aletas simples. NAR 08/2005/544 [software]. Universidad Politécnica de Cartagena.

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

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

[13] Horno J, González-Caballero F, Hayas A, González-Fernández CF., 1990. The effect of previous convective flux on the nonstationary diffusion through membranes network simulation. *Journal of Membrane Science*. 48(1): 67-77.

[14] Marín F, Alhama F, Moreno JA., 2012. Modelling of stick-slip behaviour with different hypotheses on friction forces. *International Journal of Engineering Science*. 60: 13±24. https://doi.org/10.1016/j.ijengsci.2012.06.002

[15] Morales, J.L., Moreno, J.A., Alhama, F., 2012. New additional conditions for the numerical uniqueness of the Boussines and Timpe solutions of elasticity problems. *International Journal of Computer Mathematics,* 89(13-14): 1794-1807.

[16] Morales JL, Moreno JA, Alhama F., 2011. EPSNET_10: Elasticity problems simulation network. NAR 08/2011/418 [software]. Universidad Politécnica de Cartagena.

[17] Moya, A.A., 2011. Influence of DC electric current on the electrochemical impedance of ion-exchange membrane systems. *Electrochimica Acta*, 56(8): 3015-3022.

[18] Moya, A.A., 2013. Harmonic analysis in ideal ion-exchange membrane systems. *Electrochimica Acta*, 90(15): 1-11.

[19] Nenzi, P.; Vogt, H., 2014. *Ngspice Users Manual version 26*. Ngspice home page: http://ngspice.sourceforge.net/

[20] Peusner L., 1987. The principles of network thermodynamics: theory and biophysical applications. *Entropy.* Lincoln, USA.

[21] Remer, D.S., Stokdyk, S.B. and Van Driel, M., 1993. Survey of project evaluation techniques currently used in industry. *Int. J. Prod. Econ.*, 32: 103-115.

[22] Remer, DS; Nieto, AP.,1995. A compendium and comparison of 25 project evaluation techniques 2. Ratio, payback, and accounting methods. *Int. J. Prod. Econ.,* 42(2): 101-129

[23] Sánchez, J.F., Moreno, J.A., et al. 2013. OXIPSIS_12: Oxidation processes simulation software. *N.D. Mu-751-2013* [software], Cartagena: Universidad Politécnica de Cartagena.

[24] Sánchez-Pérez, J.F., Marín, F., Morales, J.L., Cánovas, M., Alhama, F., 2018. Modeling and simulation of different and representative engineering problems using Network Simulation Method. *PLoS ONE* 13(3): e0193828.

FUNCTIONAL REGIONS DETECTION BY WALKTRAP AND CHAINS' METHODS

Samo Drobne[a], **Albero Garre**[b], **Eloy Hontoria**[c] and **Miha Konjar**[d]
[a] University of Ljubljana / Faculty of Civil and Geodetic Engineering, Jamova cesta 2, 1000 Ljubljana, Slovenia
[b] Wageningen University & Research, Laboratory of Food Microbiology, P.O. Box 17, 6700 AA, Wageningen, the Netherlands
[c] Technical University of Cartagena / Business Management, 30202 Cartagena, Spain
[d] University of Ljubljana / Faculty of Architecture, Zoisova cesta 12, 1000 Ljubljana, Slovenia
samo.drobne@fgg.uni-lj.si, alberto.garreperez@wur.nl, eloy.hontoria@upct.es, miha.konjar@fa.uni-lj.si

Abstract: In this paper, we analyse functional regions by using two graph theoretical methods, walktrap algorithm and chains' approach. The quality of the regionalization procedure is analysed by applying fuzzy set theory. The application for Slovenia shows that walktrap algorithm calculates more self-contained functional regions than chains' approach.

Keywords: functional regions, graphs, walktrap algorithm, chains' method, fuzzy sets, Slovenia.

1 INTRODUCTION

The organisation of space is a crucial factor for the understanding of and the explanation for different socio-economic phenomena. It is, therefore, necessary for spatial development, territorial planning and the implementation of different spatial policies, aimed towards a more efficient organisation [7, 12, 14]. However, many research works (e.g. [1, 4, 8, 17, 18]) have pointed out that standard administrative regions used by different governments for policy making, resource allocation, and research do not provide meaningful information on actual conditions of a particular place or region. Consequently, in last years efforts have been directed towards the identification and delineation of functional regions more meaningful than those currently in use.

A functional region (FR) is a territorial area characterised by high frequency of intra-regional interactions [17, 25, 28]. FR organisation is based on horizontal relations in space in a form of spatial flows or interactions between basic spatial units (BSUs) of the region. Functional regionalisation consists in the combination of BSUs into FRs with the goal of generalising the functional flows and spatial interactions addressed. So, FRs can be understood as generalised patterns of flows and interactions in space [10]. The literature tends to favour three approaches to FR taxonomy [14]: graph theoretical methods (e.g. [2, 11, 16, 17, 19, 24]), methods of numerical taxonomy (e.g. [20, 21, 22]) and rule-based methods (e.g. [5, 6, 15]).

In this paper, we calculate FRs for Slovenia using two methods based on graph theory: walktrap algorithm [26] and chains' method [11, 17, 19]. The quality of both regionalization procedures is analysed by applying fuzzy set theory. The analysis for Slovenia shows that the walktrap algorithm calculates more self-contained regions, i.e. regions with higher relative frequency of intra-regional flows, than chains' approach.

The remainder of this paper is structured as follows. Section 2 includes material and methods description. This is followed by results presentation and discussion in section 3. Section 4 summarizes and concludes the paper.

2 METHODOLOGY

We analysed FRs in Slovenia using two methods based on network theory. For that purpose, the complex systems of labour commuting between Slovenian municipalities in 2017 was described as a network. A network is a mathematical structure consisting of vertexes with pairwise relationships represented by edges. In our study, the municipal centroids are the vertexes of the network that are connected by an edge if inter-municipal commuting flows was recorded in the time-period analysed. Weights (w_{ij}) were assigned to each edge, according to the number of commuters registered between the two municipalities. w_{ij} represents the number of commuters from municipality i to j. Communities in the network – i.e. functional regions of the inter-municipal labour commuting flows – were identified using the walktrap algorithm [26] and chains' approach [11, 17, 19].

The walktrap algorithm [26] is a heuristic algorithm that clusters vertexes of the network based on a distance, r, that measures the connectives of two nodes, i.e. municipal centroids. The distance r_{ij} between nodes i and j is defined as shown in Equation 1, where P_{ik}^t is the transition probability from node i to k in t steps, $d(k)$ is the degree of node k and n is the number of nodes in the network.

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^t - P_{jk}^t\right)^2}{d(k)}} \tag{1}$$

In the walktrap algorithm, the transition probabilities are estimated using random walk. Briefly, Q random walks of length t are taken from randomly selected nodes. In each transition, the walker travels from node i to node j with probability $w_{ij}/\sum_k w_{ik}$. Then, the transition probabilities P_{ik}^t are calculated as the fraction of walkers who ended in node k after t steps. Once the distance matrix r_{ij} has been calculated, vertexes are aggregated using a hierarchical clustering algorithm. The walktrap algorithm has been applied using the implementation included in the *igraph* R package [9], with R version 3.4.3 [27]. The parameter t was set to four, whereas Q was increased until the results converged. In the application to the inter-municipal labour commuting data, we have also tested $t = 2$, 3, 5 and 6, but only $t = 4$ has generated realistic and compact FRs. To the knowledge of the authors, this is the first time, that walktrap algorithm has been applied in spatial science to model FRs.

As a second method to calculate FRs, we used chains' approach, introduced and applied by [17] and later improved by [19] and [11]. The first step in this approach is the identification of the centres of FRs. They are defined as most important employment centres in the analysed territory that are strongly self-sufficient. Municipality is strongly self-sufficient if most of its active population also work in that same municipality; usually, this percentage is set to 66.67% or more [11, 17, 19]. Although methods for this step are described in literature [11, 19], in order to ease comparison with the walktrap algorithm, the centres of FRs calculated by the walktrap algorithm were also used for the chains' approach. In a second step, chains of nodes are created with adding municipalities to self-sufficient municipalities, i.e. predefined centres of FRs, until the condition defined in Equation (2) is satisfied. This condition defines the border of FR_i, which is the break line, where the attraction is equal to both of the closest self-sufficient municipalities:

$$FR_i = \{x : w_i(x) \geq w_j(x)\}, \tag{2}$$

where i and j denote two FRs' centres that are connected by a line, and x stands for an intermediate point between the endpoints i and j. At a location x, the commuting frequency to the centre i is $w_i(x)$. The chains are formed for three types of municipalities (nodes): (a) the municipalities, that are directly connected with their maximum flow to the centre, are automatically placed to that centre; (b) municipalities that are not directly connected with their maximum flow to the centre, but they are connected with their maximum flow to a non self-sufficient municipality, which is then connected to a predefined centre (chains are determined iteratively); and (c) the pairs of municipalities, which present to each other the destination of their maximum flows, are connected to the region, in which the direction of the second maximum flow is oriented. As suggested by [17], the chain was allowed to have three links in our application for Slovenia. If more links existed, the link was broken at the weakest point. Moreover, we tested the approach by allowing three and four links, without any impact on the results. The chains were calculated automatically, using our own software based on Java platform [19].

After FRs were identified using both algorithms, we compared the quality of both regionalization procedures applying fuzzy set theory (FST), as suggested by [13] and [29, 30]. FST extends crisp set theory, enabling that an element (BSU, in this study) can partially belong to a group (FR, in this study). Therefore, it can also simultaneously belong to more than one group. By using FST approach, we can identify potential misallocations of BSUs across the FRs by the measurement of a membership function, so that each BSU can be partially assigned to a series of fuzzy FRs. A membership function for BSU i with respect to fuzzy residential FR m is defined as

$$M'_{im} = \sum_{j\in(g)m} w_{ji}/w_{\cdot i} \,, \quad (3)$$

where BSU i belongs to FR m on the basis of a regionalization method. On the other hand, the membership function with respect to fuzzy local employment FR m is defined as:

$$M''_{im} = \sum_{j\in(g)m} w_{ij}/w_{i\cdot} \,. \quad (4)$$

The membership function with respect to a fuzzy FR, m, M_{im}, was calculated as the average of M'_{im} and M''_{im}:

$$M_{im} = (M'_{im} + M''_{im})/2 \,. \quad (5)$$

To compare the quality of each regionalization, we calculated average membership values for each FR, and for the whole system of FRs of Slovenia.

3 RESULTS

Figure 1 shows the results of two applied regionalization procedures, both based on network theory methods. The walktrap algorithm (left) generated eight FRs of Slovenia. Seven of them are expected and in line with previous research work (see, e.g., [10, 11, 19]), those FRs are: FR Murska Sobota (80), FR Maribor (70), FR Celje (11), FR Slovenj Gradec (112), FR Ljubljana (61), FR Novo mesto (85), and FR Nova Gorica (84). However, FR Tolmin (128), which consists of only three, relatively big municipalities surrounded by high mountains, has never been modelled at macro level of Slovenia; instead of FR Tolmin, literature [10, 11, 19] show FR Koper on the south-west costal part of Slovenia. Further analyses also showed that FR Tolmin (128) relationship with other FRs is much weaker than the relationships analysed between defined FRs with more central location in Slovenia.

The result of the chains' method (Figure 1, right) is very similar to the one of walktrap. It estimates eight FRs, but there are differences in their sizes. FR Ljubljana (61) is much bigger mostly on the account of FRs Celje (11) and Novo mesto (85), that are consequently smaller. Other FRs are (almost) identical for both algorithms.

The quality analysis of the regionalization procedures, i.e. the comparative analysis of the general membership values of FRs calculated by FST approach, shows that walktrap algorithm generates FRs Ljubljana (61), Celje (11) and Novo mesto (85) with higher average membership values than chains' approach (see Table 1). Moreover, mean membership values of almost all FRs generated by walktrap algorithm are higher or equal to mean membership values of FRs calculated by chains' method; the only exception is Slovenj Gradec (112).

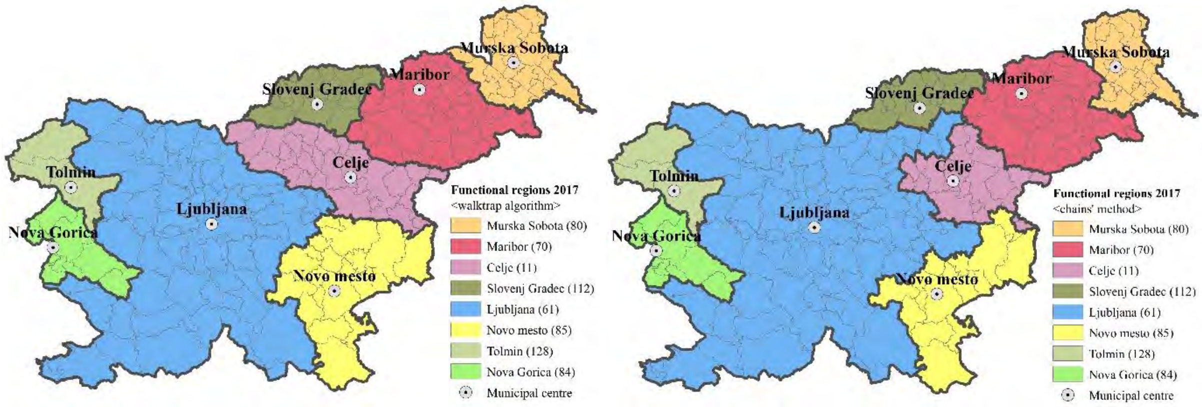


Figure 1: Eight functional regions in Slovenia in 2017 defined by inter-municipal labour commuting flows in Slovenia in 2017 and generated by walktrap algorithm (left) and chains' method (right)

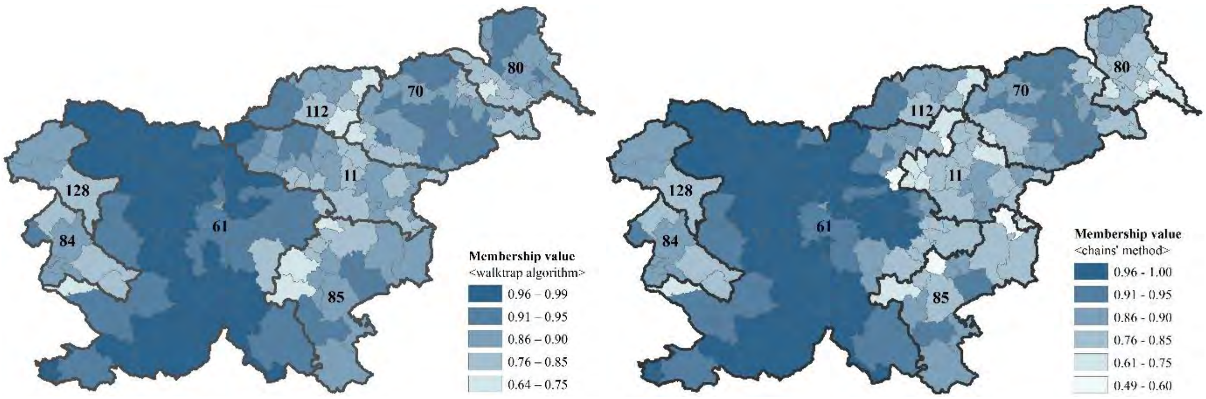


Figure 2: Membership values of Slovenian municipalities (year 2017) in the functional region to which they were located by using walktrap algorithm (left) and chains' method (right)

Table 1: Mean membership values of the functional regions / Slovenia

Functional region / Slovenia	Mean membership value of the functional region / Slovenia	
	walktrap algorithm	chains' method
Slovenia	0.895	0.874
Celje (11)	0.857	0.787
Ljubljana (61)	0.955	0.919
Maribor (70)	0.886	0.884
Murska Sobota (80)	0.869	0.801
Nova Gorica (84)	0.861	0.861
Novo mesto (85)	0.855	0.808
Slovenj Gradec (112)	0.845	0.854
Tolmin (128)	0.843	0.843

Generally, municipalities with the highest membership values are located in the centres of FRs, whereas municipalities with the lowest membership values are located on the periphery of FRs. Regarding FRs calculated by chains' method, most of the municipalities with a low membership values lie on the border between FRs. In most cases, these municipalities are also those that are assigned to a different FR by the walktrap algorithm (i.e. on the border areas between FRs Ljubljana (61), Celje (11) and Novo mesto (85)). Hence, the chains' algorithm potentially misallocates these municipalities. On the other hand, the FRs calculated by the walktrap algorithm also have some municipalities with low membership values. Those municipalities are located on the border between FRs Ljubljana (61) and Novo mesto (85), and Ljubljana (61) and Nova Gorica (84), and in the FR Slovenj Gradec (112), which is the only one with lower average membership value comparing both systems of FRs. These results, together with the higher mean membership values of the walktrap algorithm, points out that for the case studied the walktrap algorithm provides a better classification.

4 CONCLUSIONS

In the paper, we analysed FRs by using walktrap algorithm and chains' approach. The results of both methods based on graph theory have been compared and analysed in depth using fuzzy sets. In the case studied (Slovenia for year 2017), the walktrap algorithm identifies more meaningful FRs than the chains' approach: walktrap algorithm calculates FRs with lower level of potentially misallocated BSUs.

As a direction for future work, the algorithms analysed here could be compared to other graph-based methods for FR identification (e.g. [2]), as well as with other methods (e.g. the most popular rule-based regionalization procedure, i.e. CURDS's method [6]). Furthermore, network theory could be applied to analyse in depth the structure of FRs at different levels (micro, mezzo and macro).

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0406 Earth observation and geoinformatics) and from the Ministry of Economy and Competitiveness (MINECO) of the Spanish Government and European Regional Development Fund (project AGL2013-48993-C2-1-R). Alberto Garre (BES-2014-070946) is grateful to the MINECO for awarding him a pre-doctoral grant.

References

[1] Ball, R. M. 1980. The use and definition of Travel-to-Work areas in Great Britain: Some problems. Regional Studies, 14(2): 125–139.

[2] Benassi, F. Deva, M., Zindato, D. 2015. Graph regionalization with clustering and partitioning: an application for daily commuting flows in Albania. Regional Statistics, 5(1): 25–43.

[3] Berry, B. J. L., Garrison, W. L. 1958. The functional bases of the central place hierarchy. Economic Geography, 34(2): 145–154.

[4] Casado-Diaz, J. M. 2000. Local labour market areas in Spain: a case study. Regional Studies, 34: 843–856.

[5] Coombes, M. G., Green, A. E., Openshaw, S. 1986. An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work-areas revision in Britain. Journal of the Operational Research Society, 37(10): 943–953.

[6] Coombes, M., Bond, S. 2008. Travel-to-Work Areas: the 2007 review Office for National Statistics, London.

[7] Coombes, M., Casado-Díaz, J. M., Martínez-Bernabeu, L., Carausu, F. 2012. Study on comparable labour market areas – Final research report. Rome, Istat – Istituto nazionale di statistica: 146 pp. http://www.istat.it/it/files/2014/12/Final-Report_LMA-v1-0-17102012.pdf (Accessed 15.5.2018).

[8] Cörvers, F., Hensen, M., Bongaerts, D. 2009. Delimitation and coherence of functional and administrative regions. Regional Studies, 43(1): 19–31.

[9] Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Systems, 1695.

[10] Drobne, S. 2016. Model vrednotenja števila in območij funkcionalnih regij (=A model evaluating the number and areas of functional regions). PhD thesis. Ljubljana.

[11] Drobne, S., Konjar, M., Lisec, A. 2010. Razmejitev funkcionalnih regij Slovenije na podlagi analize trga dela = Delimitation of functional regions of Slovenia based on labour market analysis. Geodetski vestnik, 54(3): 481–500.

[12] Erlebach, M., Tomáš, M., Tonev, P. 2016. A functional interaction approach to the definition of meso regions: the case of the Czech Republic. Moravian Geographical Reports, 24(2): 37–46.

[13] Feng, Z. 2009. Fuzziness of travel to work areas. Regional Studies 43, 5: 707–720.

[14] Halás, M., Klapka, P., Hurbánek, P., Bleha, B., Pénzes, J., Pálóczi, G. 2018. A definition of relevant functional regions for international comparisons: The case of Central Europe. Area 00: 1–11.

[15] Halás, M., Klapka, P., Tonev, P., Bednář, M. 2015. An alternative definition and use for the constraint function for rule-based methods of functional regionalisation. Environment and Planning A, 47: 1175–1191.

[16] Holmes, J. H., Haggett, P. 1977. Graph theory interpretation of flow matrices: a note on maximization procedures for identifying significant links. Geographical Analysis, 9: 388–399.

[17] Karlsson, C., Olsson, M. 2006. The identification of functional regions: theory, methods, and applications. The Annals of Regional Science, 40(1): 1–18.

[18] Klapka, P., Halás, M., Netrdová, P., Nosek, V. 2016. The efficiency of areal units in spatial analysis: Assessing the performance of functional and administrative regions. Moravian Geographical Reports, 24(2): 47–59.

[19] Konjar, M., Lisec, A., Drobne, S. 2010. Methods for delineation of functional regions using data on commuters. V: Painho, M. (ur.), Santos, M. Y. (ur.), Pundt, H. (ur.). Geospatial thinking: proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal, May 10-14, 2010. Guimarães, Springer-Verlag: 1-10.

[20] Masser, I., Brown, P. J. B. 1975. Hierarchical aggregation procedures for interaction data. Environment and Planning A, 7(5): 509–523.

[21] Masser, I., Scheurwater, J. 1978. The specification of multi-level systems for spatial analysis. In: Masser, I. (Ed.), Brown, P. J. B. (Ed.). Spatial representation and spatial interaction. Volume 10 of the series Studies in Applied Regional Science. Leiden and Boston: Springer US: 151–172.

[22] Masser, I., Scheurwater, J. 1980. Functional regionalisation of spatial interaction data: an evaluation of some suggested strategies. Environment and Planning A, 12(12): 1357–1382.

[23] Newman, M., 2010. Networks: an introduction. Oxford university press.

[24] Nystuen, J. D., Dacey, M. F. 1961. A graph theory interpretation of nodal regions. Papers of the Regional Science Association, 7: 29–42.

[25] Philbrick, A. K. 1957. Principles of areal functional organization in regional human geography. Economic Geography, 33(4): 299–336.

[26] Pons, P., Latapy, M., 2006. Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications, 10(2): 191–218.

[27] R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[28] Ullman, E. L. 1980. Geography as spatial interaction. Seattle and London, University of Washington Press: 252 pp.

[29] Watts, M. 2009. Rules versus hierarchy: an application of fuzzy set theory to the assessment of spatial grouping techniques. In: Kolehmainen, M. et al. [Eds.]: Adaptive and naturals computing algorithms. Berlin Heidelberg, Springer-Verlag, 517–526.

[30] Watts, M. 2013. Different spatial grouping algorithms: an application to the design of Australia's new statistical geography. Spatial Economic Analysis, 8(1): 92-112.

DATA CONVERSION AND EXACT APPROACH TO OVERHEAD WIRES NETWORK MINIMISATION FOR THE BATTERY ASSISTED TROLLEYBUS FLEET

Dobroslav Grygar, Michal Kohani
Department of Mathematical Methods and Operations Research
University of Zilina, Univerzitna 8215/1, 010 26 Zilina, Slovakia
E-mail: dobroslav.grygar@fri.uniza.sk, michal.kohani@fri.uniza.sk

Abstract: The technology of battery assisted trolleybus combines benefits of classical trolleybus with freedom movement of standard buses. It can be assumed that we will experience growing trend of implementation of battery assisted trolleybuses in cities. Therefore, in our research we are trying to develop the mathematical model to create a minimal network of overhead wires that would be sufficient for deployment and operation of such battery assisted trolleybuses. Currently, there are no accurate ways how to deploy such technology to cities. This paper mainly deals with problem of developing effective way to prepare data for finding exact solutions of the problem with obtained data. We will describe the process of data preparation and introduce a linear model for exact solution of this problem to this problem. Proposed model is then tested on small networks for validation created from the real network of public transport lines in the town of Zilina.

Keywords: assisted trolleybus, battery, overhead wires, optimization, data conversion, exact approach

1 INTRODUCTION

Majority of EU member countries are calling for reduction of the environmental impact caused by carbon-dioxide emissions currently. Especially in the metropoles there are many vehicles using fossil fuels. For these reasons cities are implementing emission free zones in their city centres. There is need to introduce electric vehicles with zero carbon footprint in public transport. In our previous research we explored options in vehicle types. We performed detailed analysis of limiting factors of battery assisted trolleybuses [6]. This type of vehicles seems to be the most advantageous from the currently available rechargeable green technologies by some economic studies [1].

In our current research we are focusing to create a minimal network of overhead wires that would be enough for deployment and operation of such battery assisted trolleybuses. Ideally, we should be able to cover all the lines in selected area. Previously acquired knowledge about this technology is valuable for the present research progress.

The organisation of paper is as follows. In the section 2 we will describe related works, then we will describe the algorithm for the data conversion. The article presents the process of preparation of data for exact solver. The process we use is based on the article [16]. The authors were dealing with problem of deployment charging lines for electric cars. On their principles, we have built our own algorithm that converts the data into the desired form.

In the next part of paper, we introduce the linear model of the problem. Using the mathematical programming approach we will get the exact solution of the problem. We also performed basic validation of the model using small size problems created from the real data provided by the public transport operator DPMZ in the town of Zilina.

2 RELATED WORK

From the economical point of view the battery assisted trolleybus is considered as the most cost-effective bus system from electric powered public transport solutions. Many studies are based on real word deployment observation on this topic [3] or [13] and more.

Traction batteries are very important component of battery assisted trolleybus. In Polish town Gdynia battery assisted trolleybuses were deployed successfully. Authors from local university published valuable studies [1] and [2]. Battery life is sensitive to the choice of the right charging strategy and battery charging speed is not linear. So, the state of charge of vehicle should fluctuate between 20% and 80% of actual battery capacity. These facts were published and demonstrated by following papers [12] and [15].

In relation to climatic conditions, air-conditioning or heating must be used in vehicles. Thus, significantly higher energy consumption for extreme weather conditions in mentioned in [11]. A simulation study analysed the power consumption of an electric public transport vehicle in [5]. The profile of the route on which the battery assisted trolleybus moves also significantly affects its energy consumption. There are available calculations and simulations [10], [17].

The connection and disconnection of the trolleybus to the overhead contact wires does not affect the duration of the journey because it is realized directly on the adjusted bus stops during the boarding or exiting of the passengers. This process is automated and requires the installation of an auxiliary device to ensure proper connection of the collectors to the wires [6].

At current research, possibilities of deploying the inductive charging installation in the road are explored [7], [8], and [16]. However, building of overhead contact wires construction is easier and the maintenance is also simpler than in the inductive approach. Mentioned works solve a similar task, which is the design of a minimal network of inductive conduction. Therefore, after some adaptations, some of their findings can be used in our research.

It is known that many optimisation problems, on networks may be NP-hard. However, it seems that the real transport networks have some interesting properties which allow us to find a "good" solution in reasonable time [4]. The methods for constructing the robust schedules using agent-based simulation are presented in interesting paper [9].

3 THE PROBLEM FORMULATION

As mentioned above, in our research we are trying to create a minimal network of overhead wires that would be sufficient for deployment and operation of such battery assisted trolleybuses. Overhead contact wires and its maintenance create a significant part of the cost of the entire transport system. Therefore, there is an effort to minimize it [6].

This task can be divided into several parts. Firstly, we need to collect and convert the data from a real public transport application. The second part consists from finding exact solutions for different datasets. It is also necessary to find out boundaries for the maximum scale of tasks. This is done using Xpress IVE solver.

3.1 Data conversion

Data conversion is an important part of solving this problem. Firstly, bus lines need to be selected for optimization. Each line passes through several bus stops. These stops are the nodes of the road network graph. Only a used road segments between stops are selected as edges in road network graph.

In our approach, we want to use the location problem approach to solve the problem. In classical location problem, the set of possible locations of service centres are nodes. Due to this fact, we need to do the transformation of the network to a different form. We need to convert road network graph to road segment graph for each route. In this process we convert edges form road segment graph to nodes in road segment graph and vice versa. This transformation is illustrated in Figure 1.

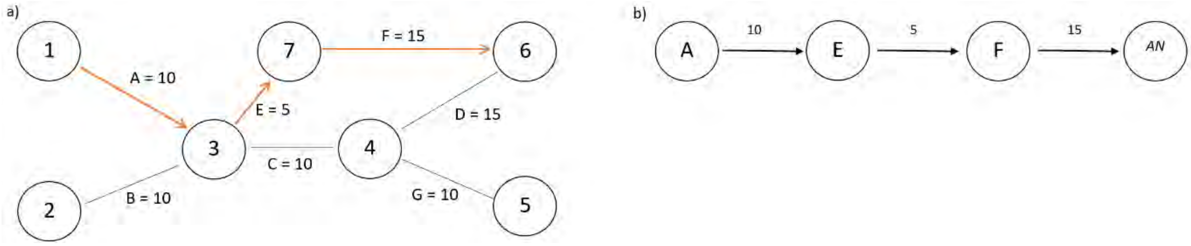


Figure 1: Creation of road segments graph from road network graph. a) Example of road network graph. Bus line (with stops 1, 3, 7, 6) is marked with orange line. Nodes are bus stops and edges are road segments. b) road segment graph created for mentioned bus line. Nodes are road segments, AN is artifical node representing SOC of EV after compleating the trip.

State of charge (SOC) represents actual state energy in vehicle's battery pack in %. Battery assisted trolleybus starts journey with an initial SOC. SOC of vehicle is calculated depending on whether a charging wire is built on the road segment.

In next phase we need to create the state of charge graph for each line. Illustration of vehicle SOC graph is in the Figure 2. Mentioned graph represents all possible combinations of vehicle's SOC. These options are created by combinations of coverage or non-coverage of individual segments of a bus line by charging wires. State of charge graph contains only feasible combinations where vehicle can pass whole route.

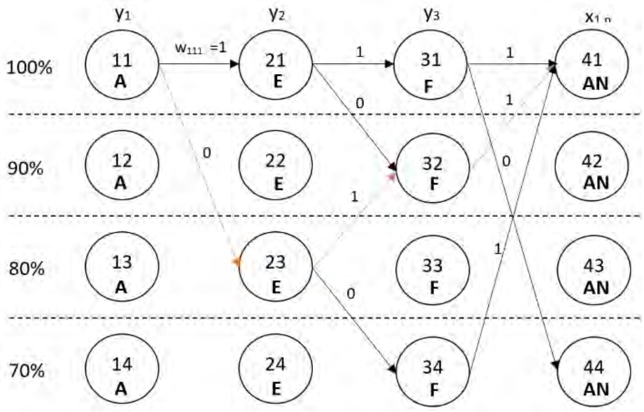


Figure 2: State of charge graph. Rows represents individual charge level and nodes are road segments. Transitions between charge level are marked with 1 of vehicle is charged from wires and with 0 if not. The y_s and $x_{r,i}$ are decision variables, and $w_{r,s,i}$ input data (individual feasible combinations of charging).

Algorithmically the process is slightly complicated. A linear list of all sections is created from the road segments and a unique ID is assigned to each. For each route a list of road segments that the route uses are created. This list may also contain used segments multiple times. Subsequently, all possible combinations of road segments (with or without contact lines) are created from the list of sections. From these combinations are selected those where a vehicle can pass whole route. This is a complex process involving various external parameters and factors. We have a good overview of these factors thanks to the results of our previous research in this area [6]. The selected combinations are stored in output files.

For mentioned data conversion purpose, we have created data conversion software. Which was developed with emphasis on algorithm efficiency, source code clarity and maintainability. This is mainly since this software will be gradually extended with new features and requirements. We used C# programming language and Visual Studio 2017 IDE.

3.2 Mathematical model of the problem

In this section, we describe the linear model of our problem. Firstly, R is set of routes, I_r is set of feasible alternatives for route r and S is set of all used segments. Then n is number of segments and m is the number of routes. The decision variables are following. $x_{r,i} \in \{0,1\}$ is selection of alternative i for route r. $y_s \in \{0,1\}$ segment s in graph will/won't be covered by charging wire.

Finally, the other input constants for the model are following. Combinations are represented by $\boldsymbol{w_{r,s,i}} \in \{\mathbf{0},\mathbf{1}\}$ where for route $\boldsymbol{r}$ on segment $\boldsymbol{s}$ for alternative $\boldsymbol{i}$ is charging line needed. $\boldsymbol{a_r}$ number of feasible alternatives of route $\boldsymbol{r}$. $\boldsymbol{c_s}$ cost of building charging wires for segment $\boldsymbol{s}$. Figure 2 also illustrates relation between decision variables and input data.

$$\min \sum_{s=1}^{n} c_s \cdot y_s \tag{1}$$

$$\sum_{i=1}^{a(r)} x_{r,i} = 1 \qquad for\ r = 1..m \tag{2}$$

$$M \cdot y_s \geq \sum_{r=1}^{n_r} \sum_{i=1}^{a(r)} x_{r,i} \cdot w_{r,s,i} \qquad for\ s = 1..n \tag{3}$$

$$y_s \in \{0,1\} \qquad for\ s = 1..n \tag{4}$$

$$x_{r,i} \in \{0,1\} \qquad for\ r = 1..m\ ; \quad i = 1..a(r) \tag{5}$$

The objective function (1) for the problem of minimizing the total building cost. The selection constraint (2) for choosing one combination for all routes on the line. The building constraint ensures (3) that we install a charging line if at least one route requires an installation. Finally, the obligatory constraints are (4) and (5).

4 THE BENCHMARKS AND COMPUTATIONAL TESTS

Initial testing was aimed to verify the functionality of the model. Several tests were done on short purpose-made bus routes, which was easy to calculate also manually. In next phase, we created two short routes with one of sections common, so that this section had to be covered with overhead line.

Current research was focused to verification of model basic functionality. For this purpose, we used real data from public transport system. We obtained data from the public transport company (DPMZ) in Zilina. Several bus lines were selected. Road sections grouping or other data optimization was not necessary to validate the model at this stage of research. Data was converted using process mentioned in Section 3.1. The maximal number of stations on the line was 21. The test data are presented in Table 1. Used bus routes are also illustrated schematically in Figure 3.

Table 1: Used bus routes in Zilina

Line number	***Stops count***	***Start***	***End***
14	20	Matice slovenskej	Fatranská
3	21	Jaseňová	Jaseňová
4	21	Fatranská	Matice slovenskej
5	15	Fatranská	Jaseňová
50	21	Železničná stanica	Stodolova
6	18	Stodolova	Matice slovenskej
7	12	Stodolova	Sv. Cyrila a Metoda

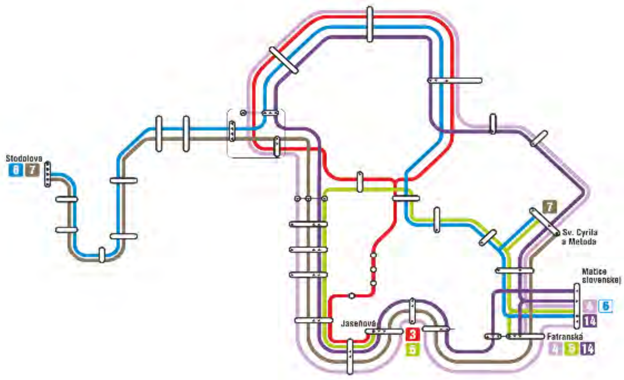


Figure 3: Used bus routes in Zilina illustrated schematically.

All our computational tests were performed on following hardware. Processor Intel Core i5-7200U 2.50GHz with 3.10 GHz turbo boost, paired with 16GB DDR4 2133MHz RAM. Mathematical model was solved using IP-solver XPRESS.

The next testing phase was focused on finding out what maximum route lengths could be processed. For this reason, we chose seven bus routes in Zilina with different lengths, partially overlapped. The situation is described on Figure 3. The longest of these routs have been shortened initially. For test runs we have been gradually adding segments to find a state where the task could not be resolved, whether because of lack of time or memory. The results of these tests can be found in Table 2.

Table 2: Data conversion and optimization time consumption

Routes count	***Maximum segments count***	***Data conversion time (s)***	***Optimization time (s)***
7	15	1.30	8.20
7	18	5.10	142.40
7	19	9.10	525.60
7	20	16.80	134568.40
7	21	out of memory	-

The results show that we can solve exactly tasks that have a maximum of 21 vertices, or 20 unique segments. What is enough for most bus routes in test networks.

5 CONCLUSIONS

We expect that the interest in the inclusion of assisted trolleybuses in urban transport will increase. Therefore, complex ways of implementing this technology need to be researched. Our article explains required data conversion approach. Data are obtained from DPMŽ.

This article also introduces mathematical model required for creating a minimal network of overhead wires that would be sufficient for deployment and operation of such battery assisted trolleybuses. Performed computational tests demonstrate that model is correct. At the present stage, we are able to solve tasks to a certain size for individual bus routes.

The number of route's bus stops has significant impact on the solvability of the problem. This number would be inadequate when working with the vehicle schedule. Schedules tend to be longer, although the segments are often used multiple times. Therefore, we will try to combine following segments or cut the peripheral edges. It will also be possible to add rules to prohibit the construction of overhead contact lines on certain sections, such as historic centres, etc.

This data conversion process will need more research in the future. We will focus on developing optimization methods for data. In the future, we would like optimize not bus routes but rather individual vehicle turnovers in selected area. The maximum scope of input data will be further investigated in future. Heuristic approach options may need to be considered for large scalled tasks. Traffic simulation software can be usefud for this verification of results.

Acknowledgement
The authors would like to thank VEGA 1/0689/19 "Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities" and APVV-15-0179 "Reliability of emergency systems on infrastructure with uncertain functionality of critical elements"

References

[1] Bartłomiejczyk M.: Practical application of in motion charging: Trolley- buses service on bus lines, 2017 18th International Scientific Conference on Electric Power Engineering (EPE), Kouty and Desnou,

[2] Bartłomiejczyk M, Styskala V, Hrbac R, Połom M. (2013) Trolleybus with traction batteries for autonomous running. 7th International Scientific Symposium on Electrical Power Engineering (ELEKTROENERGETIKA), At Stara Lesna, Slovak Republic

[3] Bergk F., Biemann K., Lambrecht U., Putz R., Landinger H.: (2016) Potential of In-Motion Charging Buses for the Electrification of Urban Bus Lines Journal of Earth Sciences and Geotechnical Eng.

[4] Czimmermann, P. (2016). Location Problems in Transportation Networks. Communications - Scientific Letters of the University of Zilina, 18(3), 50-53.

[5] Gohlich D., Kunith A., Ly T., Technology assessment of an electric urban bus system for Berlin. 2014 WIT Transactions on the Built Environment. 138. 10.2495/UT140121.

[6] Grygar D., Kohani M., Stefun R., Drgona P.: Analysis of limiting factors of battery assisted trolleybuses, TRANSCOM 2019, High Tatras, Novy Smokovec, Slovak Republic, 2019.

[7] Hwang I., Jang Y., J., Ko Y., D., Lee M. S.: System Optimization for Dynamic Wireless Charging Electric Vehicles Operating in a Multiple-Route Environment. IEEE Transactions on Intelligent TS

[8] Chen Z., He F., Yin Y.: Optimal deployment of charging lanes for electric vehicles in transportation networks. Transportation Research Part B: Methodological, 91, 2016, 344-365,

[9] Jankovic P.: Agent-based Simulation for Creating Robust Plans and Schedules, 2015, Procedia Computer Science, Volume 51, Pages 895-904, ISSN 1877-0509.

[10] Jenkins J.: A closer look at energy consumption in EVs, 2018. CHARGED, Electric Vehicles Magazine, under Features, Tech Features.

[11] Kunith A., Mendelevitch R., Goehlich D.: Electrification of a city bus network - An optimization model for cost-effective placing of charging infrastructure and battery sizing of fast-charging electric bus systems. 2017. International Journal of Sustainable Transportation, 11:10, 707-720

[12] Montoya A., Gueret Ch., Mendoza J., Villegas J.: The electric vehicle routing problem with nonlinear charging function. Transportation Research Part B: Methodological. 2017.

[13] Rice D., Søderholm T., Mortensen O.: Teknisk Forprosjekt - Trolleybus Extension to Laksevag, 2017

[14] Rogge, M.; Wollny, S.; Sauer, D..: (2015) Fast Charging Battery Buses for the Electrification of Urban Public Transport - Feasibility Study Focusing on Charging Infrastructure and Energy Storage Requir. DOI: 10.3390/en8054587

[15] Ushijima-Mwesigwa H., Khan Z., Chowdhury M. A., Safro I.: Optimal Installation for Electric Vehicle Wireless Charging Lanes 2017, in Transportation Research Part B Methodological 91:344-365

[16] Yang S.C., Li M., Lin Y., Tang T.Q., Electric vehicle's electricity consumption on a road with different slope, 2014. Physica A: Statistical Mechanics and its Applications, 402, Pages 41-48.

RealForAll pollen semaphore: A short-term prediction system for airborne pollen concentrations based on Neural Nets*

Slobodan Jelić†

J. J. Strossmayer University of Osijek - Department of Mathematics, Trg Ljudevita Gaja 6, 31000 Osijek, Croatia

Abstract

In this paper we give a short description of a RealForAll pollen semaphore for the classification of categories of airborne pollen concentrations. This approach combines several binary classifiers based on neural networks which enable short term-predictions by using meteorological predictors and realized targets in past time units. A fine tuning with the grid-search of optimal hyper-parameters through the time-series cross-validation makes this approach acceptable for practical applications.

Keywords— neural network, multilayer perceptron, binary classifier, airborne pollen concentration

1 Introduction

A short-term prediction of daily airborne pollen concentrations becomes one of the most important factors in the public health system. Forecasting of these concentrations, along with real-time measurements, substantially improves life quality of persons who suffer from allergic rhinitis. Good forecasts, for a few days in advance, help people to plan their activities in order to avoid exposures to the high concentrations. One of the most challenging tasks is to develop a prediction system that gives useful and intuitive information to end-user about the levels of airborne concentrations that are expected in the following few days. This paper gives a short description of such a prediction system that produces semaphores for predefined intervals of pollen concentrations. We focus on three types of pollen: *Ambrosia*, *Betula* and *Poaceae*. Indeed, according to Bousquet et al., Betula and Poacea cause the majority of allergic reactions in Europe [2], while Ambrosia is the second most important allergen with the significantly increased clinical relevance [10]. It is important to emphasize that majority of results deal

*This research is supported by Interreg IPA CBC project: RealForAll - Real-time measurements and forecasting for successful prevention and management of seasonal allergies in Croatia-Serbia cross-border region, HR-RS151, 15.07.2017. - 14.01.2020.

†E-mail address: `sjelic@mathos.hr`

with a prediction of daily/hourly concentrations of airborne pollen. One of the earliest results of Start et. al. [11] employs Poisson regression for prediction of high daily ragweed pollen concentrations using rainfall, wind speed, temperature, and the time measured from the start of the season. Cotos-Yanez et al. [4] used meteorological variables for short-term prediction of birch with logistic additive and partially linear models. Matyasovszky et al. [7] proposed autoregressive models for daily ragweed pollen concentrations. Csepe at al. [5] proposed one of the first models that use Multi-layer Perceptron (MLP) regression for prediction of daily ragweed concentrations in Hungary and France. One of the simplest, and the most common approach in practice, is a long-term method that presumes a preparation of predictions based on means/medians of concentrations in previous years. This method is commonly known as a *pollen calendar*. A standard approach gives the prediction for the day in the season as a mean/median of concentrations on the same day in previous years that are available in the dataset. Spieksma's model introduces a preprocessing by 10-day means that remove short-term shifts during the pollen season. Sikoparija et al. [9] improves the standard and Spieksma's approach by introducing moving mean/median in the preprocessing phase, and calculating of means/medians on daily and 10-day resolutions. Very recent result by Zewdie et al. [12] proposes machine learning models for prediction of daily ragweed concentrations based on deep neural networks, random forest, extreme gradient boosting and Bayesian ridge regression. Čorić et al. [3] presents a clustering model for the time series forecasting that can be applied to the airborne pollen concentration forecasting. This model aims to find a set of patterns which represent similar situations in the time series. In order to predict target variable, different types of fitting methods can be applied to the set of data that belongs to the same pattern. These models give prediction of concentrations rather that semaphores. In this paper we present neural network model that relatively well predicts semaphores which are more interpretable to the end-user. The paper is organized as follows. In section 2 we describe datasets, predictors and preprocessing steps necessary for training phase. Third section gives an overview of machine learning models, optimization, model selection and implementation. In the last section we give some performance results about predictions on ragweed pollen concentration categories in Novi Sad between 2001 and 2017.

2 Data and predictors

The dataset contains daily average concentrations of three types of pollen: *Ambrosia*, *Betula* and *Poaceae* in period from 2001 to 2017 in Osijek and Novi Sad that are collected using the 7-day volumetric spore trap of the Hirst design. Concentrations are measured as the number of particles per cubic meter of air (P/m^3). Daily concentrations of all three types exhibit seasonality. The flowering period of Ambrosia is July–September, for Betula March–May and for Poaceae April–October. We shrunk our datasets to these periods in order to make quality measures more realistic since zero concentrations outside of these periods are easily predicted. Beside the daily pollen concentrations, we used the following meteorological data in predictor construction: minimum (MNT), maximum (MKT) and average daily temperatures (SRT), precipitation (PAD), relative humidity (VLZ) and maximum daily wind speed (MBV).

All data are differentiated on the daily basis, which means that prediction is performed by using the relative change in the meteorological variable with respect to the previous day. All predictors are standardized by mean and standard deviation in order to make training process numerically stable. Some relevant researches regarding pollen calendar [9] concluded that seasonal shifts are caused by meteorological conditions. In order to successfully model these shifts, we used meteorology and pollen concentration categories k days before the given day and l days after the given day. Real-time prediction system uses also l meteorological forecasts that are available on the day of the prediction. Some pollen types, such as Betula, have seasonal regularities on the yearly basis, which motivate us to include predictors from two previous seasons on the same days. In all experiments k and l are fixed to 10 and 5, respectively. Categories of pollen concentration can be arbitrarily defined. Finding the best predictable categories remains an interesting task in further research. In this paper we consider six categories: first category contains all daily samples with 0 or 1 P/m^3, second category contains samples between 1 and 4 P/m^3, third between 4 and 16 P/m^3, fourth between 16 and 31 P/m^3, fifth between 31 and 100 P/m^3 and sixth all samples with concentration larger than 100 P/m^3.

3 Prediction system

3.1 Model

Since we must predict six different categories, we developed prediction system that incorporates six binary classifiers, one for each category. Binary classifier is modelled with MLP with several hidden layers. Each layer contains predefined number of units that is determined through hyperparameter search procedure. Output layer contains two softmax units that transform output from last hidden layer to probabilities for 0 and 1, respectively. A target variable is treated as 1 if it is realized in a given category, and 0 otherwise. Each binary classifier outputs the probability that the target variable equals 1. If probability is larger than some cut-off value, daily sample is classified to that category. Since it is possible that one sample is classified to more than one category, we propose simple voting procedure that decides final category. The category with the highest ordinal number (the category of the highest pollen concentration) that received a vote from some of classifiers is selected as the final decision. Cut-off selection is performed by Receiver Operating Characteristics (ROC) analysis maximizing Youden's statistics [8]. Training process is performed by using Adam optimizer [6] with cyclic learning rate policy. Maximum number of training epochs is set to 500, while the model and optimizer parameters are saved in the training epoch with the smallest value of the criterion function on validation set. These parameters are loaded to the model and used for prediction on validation set. The criterion function is averaged weighted cross-entropy loss regularized with the ridge regularization term. Weights for the criterion function are defined to be inversely proportional to the category frequency in the training set. We also added lasso regularization in order to perform feature selection during the training process. Since our dataset contains more samples than predictors, lasso regularization did not take an effect on model quality and we excluded it from the final model. Ridge regularization is set to prevent over-fitting of the model in the case of large number of predictors.

3.2 Model selection and evaluation

In order to select the most appropriate architecture of MLP, we used a grid search procedure for hyper-parameter tuning. The number and size of hidden layers, together with ridge regularization parameter are validated on separate datasets. More precisely, we use a time series cross-validation technique [1] for time series in order to avoid biased estimates for model quality. The training dataset of k consecutive seasons is split to $k-1$ seasons that are used for training, while the k-th season is used for evaluating trained model on selected combination of hyper-parameters. The model defined by a combination of hyperparameters that minimized the objective function on the validation set is selected as the final model on k consecutive seasons. An evaluation of the selected model is performed on $k+1$-th season. Pollen semaphores should not underestimate real category of the pollen concentration since it seems to be more dangerous for the end-users who suffer from allergic rhinitis. In our case, the most reasonable accuracy measure of the model is a number of predictions where predicted level is at most two higher than the real one, divided by total number of cases. This accuracy measure is called *custom accuracy.*

3.3 Implementation

All experiments are driven on the Ubuntu 16.04.4 LTS machine with Intel Xeon E7-4850 v3 processor operating at stock frequency of 2.2 GHz, 8 GB of RAM, 16 cores. We used Python machine learning libraries such as Scikit-learn[1], PyTorch[2] and Skorch[3], along with Numpy, Scipy and Pandas for data manipulation. Skorch library is used for grid-search hyper-parameter selection on PyTorch model.

4 Performance of the prediction system

In this section we provide some experimental results about RealForAll pollen semaphore. Due to the limitations on the paper length, we present results about Ambrosia in Novi Sad in period from 2001 to 2017. Full set of results will be presented in the journal version of the paper.

Cross-validation results for binary classifiers and final RealForAll pollen semaphore are given in Table 1. These results are obtained in the evaluation phase on seasons from 2007 to 2017 on the selected models. Each row in the table corresponds to the given season that is treated as the test set. We run time series cross-validation with rolling origin in order to obtain unbiased estimates for Log Loss and custom accuracy for the final predictor. Low standard deviations confirm that classifiers are not sensitive to the dataset augmentation. Even more, the optimized loss is smaller as the length of the category is larger.

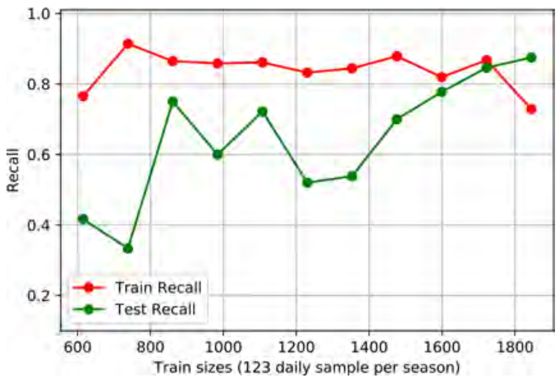
Year (Fold)	Logarithmic Loss						Custom accuracy
	BC-1	BC-2	BC-3	BC-4	BC-5	BC-6	Final Predictor
2007	0.44	0.52	0.43	0.26	0.39	0.29	**0.66**
2008	0.46	0.52	0.44	0.23	0.36	0.29	**0.78**
2009	0.45	0.52	0.41	0.20	0.36	0.28	**0.83**
2010	0.46	0.55	0.43	0.21	0.35	0.25	**0.80**
2011	0.44	0.53	0.41	0.21	0.32	0.28	**0.86**
2012	0.45	0.51	0.41	0.19	0.33	0.29	**0.81**
2013	0.46	0.51	0.42	0.21	0.34	0.27	**0.80**
2014	0.46	0.52	0.41	0.21	0.35	0.28	**0.90**
2015	0.44	0.51	0.41	0.21	0.35	0.29	**0.85**
2016	0.45	0.50	0.39	0.19	0.34	0.27	**0.84**
2017	0.45	0.51	0.41	0.19	0.33	0.27	**0.75**
Mean	**0.45**	**0.52**	**0.42**	**0.21**	**0.35**	**0.28**	0.81
Std.	**0.01**	**0.01**	**0.01**	**0.02**	**0.02**	**0.01**	0.06

Table 1: Cross-validation results on selected binary classifier and prediction system on ragweed pollen data for Novi Sad in period from 2001 to 2017. A BC-i, $i \in \{1, 2, 3, 4, 5, 6\}$ denotes the logarithmic loss of i-th binary classifier.

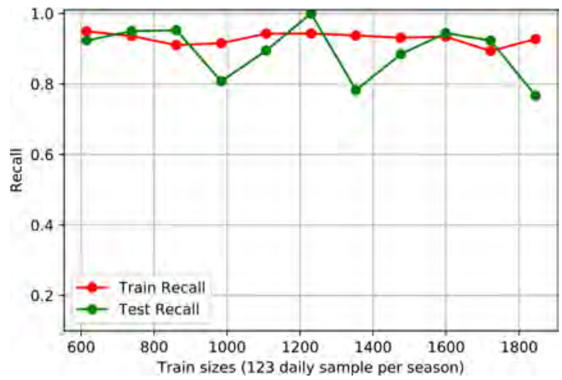
Categories with higher concentrations of pollen are better learnt and classified. In Figure 1a and 1b we observe the increase of the true positive rate on the test set as the size of

[1]https://scikit-learn.org/stable/
[2]https://pytorch.org/docs/stable/index.html
[3]https://skorch.readthedocs.io/en/stable/

training set increases. Fluctuations in the true positive rates on test set are consequence of unseen phenomena that are not learnt yet. Prediction of the fifth and sixth categories are the most important in RealForAll pollen semaphore because these categories alarm high pollen concentrations. Positive trends in learning curve for the ability to recognize these two categories are important for practical use of this prediction system.

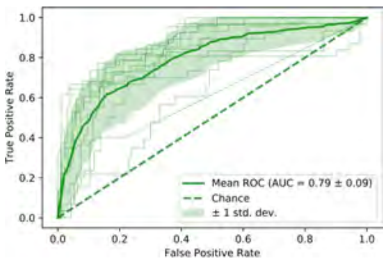


(a) Learning curve of binary classifier for category between 30 and 100 Pm^{-3}

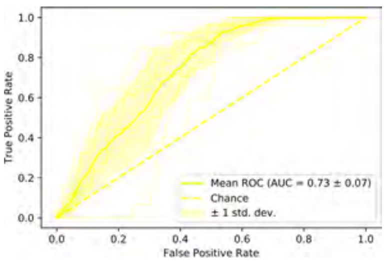


(b) Learning curve of binary classifier for category with more than 100 Pm^{-3}

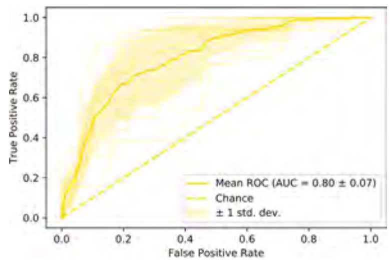
ROC curve analysis is a second measure that can reveal discriminator power of binary classifier. Cross-validation ROC curve (CV-ROC) for six binary classifiers are given in figures 1a-1f. We computed ROC curve for each cross-validation fold on test sets from table 1. Bold curve represents a linear spline on average values of true positive rates in each fold, while the shaded region around the bold curve denotes the standard deviation that measures sensitivity of the classifier to new unseen data. Area under the curve (AUC) confirms that last two categories perform best on unseen data.



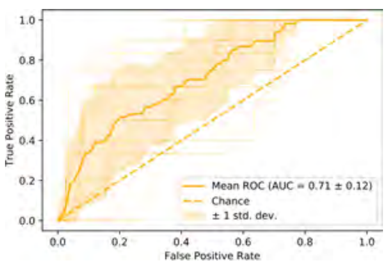
(a) CV-ROC for BC-1



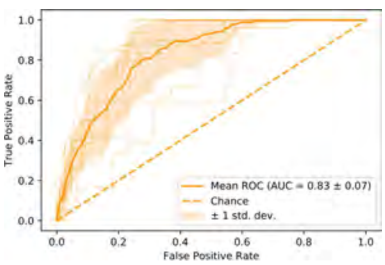
(b) CV-ROC for BC-2



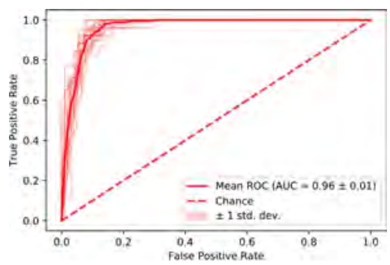
(c) CV-ROC for BC-3



(d) CV-ROC for BC-4



(e) CV-ROC for BC-5



(f) CV-ROC for BC-6

Conclusion

In this paper we give a short report about RealForAll pollen semaphore that is a part of the prediction system for airborne pollen concentration. The aim of this research is to reveal ability of simple machine learning models, such as MLP, to predict categories instead of concentrations. Most of the existing papers deal with the regression models that predict concentrations of airborne pollen in Pm^{-3}. Weakness of this approach is a prediction of low categories that do not differ, significantly. More extensive approach in hyper-parameter tuning would be to find categorization of pollen concentrations that can be the most successfully predicted.

References

[1] Christoph Bergmeir and José M. Benítez. "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191 (May 2012), pp. 192–213.

[2] P.-J. Bousquet et al. "Geographical variation in the prevalence of positive skin tests to environmental aeroallergens in the European Community Respiratory Health Survey I". In: *Allergy* 62.3 (Feb. 2007), pp. 301–309.

[3] Rebeka Čorić, Mateja Đumić, and Slobodan Jelić. "A clustering model for time-series forecasting". In: *MIPRO 2019 - 42th International Convention on Information and Communication Technology, Electronics and Microelectronics*. 2019, pp. 1295–1299.

[4] Tomas R. Cotos-Yáñez, F. J. Rodríguez-Rajo, and M. V. Jato. "Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models". In: *International Journal of Biometeorology* 48.4 (2004), pp. 179–185.

[5] Zoltán Csépe et al. "Predicting daily ragweed pollen concentrations using Computational Intelligence techniques over two heavily polluted areas in Europe". In: *Science of the Total Environment* 476-477 (2014), pp. 542–552.

[6] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (Dec. 2014), pp. 1–15. arXiv: 1412.6980.

[7] István Matyasovszky and L. Makra. "Autoregressive modelling of daily ragweed pollen concentrations for Szeged in Hungary". In: *Theoretical and Applied Climatology* 104.1-2 (2011), pp. 277–283.

[8] Enrique F. Schisterman et al. "Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples". In: *Epidemiology* 16.1 (Jan. 2005), pp. 73–81.

[9] B. Šikoparija et al. "How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae?" In: *Aerobiologia* (Jan. 2018).

[10] M. Smith et al. "Common ragweed: A threat to environmental health in Europe". In: *Environment International* 61 (Nov. 2013), pp. 115–126.

[11] P C Stark et al. "Using meterologic data to predict daily ragweed pollen levels". In: *Aerobiologia* 13 (1997), pp. 177–184.

[12] Gebreab K. Zewdie et al. "Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen". In: *International Journal of Environmental Research and Public Health* 16.11 (June 2019), p. 1992.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 5:

Mathematical Programming and Optimization

AN ARTIFICIAL-VARIABLE-FREE SIMPLEX METHOD INVOLVING THE CHOICES OF INITIAL SOLUTIONS

Aua-aree Boonperm
Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University Rangsit Center,
Pathumthani 12120, Thailand
E-mail: aua-aree@mathstat.sci.tu.ac.th

Wutiphol Sintunavarat
Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University Rangsit Center,
Pathumthani 12120, Thailand
E-mail: wutiphol@mathstat.sci.tu.ac.th

Abstract: The main goal of this research is to introduce the new technique for constructing the relaxed problem of the linear programming problem **(**LPP**)** when the basic solution is known**.** In the proposed relaxed problem, the constraints from the LPP in which the basic solution does not satisfy its conditions are relaxed and then the simplex method can be used for finding the optimal solution of the relaxed problem**.** After that, the relaxed constraints will be restored and so the simplex method can be performed for seeking the optimal solution of the LPP**.** Since the unsatisfied constraints are relaxed before the simplex method starts, the artificial variable is needless to introduce for finding the basic feasible solution**.** So, the proposed algorithm is the improvement of the simplex method without using the artificial variable. Furthermore, we also provide the example to illustrate the algorithm presented herein**.** Finally, the conclusions including the advantage and the limitation of the proposed algorithm in this work are described**.**

Keywords: linear programming problem, simplex method, artificial variable, relaxed problem.

1 INTRODUCTION

The mathematical optimization problem is concerned with the determination of the best element on some suitable set of available possibility. The one of branches in mathematical optimization problem is the linear programming problem (LPP) which is a method to achieve the best outcome in a mathematical model satisfying the linear relationships. It is well-known that the simplex algorithm presented by Dantzig [5] is quite the popular tool for solving a linear programming problem in nowadays. This algorithm is an iterative process that begins at a starting vertex in the feasible region of the problem, called initial basic feasible solution, and moves along the edges of the polytope to a neighbouring vertex until it achieves the optimal vertex at which the objective function is optimized. However, it is not easy in practice to seek an initial basic feasible solution or an initial basis. There are various existing methods in the literature for initializing the simplex algorithm. The most well-known methods for finding an initial feasible basis are the two-phase method and the big-M method. These methods require artificial variables to get the origin point as the initial basic feasible solution. The addition of artificial variables brings to increase the size of the problem and so the simplex method without artificial variables has evoked much interest to many researchers.

In 1997, Arsham [1] presented the new simplex method without using artificial variables consists of two phases. In this method, the use of artificial variables in Phase 1 can be omitted while the second phase uses exactly the classical simplex rules to reach optimality. One year later, the mistake in the proposed method of Arsham [1] is reported by Enge and Huhn [6]. Furthermore, they gave a counterexample claiming Arsham's algorithm declares the infeasibility of a feasible problem.

In 2006, Corley et.al [4] introduced the new simplex method without using artificial variables. In the first step of this method, the cosine criterion is used for choosing the suitable constraints from the LPP to construct the new relaxed problem and then it can be solved by the simplex method until the optimal solution of the relaxed problem is sought. In the next step, the relaxed constraints will be restored into the current tableau and so the dual simplex method will be performed until the optimal solution of the LPP is found. Although this algorithm does not require artificial variables, but it can be solved only the problems in which all coefficients are positive numbers.

In 2007, Arsham [2] proposed the new relaxed problem in the algorithm for solving the LPP without using the artificial variables. The greater-than or equal to constraints from the LPP is relaxed for making the relaxed problem in order to avoid the use of artificial variables. The relaxed constraints are restored whenever the optimal solution of the relaxed problem is found, and the simplex method is performed until the optimal solution of the LPP is achieved. In the case of the LPP contains only greater-than or equal to constraints, the perturbation simplex method will be used.

Seven years ago, Boonperm and Sinapiromsaran [3] first presented the non-acute constraint relaxation technique for improving the simplex method without using artificial variables. The proposed algorithm starts by relaxing the non-acute constraints. This yields that the relaxed problem is always feasible, and it can be solved without using artificial variables. The relaxed constraints are restored whenever the optimal solution of the relaxed problem is found, and the dual simplex method is performed for solving the optimal solution. One of the advantages of the non-acute constraint relaxation technique is the reducing start-up time to solve the initial relaxation problem. However, if the relaxed problem is unbounded, the proposed algorithm of Boonperm and Sinapiromsaran [3] is slow.

Recently, Prayonghom and Boonperm [9] presented the new idea for relaxing the variables from the LPP which is called the artificial-variable-free simplex method based on negative relaxation of dual. In this method, if the chosen initial basis gives a dual infeasible solution, then the primal variables which cause its dual infeasible are relaxed and then the dual simplex method can be used for seeking the primal feasible solution. Next, the relaxed variables will be restored and so the simplex method can be performed. They also gave the comparison of the average number of iterations and the CPU time from the proposed method with the two-phase method.

On the other hand, Junior and Lins [7] pointed out that the original simplex method should be started at the origin point which is far from the optimal solution for some problems. From this observation, they proposed the new basic feasible solution which forms a vertex that is much closer to the optimal vertex than the initial solution adopted by the original simplex. However, artificial variables are required in some cases of problems using this method.

Based on all above literatures, the main aim of this work is to present the new technique for constructing the relaxed problem of the LPP after the basic solution is known. This algorithm is the modification of the simplex method without using the artificial variable. In the proposed algorithm, the constraints in which the basic solution does not satisfy its conditions are relaxed and so the simplex method can be used for finding the optimal solution of the relaxed problem. After that, the relaxed constraints will be restored and then the simplex method can be performed. The illustrative example is furnished which demonstrate the validity and degree of utility of the proposed algorithm. In this case, the process of Junior and Lins [7] is considered for choosing the basic solution because it is closed to the optimal vertex.

The paper is organized as follows. Fundamental concepts used in this work are given in Section 2. In Section 3, our new proposed algorithm for solving the LPP without using artificial variables is presented. In Section 4, we present an illustrative example where our proposed algorithm is used. In the final section, we comprehend our conclusion.

2 FUNDAMENTAL CONCEPTS

Consider a linear programming problem in the standard form:

$$\begin{aligned} \text{maximize} \quad & z = \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{2.1}$$

where $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $\text{rank}(A) = m.$

Let $A = [A_{:1}, A_{:2}, \ldots, A_{:n}]$ where $A_{:j}$ is the j^{th} column of matrix A and A = [B, N] where $B \in \mathbb{R}^{m\times m}$ is a nonsingular basic matrix or basis, and $N \in \mathbb{R}^{m\times(n-m)}$ is a nonbasic matrix. Let I_B be an index set of the basic variables and I_N be an index set of the nonbasic variables.

For any basis B, the problem (*) can be written as follows:

$$\begin{aligned} \text{maximize} \quad & z + (\mathbf{c}_B^{\mathrm{T}}B^{-1}N - \mathbf{c}_N^{\mathrm{T}})\mathbf{x}_N = \mathbf{c}_B^{\mathrm{T}}B^{-1}\mathbf{b} \\ \text{subject to} \quad & \mathrm{I}\mathbf{x}_B + B^{-1}N\mathbf{x}_N = B^{-1}\mathbf{b}, \\ & \mathbf{x}_B \geq \mathbf{0}, \quad \mathbf{x}_N \geq \mathbf{0}. \end{aligned}$$

The initial tableau can be written as follows:

	z	$\mathbf{x}_B$	$\mathbf{x}_N$	RHS
z	1	0	$\mathbf{c}_B^{\mathrm{T}}B^{-1}N - \mathbf{c}_N^{\mathrm{T}}$	$\mathbf{c}_B^{\mathrm{T}}B^{-1}\mathbf{b}$
$\mathbf{x}_B$	0	I	$B^{-1}N$	$B^{-1}\mathbf{b}$

From the initial simplex tableau, it can indicate that a basis gives the primal or dual infeasible solution, that is, there exist some negative reduced cost or some negative right-hand-side value.

From the initial simplex tableau, we will use the following symbols:

$$R = \{j \mid z_j - c_j < 0, j \in I_N\} \text{ and } P = \{i \mid \bar{b}_i < 0, i = 1, \ldots, m\}, \tag{2.2}$$

where $z_j - c_j = \mathbf{c}_B^{\mathrm{T}}B^{-1}A_{:j} - c_j$, $\forall j \in I_N$, and $\bar{\mathbf{b}} = B^{-1}\mathbf{b}$.

If $P = \varnothing$, then $\mathbf{x} = \begin{bmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} B^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix}$ is a feasible solution of the problem (2.1).

If $R = \varnothing$, then $\mathbf{w}^{\mathrm{T}} = \mathbf{c}_B^{\mathrm{T}}B^{-1}$ is a feasible solution of the dual problem.

Therefore, there are four cases examined for finding the optimal solution as follows:

Case 1: If $R = \varnothing$ and $P = \varnothing$, then the primal and dual solutions are feasible. Therefore, we get the optimal solution.

Case 2: If $R = \varnothing$ and $P \neq \varnothing$, then the dual solution is feasible while the primal solution is infeasible. So, the dual simplex method can be performed for searching the optimal solution.

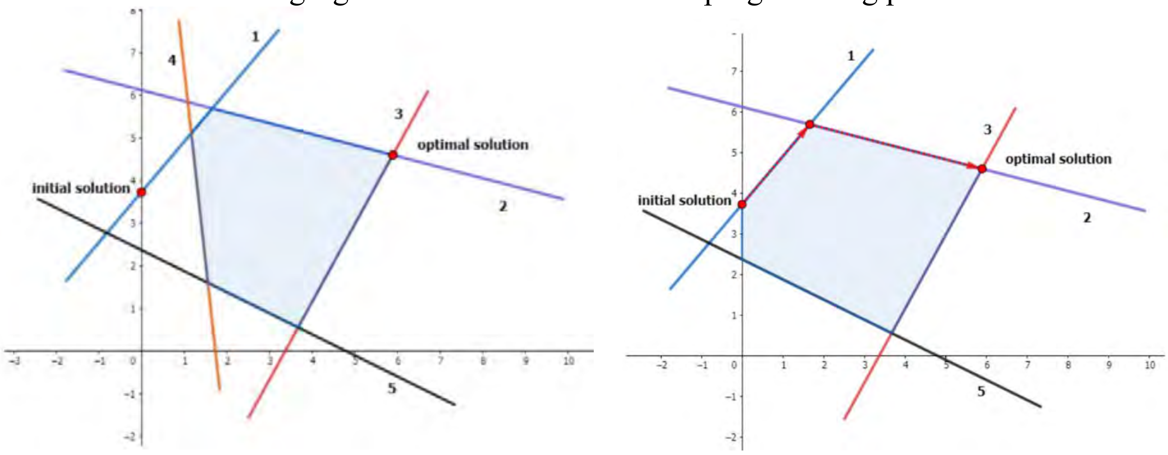
Case 3: If $R \neq \varnothing$ and $P = \varnothing$, then the dual solution is an infeasible solution while the primal solution is feasible. Therefore, the primal simplex method can be used to solve it.

Case 4: If $R \neq \varnothing$ and $P \neq \varnothing$, then both of them are infeasible solutions. In this case, neither the simplex method nor the dual simplex method could start.

For Case 4, the simplex method can start when artificial variables are added. So, the size of the problem will be expanded, and it may waste some computational time to deal with it. Moreover, the original simplex method starts at the origin point which is far from the optimal solution for some problems. In 2005, Junior and Lins [7] proposed the new initial basis which forms a vertex that is much closer to the optimal vertex than the initial solution adopted by the original simplex. However, Case 4 can still occur when the proposed basis is used, and artificial variables will be added. Therefore, we propose the improvement of the simplex method without using artificial variables and start from a vertex that is much closer to the optimal vertex than the original initial solution by constructing the relaxed problem.

Consider the problem when $R \neq \varnothing$ and $P \neq \varnothing$. Since P is a set of constraints that a primal solution is not satisfied, we will relax the constraints in P for making the primal feasible. Then, the simplex method can start.

Consider the following figure of 2-dimensional linear programming problem.



(a) (b)

Figure 2.1: Feasible region of the original problem (a) and the relaxed problem (b)

From Figure 2.1 (a), if we choose the initial solution as the marked point, we can see that the initial solution does not satisfied the constraint 4. Therefore, if the constraint 4 is relaxed, then the primal is feasible as Figure 2.1 (b) which the relaxed problem can be solved by starting at this point without using artificial variables.

3 THE PROPOSED ALGORITHM

For any basis B in (2.1) with R and P defined by (2.2), if $R = \varnothing$ or $P = \varnothing$, then the simplex method or the dual simplex method could start without using artificial variables. Therefore, we will consider the specific problem when $R \neq \varnothing$ and $P \neq \varnothing$. However, if $P = \{1, \ldots, m\}$, we could not relax all constraints. So, we will consider only the case $P \neq \{1, \ldots, m\}$.

The step of the algorithm can be summarized as follows:

Initial step: Choose the initial basis B and construct the initial simplex tableau and let

$$P = \{ i \mid \bar{b}_i < 0, i = 1, \ldots, m \} \neq \varnothing \text{ and } P \neq \{1, \ldots, m\}$$

where $\bar{\mathbf{b}} = B^{-1}\mathbf{b}$.

Step 1: Relax constraints in P and perform the simplex method.
If the optimal solution of the relaxed problem is found, then restore the constraints in P into the current tableau and go to Step 2.
Else restore the constraints in P into the current tableau and go to Step 3.

Step 2: If $\bar{\mathbf{b}} \geq \mathbf{0}$, then stop and the optimal solution is found.
Else perform the dual simplex method until the optimal solution is found.

Step 3: If $\bar{\mathbf{b}} \geq \mathbf{0}$, then performed the simplex method.
Else perform the perturbation simplex method proposed by Pan [8].

4 AN ILLUSTRATIVE EXAMPLE

The following example is illustrative the step by step of the proposed algorithm.

Example 4.1. Consider the following linear programming problem:

$$\begin{aligned} \text{maximize} \quad & z = 2x_1 + 2x_3 - 5x_4 \\ \text{subject to} \quad & x_1 + x_2 + x_3 + x_5 = 2 \\ & 2x_1 - x_2 + 3x_3 - 5x_4 - x_6 = 5 \\ & x_1 - 2x_2 - x_3 - x_7 = 6 \\ & 3x_1 - x_2 + 2x_3 - 5x_4 - x_8 = 4 \\ & x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8 \geq 0. \end{aligned}$$

First, the initial basis B will be chosen. From the proposed initial basis by Junior and Lin [7], we choose the index of basic feasible solution as $I_B = \{1,5,3,8\}$. Then, the initial tableau can be written below.

	z	x_1	x_5	x_3	x_8	x_2	x_6	x_7	x_4	RHS
z	1	0	0	0	0	-1.6	-0.8	-0.4	1	6.4
x_1	0	1	0	0	0	-1.4	-0.2	-0.6	-1	4.6
x_5	0	0	1	0	0	1.8	0.4	0.2	2	4.8
x_3	0	0	0	1	0	0.6	-0.2	0.4	-1	-1.4
x_8	0	0	0	0	1	-2	-1	-1	0	7

From the initial tableau, $R = \{2,6,7\}$ and $P = \{3\}$. So, the constraint 3 is relaxed, and the relaxed initial tableau can be rewritten as follows:

	z	x_1	x_5	x_3	x_8	x_2	x_6	x_7	x_4	RHS
z	1	0	0	0	0	-1.6	-0.8	-0.4	1	6.4
x_1	0	1	0	0	0	-1.4	-0.2	-0.6	-1	4.6
x_5	0	0	1	0	0	1.8	0.4	0.2	2	4.8
x_8	0	0	0	0	1	-2	-1	-1	0	7

Since this primal solution is feasible, the simplex method can start. After using 2 iterations to solve it, we get the optimal tableau of the relaxed problem as follows:

	z	x_1	x_5	x_3	x_8	x_2	x_6	x_7	x_4	RHS
z	1	0	2	0	0	2	0	0	5	16
x_1	0	1	0.5	0	0	-0.5	0	-0.5	0	7
x_6	0	0	2.5	0	0	4.5	1	0.5	5	12
x_8	0	0	2.5	0	1	2.5	0	-0.5	5	19

Then, the constraint 3 which is in P is restored to check the solution. After the constraint 3 is restored, we get the optimal tableau. Therefore, the proposed method uses 2 iterations to get the optimal solution without using artificial variables.

For this problem, the two-phase simplex method uses 6 iterations with three added artificial variables.

Next, the number of iterations and size of matrix are compared with the two-phase simplex method are shown as below.

	The proposed method		Two-Phase method	
	Relaxed Problem	Restored Problem	Phase I	Phase II
Number of iterations	2	0	3	3
Size of matrix	3×8	4×8	4×11	4×8

From the above table, we found that the proposed method can reduce the number of iterations. Additionally, the matrix size solved by our method is smaller than the matrix size solved by two-phase simplex method.

5 CONCLUSIONS

The algorithm for searching the solution of the LPP via the new technique for constructing the relaxed problem for the LPP have been presented. This algorithm is one of the improvements of the simplex method without using the artificial variable. In this algorithm, we start the simplex method with only constraints including the basic solution. After the optimal solution of the relaxed is found, the relaxed constraints will be restored and so the simplex method can be used for finding the optimal solution of the LPP. The proposed algorithm frequently reduces computational effort in the execution**.** In addition, the proposed algorithm has simplicity, potential for wide adaptation**.** It points out that the process of Junior and Lins [7] is considered for choosing the basic solution of the proposed algorithm in the illustrative example in Section 4**.** The efficacy of the proposed algorithm is increasing if we use the process for choosing the basic solution which is better than the process of Junior and Lins [7]. The limitation of the proposed algorithm concerns with the relaxing constraints from the LPP. If the basic solusion does not satisfy all constraints from the LPP, the proposed algorithm can not be used in this case**.**

References

[1] Arsham, H. 1997. An artificial-free simplex-type algorithm for general LP problem. Mathematical and Computer Modelling, 25: 107–123.

[2] Arsham, H. 2007. A computationally stable solution algorithm for linear programs. Applied Mathematics and Computation, 188: 1549–1561.

[3] Boonperm, A., Sinapiromsaran, K. 2014. Artificial–free simplex algorithm based on the non-acute constraint relaxation. Applied Mathematics and Computation, 234: 385–401.

[4] Corley, H. W., Rosenberger, J., Yeh, W.-C., Sung, T.K. 2006. The cosine simplex algorithm. The International Journal of Advanced Manufacturing Technology, 27: 1047–1050.

[5] Dantzig, G.B. 1963. Linear programming and extensions. New Jersey: Princeton University Press.

[6] Enge, A., Huhn, P. 1998. A counterexample to H. Arsham Initialization of the simplex algorithm: an artificial-free approach, SIAM Review, 40 1–6.

[7] Junior, H. V., Lins, M. P. E. 2005. An improved initial basis for the simplex algorithm. Computers & Operations Research 32: 1983–1993.

[8] Pan, P. Q. 2000. Primal perturbation simplex algorithms for linear programming. Journal of Computational Mathematics, 18: 587–596.

[9] Prayonghom, C., Boonperm, A. 2017. An artificial-variable-free simplex method based on negative relaxation of dual problem (Master' s thesis). Thammasat University.

INFEASIBLE INTERIOR-POINT ALGORITHM FOR LINEAR OPTIMIZATION BASED ON A NEW SEARCH DIRECTION

Zsolt Darvay[a]
[a]Babeş-Bolyai University,
Faculty of Mathematics and Computer Science, Cluj-Napoca, Romania
Petra Renáta Rigó[ab]
[b]Budapest University of Technology and Economics,
Department of Differential Equations, Hungary and
[a]Babeş-Bolyai University,
Faculty of Mathematics and Computer Science, Cluj-Napoca, Romania
Eszter Szénási[b]
[b]Budapest University of Technology and Economics,
Department of Differential Equations, Hungary

Abstract: We study the technique of algebraic equivalent transformation (AET) of the central path in order to obtain new search directions of interior-point algorithms (IPAs) for solving linear optimization (LO) problems. Using this well-known method, we present a new IPA with polynomial iteration complexity. We discuss the possibility of analyzing the case for infeasible starting points as well.

Keywords: interior-point algorithm, search direction, algebraic equivalent transformation, infeasible starting point

1 INTRODUCTION

Linear optimization problems are widely used in industry and economics to solve several specific problems. One of the most popular algorithms is the simplex method proposed by Dantzig [1] in 1947. However, IPAs have in general polynomial iteration complexity, and therefore one can expect that these algorithms will perform better in practice as well. A comparison of pivot and interior-point algorithms was published by Illés and Terlaky [5]. The most important results on IPAs were summarized in the monographs of Roos, Terlaky and Vial [10], Wright [12] and Ye [13].

1.1 Linear optimization problem

Suppose that $A \in \mathbb{R}^{m\times n}$ is a matrix with $rank(A) = m$ and $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^n$ are given vectors. The linear optimization problem, and its dual can be formulated in the following way:

$$\min\{\mathbf{c}^T\mathbf{x} \mid A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}\}, \tag{1}$$

$$\max\{\mathbf{b}^T\mathbf{y} \mid A^T\mathbf{y} + \mathbf{s} = \mathbf{c}, \quad \mathbf{s} \geq \mathbf{0}\}. \tag{2}$$

A wide class of IPAs was analyzed assuming that there are strictly feasible starting points for both problems, i.e. the *interior-point condition* holds. In the first part of this paper we make this assumption as well, thus we suppose that there is a triple $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ so that

$$\begin{aligned} A\mathbf{x}^0 &= \mathbf{b}, \quad \mathbf{x}^0 > \mathbf{0}, \\ A^T\mathbf{y}^0 + \mathbf{s}^0 &= \mathbf{c}, \quad \mathbf{s}^0 > \mathbf{0}. \end{aligned} \tag{3}$$

In general, primal-dual IPAs follow the so-called *central path*, which can be described by the following parameterized equation:

$$\begin{aligned} A\mathbf{x} &= \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}, \\ A^T\mathbf{y} + \mathbf{s} &= \mathbf{c}, \quad \mathbf{s} \geq \mathbf{0}, \\ \mathbf{x}\mathbf{s} &= \mu\mathbf{e}, \end{aligned} \tag{4}$$

where $\mu > 0$ is fixed, and $\mathbf{e}$ is the n-dimensional all-one vector. Note that if (3) holds, then for each positive real number μ, system (4) has unique solution that is called the *analytic center* of the polyhedron (Sonnevend [11]). The optimality criteria is expressed by system (4), with $\mu = 0$. Therefore, if μ tends to zero, then the central path converges to the optimal solution of the primal-dual pair.

1.2 Transformation of the central path

The AET technique was first presented by Darvay [2, 3] for solving LO problems. Suppose that $0 \leq \kappa < 1$ and let us consider the continously differentiable and invertible function $\varphi : (\kappa, +\infty) \to \mathbb{R}$. Observe that if the inequality $\frac{\mathbf{x}\mathbf{s}}{\mu} > \kappa\mathbf{e}$ holds, then system (4) can be transformed in the following equivalent form:

$$\begin{aligned} A\mathbf{x} &= \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}, \\ A^T\mathbf{y} + \mathbf{s} &= \mathbf{c}, \quad \mathbf{s} \geq \mathbf{0}, \\ \varphi\left(\frac{\mathbf{x}\mathbf{s}}{\mu}\right) &= \varphi(\mathbf{e}). \end{aligned} \tag{5}$$

Applying Newton's method leads to the following linear system:

$$\begin{aligned} A\Delta\mathbf{x} &= \mathbf{0}, \\ A^T\Delta\mathbf{y} + \Delta\mathbf{s} &= \mathbf{0}, \\ \mathbf{s}\Delta\mathbf{x} + \mathbf{x}\Delta\mathbf{s} &= \mu\frac{\varphi(\mathbf{e}) - \varphi\left(\frac{\mathbf{x}\mathbf{s}}{\mu}\right)}{\varphi'\left(\frac{\mathbf{x}\mathbf{s}}{\mu}\right)}. \end{aligned} \tag{6}$$

In order to obtain a scaled version of this system, we introduce the notations:

$$\mathbf{v} = \sqrt{\frac{\mathbf{x}\mathbf{s}}{\mu}}, \qquad \mathbf{d}_{\mathbf{x}} = \frac{\mathbf{v}\Delta\mathbf{x}}{\mathbf{x}}, \qquad \mathbf{d}_{\mathbf{s}} = \frac{\mathbf{v}\Delta\mathbf{s}}{\mathbf{s}}. \tag{7}$$

Using these, system (6) can be written in the following scaled form:

$$\begin{aligned} \bar{A}\mathbf{d}_{\mathbf{x}} &= \mathbf{0}, \\ \bar{A}^T\Delta\mathbf{y} + \mathbf{d}_{\mathbf{s}} &= \mathbf{0}, \\ \mathbf{d}_{\mathbf{x}} + \mathbf{d}_{\mathbf{s}} &= \mathbf{p}_{\mathbf{v}}, \end{aligned} \tag{8}$$

where

$$\mathbf{p}_{\mathbf{v}} = \frac{\varphi(\mathbf{e}) - \varphi(\mathbf{v}^2)}{\mathbf{v}\varphi'(\mathbf{v}^2)}, \tag{9}$$

and $\bar{A} = \frac{1}{\mu} A\, diag\left(\frac{\mathbf{x}}{\mathbf{v}}\right)$. Note that for different functions φ the value of the right-hand side vector $\mathbf{p_v}$ changes, therefore this leads to new search directions, and as a consequence we obtain new IPAs.

For example, for $\varphi(t) = t$ we obtain the most widely discussed direction $\mathbf{p_v} = \mathbf{v}^{-1} - \mathbf{v}$ (Roos, Terlaky and Vial [10]) and for $\varphi(t) = \sqrt{t}$ we obtain $\mathbf{p_v} = 2(\mathbf{e} - \mathbf{v})$, which yields to Darvay's direction [2, 3]. Recently, Kheirfam and Haghighi [6] considered the function $\varphi(t) = \frac{\sqrt{t}}{2(1+\sqrt{t})}$ for sufficient linear complementarity problems, which produces $\mathbf{p_v} = \mathbf{e} - \mathbf{v}^2$. Pirhaji, Zangiabadi and Mansouri [7] applied the same function for linear complementarity problems over circular cones. Rigó and Szénási [8] analyzed the method for LO.

The novelty of this paper consists in the following. We consider for the first time a new function φ, namely $\varphi(t) = \frac{1}{1+\sqrt{t}}$, in order to use it in the process of applying the AET method. We prove that this new function leads to the same Newton direction as the one obtained in [8]. We also consider the case of infeasible starting points and we present a new infeasible IPA for LO. This algorithm performs two inner iterations, called feasibility step and centering step, in each main iteration. In our algorithm both steps are obtained by appying the AET method together with the newly introduced function φ. Polynomial complexity can be proved for the infeasible algorithm as well. This is the first infeasible IPA for solving LO problems which determines the search direction by using the function $\varphi(t) = \frac{1}{1+\sqrt{t}}$ in the AET technique.

2 THE FEASIBLE ALGORITHM

Let us define the function $\varphi : (0, +\infty) \to \mathbb{R}$ as

$$\varphi(t) = \frac{1}{1+\sqrt{t}}. \tag{10}$$

Observe that we have $\varphi'(t) = -\frac{1}{2\sqrt{t}(1+\sqrt{t})^2}$. Therefore, a simple calculus yields

$$\mathbf{p_v} = \mathbf{e} - \mathbf{v}^2, \tag{11}$$

and thus system (8) can be written as:

$$\begin{aligned} \bar{A}\mathbf{d_x} &= \mathbf{0}, \\ \bar{A}^T \Delta\mathbf{y} + \mathbf{d_s} &= \mathbf{0}, \\ \mathbf{d_x} + \mathbf{d_s} &= \mathbf{e} - \mathbf{v}^2. \end{aligned} \tag{12}$$

Moreover, system (6) can be written in the form:

$$\begin{aligned} A\Delta\mathbf{x} &= \mathbf{0}, \\ A^T \Delta\mathbf{y} + \Delta\mathbf{s} &= \mathbf{0}, \\ \mathbf{s}\Delta\mathbf{x} + \mathbf{x}\Delta\mathbf{s} &= \sqrt{\frac{\mathbf{xs}}{\mu}}\left(\mu\mathbf{e} - \mathbf{xs}\right). \end{aligned} \tag{13}$$

A proximity measure to the central path can be defined in the following way:

$$\delta(\mathbf{x}, \mathbf{s}; \mu) = \|\mathbf{p_v}\| = \|\mathbf{e} - \mathbf{v}^2\|. \tag{14}$$

Now, a generic primal-dual algorithm can be summarized as in Fig. 1. Rigó and Szénási [8] proved the following theorem.

Primal-dual IPA for LO using a new search direction

Let $\epsilon > 0$ be the accuracy parameter, $0 < \theta < 1$ the update parameter and $\tau > 0$ the proximity parameter. Assume that for $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0)$ the IPC holds and $\mu^0 = \frac{\mathbf{x}^{0T}\mathbf{s}^0}{n}$.
Moreover, suppose that $\delta(\mathbf{x}^0, \mathbf{s}^0; \mu^0) < \tau$.

begin
  $\mathbf{x} := \mathbf{x}^0; \quad \mathbf{y} := \mathbf{y}^0; \quad \mathbf{s} := \mathbf{s}^0; \quad \mu := \mu^0;$
  while $\mathbf{x}^T\mathbf{s} > \epsilon$ **do begin**
    calculate $(\Delta\mathbf{x}, \Delta\mathbf{y}, \Delta\mathbf{s})$ *from (13);*
    $\mathbf{x} := \mathbf{x} + \boldsymbol{\Delta}\mathbf{x};$
    $\mathbf{y} := \mathbf{y} + \boldsymbol{\Delta}\mathbf{y};$
    $\mathbf{s} := \mathbf{s} + \boldsymbol{\Delta}\mathbf{s};$
    $\mu := (1-\theta)\mu;$
  end
end.

Figure 1: The feasible algorithm

Theorem 2.1. *Let $\mathbf{x}^0 = \mathbf{s}^0 = \mathbf{e}$. If $\theta = \frac{1}{5\sqrt{n}}$ and $\tau = \frac{2}{3}$, then the algorithm given in Fig. 1 requires no more than*

$$\mathcal{O}\left(\sqrt{n} \cdot \log \frac{n}{\varepsilon}\right)$$

interior-point iterations. The resulting vectors satisfy $\mathbf{x}^T\mathbf{s} < \varepsilon$.

3 INFEASIBLE STARTING POINTS

In 2006, Roos [9] proposed an infeasible algorithm for LO with full-Newton steps. Darvay, Papp and Takács [4] presented the first infeasible interior-point algorithm with full-Newton steps and one centering step in a major iteration. They applied the AET technique with $\varphi(t) = \sqrt{t}$. Now we consider the case when the function φ is defined as in (10). In this section we don't suppose that (3) is satisfied, but we assume that the primal-dual pair has optimal solutions. These will be denoted as $\bar{\mathbf{x}}$ and $(\bar{\mathbf{y}}, \bar{\mathbf{s}})$. Let $\zeta > 0$ be given so that

$$\|\bar{\mathbf{x}} + \bar{\mathbf{s}}\|_\infty \leq \zeta. \tag{15}$$

Then, the starting points of the infeasible algorithm can be defined as:

$$\mathbf{x}^0 = \zeta\mathbf{e}, \quad \mathbf{y}^0 = \mathbf{0}, \quad \mathbf{s}^0 = \zeta\mathbf{e}, \quad \mu^0 = \zeta^2. \tag{16}$$

Moreover, let us introduce the following notations:

$$\mathbf{r}_b^0 = \mathbf{b} - A\mathbf{x}^0, \quad \mathbf{r}_c^0 = \mathbf{c} - A^T\mathbf{y}^0 - \mathbf{s}^0.$$

Suppose that $\nu = \frac{\mu}{\mu^0}$ in each iteration of the algorithm. Then, using our new search direction, the feasibility step will be defined by the following system:

$$
\begin{aligned}
A\Delta^f \mathbf{x} &= \theta\nu \mathbf{r}_b^0, \\
A^T\Delta^f \mathbf{y} + \Delta^f \mathbf{s} &= \theta\nu \mathbf{r}_c^0, \\
\mathbf{s}\Delta^f \mathbf{x} + \mathbf{x}\Delta^f \mathbf{s} &= \sqrt{\frac{\mathbf{xs}}{\mu}}\left(\mu\mathbf{e} - \mathbf{xs}\right),
\end{aligned} \tag{17}
$$

where $0 < \theta < 1$ is the update parameter. Using this system, the algorithm can be presented as in Fig. 2. Let $\epsilon > 0$ be given. The triple $(\mathbf{x}, \mathbf{y}, \mathbf{s})$ is called an ϵ-solution if the inequality

---

Infeasible IPA with new search direction

Let $\epsilon > 0$ be the accuracy parameter and $0 < \theta < 1$ (default value $\theta = \frac{1}{8n}$) the update parameter. Moreover, assume that the starting points $(\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0)$ are defined as in (16).

begin
  $(\mathbf{x}, \mathbf{y}, \mathbf{s}) := (\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0)$;
  $\mu := \mu^0$; $\nu := 1$;
  while $max(\mathbf{x}^T\mathbf{s}, \|\mathbf{b} - A\mathbf{x}\|, \|\mathbf{c} - A^T\mathbf{y} - \mathbf{s}\|) \geq \epsilon$ **do begin**
    Calculate $(\Delta^f\mathbf{x}, \Delta^f\mathbf{y}, \Delta^f\mathbf{s})$ *from (17);*
    $(\mathbf{x}, \mathbf{y}, \mathbf{s}) := (\mathbf{x}, \mathbf{y}, \mathbf{s}) + (\Delta^f\mathbf{x}, \Delta^f\mathbf{y}, \Delta^f\mathbf{s})$;
    $\mu := (1 - \theta)\mu$;
    $\nu := (1 - \theta)\nu$;
    Calculate $(\Delta\mathbf{x}, \Delta\mathbf{y}, \Delta\mathbf{s})$ *from (13);*
    $(\mathbf{x}, \mathbf{y}, \mathbf{s}) := (\mathbf{x}, \mathbf{y}, \mathbf{s}) + (\Delta\mathbf{x}, \Delta\mathbf{y}, \Delta\mathbf{s})$;
  end
end.

Figure 2: The infeasible algorithm

$max(\mathbf{x}^T\mathbf{s}, \|\mathbf{b} - A\mathbf{x}\|, \|\mathbf{c} - A^T\mathbf{y} - \mathbf{s}\|) < \epsilon$ holds. Then, the following theorem can be proved.

Theorem 3.1. *If the primal and dual problems are both feasible and $\zeta > 0$ is defined as in* (15)*, then after at most*

$$\left\lceil 16n \log \frac{max\{n\zeta^2, \|\mathbf{r}_b^0\|, \|\mathbf{r}_c^0\|\}}{\epsilon} \right\rceil$$

interior-point iterations the algorithm finds an ϵ-solution of (1) *and* (2).

4 CONCLUSION

In this paper we presented a feasible and an infeasible IPA for LO based on a new search direction obtained by using the AET technique with the function given in (10). Further research may prove that these methods are efficient in practice as well.

5 Acknowledgements

The research of P.R. Rigó has been partially supported by the Hungarian Research Fund, OTKA (grant no. NKFIH 125700) and by the Higher Education Excellence Program of the

Ministry of Human Capacities in the frame of Artificial Intelligence research area of Budapest University of Technology and Economics (BME FIKP-MI/FM).

The research of Zs. Darvay and P.R. Rigó was supported by a grant of Romanian Ministry of Research and Innovation, CNCS - UEFISCDI, project number PN-III-P4-ID-PCE-2016-0190, within PNCDI III.

References

[1] G.B. Dantzig. *Linear Programming and Extension.* Princeton University Press, Princeton, NJ, 1963.

[2] Zs. Darvay. A new algorithm for solving self-dual linear optimization problems. *Studia Univ. Babeş-Bolyai, Ser. Informatica*, 47(1):15–26, 2002.

[3] Zs. Darvay. New interior point algorithms in linear programming. *Adv. Model. Optim.*, 5(1):51–92, 2003.

[4] Zs. Darvay, I.-M. Papp, and P.-R. Takács. An infeasible full-Newton step algorithm for linear optimization with one centering step in major iteration. *Studia Univ. Babeş-Bolyai, Ser. Informatica*, 59(1):28–45, 2014.

[5] T. Illés and T. Terlaky. Pivot versus interior point methods: Pros and Cons. *Eur. J. Oper. Res.*, 140:6–26, 2002.

[6] B. Kheirfam and M. Haghighi. A full-newton step feasible interior-point algorithm for $P_*(\kappa)$-LCP based on a new search direction. *CRORR*, 7(2), 2016.

[7] M. Pirhaji, M. Zangiabadi, and H. Mansouri. A path following interior-point method for linear complementarity problems over circular cones. *Japan Journal of Industrial and Applied Mathematics*, 35(3):1103–1121, Nov 2018.

[8] P.R. Rigó and Eszter Szénási. Interior-point algorithm for linear optimization based on a new search direction. Technical Report Operations Research Report 2019-01, Eötvös Loránd University of Sciences, Budapest, 2019.

[9] C. Roos. A full-Newton step O(n) infeasible interior-point algorithm for linear optimization. *SIAM J. Optim.*, 16(4):1110–1136, 2006.

[10] C. Roos, T. Terlaky, and J.-Ph. Vial. *Theory and Algorithms for Linear Optimization.* Springer, New York, USA, 2005.

[11] Gy. Sonnevend. An "analytic center" for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming. In A. Prékopa, J. Szelezsán, and B. Strazicky, editors, *System Modelling and Optimization: Proceedings of the 12th IFIP-Conference held in Budapest, Hungary, September 1985*, volume 84 of *Lecture Notes in Control and Information Sciences*, pages 866–876. Springer Verlag, Berlin, West-Germany, 1986.

[12] S.J. Wright. *Primal-Dual Interior-Point Methods.* SIAM, Philadelphia, USA, 1997.

[13] Y. Ye. *Interior Point Algorithms, Theory and Analysis.* John Wiley & Sons, Chichester, UK, 1997.

A TABU SEARCH METHOD FOR OPTIMIZING HETEROGENEOUS STRUCTURAL FRAMES

Balázs Dávid
InnoRenew CoE
Livade 6, 6310 Izola, Slovenia
E-mail: balazs.david@innorenew.eu
University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies
Glagoljaška ulica 8, 6000 Koper, Slovenia
E-mail: balazs.david@famnit.upr.si

Abstract: Structural design is responsible for the design and development of structural plans. It consists of multiple steps that have to be performed sequentially. As the output of a step acts as an input for the next one, using optimization tools to aid the design process can help in achieving a better overall quality or a particular design. In this paper, we present a heuristic method that can provide useful suggestions in a short time for the structural design of a building based on preliminary plans. We also introduce a mathematical model that can be used as quality control of the results.

Keywords: structural frames, structural design, tabu search, mathematical model

1 INTRODUCTION

The design and construction of a building is a cost intensive process regarding both operational and capital costs and environmental impacts. The structural design process consists of several stages that have to be performed sequentially: planning, design and detailing. Even separately, these are all complex problems, thus efficient optimization tools can help civil engineers with their decision-making process in order to create cost efficient and sustainable building designs. The *planning* phase is responsible for the development of the initial layout of the structure, positioning and orienting all the required element types within the frame itself. The result of this step is a preliminary design of the building, which contains the positions of all elements, but usually does not specify their important characteristics. The field dealing with this phase is called layout/topology optimization. One of the first introductions of the problem was by Dorne et al. [7] in 1964, and soon more exact approaches followed for different types of structural topologies (see Bendsøe and Kikuchi [3] and Kirsch [9] as some examples). Several different aspects of the problem have been considered over the years. Different types of solution methods have been proposed, like metaheuristics (such as the ant colony optimization of Camp [4]), genetic algorithm (e.g. Deb [6]), or an optimization framework that considers instabilities (Changizi and Jalalpour [5]). For a thorough review of this field, see Rozvany [12].

The outcome of the planning phase is a preliminary design, that is still missing many important structural properties of its elements (e.g exact shape/size, material used). The *design* phase uses this plan as an input, and creates a structural plan that can be passed on the *detailing* phase for the preparation of the construction schedule. Because of this, the goal of the design phase is to prepare a plan that optimizes all arising costs (be it capital, operational or environmental). Similar solution approaches exist for this phase as for *planning*, and the two are often considered together. Again, different genetic algorithms were proposed (for an example, see Baumann and Kost [2] or Fedelinski and Gorski [8]), and metaheuristics are also used (the simulated annealing approach of Lamberti and Pappalettere [10]). For detailed reviews of both the design phase and the combination of planning and design, refer to Lamberti and Pappalettere [11] or Azad and Hasançebi [1].

In this paper, we propose a Tabu search heuristic that can provide quick suggestions for a structure design based on preliminary plans, specifying the material and cross section of the elements in the structure to achieve a low-cost solution. As the method provides solutions in a short time, the results of this algorithm can be used by civil engineers to help with their decision-making processes during the structural design process. A mathematical model is also developed for the same problem so that we can monitor the quality of the heuristic.

First, we introduce the problem itself, and formalize it as a mathematical model. We then present the Tabu search heuristic for its solution, and present preliminary test results on small input instances.

2 PROBLEM DEFINITION

A structural frame can be regarded as the skeleton of a modern building. It consists of separate levels, and each of these can typically contain three different element types: slabs, columns and beams.

Slabs are the plate elements of the structure, which are usually used as the base, roof or ceiling. Beams are horizontal elements that support the slabs, while columns are vertical elements that support the beams and other columns above them. A simplified representation of a structural frame can be seen in Figure 1.

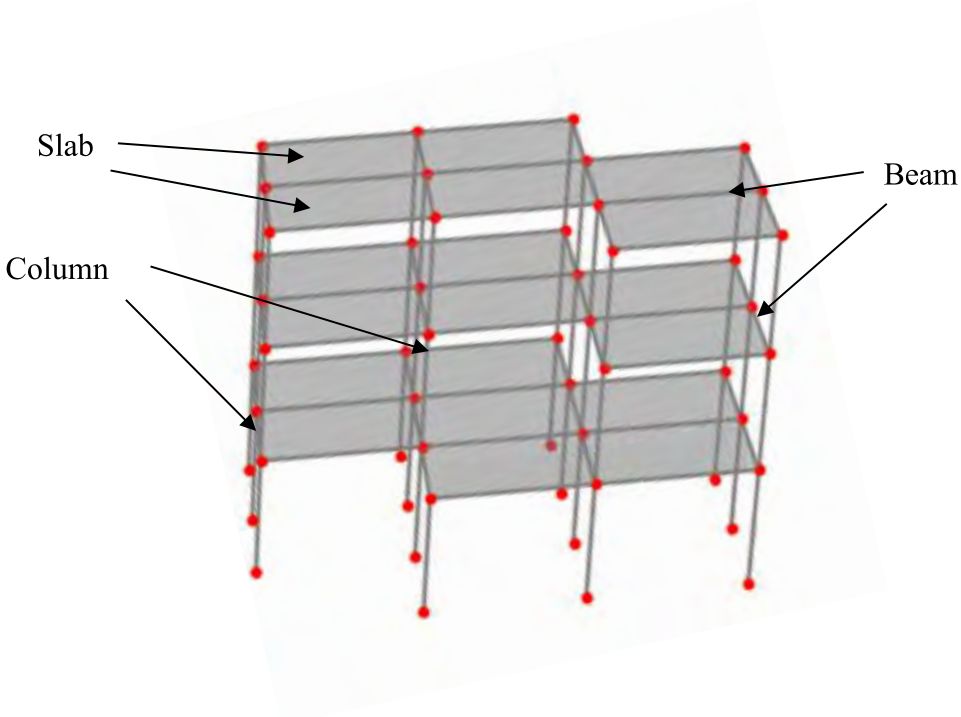


Figure 1: An abstract example of a structural frame

Different materials can be assigned to different elements. However, every element of this frame is affected by certain structural forces, and their cross-section has to be chosen accordingly when the desired material for the element is selected. Another important requirement is that certain elements have to be identical, meaning that they both have to share the same material and have the same cross-section. Such elements will be referred to as '*belonging to the same group*'.

Given all required information about the elements of the structure, and a preliminary conceptual design, we would like to use them to create a design plan that minimizes all arising costs.

In the next section, first we will formally define the above problem, and introduce a mathematical model based on the presented constraints.

2.2 Mathematical model

Let E be the set of elements in our frame and let set G denote the different groups that these elements can belong to. For each $e \in E$, let $g(e) \in G$ be the group of that element. Let M be the set of the available building materials.

The mathematical model of the problem can be formalized the following way:

$$minimize \sum_{e \in E} \sum_{j=1}^{|M|} c_{ej}\, x_{ej}$$

$s.t.$

$$\sum_{j=1}^{|M|} x_{ej} = 1, \forall e \in E \tag{1}$$

$$x_{im} - x_{jm} = 0, \forall i \in E, j \in g(i), 1 \le m \le |M| \tag{2}$$

$$y_{im} = max\{s_{jm}x_{jm} \mid j \in g(i), 1 \le m \le |M|\}, \forall i \in E \tag{3}$$

$$x_{im} \in \{0,1\}, \forall i \in E, 1 \le m \le |M| \tag{4}$$

$$y_i > 0, \forall i \in E \tag{5}$$

The binary variable x_{em} represents the decision of using material m for element e. Variable y_e gives the cross-section of element e in the final plan. Constraint (1) specifies that any given element will have exactly one material. Constraint (2) ensures that elements belonging to the same group will have the same material. Constraint (3) selects the proper cross-section of every element, making sure that elements belonging to the same group should have the same cross-section. Constraints (4) and (5) are the binary and nonnegativity constraints for the two variables respectively. The objective of the model is to minimize the arising costs of the elements in the structure. The cost largely depends on the type of material chosen for the element (the value of x_{em}), an also its cross section. The cross section is included in the objective through the cost coefficients, as they are a factor of the cross section: $c_{em} = \alpha_m y_e$. This cost coefficient includes both the capital and operational costs of the building, as well as its environmental impact.

The above model itself is non-linear because of constraint (3) and the objective. However, since they include the binary variables x_{em}, they both can be linearized easily with the introduction of extra variables.

2.3 A Tabu search heuristic

We developed a Tabu search algorithm for the solution of the problem. This heuristic was chosen for the solution of the problem, as it is able to produce multiple good quality solutions with a short running time. Having multiple solutions can help with the decision-making processes of engineers, as they are provided multiple options to evaluate and use to carry out modifications to the preliminary design.

The pseudo code of the method can be seen in Algorithm 1. The heuristic uses a preliminary design as its initial solution. A single neighborhood transformation is considered: assigning a new material to a group of elements, while the others are left unchanged. The cross section of these elements is also adjusted in accordance with the new material. The material change that results in a structure with the lowest cost is chosen as the best neighbor. The algorithm stores two different solutions. A local solution (s) is used to track the progress of the search, and it changes to the neighbor with the best cost in every iteration, while a best solution (o) is also saved. The *(g,m)* pair is added to the Tabu list, where *g* is the group of elements that was changed in this step, and *m* is the old material that these elements had.

The algorithm iterates until a terminating condition is reached (e,g. a fixed number of steps without any improvement to the best solution, or set iteration limit), and the solution stored in *o* is returned as the final result.

```
s = preliminary solution
o = s
TL = ∅
while no terminating condition reached
        manage TL
        for all groups g
                for all materials m
                        if (g, m) ∈ TL
                                continue
                        p = change all elements in g to material m
                        if cost(p) < cost(s)
                                candidate = p
                                cl = (g, old material of g)
        s = candidate
        TL = TL ∪ cl
        if cost(s) < cost(o)
                o = s
return s
```

Algorithm 1: A Tabu search heuristic

In the case of our problem, the heuristic terminated if the value of the best solution *o* has not improved after a given number of iterations. As the size of the neighbourhood is defined by the number of groups and materials, we chose this iteration number to be the function (more specifically, the product) of the number of groups and materials.

3 PRELIMINARY RESULTS

We tested the Tabu search heuristic on three different instance sets. Table 1 presents the important characteristics (number of beams, columns and slabs) of these.

Table 1: Properties of the test instances

	beams	*columns*	*slabs*	*running time (s)*		
				sc1	*sc2*	*sc3*
Instance 1	89	58	27	23	21	2
Instance 2	72	48	21	6	6	1
Instance 3	72	48	21	8	7	2

Instance 2 and 3 might contain the exact same number of elements, but the layout of these structures is different.

Structures of these sizes would more or less correspond to smaller multi-storey family houses, which might qualify as real-world input, but larger buildings would be more challenging to optimize.

Table 1 also contains the solution times of the mathematical model using Gurobi. These values are presented in three columns (sc1, sc2 and sc3) for every instance, each column representing a different test scenarios:

- *Scenario 1*: elements of the same type belong to the same group.
- *Scenario 2*: elements of the same type on the same floor belong to the same group.
- *Scenario 3*: each element is a single group of its own, and is optimized separately. While this scenario is not realistic, it shows the performance of the algorithm in an extremely large search space.

A preliminary list of cross-sectional data was compiled for each element-material pair based on the different forces affecting the given element. Using this list, the required cross section of an element can easily be decided for any material type. This list was used by both the heuristic and the mathematical model.

We performed 20 test runs for each scenario. Ten of these were using the above list cross-sectional data, while random cross-section values were generated for the other ten. The results of the above scenarios can be seen in Table 2.

Table 2: Average results of the instances

	Scenario 1		*Scenario 2*		*Scenario 3*	
	running time (s)	*gap (%)*	*running time (s)*	*gap (%)*	*running time (s)*	*gap (%)*
Instance 1	0.95	0.00	8.98	0.01	159.54	0.44
Instance 2	1.27	0.00	6.84	0.04	315.60	0.83
Instance 3	1.25	0.00	7.86	0.00	122.90	0.60

Two columns belong to each scenario in the table: the first gives the running time in seconds, while the other presents the gap from the cost of the optimal solution given by the mathematical model. As it can be seen from the results, the algorithm performed well for the first two scenarios with regards to both running time and costs. However, solving the instances of scenario 3 takes a long time, and result in poor quality solutions. It is interesting to note, that because Scenario 3 considered every element as its own group, their cross-sections have been fixed by constraint (3) of the mathematical model. This resulted in the short solution time of the model, while the Tabu search heuristic performed exceptionally poor due to the significantly increased search space.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we presented an optimization problem concerning heterogeneous structural frames. Using a preliminary design as an input, we developed a heuristic that is able to provide good quality solutions with a short running time. This is important, as such a method can help with decision-making process of civil engineers when designing structural plans, as they can consider the outcomes of several scenarios by running such a fast algorithm multiple times.
To measure the quality of this heuristic, we also developed a mathematical model for the problem, that can provide the optimal solutions for the given instances.

While the performance of the heuristic algorithm was satisfactory for smaller instances, it performed poorly for larger instance sets. The neighborhood selection of the algorithm should be modified both to speed up solution process and to find better quality solutions. Another aspect of the problem that should be considered is the multi-objective nature of its cost function. While presently we consider all costs as a linear combination of the different factors, environmental impacts and capital costs should actually be optimized as separate objectives affecting each other.

Acknowledgement
The author gratefully acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European regional Development Fund), and is grateful for the support of the National Research, Development and Innovation Office - NKFIH Fund No. SNN-117879.

References

[1] Azad, S.K., Hasançebi, O. 2014. Optimum design of skeletal structures using metaheuristics: a survey of the state-of-the-art. Journal of Engineering & Applied Sciences, 6(3): 1–11.

[2] Baumann, B., Kost, B. 2005. Structure assembly by stochastic topology optimization. Computers and Structures, 83: 2175–2184.

[3] Bendsøe, M.P., Kikuchi, N. 1988. Generating optimal topologies in structural design using a homogenization method. Computer methods in applied mechanics and engineering, 71(2): 197–224.

[4] Camp, C.V., Bichon, B.J., Stovall, S.P. 2005. Design of Steel Frames Using Ant Colony Optimization. Journal of Structural Engineering, 131(3): page–page.

[5] Changizi, N., Jalalpour, M. 2018. Topology optimization of steel frame structures with constraints on overall and individual member instabilities. Finite Elements in Analysis and Design, 141: 119–134.

[6] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2): 182–197.

[7] Dorn, W.S., Gomory, R.E., Greenberg, H.J. 1964. Automatic design of optimal structures. Journal de Mecanique, 3: 25-52.

[8] Fedelinski, P., Gorski, R. 2006. Analysis and optimization of dynamically loaded reinforced plates by the coupled boundary and finite element method. CMES: Computer Modeling in Engineering and Sciences, 15: 31–40.

[9] Kirsch, U. 1989. Optimal topologies of truss structures. Computer Methods in Applied Mechanics and Engineering, 72(1): 15–28.

[10] Lamberti, L., Pappalettere, C. 2007. Weight Optimization of Skeletal Structures with Multi-Point Simulated Annealing. BIOCELL, 18(3): 183–222.

[11] Lamberti, L., Pappalettere, C. 2011. Metaheuristic Design Optimization of Skeletal Structures: A Review. Computational Technology Reviews, 4: 1–32.

[12] Rozvany, G.I. 2009. A critical review of established methods of structural topology optimization. Structural and Multidisciplinary Optimization, 37(3): 217–237.

LINEAR COMPLEMENTARITY PROBLEM AND SUFFICIENT MATRIX CLASS

Marianna E.-Nagy
Budapest University of Technology and Economics
Muegyetem rkpt. 3., 1111 Budapest, Hungary
E-mail: enagym@math.bme.hu

Abstract: Several matrix classes have been already defined related to the linear complementarity problem. The sufficient matrix class is the widest class for which the criss-cross and the interior point algorithms for solving linear complementarity problems are efficient in some sense. It is an NP-complete problem to decide whether a matrix belongs to this class. Therefore, we discover new special properties of these matrices and develop different techniques to generate such matrices. We build a library of sufficient matrices to provide a test set for algorithms solving such linear complementarity problems.

Keywords: linear complementarity problem, sufficient matrix, interior point algorithm

INTERVAL ROBUSTNESS OF MATRIX PROPERTIES FOR THE LINEAR COMPLEMENTARITY PROBLEM

Milan Hladík
Charles University, Faculty of Mathematics and Physics,
Department of Applied Mathematics,
Malostranské nám. 25, 11800, Prague, Czech Republic,
e-mail: hladik@kam.mff.cuni.cz

Abstract: We consider the linear complementarity problem with uncertain data, where uncertainty is modeled by interval ranges of possible values. Many properties of the problem (such as solvability, uniqueness, convexity, finite number of solutions etc.) are reflected by the properties of the constraint matrix. In order that the problem has desired properties even in the uncertain environment, we have to be able to check them for all possible realizations of interval data. In particular, we will discuss S-matrix, Z-matrix, copositivity, semimonotonicity, column sufficiency and R_0-matrix. We characterize the robust versions of these properties and also suggest several efficiently recognizable subclasses.

Keywords: linear complementarity, interval analysis, special matrices, NP-hardness.

1 INTRODUCTION

Linear complementarity problem. The linear complementarity problem (LCP) appears in many optimization and operations research models such as quadratic programming, bimatrix games, or equilibria in specific economies. Its mathematical formulation reads

$$y = Mz + q, \quad y, z \geq 0, \tag{1}$$

$$y^T z = 0, \tag{2}$$

where $M \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$ are given, and $y, z \in \mathbb{R}^n$ are unknown vectors. Condition (1) is linear, but the complementarity condition (2) is nonlinear and makes the problem computationally hard. The LCP is called *feasible* if (1) is feasible, and is called *solvable* if (1)–(2) is feasible. Basic properties and algorithms for the LCP are described, e.g., in the books [1, 10].

Interval uncertainty. Properties of the solution set of LCP relate with properties of matrix M. In this paper, we study properties of M when its entries are not precisely known, but we have interval ranges covering the exact values. Formally, a square interval matrix is a set

$$\boldsymbol{M} := \{M \in \mathbb{R}^{n\times n};\ \underline{M} \leq M \leq \overline{M}\},$$

where $\underline{M}, \overline{M} \in \mathbb{R}^{n\times n}$, $\underline{M} \leq \overline{M}$, are given matrices and the inequality is understood entrywise. The corresponding midpoint and radius matrices are defined as

$$M_c := \frac{1}{2}(\underline{M} + \overline{M}), \quad M_\Delta := \frac{1}{2}(\overline{M} - \underline{M}).$$

Problem statement. Throughout this paper we consider a class of the LCP problems with $M \in \boldsymbol{M}$, where $\boldsymbol{M}$ is a given interval matrix. Let $\mathcal{P}$ be a matrix property. We say that $\mathcal{P}$ holds *strongly* for $\boldsymbol{M}$ if it holds for each $M \in \boldsymbol{M}$.

Our aim is to characterize strong versions of several fundamental matrix classes appearing in the context of the LCP. If property $\mathcal{P}$ holds strongly for an interval matrix $\boldsymbol{M}$, then we are sure that $\mathcal{P}$ is provably valid whatever are the true values of the uncertain entries. Therefore, the property holds in a robust sense for the LCP problem.

Notation. Given a matrix $M \in \mathbb{R}^{n\times n}$ and index sets $I, J \subseteq \{1,\dots,n\}$, $M_{I,J}$ denotes the restriction of M to the rows indexed by I and the columns indexed by J. The identity matrix of size n is denoted by I_n, and the spectral radius of a matrix M by $\rho(M)$. The symbol D_s stands for the diagonal matrix with entries $s_1,\dots,s_n$.

2 PARTICULAR MATRIX CLASSES

In the following sections, we consider several classes of matrices appearing in the context of the linear complementarity problem. We characterize their strong counterparts when entries are interval valued. Other matrix properties were discussed, e.g., in [3, 5, 6, 8]. In particular, we leave aside P-matrices, positive definite and positive semidefinite matrices, which were already studied by many scholars in the interval data context.

We start with two classes that are simple to characterize both in the real and interval case, and then we discuss the computationally harder classes.

2.1 S-matrix

A matrix $M \in \mathbb{R}^{n\times n}$ is called an *S-matrix* if there is $x > 0$ such that $Mx > 0$. The significance of this class is that the LCP is feasible for each $q \in \mathbb{R}^n$ if and only if M is an S-matrix.

Strong S-matrix property of an interval matrix $\boldsymbol{M} \in \mathbb{IR}^{n\times n}$ is easy to characterize by the reduction to the lower bound matrix $\underline{M}$.

Proposition 2.1 *$\boldsymbol{M}$ is strongly an S-matrix if and only if system $\underline{M}x > 0$, $x > 0$ is feasible.*

Proof. If $\underline{M}x > 0$, $x > 0$ has a solution x^*, then $Mx^* \geq \underline{M}x^* > 0$ for each $M \in \boldsymbol{M}$. Therefore, every $M \in \boldsymbol{M}$ is an S-matrix. □

2.2 Z-matrix

A matrix $M \in \mathbb{R}^{n\times n}$ is called a *Z-matrix* if $M_{ij} \leq 0$ for each $i \neq j$. Z-matrices emerge in the context of Lemke's complementary pivot algorithm, because it processes any LCP with a Z-matrix.

It is easy to see that the strong Z-matrix property reduces to Z-matrix property of the upper bound matrix $\overline{M}$.

Proposition 2.2 *$\boldsymbol{M}$ is strongly a Z-matrix if and only if $\overline{M}$ is a Z-matrix.*

2.3 Copositive matrix

A matrix $M \in \mathbb{R}^{n\times n}$ is called *copositive* if $x^T Mx \geq 0$ for each $x \geq 0$. It is *strictly copositive* if $x^T Mx > 0$ for each $x \geq 0$, $x \neq 0$. A copositive matrix ensures that the complementary pivot algorithm for solving the LCP works. A strictly copositive matrix in addition implies that the LCP has a solution for each $q \in \mathbb{R}^n$. Checking whether M is copositive is a co-NP-hard problem [9].

Proposition 2.3 *$\boldsymbol{M}$ is strongly (strictly) copositive if and only if $\underline{M}$ is (strictly) copositive.*

Proof. Suppose that $M \in \boldsymbol{M}$ is not copositive. Then $x^T Mx < 0$ for some $x \geq 0$. However, $x^T \underline{M}x \leq x^T Mx < 0$, so $\underline{M}$ is not copositive, too.

Similarly for strict copositivity. □

Since checking copositivity is co-NP-hard, it is desirable to inspect some polynomially solvable classes of problems. The following is stated for the case $M_c = I_n$, but it can be easily extended to the case when M_c is a diagonal matrix with positive entries since copositivity is closed under transformation $M \mapsto DMD$, where D is a diagonal matrix with positive entries.

Proposition 2.4 *Let $M_c = I_n$. Then*

(1) $\boldsymbol{M}$ is strongly copositive if and only if $\rho(M_\Delta + M_\Delta^T) \leq 2$,

(2) $\boldsymbol{M}$ is strongly strictly copositive if and only if $\rho(M_\Delta + M_\Delta^T) < 2$.

Proof. (1) By Proposition 2.3, strong copositivity of $\boldsymbol{M}$ is equivalent to $0 \leq x^T \underline{M} x$ for each $x \geq 0$, $x \neq 0$. This inequality draws

$$0 \leq x^T \underline{M} x = x^T (I_n - M_\Delta) x = x^T x - x^T M_\Delta x,$$

from which

$$x^T x \geq x^T M_\Delta x = \frac{1}{2} x^T (M_\Delta + M_\Delta^T) x,$$

and so

$$2 \geq \frac{x^T (M_\Delta + M_\Delta^T) x}{x^T x}.$$

Now, we take the worst case of the right-hand side

$$2 \geq \max_{x \geq 0,\, x \neq 0} \frac{x^T (M_\Delta + M_\Delta^T) x}{x^T x} = \max_{x \neq 0} \frac{x^T (M_\Delta + M_\Delta^T) x}{x^T x} = \rho(M_\Delta + M_\Delta^T),$$

where we used the Rayleigh–Ritz formula for computing the largest eigenvalue of a symmetric matrix $M_\Delta + M_\Delta^T$. Since the matrix is nonnegative as well, the largest eigenvalue is equal to the spectral radius [7].

(2) For strict copositivity we proceed analogously. □

2.4 Semimonotone matrix

A matrix $M \in \mathbb{R}^{n \times n}$ is called *semimonotone (an E_0-matrix)* if the LCP has a unique solution for each $q > 0$. Equivalently, for each index set $\emptyset \neq I \subseteq \{1, \ldots, n\}$ the system

$$M_{I,I} x < 0, \quad x \geq 0 \tag{3}$$

is infeasible. By [13], checking whether M is semimonotone is a co-NP-hard problem.

Proposition 2.5 *$\boldsymbol{M}$ is strongly semimonotone if and only if $\underline{M}$ is semimonotone.*

Proof. Let $M \in \boldsymbol{M}$ be not semimonotone, so (3) has a solution x^* for some I. Then $\underline{M}_{I,I} x^* \leq M_{I,I} x^* < 0$, showing that $\underline{M}$ is not semimonotone, too. The converse direction is obvious. □

The next result shows a class of interval matrices, for which checking strong semimonotonicity can be performed effectively in polynomial time. Notice that semimonotone matrices are closed under positive row or column scaling, so the result is immediately extended to an interval matrix $\boldsymbol{M}$ such that M_c is diagonal with positive entries.

Proposition 2.6 *Let $M_c = I_n$ and $M_\Delta > 0$. Then $\boldsymbol{M}$ is strongly semimonotone if and only if $\rho(M_\Delta) \leq 1$.*

Proof. Let $\rho(M_\Delta) \leq 1$. Suppose to the contrary that (3) has a solution $x^* \geq 0$ for some I and $M \in \boldsymbol{M}$. Denote $m := |I|$, $B := I_m - M_{I,I}$ and $\Delta := (M_{I,I})_\Delta > 0$. By the Perron theory of nonnegative matrices [7], we have $\rho(\Delta) \leq \rho(M_\Delta) \leq 1$. The first inequality in (3) then reads $(I_m - B)x^* < 0$, where $|B| \leq \Delta$. From this, we derive $x^* < Bx^* \leq \Delta x^*$. Since $x^* \geq 0$, but not necessarily positive in all components, we introduce $\tilde{x} := x^* + \Delta x^* > 0$. Since $x^* < \Delta x^*$, also $\Delta x^* < \Delta^2 x^*$, whence

$$\tilde{x} = x^* + \Delta x^* < \Delta x^* + \Delta^2 x^* = \Delta \tilde{x}.$$

By the Perron theory [7], $\rho(\Delta) > 1$; a contradiction.

Conversely, suppose that $\rho(M_\Delta) > 1$. Put $I := \{1, \ldots, n\}$ and $M := I_n - M_\Delta$. Let $x > 0$ be the Perron vector for M_Δ, that is, $M_\Delta x = \rho(M_\Delta) x > x$. Then $Mx = (I_n - M_\Delta)x < 0$, so x is a solution to (3). □

2.5 Column sufficient matrix

A matrix $M \in \mathbb{R}^{n\times n}$ is *column sufficient* if for each pair of disjoint index sets $I, J \subseteq \{1,\dots,n\}$, $I \cup J \neq \emptyset$, the system

$$0 \neq \begin{pmatrix} M_{I,I} & -M_{I,J} \\ -M_{J,I} & M_{J,J} \end{pmatrix} x \leq 0, \quad x > 0 \tag{4}$$

is infeasible. Checking this condition is co-NP-hard [13], which justifies necessity of inspecting all index sets I, J. Among other properties, column sufficiency implies that for any $q \in \mathbb{R}^n$ the solution set of the LCP is a convex set (including the empty set).

Proposition 2.7 *$\boldsymbol{M}$ is strongly column sufficient if and only if system*

$$0 \neq \begin{pmatrix} \underline{M}_{I,I} & -\overline{M}_{I,J} \\ -\overline{M}_{J,I} & \underline{M}_{J,J} \end{pmatrix} x \leq 0, \quad x > 0 \tag{5}$$

is infeasible for each admissible I, J.

Proof. If $\boldsymbol{M}$ is strongly column sufficient, then (5) must be infeasible, because the matrix there comes from $\boldsymbol{M}$. Conversely, if some $M \in \boldsymbol{M}$ is not column sufficient, then (4) has a solution x^* for certain I, J. Since

$$\begin{pmatrix} \underline{M}_{I,I} & -\overline{M}_{I,J} \\ -\overline{M}_{J,I} & \underline{M}_{J,J} \end{pmatrix} x^* \leq \begin{pmatrix} M_{I,I} & -M_{I,J} \\ -M_{J,I} & M_{J,J} \end{pmatrix} x^*,$$

we have that x^* is a solution to (5); a contradiction. □

The above results also suggests a reduction to finitely many (namely, 2^n) instances.

Proposition 2.8 *$\boldsymbol{M}$ is strongly column sufficient if and only if matrices of the form $M_{ss} = M_c - D_s M_\Delta D_s$ are column sufficient for each $s \in \{\pm 1\}^n$.*

Proof. If $\boldsymbol{M}$ is strongly column sufficient, then M_{ss} is column sufficient since $M_{ss} \in \boldsymbol{M}$. Conversely, if $\boldsymbol{M}$ is not strongly column sufficient, then (5) has a solution. However, feasibility of (5) implies that M_{ss} is not column sufficient for $s \in \{\pm 1\}^n$ defined as follows: $s_i := 1$ if $i \in I$ and $s_i := -1$ otherwise, because

$$(M_{ss})_{I,I} = \underline{M}_{I,I}, \quad (M_{ss})_{J,J} = \underline{M}_{J,J}, \quad (M_{ss})_{I,J} = \overline{M}_{I,J}.$$

□

Now we state a polynomially recognizable class. Since column sufficient matrices are closed under positive row or column scaling, it can be easily extended to an interval matrix $\boldsymbol{M}$ such that M_c is diagonal with positive entries.

Proposition 2.9 *Let $M_c = I_n$ and $M_\Delta > 0$. Then $\boldsymbol{M}$ is strongly column sufficient if and only if $\rho(M_\Delta) \leq 1$.*

Proof. Let $\rho(M_\Delta) \leq 1$. Suppose to the contrary that (4) has a solution $x^* > 0$ for some I, J and $M \in \boldsymbol{M}$. Denote $K := I \cup J$, $m := |K|$ and $\Delta := (M_{K,K})_\Delta$. By the Perron theory of nonnegative matrices [7], we have $\rho(\Delta) \leq \rho(M_\Delta) \leq 1$. Next, (4) can be written as $(I_m - B)x^* \leq 0$, where $|B| \leq \Delta$. From this $x^* \leq Bx^* \leq \Delta x^*$, but $x^* \neq \Delta x^*$. Denote $\tilde{x} := x^* + \Delta x^* > 0$. Since $\Delta x^* < \Delta^2 x^*$ in view of $\Delta > 0$, we have

$$\tilde{x} = x^* + \Delta x^* < \Delta x^* + \Delta^2 x^* = \Delta \tilde{x}.$$

By the Perron theory again, $\rho(\Delta) > 1$; a contradiction.

Conversely, suppose that $\rho(M_\Delta) > 1$. Consider $I := \{1,\dots,n\}$, $J := \emptyset$ and $M := I_n - M_\Delta \in \boldsymbol{M}$. Let $x > 0$ be the Perron vector for M_Δ, that is, $M_\Delta x = \rho(M_\Delta)x > x$. Then $Mx = (I_n - M_\Delta)x < 0$, so x is a solution to (4). □

2.6 R_0-matrix

A matrix $M \in \mathbb{R}^{n\times n}$ is an R_0-*matrix* if the LCP with $q = 0$ has the only solution $x = 0$. Equivalently, for each index set $\emptyset \neq I \subseteq \{1,\dots,n\}$, the system

$$M_{I,I}x = 0, \quad M_{J,I}x \geq 0, \quad x > 0 \tag{6}$$

is infeasible, where $J := \{1,\dots,n\} \setminus I$. Checking R_0-matrix property is co-NP-hard [13]. If M is an R_0-matrix, then for any $q \in \mathbb{R}^n$ the LCP has a bounded solution set.

Proposition 2.10 *$\boldsymbol{M}$ is strongly R_0-matrix if and only if system*

$$\underline{M}_{I,I}x \leq 0, \quad \overline{M}_{I,I}x \geq 0, \quad \underline{M}_{J,I}x \geq 0, \quad x > 0 \tag{7}$$

is infeasible for each admissible I, J.

Proof. $\boldsymbol{M}$ is not strongly an R_0-matrix if and only if there are I, J and $M \in \boldsymbol{M}$ such that (6) is feasible. It is known [2, 4] that (6) is feasible for some $M \in \boldsymbol{M}$ if and only if (7) is feasible, from which the statement follows. □

Despite intractability in the general case, we can formulate a polynomial time recognizable sub-class. Since R_0-matrices are closed under positive row scaling, so the result analogously holds for an interval matrix $\boldsymbol{M}$ such that M_c is diagonal with positive entries.

Proposition 2.11 *Let $M_c = I_n$ and $M_\Delta > 0$. Then $\boldsymbol{M}$ is strongly an R_0-matrix if and only if $\rho(M_\Delta) < 1$.*

Proof. If $\rho(M_\Delta) < 1$, then by [7] also $\rho((M_{I,I})_\Delta) \leq \rho(M_\Delta) < 1$ for each I. By the condition from [11], every $M_{I,I} \in \boldsymbol{M}_{I,I}$ is nonsingular, so the system $M_{I,I}x = 0$ has the only solution $x = 0$. Thus (6) is infeasible.

Conversely, suppose that $\rho(M_\Delta) \geq 1$ and define $M := I_n - \frac{1}{\rho(M_\Delta)}M_\Delta$. By [12], M is singular and there is a (Perron) vector $x > 0$ such that $Mx = 0$. Putting $I := \{1,\dots,n\}$ and $J := \emptyset$, we have that (6) is feasible. □

3 CONCLUSION

We analysed important classes of matrices, which guarantee that the linear complementarity problem has convenient properties related to the structure of the solution set. We characterized the matrix properties in the situation that input coefficients have the form of compact intervals. As a consequence, we obtained robust properties for the linear complementarity problem: whatever are the true values from the interval data, we are sure that the corresponding property is satisfied.

Since many problems are hard to check even in the real case, it is desirable to investigate some easy-to-recognize cases. We proposed several such cases, but it is still a challenging problem to explore new ones.

4 Acknowledgements

The author was supported by the Czech Science Foundation Grant P403-18-04735S.

References

[1] R. W. Cottle, J.-S. Pang, and R. E. Stone. *The Linear Complementarity Problem*. SIAM, 2009.

[2] M. Fiedler, J. Nedoma, J. Ramík, J. Rohn, and K. Zimmermann. *Linear Optimization Problems with Inexact Data*. Springer, New York, 2006.

[3] J. Garloff, M. Adm, and J. Titi. A survey of classes of matrices possessing the interval property and related properties. *Reliab. Comput.*, 22:1–10, 2016.

[4] M. Hladík. Weak and strong solvability of interval linear systems of equations and inequalities. *Linear Algebra Appl.*, 438(11):4156–4165, 2013.

[5] M. Hladík. An overview of polynomially computable characteristics of special interval matrices. preprint arXiv: 1711.08732, `http://arxiv.org/abs/1711.08732`, 2017.

[6] M. Hladík. Tolerances, robustness and parametrization of matrix properties related to optimization problems. *Optim.*, 68(2-3):667–690, 2019.

[7] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[8] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl. *Computational Complexity and Feasibility of Data Processing and Interval Computations*. Kluwer, Dordrecht, 1998.

[9] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Math. Program.*, 39(2):117–129, 1987.

[10] K. G. Murty and F.-T. Yu. *Linear Complementarity, Linear and Nonlinear Programming*. Internet edition, 1997.

[11] G. Rex and J. Rohn. Sufficient conditions for regularity and singularity of interval matrices. *SIAM J. Matrix Anal. Appl.*, 20(2):437–445, 1998.

[12] J. Rohn. A manual of results on interval linear problems. Technical Report 1164, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2012. `http://www.library.sk/arl-cav/en/detail/?&idx=cav_un_epca*0381706`.

[13] P. Tseng. Co-NP-completeness of some matrix classification problems. *Math. Program.*, 88(1):183–192, June 2000.

USAGE OF UNIFORMLY DEPLOYED SET FOR P-LOCATION MIN-SUM PROBLEM WITH GENERALIZED DISUTILITY

Jaroslav Janáček
University of Žilina, Faculty of Management and Informatics,
Univerzitná 8215/1, 010 26 Žilina, Slovak Republic,
jaroslav.janacek@fri.uniza.sk

Marek Kvet
University of Žilina, Faculty of Management and Informatics,
Univerzitná 8215/1, 010 26 Žilina, Slovak Republic,
marek.kvet@fri.uniza.sk

Abstract: The main research goal of this paper consists in studying characteristics of an effective approximate algorithm for min-sum p-location problem, in which the concept of generalized disutility is used. The basic idea of the generalized disutility follows from the fact that the nearest located service center from any user may be temporarily unavailable due to satisfying previously arisen demand for service. Thus, an assumption is made that the service can be provided from more than one nearest center. In this paper, we present an approximate solving algorithm, which makes use of a uniformly deployed set of the p-location problem solutions. Suggested method connects the mapping approach and the incrementing exchange heuristic. Theoretical explanation of the developed approach is accompanied here with a computational study on real problem instances obtained from the road network of self-governing regions of Slovakia.

Keywords: Location problems, generalized disutility, approximate approach, uniformly deployed set of solutions

1 INTRODUCTION

Most of public service design problems with the min-sum objective are modelled using the weighted p-median problem model [3, 6, 9, 10, 14, 15]. This approach does not reflect the substantial characteristic of the public service systems with random occurring demands, which perform as queuing stochastic systems. This way of performance causes that a randomly emerged demand is not serviced from the nearest service center, but from the nearest available one, which can be much distant from the demand location than the nearest center. To comprise the new feature into associated models, notion of generalized disutility has been introduced [11, 16]. The generalized disutility is based on estimation of the probability that the nearest center, the second nearest center and so on to the r-th nearest center are the nearest available centers. Having estimated these probabilities [8], a linear integer programming model can be established and the optimal solution of the p-location problem with generalized disutility can be found by a common commercial integer programming solver (IP-solver) for medium sized instances of the problem [4, 7, 11]. Big instances of the problem represent a challenge for professionals and researchers responsible for the instance solutions. Due to unpredictable computational time of the branch-and-bound based exact method embedded in the commercial IP-solvers, the professionals turn their attention to the broad spectrum of heuristic methods [1, 2, 5]. As successfulness of heuristics often depends on suitable choice of an initial solution or an initial population, this contribution is focused on a study, how the preliminary mapping of a feasible solution set can improve performance of a simple incrementing heuristic applied to a design of the public service system with generalized disutility. The preliminary mapping will be worked up by a uniformly deployed set of p-location problem solutions, construction of which does not depend on particular instance of the problem, but it can be constructed in advance only for p-location problem parameters. The parameters of the p-locations problems are numbers m and p, where m is the number of all possible center locations and p is the number of service centers, which should be placed in the set of all possible center locations.

The remainder of the contribution is assembled in the following way. The next section comprises the formulation of the min-sum p-location problem with generalized disutility and definition of the uniformly deployed set of the p-location problem solutions. The third section deals with the algorithm connecting the mapping approach and the incrementing exchange heuristic. The fourth section contains the results of numerical experiments and a comparison of the heuristic solutions to the exact ones. Finally, the last section summarizes the obtained findings.

2 THE MIN-SUM P-LOCATION PROBLEM WITH GENERALIZED DISUTILITY

The general p-location problem in the set I of m possible center locations numbered by integers 1, ..., m can be formulated according to [1, 4, 12] in the form of (1).

$$\min\left\{F(P): P \subseteq I, |P| = p\right\} \tag{1}$$

The objective function $F(P)$ is defined by (2) for network distances d_{ij} between locations i from the set I of possible service center locations and system user location j from the set J of users locations, where the weight b_j corresponds to the user's demand. The coefficients q_k for k=1, …, r are proportional to the probability values that the nearest center, the second nearest center and so on to the r-th nearest center are the nearest available centers. In the following formula, the denotation $\min_k$ stands for the k-th minimal value of the set indicated in the brackets as introduced in [7, 11].

$$F(P) = \sum_{j \in J} b_j \sum_{k=1}^{r} q_k \min_k \left\{d_{ij} : i \in P\right\} \tag{2}$$

The min-sum problem (1) is a combinatorial problem, which can be studied as a search in a subset of m-dimensional hypercube vertices. The distance between two feasible solutions $\boldsymbol{y}$ and $\boldsymbol{x}$ represented by m dimensional zero-one vectors can be measured by so called Hamming distance defined by (3). We assume that i-th components y_i or x_i of the vectors $\boldsymbol{y}$ or $\boldsymbol{x}$ respectively equal to one, if location i belongs to the set of p selected locations, which represents the solution $\boldsymbol{y}$ or $\boldsymbol{x}$.

$$H(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{m} |y_i - x_i| \tag{3}$$

We have to realize that the distance of two feasible solutions of (1) is only even integer number ranging from 0 to $2p$ and the expression p-$H(\boldsymbol{y}, \boldsymbol{x})/2$ gives the number of possible center locations contained in both solutions.

Making use of the Hamming metric defined on the set of feasible solutions of the discussed p-location problem, the uniformly deployed set of solutions can be defined as such maximal subset S of all feasible solution that the inequality $H(\boldsymbol{y}, \boldsymbol{x}) \geq h$ holds for each $\boldsymbol{x}, \boldsymbol{y} \in S$. It must be noted that formal construction of a maximal uniformly deployed set does not depend on the specific instance of the p-location problem, but it depends only on the numbers

p, m and h and thus the set can be constructed in advance for a broad class of the p-location problems.

3 MAPPING AND INCREMENTING EXCHANGE ALGORITHM

Within this paper, we focused on exploitation of the uniformly deployed set for improvement of performance of a simple incrementing heuristic based on search through current solution neighborhood. The neighborhood of a current solution is formed by all feasible solutions, Hamming distance of which from the current solution is equal to 2. The exchange heuristic making use of the best admissible strategy is described below. In the algorithm, a solution $\mathbf{x}$ is represented by a list of p selected locations, i.e. it contains p subscripts i, for which $x_i=1$ holds. The algorithm *BA* starts with an initial solution, which inputs as a list *sP*. Further in the algorithm, the list *sP* stands for the current solution. The objective function value $F(sP)$ computed for the list *sP* of p selected service center locations is computed according to (2). The set *sC* is the complement of the set *sP* in universe I represented by m possible service center locations.

Algorithm *BA*(*sP*)

0. Initialize $F^{**}=F(sP)$, $F^{*}= F^{**}$, $sC=I\text{-}sP$.
1. {*Search through neighborhood of the current solution sP*}
   For each pair i, j for $i \in sP$ and $j \in sC$ perform inspection of the exchange I for j and after all inspections have been performed, continue with step 2.
   {*Exchange inspection*}
   Set $\underline{sP}=(sP\text{-}\{i\})\cup\{j\}$ and compute $F(\underline{sP})$.
   If $F(\underline{sP})<F^{*}$, then set $F^{*}=F(\underline{sP})$ and $sP^{*}=\underline{sP}$.
2. If $F^{**}\leq F^{*}$ then terminate with the resulting solution sP^{*} and its objective function value F^{*}.
   Otherwise, { $F^{**}> F^{*}$, *i.e. the current solution has been improved*}
   set $F^{**}= F^{*}$, $sP = sP^{*}$, $sC=I\text{-}sP$ and go to the step 1.

The time-complexity of the step 1 (Search through neighborhood of the current solution sP) is $|sP| * |sC| = p*(m\text{-}p)$. The number of searched neighborhoods in the run of algorithm cannot be estimated, as the algorithm terminates at the state, when the current solution does not enable any improvement by exchanging a pair of locations.

A uniformly deployed set S of the p-location problem solutions obtained in advance by arbitrary process can be used for improvement of the above algorithm *BA* in the following way.

a) Compute $F(s)$ for each $s \in S$ and determine the s^{*} with the lowest value of (2).
b) Set $sP=s^{*}$ and perform *BA*(*sP*).

The resulting algorithm can be easily generalized by applying the procedure *BA* to a given number of the best solutions of S or the whole algorithm can be repeated for several uniformly deployed sets, which can be obtained by permutations of the original numbering of the possible service center location set I.

4 NUMERICAL EXPERIMENTS

This section is devoted to the computational study aimed at studying the suggested approach from the viewpoint of computational time demands and the solution accuracy.

The set of used benchmarks was derived from real emergency medical service system operating on the road network within the self-governing regions of Slovakia. For each self-governing region, i.e. Bratislava (BA), Banská Bystrica (BB), Košice (KE), Nitra (NR), Prešov (PO), Trenčín (TN), Trnava (TT) and Žilina (ZA), all cities and villages with corresponding number of inhabitants b_j were taken into account. The coefficients b_j were rounded to hundreds. The set of communities represents both the set of users' locations and the set of possible center locations as well. The size of used problem instances is reported in the left part of Table 1. The number of possible service center locations is reported in the column denoted by $|I|$ and p denotes the number of centers to be located.

An individual experiment was organized so that the exact solution of the underlying p-location problem was computed using the radial approach described in [13], first. To obtain the exact solution of the problem, the optimization software FICO Xpress 7.3 was used and the experiments were run on a PC equipped with the Intel® Core™ i7 5500U processor with the parameters: 2.4 GHz and 16 GB RAM. The obtained results are summarized in the columns denoted by "Optimal solution". The column denoted by *minSum* contains the value of objective function (2). Here, the problem (1) was solved for $r = 3$. As discussed besides in [8, 11], three nearest service centers are enough to model real emergency medical service system with satisfactory solution accuracy. The associated coefficients q_k for $k=1 \ldots r$ were set in percentage in the following way: $q_1 = 77.063$, $q_2 = 16.476$ and $q_3 = 100 - q_1 - q_2$. These values were obtained from a simulation model of existing system in Slovakia as described in [8]. The computational time in seconds is given in the column denoted by *CT* [s].

The right part of Table 1 is dedicated to the results of suggested approximate solving method, which makes use of the uniformly deployed set of solutions. The uniformly deployed sets were obtained by previously performed process for the individual problems. The process consists of creating an initial uniformly deployed set and then the process continues with series of optimization problem solving procedures, when each step either adds a new solution to the set or declares that the set is maximal. The cardinalities of the resulting individual sets are reported in Table 1 in the column denoted by $|S|$. The following column denoted by *bestFit input* contains the smallest objective function value computed according to (2) for each solution of the uniformly deployed set separately. The resulting objective value of the whole approximate solving method, i.e., after the procedure *BA* was applied on the best solution of the set, is reported in the column denoted by *bestFit improved*. This objective function value was compared to the optimal solution by so called *gap*, which is defined as difference of two objective function values and it is usually expressed in percentage of the optimal objective function value. Finally, the obtained system design was compared to the optimal solution of the p-location problem by the Hamming distance *HD* of associated vectors of location variables. The value of *HD* was computed according to (3) and it is reported to demonstrate the difference between the optimal and heuristic solutions. The *HD* value gives doubled number of service center locations, in which the solutions differ. The last column of the table contains the computational time of the heuristic in seconds.

Table 1: Results of numerical experiments for the self-governing regions of Slovakia

region	$\|I\|$	p	**Optimal solution**		**Usage of even deployment for generalized disutility**					
			minSum	*CT* [s]	$\|S\|$	*bestFit input*	*bestFit improved*	*gap* [%]	*HD*	*CT* [s]
BA	87	14	26650	0.35	23	40057	26649	0.00	0	0.07
BB	515	36	44752	10.57	172	69435	44751	0.00	0	6.25
KE	460	32	45588	7.58	60	66219	45719	0.29	10	3.55
NR	350	27	48940	19.21	83	71869	48940	0.00	2	1.58
PO	664	32	56704	76.53	232	94706	56847	0.25	12	8.09
TN	276	21	35275	4.04	137	55657	35274	0.00	0	0.47
TT	249	18	41338	2.79	212	57510	41338	0.00	0	0.46
ZA	315	29	42110	2.70	112	62111	42201	0.22	12	1.31

5 CONCLUSIONS

The main goal of this paper was to introduce and explore the usage of a uniformly deployed set for *p*-location min-sum problem with generalized disutility. The concept of generalized disutility assumes that the service does not have to be necessarily provided from the nearest located service center to the system user, but it takes into account more centers. Presented results of performed computational study confirm the usefulness of suggested approach. The usage of the uniformly deployed set can bring a resulting solution of a satisfactory accuracy in very short computational time. Therefore, we can conclude that we have constructed a very useful heuristic method for effective and fast *p*-location problem solving.

Future research in this field may be aimed at other forms of uniformly deployed set usage or better forms of obtaining the initial set of solutions.

Acknowledgement
This work was supported by the research grants VEGA 1/0342/18 "Optimal dimensioning of service systems", VEGA1/0089/19 "Data analysis methods and decisions support tools for service systems supporting electric vehicles", and VEGA 1/0689/19 "Optimal design and economically efficient charging infrastructure deployment for electric buses in public transportation of smart cities" and APVV-15-0179 "Reliability of emergency systems on infrastructure with uncertain functionality of critical elements".

References

[1] Avella, P., Sassano, A. and Vasil'ev, I. (2007). Computational study of large scale p-median problems. Mathematical Programming, 109: pp. 89-114.

[2] Brotcorne, L., Laporte, G. and Semet, F. (2003). Ambulance location and relocation models. European Journal of Operational Research, 147: pp. 451–463.

[3] Czimmermann, P., Koháni, M. (2018). Computation of transportation performance in public service systems. In IEEE workshop on complexity in engineering, Firenze, pp. 1-5.

[4] García, S., Labbé, M. and Marín, A. (2011). Solving large p-median problems with a radius formulation. INFORMS Journal on Computing, 23(4): pp. 546-556.

[5] Chanta, S., Mayorga, M.E., and McLay, L.A. (2011). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. Annals of Operations Research [online] DOI 10.1007/s10479-011-0972-6.

[6] Ingolfsson, A., Budge, S. and Erkut, E. (2008). Optimal ambulance location with random delays and travel times. Health Care management science, 11(3): pp. 262-274.

[7] Janáček, J., Kvet, M. (2015). Emergency System Design with Temporarily Failing Centers. In SOR 15: Proceedings of the 13th International Symposium on Operational Research, Ljubljana: Slovenian Society Informatika, Section for Operational Research, pp. 490-495.

[8] Jankovič P. (2016). Calculating Reduction Coefficients for Optimization of Emergency Service System Using Microscopic Simulation Model. In 17th International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, pp. 163-167.

[9] Jánošíková, Ľ. (2007). Emergency medical service planning. Communications: Scientific Letters of the University of Zilina, 9(2): pp. 64-68.

[10] Jánošíková, Ľ., Žarnay, M. (2014). Location of emergency stations as the capacitated p-median problem. International scientific conference: Quantitative Methods in Economics-Multiple Criteria Decision Making XVII, Virt, Slovak Republic.

[11] Kvet, M. (2014). Computational study of radial approach to public service system design with generalized utility. Digital Technologies 2014: the 10th International IEEE conference, Zilina, Slovak Republic.

[12] Kvet, M. (2015). Exact and heuristic radial approach to fair public service system design. Information and Digital Technologies 2015: IEEE catalog number CFP15CDT-USB, Zilina: Slovak Republic.

[13] Kvet, M., Janáček, J. (2019). Identification of the Maximal Relevant Distance in Emergency System Designing. Mathematical methods in economics 2019, České Budějovice, in print

[14] Marianov, V., Serra, D. (2004). Location problems in the public sector, Facility location. Applications and theory (by Drezner Z (ed.) et al.), Berlin, Springer: pp. 119-150.

[15] Matiaško, K., Kvet, M. (2017). Medical data management. Proceedings of IEEE conference Informatics 2017, Poprad, Slovakia, pp. 253-258.

[16] Snyder, L. V., Daskin, M. S. (2005). Reliability Models for Facility Location; the Expected Failure Cost Case. Transport Science, 39 (3), pp. 400-416.

An existence criterion for the sum of squares*

Dragan Jukić[†] Kristian Sabo [‡]

Abstract

In this paper, we give a necessary and sufficient criterion for the existence of the least squares estimate for the nonlinear sum of squares. Our criterion is based on the existence level that describes the behavior of the sum of squares as its argument approaches the extended boundary of the parameter space.

Keywords: Nonlinear regression; Least squares estimate; Existence level

AMS Subject Classification: 62J02; 65D10; 65C20

1 Introduction

The general regression model is of the form:

$$y_i = f(t_i; \boldsymbol{\vartheta}) + \varepsilon_i, \quad \boldsymbol{\vartheta} \in \boldsymbol{\Theta}, \qquad i = 1, \dots, n,$$

where f is a model-function given in advance which is defined on parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^m$, t_i are independent variables, y_i are observations and ε_i are random errors, usually assumed to be normally distributed with mean zero and constant variance. As is usual in statistics, the unknown parameters are denoted by Greek letters. In order to simplify the notation, denote $f_i(\boldsymbol{\vartheta}) := f(\boldsymbol{t}_i; \boldsymbol{\vartheta})$, $i = 1, \dots, n$. The least squares estimates (LSE) are those that minimize (on the set $\boldsymbol{\Theta}$) the following sum of squares (see [1, 13]):

$$S(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} (f_i(\boldsymbol{\vartheta}) - y_i)^2.$$

In matrix form, we can write

$$S(\boldsymbol{\vartheta}) = \|\mathbf{f}(\boldsymbol{\vartheta}) - \mathbf{y}\|^2,$$

where $\mathbf{f}(\boldsymbol{\vartheta}) = (f_1(\boldsymbol{\vartheta}), \dots, f_n(\boldsymbol{\vartheta}))^T$ and $\mathbf{y} = (y_1, \dots, y_n)^T$. The set of all n-vectors $\mathbf{f}(\boldsymbol{\vartheta})$, i.e., the set

$$\mathcal{E} = \{\mathbf{f}(\boldsymbol{\vartheta}) : \boldsymbol{\vartheta} \in \boldsymbol{\Theta}\},$$

defines an m-dimensional surface, the so-called *expectation surface* or *solution locus*. The LSE $\hat{\boldsymbol{\vartheta}}$, if it exists, corresponds to the point $\mathbf{f}(\hat{\boldsymbol{\vartheta}})$ on $\mathcal{E}$ which is closest to $\mathbf{y}$.

In this paper, we give a necessary and sufficient criterion for the existence of the LSE for the nonlinear sum of squares. The structure of the paper is as follows. In Section 2, firstly according to [7], we define the *existence level* that describes the behavior of the sum of squares as its argument approaches the extended boundary of the parameter space. The main existence result closely related to the existence level is given in Theorem 1. Finally, throughout the three examples in Section 3, we illustrate the practical usability of this theorem.

2 The existence criterion

In this section, we assume that all regression functions f_i, $i = 1, \dots, n$, are continuous on the parameter space $\boldsymbol{\Theta}$. We also assume that the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^m$ is noncompact, because otherwise the LSE exists.

Lemma 1 *The least squares estimate exists if and only if there exists a point* $\mathbf{f}(\boldsymbol{\vartheta}_0) \in \mathcal{E}$ *such that*

$$\|\mathbf{f}(\boldsymbol{\vartheta}_0) - \mathbf{y}\| \le \|\mathbf{e} - \mathbf{y}\| \quad \textit{for all } \mathbf{e} \in \operatorname{Cl}\mathcal{E}\backslash\mathcal{E}, \tag{1}$$

or, equivalently,

$$\|\mathbf{f}(\boldsymbol{\vartheta}_0) - \mathbf{y}\|^2 \le \inf_{\mathbf{e} \in \operatorname{Cl}\mathcal{E}\backslash\mathcal{E}} \|\mathbf{e} - \mathbf{y}\|^2. \tag{2}$$

*This work was supported by the Croatian Science Foundation through research grant IP-2016-06-6545

[†]Department of Mathematics, J.J. Strossmayer University of Osijek, Trg Ljudevita Gaja 6, HR-31 000 Osijek, Croatia, E-mail address: `jukicd@mathos.hr`

[‡]Department of Mathematics, J.J. Strossmayer University of Osijek, Trg Ljudevita Gaja 6, HR-31 000 Osijek, Croatia, E-mail address: `ksabo@mathos.hr`

Proof. Suppose that $\hat{\boldsymbol{\vartheta}}$ is an LSE. Let $\mathbf{e} \in \mathrm{Cl}\,\mathcal{E}\backslash\mathcal{E}$. Choose a sequence $\{\mathbf{f}(\boldsymbol{\vartheta}_k)\} \subset \mathcal{E}$ converging to $\mathbf{e}$. Since $\|\mathbf{f}(\hat{\boldsymbol{\vartheta}}) - \mathbf{y}\| \leq \|\mathbf{f}(\boldsymbol{\vartheta}_k) - \mathbf{y}\|$ for all k, it follows, by continuity of the norm, that $\|\mathbf{f}(\hat{\boldsymbol{\vartheta}}) - \mathbf{y}\| \leq \|\mathbf{e} - \mathbf{y}\|$, so that is enough to set $\mathbf{f}(\boldsymbol{\vartheta}_0) = \mathbf{f}(\hat{\boldsymbol{\vartheta}})$.

Conversely, assume that there exists $\mathbf{f}(\boldsymbol{\vartheta}_0) \in \mathcal{E}$ such that (1) holds. Let $\mathbf{e}_\mathbf{y} \in \mathrm{Cl}\,\mathcal{E}$ be the projection of $\mathbf{y}$ into $\mathrm{Cl}\,\mathcal{E}$, i.e.,

$$\|\mathbf{e}_\mathbf{y} - \mathbf{y}\| \leq \|\mathbf{e} - \mathbf{y}\| \quad \text{for all } \mathbf{e} \in \mathrm{Cl}\,\mathcal{E}. \tag{3}$$

If $\mathbf{e}_\mathbf{y} \in \mathcal{E}$, then there exists $\hat{\vartheta} \in \boldsymbol{\Theta}$ such that $\mathbf{e}_\mathbf{y} = \mathbf{f}(\hat{\vartheta})$, and therefore, from the last inequality it follows that $\hat{\vartheta}$ is an LSE; the proof is complete. Otherwise, if $\mathbf{e}_\mathbf{y} \in \mathrm{Cl}\,\mathcal{E}\backslash\mathcal{E}$, then from (1) and (3) we obtain

$$\|\mathbf{f}(\boldsymbol{\vartheta}_0) - \mathbf{y}\| \overset{(1)}{\leq} \|\mathbf{e}_\mathbf{y} - \mathbf{y}\| \overset{(3)}{\leq} \|\mathbf{e} - \mathbf{y}\| \quad \text{for all } \mathbf{e} \in \mathrm{Cl}\,\mathcal{E},$$

which means that $\boldsymbol{\vartheta}_0$ is an LSE. □

In what follows, for the sake of simplicity, we adopt the following convention: whenever we write $\boldsymbol{\vartheta}_k \to \infty$, we will mean that $\|\boldsymbol{\vartheta}_k\| \to \infty$. With this convention, let us define the so-called *existence level* $S_E^\star$ as

$$S_E^\star = \inf_{\substack{\{\boldsymbol{\vartheta}_k\}\subseteq\boldsymbol{\Theta} \\ \boldsymbol{\vartheta}_k \to \boldsymbol{x}_0 \in \partial_\infty\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}}} \varliminf_k S(\boldsymbol{\vartheta}_k), \tag{4}$$

where

$$\partial_\infty\boldsymbol{\Theta} = \begin{cases} \partial\boldsymbol{\Theta}, & \text{if } \boldsymbol{\Theta} \text{ is bounded} \\ \partial\boldsymbol{\Theta} \cup \{\infty\}, & \text{if } \boldsymbol{\Theta} \text{ is unbounded} \end{cases}$$

is the so-called *extended boundary* of the set $\boldsymbol{\Theta}$. Since the parameter space $\boldsymbol{\Theta}$ is noncompact, there exists a sequence $\{\boldsymbol{\vartheta}_k\} \subset \boldsymbol{\Theta}$ such that $\lim_{k\to\infty} \|\boldsymbol{\vartheta}_k\| = \infty$ (if not bounded), or such that $\lim_{k\to\infty} \boldsymbol{\vartheta}_k \in \partial\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}$ (if not closed). In other words, the set of sequences over which the infimum in (4) is taken is nonempty, i.e., $S_E^\star$ is well-defined.

To the best of our knowledge, the existence level defined by (4) was for the first time introduced by Jukić in [7], where he gave one necessary and sufficient criterion for the existence of the global minima of a continuous lower bounded function on noncompact set. Although our next Theorem 1 follows from his criterion, it should be emphasized that our proof uses different techniques based on the expectation surface. Some similar concepts of the existence level can be found in, e.g., Demidenko [3, 4, 5], Nakumara [11] and Nakamura and Lee [12]. The following lemma will be needed in the proof of Theorem 1.

Lemma 2

$$S_E^\star \leq \inf_{\mathbf{e}\in\mathrm{Cl}\,\mathcal{E}\backslash\mathcal{E}} \|\mathbf{e} - \mathbf{y}\|^2. \tag{5}$$

Proof. Denote

$$\begin{aligned} A &:= \{\|\mathbf{e} - \mathbf{y}\|^2 : \mathbf{e} \in \mathrm{Cl}\,\mathcal{E}\backslash\mathcal{E}\} \\ B &:= \Big\{\varliminf_k S(\boldsymbol{\vartheta}_k) : \{\boldsymbol{\vartheta}_k\} \subset \boldsymbol{\Theta}, \boldsymbol{\vartheta}_k \to \boldsymbol{\vartheta}_0 \in \partial\boldsymbol{\Theta}\backslash\boldsymbol{\Theta} \text{ or } \|\boldsymbol{\vartheta}_k\| \to \infty\Big\}, \end{aligned}$$

and note that $S_E^\star = \inf B$.

It is enough to prove that $A \subseteq B$. Let $\mathbf{e} \in \mathrm{Cl}\,\mathcal{E}\backslash\mathcal{E}$. Since $\mathbf{e} \in \mathrm{Cl}\,\mathcal{E}$, there exists a sequence $\{\mathbf{f}(\boldsymbol{\vartheta}_k)\} \subset \mathcal{E}$ converging to $\mathbf{e}$. The sequence $\{\boldsymbol{\vartheta}_k\}$ has a subsequence $\{\boldsymbol{\vartheta}_{u_k}\}$ such that all its component sequences are monotone. Therefore, if the subsequence $\{\boldsymbol{\vartheta}_{u_k}\}$ is bounded, then it converges to some $\boldsymbol{\vartheta}_0 \in \mathrm{Cl}\,\boldsymbol{\Theta}$; otherwise $\|\boldsymbol{\vartheta}_{u_k}\| \to \infty$. Note that since $\mathbf{e} \notin \mathcal{E}$, then $\boldsymbol{\vartheta}_0 \notin \boldsymbol{\Theta}$, because otherwise the continuity of $\mathbf{f}$ would imply that $\mathbf{e} = \lim_{k\to\infty} \mathbf{f}(\boldsymbol{\vartheta}_{u_k}) = \mathbf{f}(\boldsymbol{\vartheta}_0) \in \mathcal{E}$. Thus, $\boldsymbol{\vartheta}_0 \in \mathrm{Cl}\,\boldsymbol{\Theta}\backslash\boldsymbol{\Theta} = \partial\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}$. Furthermore, since $\lim_{k\to\infty} \mathbf{f}(\boldsymbol{\vartheta}_{u_k}) = \lim_{k\to\infty} \mathbf{f}(\boldsymbol{\vartheta}_k) = \mathbf{e}$,

$$\|\mathbf{e} - \mathbf{y}\|^2 = \lim_{k\to\infty} \|\mathbf{f}(\boldsymbol{\vartheta}_{u_k}) - \mathbf{y}\|^2 = \lim_{k\to\infty} S(\boldsymbol{\vartheta}_{u_k}) = \varliminf_k S(\boldsymbol{\vartheta}_{u_k}).$$

This proves that $A \subseteq B$. □

Theorem 1 (Necessary and sufficient criterion for the existence of the LSE) *The LSE exists if and only if there exists* $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$ *such that*

$$S(\boldsymbol{\vartheta}_0) \leq S_E^\star. \tag{6}$$

Proof. Suppose that $\hat{\boldsymbol{\vartheta}} \in \boldsymbol{\Theta}$ is the LSE. Then, for any sequence $\{\boldsymbol{\vartheta}_k\} \subset \boldsymbol{\Theta}$ converging to $\boldsymbol{\vartheta}_0 \in \partial_\infty\boldsymbol{\Theta}\backslash\boldsymbol{\Theta} = (\partial\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}) \cup \{\infty\}$, we have $S(\hat{\boldsymbol{\vartheta}}) \le S(\boldsymbol{\vartheta}_k)$. Therefore, $S(\hat{\boldsymbol{\vartheta}}) \le \underline{\lim}_k S(\boldsymbol{\vartheta}_k)$, and, consequently,

$$S(\hat{\boldsymbol{\vartheta}}) \le \inf_{\boldsymbol{\vartheta}_k \to \boldsymbol{\vartheta}_0 \in \partial_\infty\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}} \underline{\lim_k} S(\boldsymbol{\vartheta}_k) = S_E^\star.$$

Conversely, suppose that $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$ is sucht that $S(\boldsymbol{\vartheta}_0) \le S_E^\star$. Then, the existence of the LSE follows from Lemmas 1 and 2. □

We will say that $\bar{\mathbf{f}}(\boldsymbol{\beta})$ is *limit regression* of $\mathbf{f}(\boldsymbol{\vartheta})$ if

$$\bar{\mathbf{f}}(\boldsymbol{\beta}) = \lim_{k\to\infty} \mathbf{f}(\boldsymbol{\vartheta})$$

for some sequence $\{\boldsymbol{\vartheta}_k\} \subset \boldsymbol{\Theta}$ with a limit in $\partial_\infty\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}$. If we denote by LR the set of all limit regressions, then, by definition of $S_E^\star$, it is easy to check that

$$S_E^\star = \inf_{\bar{\mathbf{f}}(\boldsymbol{\beta}) \in LR} \|\bar{\mathbf{f}}(\boldsymbol{\beta}) - \mathbf{y}\|^2.$$

Therefore, the statement of Theorem 1 can be reformulated as follows:
The LSE exists if and only if there is at least one point $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$ with the sum of squares less than or equal to the corresponding weighted sum of squares for any limit regression.

3 Some illustrative examples

In this section, we will give three illustrative examples that show the practical usability of Theorem 1. Functions in the first two examples are used very often to test optimization algorithms.

Example 1 *(Modified Biggs EXP2 Function) (see [2]) Let*

$$f_i(\boldsymbol{\vartheta}) = e^{at_i} + 5e^{bt_i},\ \boldsymbol{\vartheta} = (a,b) \in \boldsymbol{\Theta} = \{(a,b) \in \mathbb{R}^2 : a, b \ge 0\},$$

and $t_i = 0.1\,i$, $\quad y_i = e^{t_i} + 5e^{10t_i}$, $i = 1,\dots,10$. *The corresponding sum of squares reads* $S(\boldsymbol{\vartheta}) = \sum_{i=1}^{10}\left(e^{at_i} + 5e^{bt_i} - y_i\right)^2$. *It is easy to see*

$$\min_{\boldsymbol{\vartheta}\in\boldsymbol{\Theta}} S(\boldsymbol{\vartheta}) = S(\hat{\boldsymbol{\vartheta}}) = 0, \quad \hat{\boldsymbol{\vartheta}} = (1, 10).$$

Note that $\partial_\infty\boldsymbol{\Theta} \setminus \boldsymbol{\Theta} = \{\infty\}$. *Let* $\{\boldsymbol{\vartheta_k}\} = \{(a_k, b_k)\} \subset \boldsymbol{\Theta}$ *be any sequence such that* $\boldsymbol{\vartheta_k} \to \infty$, *i.e., such that* $a_k^2 + b_k^2 \to \infty$. *Then obviously* $S(a_k, b_k) \to \infty$, *and therefore* $S_E^\star = \infty$. *Since* $S(\boldsymbol{\vartheta_0}) < S_E^\star$, *for any* $\boldsymbol{\vartheta_0} \in \boldsymbol{\Theta}$, *in accordance with Theorem 1, it follows that the LSE exists.*

Example 2 *(Biggs EXP2 Function) (see [2]) Let*

$$f_i(\boldsymbol{\vartheta}) = e^{-at_i} - 5e^{-bt_i},\ \boldsymbol{\vartheta} = (a,b) \in \boldsymbol{\Theta} = \{(a,b) \in \mathbb{R}^2 : a, b \ge 0\},$$

and $t_i = 0.1\,i$, $y_i = e^{-t_i} - 5e^{-10t_i}$, $i = 1,\dots,10$. *It is easy to see that the corresponding sum of squares* $S(\boldsymbol{\vartheta}) = \sum_{i=1}^{10}\left(e^{-at_i} - 5e^{-bt_i} - y_i\right)^2$ *attains its minimum value of* 0 *at* $\hat{\boldsymbol{\vartheta}} = (1, 10)$. *Note that* $\partial_\infty\boldsymbol{\Theta} \setminus \boldsymbol{\Theta} = \{\infty\}$. *Let* $\{\boldsymbol{\vartheta_k}\} = \{(a_k, b_k)\} \subset \boldsymbol{\Theta}$ *be any sequence such that* $\boldsymbol{\vartheta_k} \to \boldsymbol{\vartheta_0} \in \partial_\infty\boldsymbol{\Theta}\backslash\boldsymbol{\Theta}$. *Then, if* $(a_k, b_k) \to (\infty, \infty)$, *it follows* $S(a_k, b_k) \to \sum_{i=1}^{10} y_i^2$, *if* $(a_k, b_k) \to (\infty, b_0)$, $b_0 \ge 0$, *then* $S(a_k, b_k) \to \sum_{i=1}^{10}(-5e^{-b_0 t_i} - y_i)^2 \ge S_b^\star$ *and finally, if* $(a_k, b_k) \to (a_0, \infty)$, $a_0 \ge 0$, *then* $S(a_k, b_k) \to \sum_{i=1}^{10}\left(e^{-a_0 t_i} - y_i\right)^2 \ge S_a^\star$, *where* $S_b^\star = \inf_{b\ge 0} S_b(b)$, $S_b(b) = \sum_{i=1}^{10}(-5e^{-bt_i} - y_i)^2$, *and* $S_a^\star = \inf_{a\ge 0} S_a(a)$, $S_a(a) = \sum_{i=1}^{10}(e^{-at_i} - y_i)^2$. *Consequently,* $S_E^\star = \min\left\{\sum_{i=1}^{10} y_i^2, S_a^\star, S_b^\star\right\}$.

Since for $a_k \to \infty$ *and* $b_k \to \infty$ *we have* $S_a(a_k) \to \sum_{i=1}^{10} y_i^2$ *and* $S_b(b_k) \to \sum_{i=1}^{10} y_i^2$, *according to Theorem 1, the LSE* $\hat{a} \ge 0$ *for limit regression* $t \mapsto e^{-at}$ *exists if and only if there exists* $a_0 \ge 0$ *such that* $S_a(a_0) \le \sum_{i=1}^{10} y_i^2$, *and analogously, the LSE* $\hat{b} \ge 0$ *for limit regression* $t \mapsto -5e^{-bt}$ *exists if and only if there exists* $b_0 \ge 0$ *such that* $S_b(b_0) \le \sum_{i=1}^{10} y_i^2$.

Finally, we have

$$S_E^\star = \begin{cases} \min\{S_a(\hat{a}), S_b(\hat{b})\}, & \text{if both } \hat{a}, \hat{b} \ge 0 \text{ exist} \\ S_a(\hat{a}), & \text{if } \hat{a} \ge 0 \text{ exists and } \hat{b} \ge 0 \text{ does not exist} \\ S_b(\hat{b}), & \text{if } \hat{b} \ge 0 \text{ exists and } \hat{a} \ge 0 \text{ does not exist} \\ \sum_{i=1}^{10} y_i^2, & \text{if neither } \hat{a} \ge 0 \text{ nor } \hat{b} \ge 0 \text{ exists.} \end{cases}$$

From Figure 1 it is obvious that $\hat{b}$ *exists and* $\hat{a}$ *does not exist, and consequently,* $S_E^\star = S_b(\hat{b}) \approx 2.05133$, *where* $S_b(\hat{b})$ *has been estimated numerically. Since the level set* $\{\boldsymbol{\vartheta} \in \boldsymbol{\Theta} : S(\boldsymbol{\vartheta}) \le S_E^\star\}$ *is nonempty (see Figure 2), in accordance with Theorem 1, there exists the LSE for the original problem.*

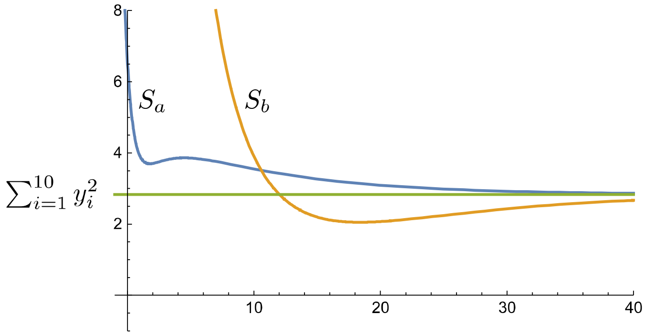


Figure 1: Graphs of objective functions S_a, S_b for limit regressions

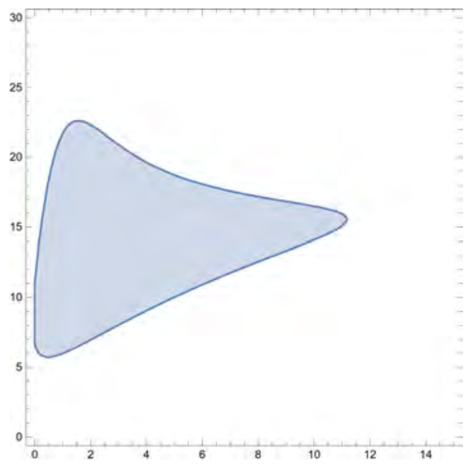


Figure 2: Level set $\{\boldsymbol{\vartheta} \in \boldsymbol{\Theta} : S(\boldsymbol{\vartheta}) \leq S_E^\star\}$ generated by the Mathematica instruction `RegionPlot`.

Example 3 *(Michaelis-Menten regression model) (see [8], [10]) Let* $\boldsymbol{D} = \{(a,b) : a > 0, b > 0\}$ *and*

$$f_i(a,b) = \frac{a\,t_i}{b + t_i},\ i = 1,\ldots,n.$$

Assume that $0 < t_1 \leq t_2 \ldots \leq t_n$ *and* $y_1,\ldots,y_n > 0$.

In order to examine the existence of the LSE, let us introduce the following reparametrization

$$f_i(\boldsymbol{\vartheta}) = \frac{\alpha\,t_i}{\beta + \gamma t_i},\ i = 1,\ldots,n, \quad \alpha := \frac{a}{\sqrt{1+b^2}},\ \beta := \frac{b}{\sqrt{1+b^2}},\ \gamma := \frac{1}{\sqrt{1+b^2}}.$$

Since the map $(a,b) \mapsto (\alpha,\beta,\gamma)$ *is a bijection of* D *onto*

$$\boldsymbol{\Theta} := \{(\alpha,\beta,\gamma) : \alpha,\beta,\gamma > 0,\ \beta^2 + \gamma^2 = 1\},$$

we will consider the following sum of squares

$$S(\boldsymbol{\vartheta}) = \sum_{i=1}^{n}(f_i(\boldsymbol{\vartheta}) - y_i)^2 = \sum_{i=1}^{n}\left(\frac{\alpha\,t_i}{\beta + \gamma t_i} - y_i\right)^2.$$

Note that $\partial_\infty \boldsymbol{\Theta} \setminus \boldsymbol{\Theta} = \{(\alpha,0,1) : \alpha \geq 0\} \cup \{(\alpha,1,0) : \alpha \geq 0\} \cup \{\infty\}$. *Let* $\{\boldsymbol{\vartheta_k}\} = \{(\alpha_k,\beta_k,\gamma_k)\} \subset \boldsymbol{\Theta}$ *be any sequence such that* $\boldsymbol{\vartheta_k} \to \boldsymbol{\vartheta_0} \in \partial_\infty \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}$, *i.e., such that (a)* $\alpha_k \to \infty$, *(b)* $(\alpha_k,\beta_k,\gamma_k) \to (\alpha_0,1,0)$ *or (c)* $(\alpha_k,\beta_k,\gamma_k) \to (\alpha_0,0,1)$. *In case (a), obviously* $S(\alpha_k,\beta_k,\gamma_k) \to \infty$. *For* $\alpha_0 = 0$, *in both cases (b) and (c) it follows* $S(\alpha_k,\beta_k,\gamma_k) \to \sum_{i=1}^n y_i^2$. *Therefore, suppose further that* $\alpha_0 > 0$. *Then, in case (b), we obtain* $S(\alpha_k,\beta_k,\gamma_k) \to \sum_{i=1}^n (a_0 t_i - y_i)^2 \geq \sum_{i=1}^n (\hat\kappa t_i - y_i)^2$, *where*

$$\hat\kappa = \operatorname*{argmin}_{\kappa>0} \sum_{i=1}^{n} (\kappa t_i - y_i)^2 = \frac{\sum_{i=1}^n t_i y_i}{\sum_{i=1}^n t_i^2}. \tag{7}$$

Finally, in case (c), we have $S(\alpha_k,\beta_k,\gamma_k) \to \sum_{i=1}^n (a_0 - y_i)^2 \geq \sum_{i=1}^n (\overline{y} - y_i)^2$, *where*

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{8}$$

From previous considerations it follows

$$S_E^\star = \min\left\{\sum_{i=1}^n y_i^2, \sum_{i=1}^n (y_i-\overline{y})^2, \sum_{i=1}^n (\hat{\kappa} t_i - y_i)^2\right\} = \min\left\{\sum_{i=1}^n (y_i-\overline{y})^2, \sum_{i=1}^n (\hat{\kappa} t_i - y_i)^2\right\},$$

where $\hat{\kappa}$ *and* $\overline{y}$ *are given by (7) and (8), respectively.*

In [6], it was shown that under some natural conditions there exists $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$ such that $S(\boldsymbol{\vartheta}_0) \leq S_E^\star$, and according to Theorem 1, there exists an LSE for the Michaelis-Menten function. Particulary, the following statement was proved:

- *If the data* (t_i, y_i), $i = 1, \ldots, n$, $n \geq 3$, *are such that* $0 < t_1 \leq t_2 \leq \ldots \leq t_n$, $y_i > 0$, *and if they fulfill one of the two sets of inequalities*

$$n\left(\sum_{i=1}^n t_i\, y_i\right)^2 \leq \sum_{i=1}^n t_i^2 \left(\sum_{i=1}^m y_i\right)^2 \tag{9}$$

$$\sum_{i=1}^n \frac{1}{t_i}\sum_{i=1}^m y_i > n \sum_{i=1}^m \frac{y_i}{t_i} \tag{10}$$

or

$$n\left(\sum_{i=1}^n t_i\, y_i\right)^2 \geq \sum_{i=1}^n t_i^2 \left(\sum_{i=1}^n y_i\right)^2 \tag{11}$$

$$\sum_{i=1}^n t_i\, y_i \sum_{i=1}^n y_i^3 > \sum_{i=1}^n t_i^2 \sum_{i=1}^n t_i^2 y_i, \tag{12}$$

then there exists $\boldsymbol{\vartheta}_0 \in \boldsymbol{\Theta}$ *such that* $S(\boldsymbol{\vartheta}_0) \leq S_E^\star$, *i.e., there exists an LSE for the Michaelis-Menten function.*

Let us briefly explain the meanings of inequalities (9)-(12). Firstly, note that one of the two inequalities (9) or (11) is always fulfilled, for any data point. It is easy to see that inequality (9) holds if and only if the sum of squares for limit regression $t \mapsto \overline{y}$ is not greater than the sum of squares for limit regression $t \mapsto \hat{\kappa} t$, and analogously, inequality (11) holds if and only if the sum of squares for limit regression $t \mapsto \hat{\kappa} t$ is not greater than the sum of squares for limit regression $t \mapsto \overline{y}$, where $\hat{\kappa}$ and $\overline{y}$ are given by (7) and (8), respectively.

Furthermore, inequalities (10) and (12) are in connection with the well-known Chebyshev's sum inequality (see e.g. [9]). Chebyshev's sum inequality is usually stated as follows. Let $0 < x_1 \leq x_2 \leq \ldots \leq x_n$ and $0 < z_1 \leq z_2 \leq \ldots \leq z_n$. Then for $p_i > 0$, $i = 1, \ldots, n$:

$$\sum_{i=1}^m p_i \sum_{i=1}^n p_i x_i z_i \geq \sum_{i=1}^n p_i x_i \sum_{i=1}^n p_i z_i. \tag{13}$$

The equality holds if and only if $x_1 = x_2 = ... = x_n$ or $z_1 = z_2 = ... = z_n$. Let us mention that inequality (13) holds if and only if the slope of linear regression for data points (p_i, x_i, z_i), $i = 1, \ldots, n$ is nonnegative, where x_i are independent variables, y_i are observations and $p_i > 0$ are corresponding data weights.

Inequality (13) can be reversed. If $0 < x_1 \leq x_2 \leq \ldots \leq x_n$ and $z_1 \geq z_2 \geq \ldots \geq z_n > 0$, then for $p_i > 0$, $i = 1, \ldots, n$:

$$\sum_{i=1}^n p_i \sum_{i=1}^n p_i x_i z_i \leq \sum_{i=1}^n p_i x_i \sum_{i=1}^n p_i z_i. \tag{14}$$

Similarly, the equality holds if and only if $x_1 = x_2 = ... = x_n$ or $z_1 = z_2 = ... = z_n$. Also, inequality (14) holds if and only if the slope of linear regression for data points (p_i, x_i, z_i), $i = 1, \ldots, n$ is nonpositive, where x_i are independent variables, y_i are observations and $p_i > 0$ are corresponding data weights. Consequently, if we suppose that data points (t_i, y_i), $i = 1, \ldots, n$ satisfy the following conditions:

$$t_1 \leq t_2 \leq \ldots \leq t_n,\ t_1 < t_n,$$

$$y_1 \leq y_2 \leq \ldots \leq y_n,\ y_1 < y_n,$$

then for

$$p_i := 1, \quad x_i := y_i, \quad z_i := \frac{1}{t_i}, \quad i = 1, \ldots, n,$$

inequality (10) follows from Chebyshev's sum inequality (14). Similarly, if we suppose that data points (t_i, y_i), $i = 1, \dots, n$ satisfy the following conditions:

$$t_1 \le t_2 \le \dots \le t_n,\ t_1 < t_n,$$

$$y_1/t_1 \ge y_2/t_2 \ge \dots \ge y_n/t_n,\ y_1/t_1 > y_n/t_n,$$

then for

$$p_i := t_i^2, \quad x_i := t_i, \quad z_i := y_i/t_i, \quad i = 1, \dots, n$$

inequality (12) follows from Chebyshev's sum inequality (13).

References

[1] D.M. Bates, D.G. Watts, Nonlinear Regression Analysis and its Applications, Wiley, New York, 1988.

[2] M. C. Biggs, A New Variable Metric Technique Taking Account of Non-Quadratic Behaviour of the Objective Function, IMA Journal of Applied Mathematics, 8 (1971), 315-327.

[3] E. Demidenko, Criteria for unconstrained global optimization, J. Optim. Theory Appl. 136 (2008) 375-395.

[4] E. Demidenko, Criteria for global minimum of sum of squares in nonlinear regression, Comput. Statist. Data Anal. 51 (2006) 1739-1753.

[5] E. Demidenko, Optimization and Regression (in Russian). Nauka, Moskva 1989.

[6] K. P. Hadeler, D. Jukić, K. Sabo, Least squares problems for Michaelis Menten kinetics, Mathematical Methods in the Applied Sciences 30 (2007), 1231–1241

[7] D. Jukić, A necessary and sufficient criterion for the existence of the global minima of a continuous lower bounded function on a noncompact set, submitted

[8] L. Michaelis, M.L. Menten, Die Kinetik der Invertinwirkung, Biochem. Z. 49 (1913) 333–369.

[9] D. S. Mitrinović, J. Pečarić, A. M. Fink, Classical and New Inequalities in Analysis, Kluwer: Dordrecht, 1993.

[10] J. D. Murray, Mathematical Biology, Springer, Berlin, 1993.

[11] T. Nakamura, Existence theorems of a maximum likelihood estimate from a generalized censored data sample, Annals of the Institute of Statistical Mathematics 36(1984) 375-393.

[12] T. Nakamura,C. Lee, On the existence of minimum contrast estimates in binary response model, Ann. Inst. Statist. Math. 45(1993), 741-758.

[13] G.J.S. Ross, Nonlinear Estimation, Springer, New York, 1990.

A NEW PRUNING TEST FOR PARAMETRIC INTERVAL LINEAR SYSTEMS

Miroslav Rada
University of Economics, Prague, Department of Financial Accounting and Auditing,
nám. W. Churchilla 4, 13067 Prague, Czech Republic
miroslav.rada@vse.cz
Elif Garajová, Jaroslav Horáček, Milan Hladík
Charles University, Faculty of Mathematics and Physics, Department of Applied Mathematics
Malostranské nám. 25, 11800 Prague, Czech Republic
{elif, horacek, hladik}@kam.mff.cuni.cz

Abstract: We deal with the weak solution set of parametric interval linear systems. We adopt the branch-and-prune `SIVIA` algorithm to obtain an outer approximation of the weak solution set and, if possible, also an inner approximation. We propose a new pruning test, based on the fact that solvability of a linear system can be viewed as a zonotope membership problem. The proposed test works efficiently in its full strength especially when pruning infeasible boxes.

Keywords: parametric interval linear systems, branch-and-prune, zonotope

1 INTRODUCTION

Solving systems of linear equations belongs to one of the essential problems addressed in linear algebra and numerical analysis. However, in practical problems, linear systems are often affected by uncertainty due to rounding errors in the computation or inherent inexactness of the input data. Such uncertainty needs to be reflected in the system, since disregarding it may lead to distorted and unusable solutions. In many applications, the uncertain coefficients depend on a set of parameters that can be perturbed within known bounds, thus leading to a *parametric interval linear system*.

Throughout the last years, parametric interval systems have received a lot of attention in the literature. Theoretical properties and characterizations of the feasible solution set have been studied by Alefeld et al. [1]. Several methods for computing enclosures of the feasible set have been proposed, mainly using an interval box as an inner or outer enclosure of the set [3, 8, 9, 11]. A parametric interval linear solver for Mathematica was introduced by Popova [6]. Another solver (for MATLAB and Octave) has been developed by Horáček [4].

Goal. In this paper, we propose a branch-and-prune method to compute an outer (and, if possible, inner, too) approximation of the weak solution set, via a union of interval boxes. The pruning part of the method has geometric foundations: the problem of testing feasibility of a given $x \in \mathbb{R}^n$ can be actually viewed as a problem whether a point is inside a polytope of a special class—inside a *zonotope*.

1.1 Parametric interval linear system

Let K be a given number of parameters, m be a given number of equations and n be a given number of variables. A *parametric interval linear system* (hereinafter also *PILS*) is the family of linear systems

$$\{A(p)x = b(p) \mid p \in [0, 1]^K\}, \tag{1}$$

where

$$A(p) = A^0 + \sum_{k=1}^{K} p_k A^k, \qquad b(p) = b^0 + \sum_{k=1}^{K} p_k b^k, \tag{2}$$

$A^0, \dots, A^K \in \mathbb{R}^{m \times n}$ are given matrices and $b^0, \dots, b^K \in \mathbb{R}^m$ are given vectors.

Definition 1.1 (Weak solutions of PILS)**.**

a) We say that a vector $x \in \mathbb{R}^n$ is a *(weak) solution* of PILS, if there exists $p \in [0,1]^K$ such that $A(p)x = b(p)$.

b) The *(weak) solution set* of PILS is the set $\Sigma := \{x \in \mathbb{R}^n \mid A(p)x = b(p),\ p \in [0,1]^K\}$.

From Definition 1.1 and (2), Observation 1.2 immediately follows:

Observation 1.2. *The test* "is a given $x^* \in \mathbb{R}^n$ a weak solution of PILS" *amounts to checking feasibility of a linear program in the form*

$$\left(A^0 + \sum_{k=1}^{K} p_k A^k\right) x^* = b^0 + \sum_{k=1}^{K} p_k b^k, \quad 0 \le p_k \le 1, \quad \forall k \in \{1, \dots, K\}. \tag{3}$$

However, it is not an easy task to describe the weak solution set of PILS: the weak solution set can be nonconvex or even disconnected. Nonconvexity is illustrated in the following example.

Example 1.3. Consider PILS with a single parameter given by the following data:

$$A(p) = \begin{pmatrix} p & 1 \\ -2 & 3p-1 \end{pmatrix}, \quad b(p) = \begin{pmatrix} 2p \\ 0 \end{pmatrix}, \quad p \in [0,1]. \tag{4}$$

The set of weak solutions Σ of system (4) is depicted in Figure 1 by the black curve. The set is nonconvex, and cannot be described as a finite union of polyhedra. The gray set is the outer approximation of Σ obtained by Algorithm 1 using the infeasibility test proposed by this paper.

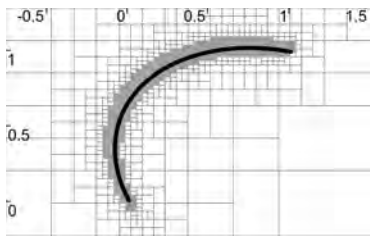


Figure 1: The weak solution set of PILS (4).

2 HULLS, ENVELOPES, UNIONS OF BOXES

As Example 1.3 suggests, it might be hard to obtain an explicit description of the solution set Σ that is easy to work with. The common approach is to approximate such a "complex" set with a "simpler" object. For general convex bounded sets, simpler objects of interests are e.g. ellipsoids, interval boxes (hypercubes), other regular polytopes (e.g. zonotopes) or random polytopes. In interval analysis, one usually approximates sets by the objects that are most natural for the intervalists: by interval boxes.

Interval boxes. Given two vectors $\underline{a}, \overline{a} \in \mathbb{R}^n$ with $\underline{a} \le \overline{a}$ (componentwise), we define an *interval box* or *interval vector* $\boldsymbol{a} = [\underline{a}, \overline{a}]$ as the set $\{a \in \mathbb{R}^n \mid \underline{a} \le a \le \overline{a}\}$. The set of all n-dimensional interval boxes is denoted by $\mathbb{IR}^n$.

Alternatively, we can specify an interval box in terms of the *center* $a^c \in \mathbb{R}^n$ and the *radius* $a^\Delta \in \mathbb{R}^n$. Obviously, the relation between a^c, a^Δ and $\underline{a}, \overline{a}$ is $a^c = \frac{1}{2}(\overline{a} + \underline{a})$, $a^\Delta = \frac{1}{2}(\overline{a} - \underline{a})$.

Interval approximations, envelopes and hulls. Given a bounded set $S \subset \mathbb{R}^n$, we define:

(a) an *(outer) interval approximation* or *interval envelope* of S is an interval vector $\boldsymbol{a} \in \mathbb{IR}^n$ satisfying $S \subseteq \boldsymbol{a}$,

(b) the *interval hull* of S, denoted by $\square S$, is the tightest interval envelope of S,

(c) an *inner interval approximation* or *inner interval envelope* of S is an interval vector $\boldsymbol{a} \in \mathbb{IR}^n$ satisfying $\boldsymbol{a} \subseteq S$.

Several methods of computing interval envelopes for PILS can be found in [3, 11, 9, 8].

Remark 2.1. The definition of an outer approximation of a set S requires S to be bounded. This can be overcome by some regularity assumption on the PILS instance itself, or by setting an initial interval box of interest in advance. In this paper, we assume that a PILS instance is constrained such that Σ is bounded.

Unions of interval boxes. An approximation of a set $S \subset \mathbb{R}^n$ using *one* interval box is not sufficient in some cases. This holds true also for PILS. Firstly, it is NP-hard to find the interval hull of the weak solution set of a *nonparametric* system (see [10]). Hence, only more or less tight envelopes can usually be computed in a limited computation time. Secondly, the approximation with just one box can be quite overestimated. For both of these reasons, the set of interest is often approximated as a *list of interval boxes*, whose union builds the sought approximation.

Typical approach to obtain an approximation via a list of boxes is to subdivide a PILS instance into instances in some sense smaller. There are basically two ways to build such smaller instances: by subdividing the interval box of parameters $\boldsymbol{p}$ into more smaller interval boxes and by subdividing the solution space to smaller regions, which are easier to say something about. This paper is an example of the latter approach.

3 BRANCH-AND-PRUNE OVER VARIABLES – SIVIA

Here, we make use of a general method for describing sets via a list of boxes: the so-called SIVIA algorithm [5]. SIVIA is an abbreviation for "set inversion via interval analysis". For a set S, SIVIA creates three lists of boxes of the following types:

- *list of infeasible boxes*; denoted by I; we have $S \cap \boldsymbol{a} = \emptyset$ for each infeasible box $\boldsymbol{a}$,
- *list of feasible* or *solution boxes*; denoted by F; we have $\boldsymbol{a} \subseteq S$ for each feasible box $\boldsymbol{a}$,
- *list of borderline boxes*; denoted by B; borderline boxes can contain both $x \in S$ and $x \notin S$.

Clearly, the union of boxes in F is an inner approximation of S, the union of boxes in $F \cup B$ is an outer approximation of S.

The basic SIVIA algorithm is formulated in Algorithm 1 in a recursive version. Every recursive call takes the following inputs:

- An interval box $\boldsymbol{x} \in \mathbb{IR}^n$ that shall be assigned to one of the lists I, B, F or subdivided to smaller boxes. At the very first call, this shall be an interval envelope of the set being described by SIVIA. For PILS, we assume that such an envelope is given as a part of the input (see Remark 2.1), however, several methods are available for computing such envelopes (see Section 2 or [3, 7, 9, 11]).
- An infeasibility test $i\colon \mathbb{IR}^n \to \{0,1\}$ and a feasibility test $f\colon \mathbb{IR}^n \to \{0,1\}$. These are used on Lines 2 and 3 to determine whether $\boldsymbol{x}$ shall be added to I or F. The infeasibility test returns 1 (true) if it finds out that $x \notin \Sigma$ for all $x \in \boldsymbol{x}$, otherwise it returns 0 (false), the feasibility test returns 1 if $x \in \Sigma$ for all $x \in \boldsymbol{x}$, otherwise it returns 0.
- A box-radius threshold $\varepsilon > 0$ determining the precision of the obtained description. If $\boldsymbol{x}$ passes neither the infeasibility nor the feasibility test, its size is checked: if the radii of all intervals of $\boldsymbol{x}$ are below the threshold, then $\boldsymbol{x}$ is considered sufficiently small and it is

Algorithm 1 The basic form of `SIVIA` algorithm based on inclusion tests.

Input: initial box $\boldsymbol{x} \in \mathbb{IR}^n$, infeasibility test i, feasibility test f, box-radius threshold $\varepsilon > 0$
Output: lists of boxes I, F, B as global variables

`SIVIA`$(\boldsymbol{x}, i, f, \varepsilon)$:
  if $i(\boldsymbol{x})$ **then** $I := I \cup \boldsymbol{x}$
    else if $f(\boldsymbol{x})$ **then** $F := F \cup \boldsymbol{x}$
      else if $\max_i x_i^\Delta < \varepsilon$ **then** $B := B \cup \boldsymbol{x}$
        else `SIVIA`$(L(\boldsymbol{x}), i, f, \varepsilon)$; `SIVIA`$(R(\boldsymbol{x}), i, f, \varepsilon)$

added to borderline boxes. Otherwise, it is subdivided into two smaller boxes $L(\boldsymbol{x})$ and $R(\boldsymbol{x})$ and `SIVIA` is recursively run for both these smaller boxes.

There are various strategies for obtaining the smaller boxes $L(\boldsymbol{x})$ and $R(\boldsymbol{x})$. One of the most natural strategy is to select the interval $\boldsymbol{x}_k$ satisfying $\max_i x_i^\Delta = x_k^\Delta$ and set

$$L(\boldsymbol{x}) = \left[\underline{x}, (\overline{x}_1, \ldots, \overline{x}_{k-1}, x_k^c, \overline{x}_{k+1}, \ldots, \overline{x}_n)\right], \quad R(\boldsymbol{x}) = \left[(\underline{x}_1, \ldots, \underline{x}_{k-1}, x_k^c, \underline{x}_{k+1}, \ldots, \underline{x}_n), \overline{x}\right].$$

Note that boxes in B have all radii below the threshold ε; such boxes are not split further in the course of the algorithm. On the other hand, splitting of some boxes (of boxes that are added to I and F) is avoided. The branch induced by such a box is *pruned*.

In Section 4, we propose infeasibility and feasibility tests that are both powerful and—especially for the infeasibility test—efficient.

4 ZONOTOPE-BASED PRUNING TESTS

Definition 4.1. Given a *generator matrix* $G \in \mathbb{R}^{m \times K}$ and a *shift* $s \in \mathbb{R}^m$, a *zonotope* is an image of the hypercube $[0,1]^K$ under an affine mapping $p \mapsto s + Gp$. Alternatively, a zonotope can be understood as the set $\{y \in \mathbb{R}^m \mid y = s + Gp,\ 0 \leq p \leq 1\}$.

Set $s(x) := A^0 x - b^0$ and let $G(x) \in \mathbb{R}^{m \times K}$ be a matrix defined as

$$G(x) := (A^1 x - b^1, A^2 x - b^2, \cdots, A^K x - b^K). \tag{5}$$

Now, system (1) can be rewritten as $\{s(x) + G(x)p = 0 \mid p \in [0,1]^K\}$. The test $x \in \Sigma$ can be performed as the test whether $0 \in Z(x)$, where $Z(x) := \{y \in \mathbb{R}^m \mid y = s(x) + G(x)p, p \in [0,1]^K\}$ is the zonotope with shift $s(x)$ and generator matrix $G(x)$. In accordance with Observation 1.2, such a test can be performed by linear programming: for a given vector $x^* \in \mathbb{R}^n$ it reduces to testing the feasibility of the linear program $s(x^*) + G(x^*)p = 0$, $0 \leq p_k \leq 1$, $\forall k \in \{1, \ldots, K\}$.

Balls and norms. The symbol $B(c, r) := \{x \mid \|x - c\| \leq r\}$ stands for the ball with radius r centered in c. For $x \in \mathbb{R}^n$, the symbol $\|x\|$ denotes a norm of x. In this paper, we assume one of the usual norms such that Euclidean, Manhattan or Chebyshev norm, however, the results in this section hold for any norm.

The idea of our pruning tests. We define

$$d^o(\boldsymbol{x}) := \min_{y \in Z(x^c)} \|y\|, \qquad d^i(\boldsymbol{x}) := \max_{r \in \mathbb{R}} \{r \mid B(0, r) \subseteq Z(x^c)\}. \tag{6}$$

Given an interval box $\boldsymbol{x}$, test whether $0 \in Z(x)$ for all $x \in \boldsymbol{x}$, or whether $0 \notin Z(x)$ for all $x \in \boldsymbol{x}$. We measure the sensitivity of points of $Z(x)$ with respect to changes of x varying over $\boldsymbol{x}$:

$$\max_{x\in\boldsymbol{x}} \max_{p\in[0,1]^K} \|s(x^c) + G(x^c)p - s(x) - G(x)p\| \tag{7}$$

$$= \max_{x\in\boldsymbol{x}} \max_{p\in[0,1]^K} \left\|A^0(x^c - x) + (A^1(x^c - x), \cdots, A^K(x^c - x))p\right\| \tag{8}$$

$$\leq \max_{p\in[0,1]^K} \left\||A^0|x^\Delta + (|A^1|x^\Delta, \cdots, |A^K|x^\Delta)p\right\| = \left\|\sum_{k=0}^{K} |A^k|x^\Delta\right\| =: \sigma(\boldsymbol{x}). \tag{9}$$

The interpretation of $\sigma(\boldsymbol{x}), d^o(\boldsymbol{x})$ and $d^i(\boldsymbol{x})$ is the following: any point of $Z(x^c)$ can move to a distance at most $\sigma(\boldsymbol{x})$ when considering all zonotopes $\{Z(x) \mid x \in \boldsymbol{x}\}$. The symbol $d^0(\boldsymbol{x})$ holds the distance of the zonotope $Z(x^c)$ from the origin. Finally, $d^i(\boldsymbol{x})$ is the distance of the origin from the boundary of $Z(x^c)$.

The idea of the forthcoming Theorem 4.2 is visualized in Figure 2.

Theorem 4.2. *Let $\boldsymbol{x} \in \mathbb{IR}^n$ be given.*

(a) Assume $0 \notin Z(x^c)$. If $d^o(\boldsymbol{x}) > \sigma(\boldsymbol{x})$, then $0 \notin Z(x)$ for all $x \in \boldsymbol{x}$.

(b) Assume $0 \in Z(x^c)$. If $d^i(\boldsymbol{x}) \geq \sigma(\boldsymbol{x})$, then $0 \in Z(x)$ for all $x \in \boldsymbol{x}$.

Proof. For (a), assume that $0 \notin Z(x^c)$ and $d^o(\boldsymbol{x}) > \sigma(\boldsymbol{x})$. Furthermore, suppose for contradiction that there exists an $x^* \in \boldsymbol{x}$ such that $0 \in Z(x^*)$. By definition of the zonotope $Z(x^*)$, we have $0 = s(x^*) + G(x^*)p^*$ for some $p^* \in [0, 1]^K$.

Then, by (7)–(9) and using the assumption $d^o(\boldsymbol{x}) > \sigma(\boldsymbol{x})$, we derive

$$\begin{aligned} &\|s(x^c) + G(x^c)p^* - s(x^*) - G(x^*)p^*\| \leq \sigma(\boldsymbol{x}) < d^o(\boldsymbol{x}) \\ &= \min_{p\in[0,1]^K} \|s(x^c) + G(x^c)p\| \leq \|s(x^c) + G(x^c)p^*\|. \end{aligned}$$

Since $s(x^*) + G(x^*)p^* = 0$ holds by the choice of p^*, this directly leads to the contradiction $\|s(x^c) + G(x^c)p^*\| < \|s(x^c) + G(x^c)p^*\|$. Proof of (a) is finished.

For (b), assume that $0 \in Z(x^c)$, $d^i(\boldsymbol{x}) \geq \sigma(\boldsymbol{x})$ and that $0 \notin Z(x^*)$ for some $x^* \in \boldsymbol{x}$. By the hyperplane separation theorem, we can strictly separate the zonotope $Z(x^*)$ and the point 0 by some hyperplane $H := \{y \in \mathbb{R}^n \mid a^T y = b\}$ with $b > 0$ and $a^T y \geq b$ for each $y \in Z(x^*)$.

Let $p' \in [0, 1]^K$ define a point of $Z(x^c)$ lying on the boundary of $B(0, d^i(\boldsymbol{x}))$ that is furthest from H. Again, we have $\|s(x^c) + G(x^c)p' - s(x^*) - G(x^*)p'\| \leq \sigma(\boldsymbol{x}) \leq d^i(\boldsymbol{x})$. However, by the choice of p' we also obtain $\|s(x^c) + G(x^c)p' - s(x^*) - G(x^*)p'\| > d^i(\boldsymbol{x})$. □

Theorem 4.2 leads to Corollary 4.3, showing how to implement the tests $i(\boldsymbol{x})$ and $f(\boldsymbol{x})$.

Corollary 4.3. *Let $\boldsymbol{x} \in \mathbb{IR}^n$ be given.*

(a) If $0 \notin Z(x^c)$ and $d^o(\boldsymbol{x}) > \sigma(\boldsymbol{x})$, then $\boldsymbol{x} \cap \Sigma = \emptyset$ and $i(\boldsymbol{x})$ shall return true.

(b) If $0 \in Z(x^c)$ and $d^i(\boldsymbol{x}) \geq \sigma(\boldsymbol{x})$, then $\boldsymbol{x} \subseteq \Sigma$ and $f(\boldsymbol{x})$ shall return true.

Complexity of the proposed tests. It remains to show how to compute $d^o(\boldsymbol{x})$ and $d^i(\boldsymbol{x})$. By (6), $d^o(\boldsymbol{x})$ can be computed as a linear program with Chebyshev or Manhattan norm and as a convex quadratic program with Euclidean norm. Hence the infeasibility test $i(\boldsymbol{x})$ can be solved efficiently in time polynomial in n, K and m. Unfortunately, computation of $d^i(\boldsymbol{x})$ seems to be much harder. It can be solved to optimality by inspection of $\binom{K}{m}$ facets of $Z(x^c)$, which, however, immediately yields exponential complexity in m. For a polynomially computable but approximative solution, an adaptation of Goffin's algorithm for zonotopes [2] can be used. The approximative solution provides only a lower bound on $d^i(\boldsymbol{x})$, which reduces the power of the feasibility test.

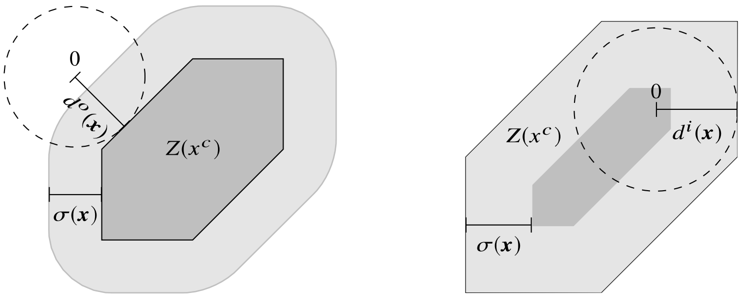


Figure 2: Visualization of the idea of the pruning tests using Euclidean norm

5 CONCLUSION

We proposed a new pruning test for the branch-and-prune algorithm for PILS, usable for pruning both the infeasible and feasible boxes. The infeasibility test can be efficiently performed in its full strength. This does not hold true for the feasibility test, where an efficient algorithm is known only for a weaker form. On the other hand, infeasibility tests are much more important for PILS, since there are instances with a solution set that is not fulldimensional, meaning that no feasible box can even exist.

Acknowledgements

The first author was supported by the Czech Science Foundation under Grant 17-13086S. The other authors were supported by the Czech Science Foundation under Grant 18-04735S. The second author was also supported by the Charles University project GA UK No. 156317.

References

[1] Alefeld, G., Kreinovich, V. and Mayer, G. (1998). The Shape of the Solution Set for Systems of Interval Linear Equations with Dependent Coefficients. Mathematische Nachrichten, 192(1): 23–36.

[2] Černý, M. and Rada, M. (2012). Polynomial time construction of ellipsoidal approximations of zonotopes given by generator descriptions. In LNCS, volume 7287, (pp. 156–163). Springer.

[3] Hladík, M. (2012). Enclosures for the solution set of parametric interval linear systems. International Journal of Applied Mathematics and Computer Science, 22(3): 561–574.

[4] Horáček, J. (2019). Interval toolbox - LIME. https://kam.mff.cuni.cz/~horacek/lime.html.

[5] Jaulin, L., Kieffer, M., Didrit, O. and Walter, É. (2001). Applied Interval Analysis. London: Springer London.

[6] Popova, E. D. (2004). Parametric interval linear solver. Numerical Algorithms, 37(1-4): 345–356.

[7] Popova, E. D. (2009). Explicit Characterization of a Class of Parametric Solution Sets.

[8] Popova, E. D. (2012). Explicit description of AE solution sets for parametric linear systems. SIAM Journal on Matrix Analysis and Applications, 33(4): 1172–1189.

[9] Popova, E. D. and Hladík, M. (2013). Outer enclosures to the parametric AE solution set. Soft Computing, 17(8): 1403–1414.

[10] Rohn, J. and Kreinovich, V. (1995). Computing Exact Componentwise Bounds on Solutions of Linear Systems with Interval Data is NP-Hard. SIAM Journal on Matrix Analysis and Applications, 16(2): 415–420.

[11] Skalna, I. (2018). Parametric Interval Algebraic Systems, volume 766 of *Studies in Computational Intelligence*. Cham: Springer International Publishing.

INTERIOR POINT HEURISTICS FOR A CLASS OF MARKET EXCHANGE MODELS

Anita Varga
Budapest University of Technology and Economics, Department of Differential Equations
H-1111 Budapest, Egry József street 1., Hungary
vanita@math.bme.hu

Marianna E.-Nagy
Budapest University of Technology and Economics, Department of Differential Equations
H-1111 Budapest, Egry József street 1., Hungary
enagym@math.bme.hu

Tibor Illés
Budapest University of Technology and Economics, Department of Differential Equations
H-1111 Budapest, Egry József street 1., Hungary
illes@math.bme.hu

Abstract: The Fisher type market exchange model (MEM) is a special case of the Arrow-Debreu type MEM. In this case, the players are divided into two groups, consumers and producers. Producers sell their products for money, and the consumers have an initial amount of money that they can use to buy a bundle of goods which maximizes their utility functions.

In the talk we present different interior point heuristics for the skew-symmetric weighted linear complementarity problem (WLCP) introduced by Potra in 2012. The Fisher type market exchange model can be considered as a special WLCP, and this way the new algorithms can also be applied to the Fisher type MEM. We also present our numerical results and compare them with the interior point algorithms introduced by Ye and Potra.

Keywords: convex optimization, market exchange models, interior point algorithms

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 6:

Multi-Criteria Decision-Making

EXPERIMENTAL EVALUATION OF MULTIPLE CRITERIA UTILITY MODELS WITH VETO RELATED PREFERENCE STRUCTURES

Andrej Bregar
Informatika d.d.
Vetrinjska ulica 2, 2000 Maribor, Slovenia
E-mail: andrej.bregar@informatika.si

Abstract: The paper investigates the properties of veto functions in multiple criteria decision models that incorporate utilities and discordance related preferential information. It completes a simulation study which aims to assess the influence of risk aversion on the decision by analysing and comparing the outcomes of risk averse, risk seeking and risk neutral veto functions in various settings that deal with ranking and sorting as well as with complete, weak and partial rank-orders. The obtained results indicate that the specification of veto can enhance the efficiency and credibility of decisions. Several quality factors improve, including the accuracy and validity of data, ability to discriminate optimal from suboptimal alternatives and robustness of judgements. It is shown that different forms of veto produce similar results as Rank Order Centroid (ROC) and Rank Sum (RS) surrogate weights, while fully compensatory utility models are unable to discriminate alternatives as efficiently.

Keywords: multiple criteria decision analysis, utility theory, veto, ranking, sorting, simulation study.

1 INTRODUCTION

A fundamental and widely applied approach to decision-making is the multi-attribute utility theory [7]. It aggregates preferences in the compensatory manner, in contrast to the school of outranking [9] which introduces the concepts of constructivism, incomparability and (partial) incompensation of preferences. These three phenomena are modelled with regard to the veto threshold and have been shown to be useful in various decision-making settings. Some ideas have therefore been expressed in the past to introduce the key concepts of outranking into the utility theory [10]. Based on these ideas, the utility theory has been extended with the concept of veto. Bregar et al. [5] have adopted the veto function from the outranking school in order to model full or partial non-compensation of unsatisfactory preferences in ranking and sorting problems. Almeida et al. have initially proposed a similar additive-veto approach only for the purpose of ranking [2], but have recently applied it to sorting as well [8].

Properties of such models have not been extensively and systematically studied yet. For this reason, it is the aim of the presented research work to study and analyse the underlying methodological foundations and approaches to express non-compensation in the utility based multi-criteria decision models, particularly with the focus on the veto criterion, veto function, aggregation operators and risk aversion of veto functions. Partial results of the experimental study have already been presented in the past [4], but the scope has been limited to ranking, additive aggregation and complete rank-orders only. In this paper, the completed research is addressed. Different problem solving problematics, aggregation models and types of rank-orders are considered. The simulation model is extended to deal with (1.) ranking and sorting, (2.) additive and multiplicative aggregation operators, and (3.) complete, weak and partial rank-orders. Based on the steepness, various shapes of utility and veto functions are assesed that exhibit different risk aversion characteristics, ranging from very/slightly risk seeking to neutral and slightly/very risk averse. Several specific scenarios are also defined to cope with mixed, uniform, conflicting and predominantly good or weak alternatives. The outcomes of risk averse, risk seeking and risk neutral veto functions are compared to two standard ordinal preference specification methods, i.e. ROC (*Rank Order Centroid*) and RS (*Rank Sum*) surrogate weights [6], because ROC weights, in particular, are considered to perform highly efficiently in terms of the selection of alternatives [1]. To ensure the representativeness of simulation results with regard to the characteristics of decision models that combine utility and veto, several evaluation factors are observed that pertain to the standard referential framework for the assessment of multi-criteria decision methods and systems [3].

The rest of the paper is organised as follows. In Section 2, the key theoretical foundations of multi-attribute utility models with veto related preference structures are summarized. In Section 3, the simulation based experimental setting is defined. Sections 4 and 5 present the results of the simulation study and analyse the properties of models for the problematics of ranking and sorting, respectively. Finally, Section 6 concludes the paper.

2 THEORETICAL FOUNDATIONS

Three fundamental principles of outranking may be applied to the utility theory [10]:

- *Constructivism* assumes the bounded rationality in which knowledge is subjected to limitations. Because judgements are neither totally defined nor stable, people are not required to obey axioms that pertain to the dogma of rational behaviour.
- *Incompensation* utilizes the veto threshold or function to eliminate alternatives that are unacceptably poor with regard to a certain criterion. The original multi-attribute utility function cannot prevent the selection of intolerable alternatives because it is fully compensatory, so that the deficiency of an alternative on one criterion can be compensated by advantages on other criteria.
- *Incomparability* assumes that it may be impossible to state for a pair of alternatives which one of them is preferred. The rationale is that evaluations can be conflicting as they are subjected to imprecision, indetermination and uncertainty.

Based on these principles, the veto criterion is modelled in accordance with the underlying concepts of the utility theory. The veto function $v_j(x)$ is specified by obeying the formal axiomatized approach of certain equivalence, so that criterion-wise values of alternatives are monotonously projected to the [0, 1] interval. The best unacceptable value is assigned the maximal veto of $v_j(x') = 1$, while the worst still acceptable value is not subjected to veto, and is hence assigned $v_j(x'') = 0$. These extreme points determine the standard preference lottery L_S. The veto degrees of values $x'' > x > x'$ are derived with a sequence of iterative steps, such that for each value x the decision-maker is indifferent between x and L_S with the probability of p_S. In this way, a linear and transitive total order is obtained as is depicted on Figure 1 in the general form for the maximized criterion and the risk seeking veto function. The domain of veto function is generally not bound by x' and x'', but by the lower and upper limits of x_L and x_U. If $v_j(x) = 1$, total incompensation occurs (strict veto). If $0 < v_j(x) < 1$, incompensation is partial (weak veto).

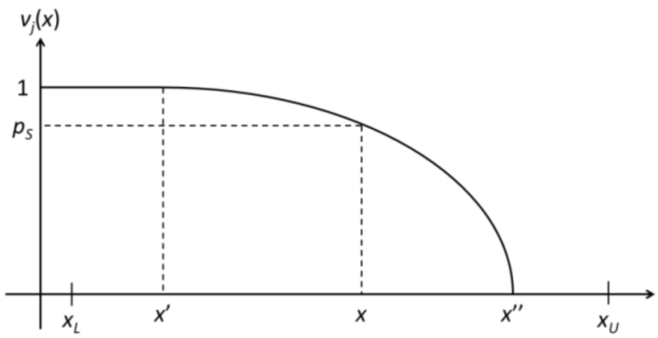


Figure 1: Specification of the veto function

The concept of synthesizing concordance and discordance in ELECTRE type methods [9] is adopted to aggregate utility and veto. This allows the effect of veto to reflect incompensation on the global level. The operator (3) that aggregates compensatory utilities from equation (1) with noncompensatory discordance information from equation (2) multiplies the total utility with the product of inverse veto degrees:

$$u(a_i) = \sum_{j=1..n} w_j u_j(a_i), \tag{1}$$

$$\tilde{v}(a_i) = \prod_{j=1..n} \left(1 - v_j(a_i)\right), \tag{2}$$

$$\sigma(a_i) = u(a_i)\tilde{v}(a_i). \tag{3}$$

Based on the $\sigma(a_i)$, $u(a_i)$ and $\tilde{v}(a_i)$ values, alternatives may be rank ordered or sorted into predefined classes. A partial rank order is obtained according to equations (4) to (6). In the case of sorting, an alternative may be either assigned to a certain class according to equation (7) or incomparable with this class with regard to equation (8):

$$aPb \Leftrightarrow \left(u(a) > u(b)\right) \wedge \left(\tilde{v}(a) > \tilde{v}(b)\right), \tag{4}$$

$$aIb \Leftrightarrow \left(u(a) = u(b)\right) \wedge \left(\tilde{v}(a) = \tilde{v}(b)\right), \tag{5}$$

$$aRb \Leftrightarrow \left(u(a) > u(b) \wedge \tilde{v}(a) < \tilde{v}(b)\right) \vee \left(u(a) < u(b) \wedge \tilde{v}(a) > \tilde{v}(b)\right), \tag{6}$$

$$a \in C \Leftrightarrow \left(\sigma(a) \geq u^-(C)\right) \wedge \left(\sigma(a) < u^+(C)\right), \tag{7}$$

$$\forall C\colon aRC \Leftrightarrow \left(u(a) \geq u^-(C)\right) \wedge \left(\sigma(a) < u^-(C)\right). \tag{8}$$

In equations (1) to (8), $\sigma(a_i)$ represents the overall evaluation of the i-th alternative, $u(a_i)$ total utility, $\tilde{v}(a_i)$ total inverse veto, $u_j(a_i)$ partial criterion-wise utility, $v_j(a_i)$ criterion-wise veto and w_j criterion weight. Alternatives are denoted with a, b and a_i, while C represents a single ordered category/class. Its upper and lower limits are $u^+(C)$ and $u^-(C)$, respectively. Finally, P, I and R denote the relations of preference, indifference and incomparability.

Several distinctions between the ordinary criterion and the veto criterion can be observed. The first has a relative compensatory effect, exhibits positive characteristics that should be maximized, and is modelled locally on various hierarchical levels of the criteria structure. The latter has an absolute noncompensatory effect, shows negative characteristics that should be minimized, and is modelled globally on the highest level. Details on the comparison are available in the literature [4, 5].

3 SIMULATION BASED EXPERIMENTAL SETTING

The experimental model extends and completes the model from our previous research study [4]. In this follow up study, different problem solving problematics, aggregation models and types of rank-orders are considered. The model is hence extended to deal with (1.) ranking and sorting, (2.) additive and multiplicative aggregation operators, and (3.) complete, weak and partial rank-orders. To ensure the representativeness of simulation results with regard to the characteristics of decision models that combine utility and veto, several evaluation factors are observed that pertain to the standard referential framework for the assessment of multi-criteria decision methods and systems [3].

The *accuracy and validity of results* is assessed based on the assumption that alternatives must be efficiently discriminated. It is therefore observed to what extent alternatives differ in preferability. The higher the differences between evaluations of alternatives are, the richer is the discriminating information. It ensures that a small subset of optimal alternatives can stand out. It is essential that evaluations are rich enough, but not too extreme. Several metrics are applied, such as minimal and maximal assessments of alternatives, distance from the best to the second best alternative and distance from the best alternative to all other alternatives.

The second observed evaluation factor is the *robustness*. The main issue is, whether the robustness deteriorates when veto functions are introduced. In addition to distances between alternatives, perturbations in rank orders of alternatives and in assignments of alternatives to categories are measured with two metrics. The weighted distance calculates cardinal ranks, while the Kemeny-Snell distance quantifies relations with the values of −1, 0 and 1.

The simulation model has the following input parameters:

- The number of utility criteria is fixed to 10, while the number of veto criteria is set to 2 or 5. Similarly, the number of alternatives may be 5 or 10.
- All criteria are maximized on the fixed interval of $[0 \dots 100]$.
- Criteria-wise values of alternatives are sampled in each simulation trial from the uniform probability distribution on the $[0 \dots 100]$ interval.
- Limits of utility and veto functions are sampled from the uniform probability distributions, such that $u_{min} \in [0 \dots 40]$, $u_{max} \in [60 \dots 100]$, $v_{min} \in [0 \dots 20]$, $v_{max} \in [30 \dots 50]$, $u_{min} < u_{max}$, $v_{min} < v_{max}$ and $v_{max} < u_{max}$.
- Shapes of utility and veto functions reflect the risk aversion characteristics. They are determined by the exponent or linear coefficient, which implies the steepness.
- The decision-making problematics may pertain to ranking or sorting.
- Complete/weak (preference and indifference only) or partial (incomparability) rank-orders are inferred.

Based on the steepness, various shapes of utility and veto functions are assessed that exhibit different risk aversion characteristics, ranging from very/slightly risk seeking to neutral and slightly/very risk averse. Figure 2 shows the exponential veto functions which are aggregated in the simulation model. These functions are scaled in each simulation trial according to the sampled v_{min} and v_{max} thresholds. Linear functions are not presented on this figure although they are used in the experiments as well.

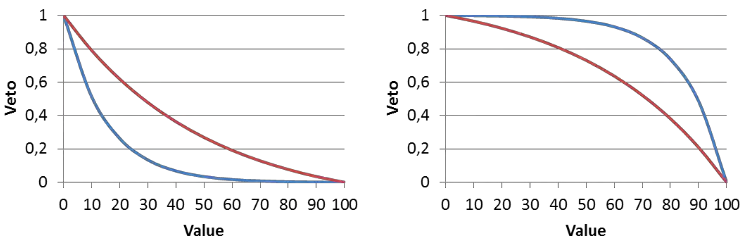


Figure 2: Risk averse (left) and risk seeking (right) veto functions

Because the simulation model is relatively complex, several specific scenarios are defined to reduce the complexity and to cope with mixed, uniform, conflicting and predominantly good or weak alternatives. The scenarios are:

- scenario 1: 5 alternatives, 10 utility criteria, 5 veto criteria;
- scenario 2: 10 uniform alternatives, 10 utility criteria, 2 veto criteria;
- scenario 3: 10 uniform alternatives, 10 utility criteria, 5 veto criteria;
- scenario 4: 10 mixed alternatives (very good, moderately good, uniform, moderately weak and conflicting), 10 utility criteria, 5 veto criteria;
- scenario 5: 10 predominantly good alternatives, 10 utility criteria, 5 veto criteria;
- scenario 6: 10 conflicting alternatives (they perform well according to a small subset of criteria and poorly according to a disjunctive subset of other criteria), 10 utility criteria, 5 veto criteria.

Since it is difficult to objectively assess the efficiency of approaches without an appropriate benchmark, the outcomes of risk averse, risk seeking and risk neutral veto functions are compared to two standard ordinal preference specification methods – ROC and RS surrogate weights. These weights are computed with equations (9) and (10), respectively [6]:

$$w_i(ROC) = \frac{1}{n}\sum_{j=i..n}\frac{1}{j}, \tag{9}$$

$$w_i(RS) = \frac{2(n+1-i)}{n(n+1)}. \tag{10}$$

4 PROPERTIES FOR THE CASE OF RANKING

Figure 3 shows the average evaluations of alternatives with regard to different ranks. The *x*-axis presents the ranks of 10 alternatives, while the *y*-axis refers to the overall evaluation of each alternative, which is either based strictly on utilities or combines (aggregates) criterion-wise utilities and degrees of veto. It can be observed that the effect of veto is reflected in the assessments of alternatives. The risk seeking veto clearly exhibits the highest discrimination power and consequently also the most intense extremeness. Particularly in scenario 3, only the best ranked alternative is good enough to be selected for implementation. Veto always causes a deterioration of alternatives compared to strict utility, regardless of the risk aversion form.

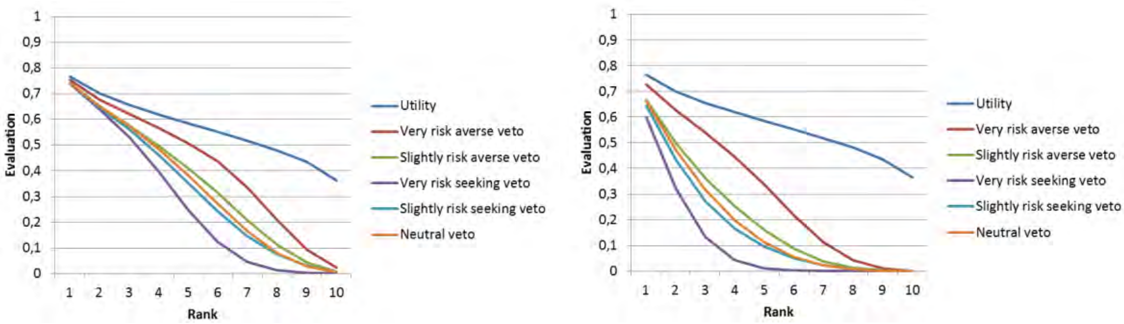


Figure 3: Average evaluations of alternatives in different ranks for scenario 2 (left) and scenario 3 (right)

A more credible interpretation of the effect of veto may be given if the results are compared to a standard benchmark. It is shown on Figure 4 that very risk averse veto produces similar results as RS weights, while other forms of veto approach ROC weights. On the other hand, fully compensatory (utility only) models are unable to discriminate alternatives as efficiently as either ROC or RS weights.

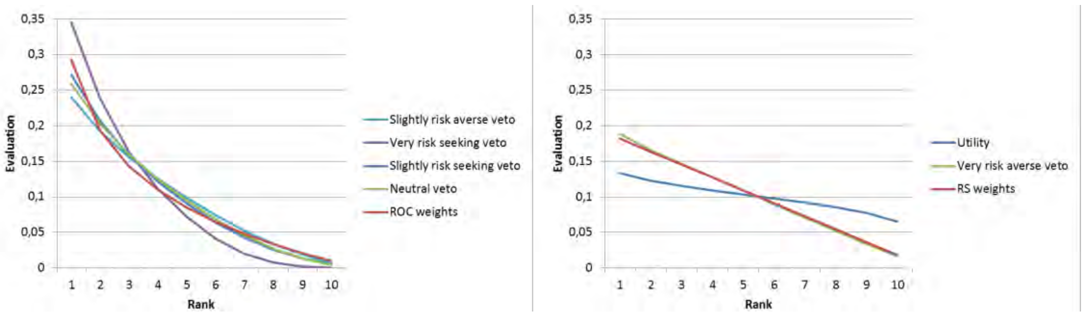


Figure 4: Comparison to ROC weights (left) and RS weights (right)

Figure 5 indicates that the evaluation of the optimal alternative is considerably more robust if the non-compensatory veto based preferential information is incorporated in the decision model than in the case when only fully compensatory utility functions are modelled. This is in accordance with the perturbations in rank orders of alternatives that occur when veto functions are introduced. The Kemeny-Snell distances between fully compensatory (utility only) and partially non-compensatory (utility and veto combined) rank orders are 0.290 and 0.350 for the risk averse and risk seeking veto, respectively. The weighted distances are similar with the values of 0.267 and 0.404. These differences appear reasonable considering the fact that new preferential information is added. As expected, they are the highest for the risk seeking veto.

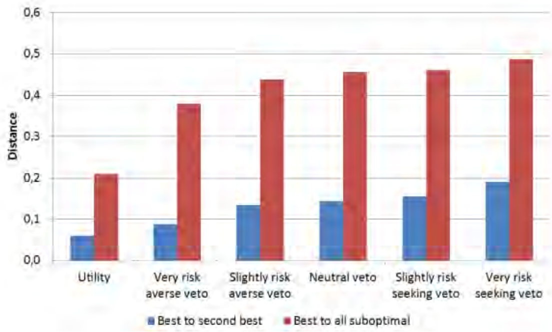


Figure 5: Average distances between evaluations of alternatives for the case of ranking

Partial rank-orders with incomparability relations are also considered in the simulation study. The results are presented on Figure 6. It can be seen from the incomparability count indicator on the left graph that the number of incomparability relations increases to a certain point as ranks deteriorate. From this point onward, the trend reverses so that incomparabilities begin to decrease again. The interpretation is that the best ranked alternative exhibits the fewest incomparabilities. This is an important strength that clearly separates the optimal alternative from the other suboptimal ones. The worst ranked alternative also appears to be in relatively few incomparability relations. This is a consequence of a high veto to which it is subjected as is evident from the right graph. It is hence strongly inferior to all other alternatives, but it is interesting to note that the worst alternative is mainly incomparable to the best ranked ones.

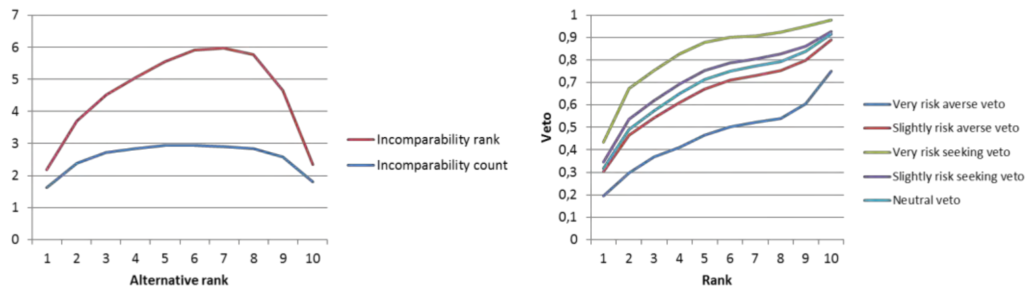


Figure 6: Average incomparability indicators (left) and average veto (right)

5 PROPERTIES FOR THE CASE OF SORTING

In the case of sorting decision-making problematics, a similar pattern may be observed as in the case of ranking. Models that incorporate veto functions with various risk attitudes have a higher discriminating power than models that are based on utility functions only. When veto is applied, one or two optimal alternatives are clearly and unambigously assigned to the best or second best category, depending on the simulation scenario. They stand out in comparison to other suboptimal alternatives, which quickly deteriorate and are mostly sorted into several worst classes. The more risk seeking the veto is, the more evident this characteristic becomes. In contrast, alternatives are sorted very similarly into a few adjacent categories if the decision model considers and aggregates only utilities. This characteristics can be derived from Figure 7. Moreover, only a couple of good or average categories are predominantly occupied by the utility based assessments according to Figure 8. On the other hand, the assignments are more uniformly distributed when veto is applied. One or two alternatives are sorted into each class in the latter case, which makes the decision considerably easier and more transparent. The exception is the worst category that contains a higher number of alternatives. However, these are unacceptable alternatives which may be objectively disregarded by the decision-maker.

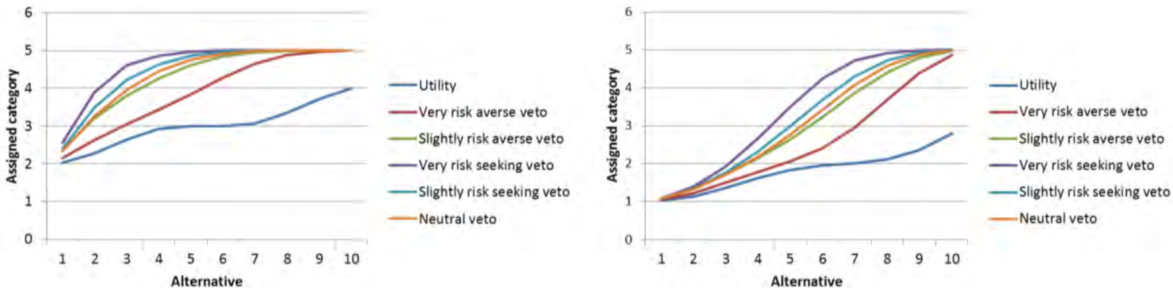


Figure 7: Categories of alternatives for scenario 3 (left) and scenario 5 (right)

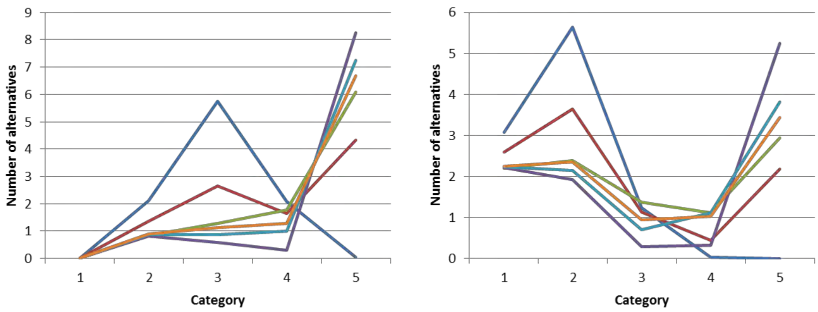


Figure 8: Average number of alternatives in different categories for scenario 3 (left) and scenario 5 (right)

Figure 9 presents the average distances between the category of the best alternative and the categories into which the other suboptimal alternatives are sorted. It can be concluded that the decision is more robust if veto is modelled and aggregated into the overall evaluations of alternatives.

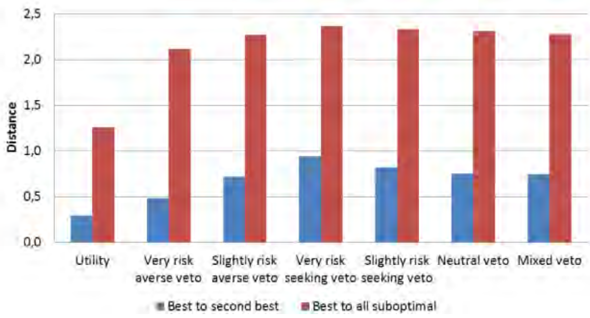


Figure 9: Average distances between categories of alternatives for the case of sorting

6 CONCLUSION

The main contribution of the presented work pertains to the investigation and evaluation of properties of veto functions in multi-criteria decision models that incorporate utilities as well as discordance related information. The obtained results indicate that the specification of veto can enhance the efficiency and credibility of decisions irrespective of the decision-making problematics. Several quality factors improve, such as the accuracy and validity of results, ability to discriminate optimal from suboptimal alternatives, and robustness of judgements. This is a consequence of the fact that additional preferential information on veto structures increases the expressiveness and completeness of quantitative models.

References

[1] Ahn, B. S. 2011. Compatible weighting method with rank order centroid: Maximum entropy ordered weighted averaging approach. European Journal of Operational Research, 212(3): 552–559.

[2] Almeida, A. T. 2013. Additive-veto models for choice and ranking multicriteria decision problems. Asia-Pacifc Journal of Operational Research, 30(6): 1–20.

[3] Bregar, A. 2014. Towards a framework for the measurement and reduction of user-perceivable complexity of group decision-making methods. International Journal of Decision Support System Technology, 6(2): 21–45.

[4] Bregar, A. 2018. Decision support on the basis of utility models with discordance-related preferential information: investigation of risk aversion properties. Journal of Decision Systems, 27(sup1): 236–243.

[5] Bregar, A., Gyorkos, J., Jurič, M. B. 2007. Applying the noncompensatory veto effect to the multiple attribute utility function. In Novaković, A., Bajec, M., Poženel, J., Indihar Štemberger, M. (Eds.). Proceedings of 14th conference Dnevi slovenske informatike (10 pp.). Portorož: Slovenian Society Informatika.

[6] Danielson, M., Ekenberg, L. 2017. A robustness study of state-of-the-art surrogate weights for MCDM. Group Decision and Negotiation, 26(4): 677–691.

[7] Keeney, R. L., Raiffa, H. 1993. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: Cambridge University Press.

[8] Palha, R. P., Almeida, A. T., Morais, D. C., Hipel, K. W. 2019. Sorting subcontractors' activities in construction projects with a novel additive-veto sorting approach. Journal of Civil Engineering and Management, 25(4): 306–321.

[9] Roy, B. 1996. Multicriteria methodology for decision aiding. Dordrecht: Kluwer Academic Publishers.

[10] Stewart, T. J., Losa, F. B. 2003. Towards reconciling outranking and value measurement practice. European Journal of Operational Research, 145(3): 645–659.

MULTI-ATTRIBUTE RISK ASSESSMENT MODEL FOR DEVELOPING VENTILATOR-ASSOCIATED PNEUMONIA

Rok Drnovšek
University Medical Centre Ljubljana,
Zaloška cesta 2, 1000 Ljubljana, Slovenia
E-mail: rok.drnovsek@kclj.si

Marija Milavec Kapun
University of Ljubljana, Faculty of Health Sciences,
Zdravstvena pot 5, 1000 Ljubljana, Slovenia
E-mail: marija.milavec@zf.uni-lj.si

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

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

Abstract: Ventilator-associated pneumonia is associated with avoidable costs and negative patient outcomes. In this paper, a qualitative multi-attribute risk assessment model for ventilator-associated pneumonia is presented. The model is based on the DEX methodology and was developed by a group of clinical experts. It was preliminarily tested by nurses to evaluate 10 patients in a hospital setting. In the future we intend to test the proposed model extensively on adequate number of practical cases. The model detects susceptible patients or poor implementation of recommended preventive measures. It helps to identify good practices and vulnerable patients. This way it may help ensure better distribution of resources and staff for effective ventilator-associated pneumonia prevention strategies.

Keywords: decision support, health care, ventilator-associated pneumonia, DEX

1 INTRODUCTION

Ventilator-associated pneumonia (VAP) is one of the most common infections in intensive care units, that is followed by urinary tract infections and bloodstream infections associated with central venous catheters [7]. The infection of the lungs that develops at least 48 hours after the intubation or beginning of mechanical ventilation is to be classified as VAP [10, 11]. Acquiring VAP is closely associated with microaspiration of oropharyngeal and gastric secretions, making the use of appropriately designed tracheal tubes, adequate cuff pressure management techniques and subglottic drainage important strategies in prevention of VAP in intensive care units [16, 20].

Besides cost associated with prolonged stay in intensive care units of about 5 to 7 days [12], VAP is also associated with increased mortality of hospitalized patients. A large-scale record analysis study that included 4479 patients estimated that in intensive care units somewhere between 1.6 – 7.0 % of deaths on day 30 and between 2.5 - 9.1% of deaths on day 60 can be attributed to VAP [3], however studies vary in results. Estimated reports of mortality attributed to VAP range from 20 – 76%, while VAP mortality related to resistant bacteria are estimated even higher. Reasons for discrepancies could be a lack of uniform methodological research approaches and rigorous diagnosis protocols [13].

Efforts for accurate and timely diagnosis of VAP are important to reduce associated mortality and additional costs. Therefore, numerous strategies have been employed for more effective VAP surveillance in clinical environments. Regular monitoring of specific biomarkers like C reactive protein (CRP) and procalcitonin (PCT) are used in diagnosing and predicting VAP [15, 18, 19]. Data analysis gathered by modern electronic documentation systems offers a novel approach in monitoring various healthcare-associated infections in intensive care units [6]. For example, a protocol developed in 2013 by a National Healthcare Safety Network work group (NHSN) for daily VAP surveillance can be electronically implemented. The NHSN protocol was designed to sidestep subjectivity and shift the focus away from VAP to include other relevant ventilator-associated complications in order to simplify the reporting process and reduce variability in VAP surveillance [21]. Parameters such as a minute-to-minute ventilator settings, antibiotic use, microbiology data, and clinical characteristics are used in electronically supported VAP surveillance [14].

Nurses in intensive care units regularly monitor parameters associated with increased risk of developing VAP. However, no significant body of research was found regarding the systemic use of nursing specific data for systematic prevention and timely discovery of VAP.

The aim of presented study was to develop a multi-attribute decision model to assess risk of VAP of an individual patient by continuously evaluating consistency of preventive measures conducted by staff and to identify risk factors for VAP acquisition. The approach also embraces nursing specific data to enable better integration of nursing science and enhance interdisciplinary cooperation needed for tackling complex modern health care problems such as VAP.

2 METHODS

A systematic review of literature was used to identify the most crucial studies and guidelines related to VAP risk factors and corresponding preventive measures. A synthesis of identified VAP care bundles was done and is in depth described elsewhere [8]. In our previous work these findings were included in a multi-attribute risk assessment model heavily reliant on consistency of preventive measures implementation [9]. In this paper we aim to improve this concept by including additional necessary attributes to identify patients that are more susceptible to VAP, even in cases when all necessary preventive measures are taken.

Parameters that were identified as relevant were included in the risk assessment model using the DEX methodology. The designed model uses qualitative – descriptive value domains to describe evaluation of patients.

DEX methodology belongs to the multi-attribute utility theory. Attributes have qualitative value domains and therefore such models classify alternatives into classes. Attributes are structured into a hierarchical tree. Options are firstly described by leaves in tree of attributes. Secondly, the values of aggregated attributes, which lie above them, are calculated according to their utility functions, which are presented as a set of simple if-then rules. This differs from the usual weighting sum models in a way, that weights of attributes are not pre-fixed, but may depend on the values of attributes. For example, a very negative value may be more important than positive values of the same attribute. This methodology is included in a MS Windows based software titled DEXi, that was used in our study. [5, 4]

The presented risk assessment model includes risk factors and preventive measures as attributes. The goal of the risk assessment process is to evaluate the overall risk of developing VAP for a specific patient.

3 RESULTS

The designed multi-attribute risk assessment model consists of two main sections for assessing preventive measures and patient related attributes. The tree of attributes has 17 final attributes and 11 aggregated attributes. Hierarchical structure of the designed model is presented in Figure 1.

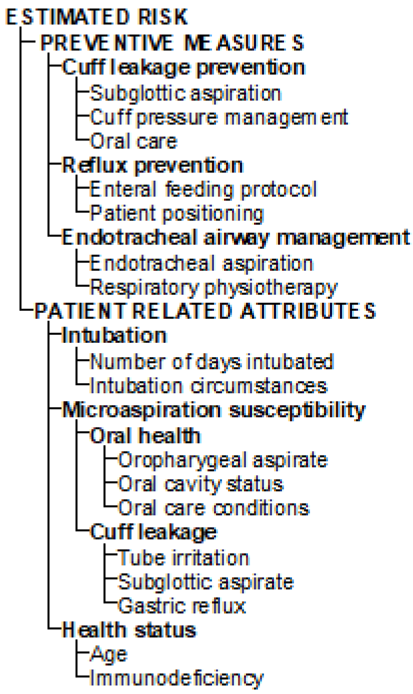


Figure 1: Tree of attributes of a multi-attribute risk assessment model for developing VAP.

Since not all attributes contribute to the estimated risk equally, utility functions for individual attributes were assigned. Overall estimated risk for development of VAP is derived from the level of preventive measures implementation and patient specific attributes that contribute to increased risk for developing VAP. Figure 2 represents how estimated risk differs if stricter preventive measures are implemented for the same patient. It considers patient-specific data and its evaluation and shows the expected added value of stricter measures.

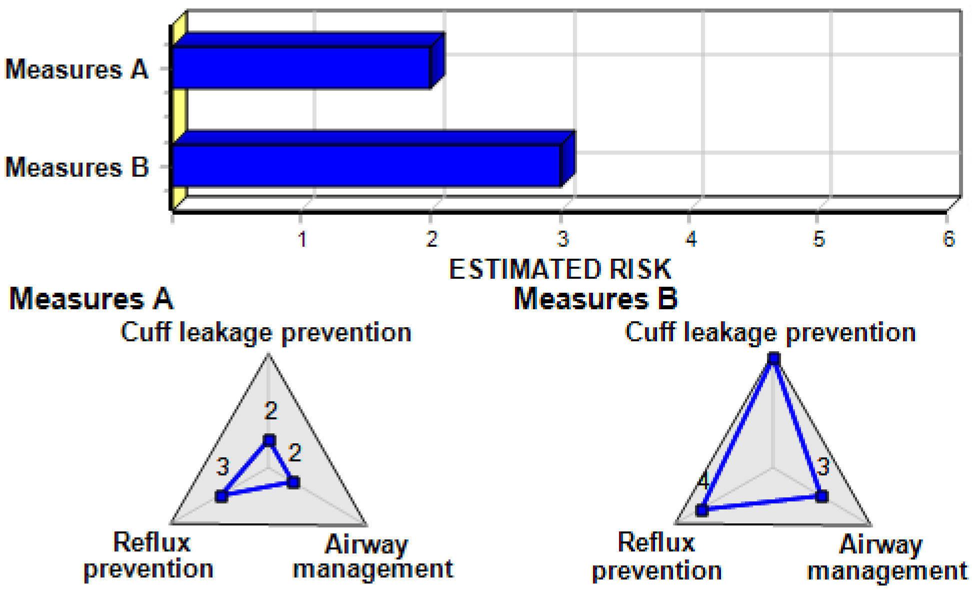


Figure2: Estimation of risk for two patients, with identical patient related attributes, that differ in implemented preventive measures attributes.

Similarly, the designed model can depict differences among patients that differ in their own specific susceptibility. Figure 3 represents how estimated risk between two patients differs based on patients' related factors, while identical preventive measures are implemented. This use of the model helps us identify patients with higher risk for developing VAP in a setting of comparable preventive measures. The designed multi-attribute risk assessment model can distinguish among different levels of estimated risk even when uniform preventive measures are implemented. It can be used to estimate differences in the risk for developing VAP and to identify most susceptible patients.

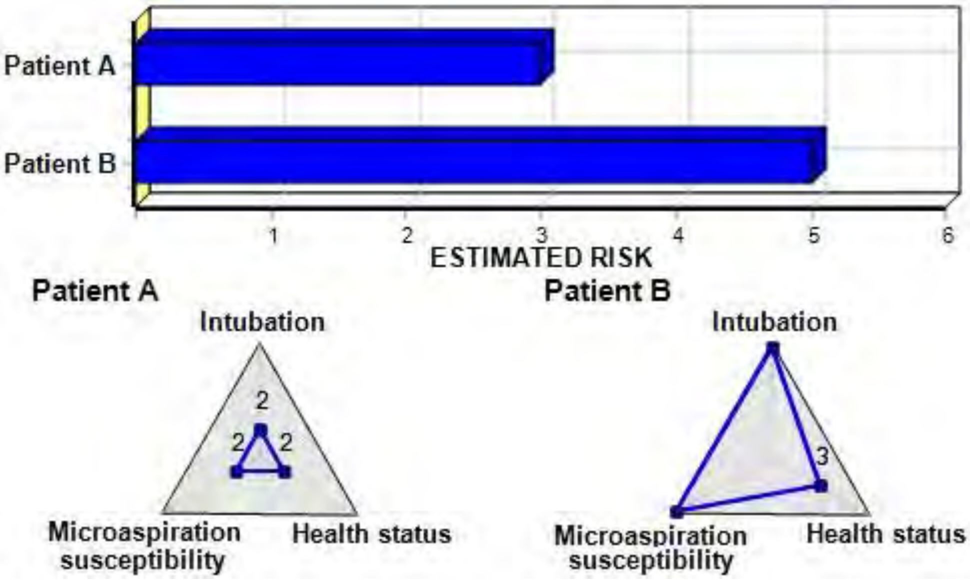


Figure 3: Estimation of risk for two patients with identical implementation of preventive measures attributes that differ in their patient related attributes.

The designed multi-attribute decision aims to distinguish different levels of estimated risk even when uniform preventive measures are implemented to identify patients most with highest risk of developing VAP.

4 DISCUSSION

Preventing and monitoring VAP in intensive care units is vital for prevention of avoidable costs and decreasing patient mortality. Significant effort is being invested in developing VAP prevention strategies that aim to maximize preventive behaviour among staff. Although necessary, this approach has limited effect on account of difficult implementation and low staff adherence. Educational activity to promote preventive measures can lead to better implementation, however some studies report limited effects and stress the importance of workload reduction [1]. Our study focuses on identifying patients with higher risk of developing VAP that could prevent infections and support more effective VAP surveillance.

Contemporary approaches in VAP preventive measures are focused on bundle implementation approach that assumes identical lists of preventive measures to be implemented universally, not specific to individual patients [2, 17]. Our previous efforts to identify susceptible patients was heavily reliant on identifying weaknesses in preventive measures implementation. This approach is subsequentially appropriate for identifying poor preventive measure practices but less sensitive in identifying high risk patients. The newly presented model considers both preventive measures implementation and patient specific attributes. The model can therefore be more beneficial for detecting susceptible patients, when identical or similar preventive measures are implemented. Furthermore, the structure of the model can be used to identify specific attributes that contribute to individuals' increased susceptibility. This information is vital for targeted activities for reducing the risk of VAP.

Data required for risk estimation may be collected daily with minimal additional effort, since the model is heavily reliant on nursing care specific observations. Although manual input of data is sufficient, automatic gathering and analysis of data in collusion with modern electronic documentation practices is optimal. This way nurses and physicians could gain access to a daily in dept assessment of risk and adjust patient's medications or planed nursing care interventions to rationally and efficiently improve patient care.

This is a novel approach that could promote a more continuous VAP surveillance and include previously poorly utilized data. At this stage of model development, empirical data was not yet analysed to determine usefulness in clinical environment, which is the main limitation of our study. Future work on this topic should therefore be focused on empirical research and decision model modifications to ensure adequate validity. It is also important to consider that VAP prevention is a uniquely interdisciplinary field that should include continuous efforts of nurses, medical professionals and respiratory physiotherapists to ensure optimal results.

5 CONCLUSION

The presented multi-attribute decision model is focused on nursing specific data and presents a novel method in VAP surveillance. Simplicity of data gathering enables continuous VAP surveillance with reasonable additional effort. Empirical research and interdisciplinary cooperation should be applied to fully develop the potential of this approach.

References

[1] Aloush, S.M. 2017. Does educating nurses with ventilator-associated pneumonia prevention guidelines improve their compliance?. American Journal of Infection Control, 45(9): 969–973.

[2] Álvarez Lerma, F., Sánchez García, M., Lorente, L., Gordo, F., Añón, J.M., Álvarez, J., Palomar, M., García, R., Arias, S., Vázquez-Calatayud, M. Jam, R. 2014. Guidelines for the prevention of ventilator-associated pneumonia and their implementation. The Spanish "Zero-VAP" bundle. Medicina Intensiva, 38(4): 226–236.

[3] Bekaert, M., Timsit, J.F., Vansteelandt, S., Depuydt, P., Vésin, A., Garrouste-Orgeas, M., Decruyenaere, J., Clec'h, C., Azoulay, E., Benoit, D. 2011. Attributable Mortality of Ventilator-Associated Pneumonia. American Journal of Respiratory and Critical Care Medicine, 184(10): 1133–1139.

[4] Bohanec, M., Žnidaršič, M., Rajkovič, V., Bratko, I., Zupan, B. 2013. DEX methodology: three decades of qualitative multi-attribute modeling. Informatica, 37: 49–54.

[5] Bohanec, M. 2017. DEXi: a program for qualitative multi-attribute decision modelling. Jožef Stefan Institute. https://kt.ijs.si/MarkoBohanec/dexi.html [3/6/2019].

[6] Cato, K.D., Cohen, B., Larson, E. 2015. Data elements and validation methods used for electronic surveillance of health care-associated infections: A systematic review. American Journal of Infection Control, 43(6): 600–605.

[7] Dasgupta, S., Das, S., Chawan, N.S., Hazra, A. .2015. Nosocomial infections in the intensive care unit: Incidence, risk factors, outcome and associated pathogens in a public tertiary teaching hospital of Eastern India. Indian Journal of Critical Care Medicine, 19(1): 14–20.

[8] Drnovšek, R., Dežman, A., Marolt, A., Starc, A., 2017. Pregled preventivnih ukrepov za preprečevanje ventilatorske pljučnice. In Majcen Dvoršak, S., Štemberger Kolnik, T., Kvas, A. (Eds.). Kongres zdravstvene in babiške nege Slovenije. Medicinske sestre in babice – ključne za zdravstveni sistem: zbornik prispevkov z recenzijo (pp. 437–444). Ljubljana: Ljubljana : Zbornica zdravstvene in babiške nege Slovenije - Zveza strokovnih društev medicinskih sester, babic in zdravstvenih tehnikov Slovenije.

[9] Drnovšek, R., Rupar, T., Milavec Kapun, M., Rajkovič, V., 2018. Multi-attribute decision model for preventing ventilator-associated pneumonia. In Mileva-Boshkoska, B., Bohanec, M., Žnidaršič, M., (Eds.). Proceedings of the 19th Open Conference of the IFIP WG 8.3 on Decision Support Systems (IFIP DSS 2018) "DSS Research Delivering High Impacts to Business and Society". (pp. 63–64). Ljubljana: Jožef Stefan Institute.

[10] Feng, D.Y., Zhou, Y.Q., Zou, X.L., Zhou, M., Zhu, J.X., Wang, Y.H., Zhang, T.T. 2019. Differences in microbial etiology between hospital-acquired pneumonia and ventilator-associated pneumonia: a single-center retrospective study in Guang Zhou. Infection and drug resistance, 12: 993–1000.

[11] Gupta, R., Sharma, S., Parwez, Saxena, S. 2018. Changing panorama for surveillance of device-associated healthcare infections: Challenges faced in implementation of current guidelines'. Indian Journal of Medical Microbiology, 36(1): 18–25.

[12] Hillier, B., Wilson, C., Chamberlain, D., King, L. 2013. Preventing Ventilator-Associated Pneumonia Through Oral Care, Product Selection, and Application Method. AACN Advanced Critical Care, 24(1): 38–58.

[13] Joseph, N.M., Sistla, S., Dutta, T.K., Badhe, A.S., Parija, S.C. 2010. Ventilator-associated pneumonia: A review. European Journal of Internal Medicine, 21(5): 360–368.

[14] Klein Klouwenberg, P.M., van Mourik, M.S., Ong, D.S., Horn, J., Schultz, M.J., Cremer, O.L., Bonten, M.J. 2014 Electronic Implementation of a Novel Surveillance Paradigm for Ventilator-associated Events. Feasibility and Validation. American Journal of Respiratory and Critical Care Medicine, 189(8): 947–955.

[15] Luyt, C.-E. et al. (2008) 'Usefulness of procalcitonin for the diagnosis of ventilator-associated pneumonia', Intensive Care Medicine, 34(8), pp. 1434–1440. doi: 10.1007/s00134-008-1112-x.

[16] Mao, Z., Gao, L., Wang, G., Liu, C., Zhao, Y., Gu, W., Kang, H., Zhou, F. 2016 Subglottic secretion suction for preventing ventilator-associated pneumonia: an updated meta-analysis and trial sequential analysis. Critical Care, 20(1): 353.

[17] Okgün Alcan, A., Demir Korkmaz, F., Uyar, M. 2016. Prevention of ventilator-associated pneumonia: Use of the care bundle approach. American Journal of Infection Control, 44(10): e173–e176.

[18] Póvoa, P. et al. (2016) 'Biomarker kinetics in the prediction of VAP diagnosis: results from the BioVAP study', Annals of Intensive Care. SpringerOpen, 6(1), p. 32. doi: 10.1186/s13613-

[19] Póvoa, P., Martin-Loeches, I., Ramirez, P., Bos, L.D., Esperatti, M., Silvestre, J., Gili, G., Goma, G., Berlanga, E., Espasa, M., Gonçalves, E., Torres, A., Artigas, A. 2015. Biomarker Kinetics in VAP. Clinical Pulmonary Medicine, 22(4): 185–191.

[20] Rouzé, A., Jaillette, E., Poissy, J., Préau, S., Nseir, S. 2017. Tracheal Tube Design and Ventilator-Associated Pneumonia. Respiratory Care, 62(10): 1316–1323.

[21] Spalding, M.C., Cripps, M.W., Minshall, C.T. 2017. Ventilator-Associated Pneumonia. Critical Care Clinics, 33(2): 277–292.

A HEURISTIC ALGORITHM APPROACH TO IMPRECISE MALMQUIST PRODUCTIVITY INDEX

Shiang-Tai Liu
Vanung University, Graduate School of Business and Management
No.1, Wanneng Rd, Zhongli District, Taoyuan 320, Taiwan
E-mail: stliu@vnu.edu.tw

Abstract: The Malmquist productivity index (MPI), which is essentially the ratio of the efficiencies in two periods, can be used to measure the performance improvement of a production unit. Under uncertain conditions, such as predicting future demand, estimating missing data, or describing human judgment, observations become imprecise, and the associated MPI is also imprecise. This paper presents a methodology for measuring the MPI under uncertainty, where the imprecise observations are represented by intervals. A pair of two-level programming models is developed for calculating the lower and upper bounds of the interval-valued MPI. By structuring the two-level programming problem as an unconstrained nonlinear programming problem with bounded variables, the interval-valued MPI can be calculated.

Keywords: Malmquist productivity index; two-level programming; interval-valued data

1 INTRODUCTION

The Malmquist productivity index (MPI) is an effective measure of changes in the efficiency of a unit in different periods. Conceptually, the MPI is the ratio of the efficiencies in two periods, which can be calculated from DEA programs. Numerous studies concerning the methodology and application of DEA have been published. In the real world, however, there are uncertain factors which prevent precise measurements. For example, when an event has not occurred yet, the data must be predicted in advance. A similar case is that the data is missing, and must be estimated. There are also cases where uncertainty is due to human judgment; such as when rating the service quality of a company. In all these cases, the observations collected are imprecise.

Several approaches have been proposed to handle such cases. The most common one is to assume the data are stochastic, obeying specific probability distributions. A relaxed version of this stochastic approach is to represent the data by intervals, without requiring knowledge of the distribution. Of all these approaches, intervals are probably the easiest, and also the most intuitive, as larger intervals imply less precision.

When the imprecise data is represented by intervals, the resulting efficiencies and MPIs have the same characteristic. Although the MPI is the ratio of efficiencies, the interval-valued MPI cannot be obtained directly as the ratio of efficiency intervals. The objective of this paper is thus to develop models to calculate the MPI with imprecise data, where the imprecise observations are represented by intervals.

2 CONVENTIONAL MPI

Let $\boldsymbol{X}_j^t=(X_{1j}^t,\ldots,X_{mj}^t)$ and $\boldsymbol{Y}_j^t=(Y_{1j}^t,\ldots,Y_{sj}^t)$ denote the input and output vectors, respectively, of the jth DMU, j=1,…, n, in period t. The output-oriented efficiency of the kth DMU under the assumption of variable returns to scale (VRS), $E^t(\boldsymbol{X}_k^t, \boldsymbol{Y}_k^t)$, can be measured via the following model (Cooper, Seiford, & Tone, 2000):

$$
\begin{aligned}
1/E^t(\boldsymbol{X}_k^t, \boldsymbol{Y}_k^t) = \max.\ & \theta + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\
\text{s.t.}\ & \sum_{j=1}^{n} \lambda_j X_{ij}^t + s_i^- = X_{ik}^t, \quad i=1,\ldots,m \\
& \sum_{j=1}^{n} \lambda_j Y_{rj}^t - s_r^+ = \theta Y_{rk}^t, \quad r=1,\ldots,s \\
& \sum_{j=1}^{n} \lambda_j = 1 \\
& \lambda_j, s_i^-, s_r^+ \geq 0, \quad j=1,\ldots,n,\ r=1,\ldots,s,\ i=1,\ldots,m
\end{aligned}
\tag{1}
$$

This model is able to identify non-dominated DMUs and the associated production frontier.

The efficiency of the kth DMU in period t+1 based on the production technology of period t is:

$$
\begin{aligned}
1/E^t(\boldsymbol{X}_k^{t+1}, \boldsymbol{Y}_k^{t+1}) = \max.\ & \theta + \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\
\text{s.t.}\ & \sum_{j=1}^{n} \lambda_j X_{ij}^t + s_i^- = X_{ik}^{t+1}, \quad i=1,\ldots,m \\
& \sum_{j=1}^{n} \lambda_j Y_{rj}^t - s_r^+ = \theta Y_{rk}^{t+1}, \quad r=1,\ldots,s \\
& \sum_{j=1}^{n} \lambda_j = 1 \\
& \lambda_j, s_i^-, s_r^+ \geq 0, \quad j=1,\ldots,n,\ r=1,\ldots,s,\ i=1,\ldots,m
\end{aligned}
\tag{2}
$$

where $\boldsymbol{X}_k^{t+1}$ and $\boldsymbol{Y}_k^{t+1}$ are the input and output vectors, respectively, of DMU k in period t+1.

Since the MPIs calculated by using periods t and t+1 as the base period are probably different, Färe et al. (1994) suggested using the geometric mean of the two measures as the MPI. That is,

$$
\text{MPI}_k = \left[\frac{E^t(\boldsymbol{X}_k^{t+1}, \boldsymbol{Y}_k^{t+1})}{E^t(\boldsymbol{X}_k^t, \boldsymbol{Y}_k^t)} \times \frac{E^{t+1}(\boldsymbol{X}_k^{t+1}, \boldsymbol{Y}_k^{t+1})}{E^{t+1}(\boldsymbol{X}_k^t, \boldsymbol{Y}_k^t)}\right]^{1/2} \tag{3}
$$

They also assumed constant returns to scale in calculating efficiencies, and the MPI was decomposed to provide economic meaning. Since the focus of this paper is not on the decomposition of MPI, the more general DEA model under the assumption of VRS is used. When the observations $\boldsymbol{X}$ and $\boldsymbol{Y}$ have interval values, the feasible region for them is a hyper-rectangle, which is a convex and compact set, and the objective function will be continuous. Therefore, the MPI will also have interval values.

3 TWO-LEVEL PROGRAMMING APPROACH

The interval MPI has the following form:

$$\widehat{\text{MPI}}_k = \left[\frac{E^t(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})}{E^t(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)} \times \frac{E^{t+1}(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})}{E^{t+1}(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)} \right]^{1/2} \tag{4}$$

The efficiencies $E^t(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})$, $E^t(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)$, $E^{t+1}(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})$, and $E^{t+1}(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)$ in Equation (4) have interval values, and the mathematical operations involved in Equation (4) for interval values, namely division and square root, are not defined. Therefore, $\widehat{\text{MPI}}_k$ cannot be calculated from $E^t(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})$, $E^t(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)$, $E^{t+1}(\widehat{\boldsymbol{X}}_k^{t+1}, \widehat{\boldsymbol{Y}}_k^{t+1})$, and $E^{t+1}(\widehat{\boldsymbol{X}}_k^t, \widehat{\boldsymbol{Y}}_k^t)$ after they are obtained from the existing methods, and other methods must be devised.

Let $\widehat{X}_{ij}^p = [(X_{ij}^p)^L, (X_{ij}^p)^U]$ and $\widehat{Y}_{rj}^p = [(Y_{rj}^p)^L, (Y_{rj}^p)^U]$, $p=t$ and $t+1$, represent interval-valued data, where the values in the square brackets are the lower and upper bounds of the interval. Different values of $x_{ij}^p \in \widehat{X}_{ij}^p$ and $y_{rj}^p \in \widehat{Y}_{rj}^p$ produce different values for $\widehat{\text{MPI}}_k$. Denote $(\text{MPI}_k)^L$ and $(\text{MPI}_k)^U$ as the lower and upper bounds, respectively, of the interval-valued $\widehat{\text{MPI}}_k$, that is, $\widehat{\text{MPI}}_k = [(\text{MPI}_k)^L, (\text{MPI}_k)^U]$. Based on Equation (3), one has,

$$(\text{MPI}_k)^L == \min_{\substack{x_{ij}^p \in \widehat{X}_{ij}^p, y_{rj}^p \in \widehat{Y}_{rj}^p \\ \forall i,r,j,p}} \left[\frac{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^t(t) \end{array}}{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^t(t+1) \end{array}} \times \frac{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^{t+1}(t) \end{array}}{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^{t+1}(t+1) \end{array}} \right]^{1/2} \tag{5a}$$

$$(\text{MPI}_k)^U == \max_{\substack{x_{ij}^p \in \widehat{X}_{ij}^p, y_{rj}^p \in \widehat{Y}_{rj}^p \\ \forall i,r,j,p}} \left[\frac{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^t(t) \end{array}}{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^t(t+1) \end{array}} \times \frac{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^{t+1}(t) \end{array}}{\begin{array}{l} \max\ \theta + \varepsilon s \\ \text{s.t. } \theta, \lambda, \boldsymbol{s}^-, \boldsymbol{s}^+ \\ \quad \in F^{t+1}(t+1) \end{array}} \right]^{1/2} \tag{5b}$$

These are a special type of two-level programming models. At the first level, the values of x_{ij}^p and y_{rj}^p which produce the minimum (or maximum) for the program at the second level are sought. For each set of x_{ij}^p and y_{rj}^p values specified at the first level, the second-level program calculates the reciprocal of four types of efficiency. Once $(\text{MPI}_k)^L$ and $(\text{MPI}_k)^U$ are calculated, the interval MPI, $\widehat{\text{MPI}}_k$, is obtained.

The second-level program of Models (5a) and (5b) is a complicated function of

x_{ij}^p and y_{rj}^p, which involves the calculation of four DEA programs. Although the form of this function is not known, its value can be calculated once a set of x_{ij}^p and y_{rj}^p values is provided. This problem can thus be considered as an unconstrained nonlinear program with bounded variables of the following form:

$$\left.\begin{matrix}(\text{MPI}_k)^L \\ (\text{MPI}_k)^U\end{matrix}\right\} = \begin{cases}\min \ \text{MPI}_k \\ \max \ \text{MPI}_k\end{cases}$$

$$\begin{aligned}\text{s.t.} \quad & (X_{ij}^p)^L \le x_{ij}^p \le (X_{ij}^p)^U, \ i=1,\ldots,m; j=1,\ldots,n; p=t, t+1 \\ & (Y_{rj}^p)^L \le y_{rj}^p \le (Y_{rj}^p)^U, \ r=1,\ldots,s; j=1,\ldots,n; p=t, t+1\end{aligned} \quad (6)$$

The function form of MPI_k is not known, yet its value can be calculated by solving four DEA programs.

To solve Model (6), one can start with a feasible point, for example, the center of the feasible region, $\{[(X_{ij}^p)^L + (X_{ij}^p)^U]/2, \ [(Y_{rj}^p)^L + (Y_{rj}^p)^U]/2\}$. This process is continued until two consecutive trial points are close enough. To accelerate the convergence, a quasi-Newton modification of the search direction, such as with the DFP or BFGS formula (Fletcher 1987), can be incorporated into the algorithm. The basic algorithm is as follows.

Step 1 Initialization

1.1 Set ($\boldsymbol{x}^{(0)}$, $\boldsymbol{y}^{(0)}$) to the center point, (($\boldsymbol{X}^L + \boldsymbol{X}^U$)/2, ($\boldsymbol{Y}^L + \boldsymbol{Y}^U$)/2), step=$c$, n=0.

1.2 Calculate $\nabla f(\mathbf{x}^{(n)}, \boldsymbol{y}^{(n)})$, approximated by the central difference.

1.3 Set $\boldsymbol{H}^{(n)}=\boldsymbol{I}$, $\boldsymbol{d}^{(n)} = \nabla f(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$ ($\boldsymbol{d}^{(n)} = -\boldsymbol{d}^{(n)}$ for minimization).

Step 2 Termination check

If $\|\boldsymbol{d}^{(n)}\| < \varepsilon$, then terminate, with MPI=$f(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$; otherwise, continue Step 3.

Step 3 Line search

3.1 Set ($\hat{\boldsymbol{x}}^{(0)}$, $\hat{\boldsymbol{y}}^{(0)}$)=($\boldsymbol{x}^{(n)}$, $\boldsymbol{y}^{(n)}$), k=0.

3.2 Set ($\hat{\boldsymbol{x}}^{(k+1)}$, $\hat{\boldsymbol{y}}^{(k+1)}$)=($\hat{\boldsymbol{x}}^{(k)}$, $\hat{\boldsymbol{y}}^{(k)}$)+step$\times$ $\boldsymbol{d}^{(n)}$

3.3 Boundary check

If $\hat{x}_{ij}^{(p)(k+1)} < (X_{ij}^{(p)})^L$, then set $\hat{x}_{ij}^{(p)(k+1)} = (X_{ij}^{(p)})^L$, $\forall\, i, j, p$

If $\hat{x}_{ij}^{(p)(k+1)} > (X_{ij}^{(p)})^U$, then set $\hat{x}_{ij}^{(p)(k+1)} = (X_{ij}^{(p)})^U$, $\forall\, i, j, p$

If $\hat{y}_{rj}^{(p)(k+1)} < (X_{ij}^{(p)})^L$, then set $\hat{y}_{rj}^{(p)(k+1)} = (X_{ij}^{(p)})^L$, $\forall\, r, j, p$

If $\hat{y}_{rj}^{(p)(k+1)} > (X_{ij}^{(p)})^U$, then set $\hat{y}_{rj}^{(p)(k+1)} = (X_{ij}^{(p)})^U$, $\forall\, r, j, p$

If any of the above conditions occur, then set ($\boldsymbol{x}^{(n+1)}$, $\boldsymbol{y}^{(n+1)}$)=($\hat{\boldsymbol{x}}^{(k+1)}$, $\hat{\boldsymbol{y}}^{(k+1)}$), n=n+1, and go to Step 1.2; otherwise, calculate $f(\hat{\boldsymbol{x}}^{(k+1)}, \hat{\boldsymbol{y}}^{(k+1)})$ and continue Step 3.4.

3.4 If $f(\hat{\boldsymbol{x}}^{(k+1)}, \hat{\boldsymbol{y}}^{(k+1)}) > f(\hat{\boldsymbol{x}}^{(k)}, \hat{\boldsymbol{y}}^{(k)})$ (replace “>” with “<” for minimization), then set step=2 × step, k=k+1 and go to Step 3.2; otherwise, set ($\boldsymbol{x}^{(n+1)}$, $\boldsymbol{y}^{(n+1)}$)=($\hat{\boldsymbol{x}}^{(k)}$, $\hat{\boldsymbol{y}}^{(k)}$) and continue Step 4.

Step 4 Direction generation

4.1 Calculate $\Delta g = \nabla f(\boldsymbol{x}^{(n+1)}, \boldsymbol{y}^{(n+1)}) - \nabla f(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$, $\Delta z = (\boldsymbol{x}^{(n+1)}, \boldsymbol{y}^{(n+1)}) - (\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})$,

$$\boldsymbol{H}^{(n+1)} = \left[\boldsymbol{I} - \frac{\Delta z^t \Delta g}{\Delta z \Delta g^t}\right] \boldsymbol{H}^{(n)} \left[\boldsymbol{I} - \frac{\Delta z^t \Delta g}{\Delta z \Delta g^t}\right] + \frac{\Delta z^t \Delta g}{\Delta z \Delta g^t}.$$

4.2 Calculate $\boldsymbol{d}^{(n+1)} = \nabla f(\boldsymbol{x}^{(n+1)}, \boldsymbol{y}^{(n+1)}) \boldsymbol{H}^{(n+1)}$ ($\boldsymbol{d}^{(n+1)} = -\boldsymbol{d}^{(n+1)}$ for minimization).

4.3 Set n=n+1 and go to Step 2.

4 CONCLUSIONS

The Malmquist productivity index (MPI) measures changes in the efficiency of DMU between two periods. Conventionally, it is applied to cases where the observations have precise values. This paper discusses imprecise cases where the observations are represented by intervals. By formulating the problem as a pair of two-level programs to represent the lower and upper bounds of the interval MPI, the problem becomes an unconstrained nonlinear program with bounded variables, and can then be solved fairly easily. This paper uses intervals to describe imprecise condition, and representing uncertain values by intervals produces results that are more informative than those obtained from observations represented by estimated precise values.

References

[1] Amouzegar, M.A. (1999). A global optimization method for nonlinear bilevel programming problems. *IEEE Transactions on Systems, Man, and Cybernetics-B*, 29, 771-776.

[2] Bard, J.F. (1988). Convex two-level programming. *Mathematical Programming*, 40, 15-27.

[3] Bialas, W.F., & Karwan, M.H. (1984). Two-level linear programming. *Management Science*, 30, 1004-1020.

[4] Caves, D.W., Christensen, L.R., & Diewert, W.E. (1982). Multilateral comparisons of output, input, and productivity using superlative index numbers. *The Economic Journal*, 92, 73-86.

[5] Charnes, A., Cooper, W.W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2, 429-444.

[6] Cooper, W.W., Seiford, L.M., & Tone, K. (2000). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA- Solver Software*. Boston: Kluwer Academic Publishers.

[7] Färe, R., Grosskopf, S., Norris, M., & Zhang, Z. (1994). Productivity growth, technical progress and efficiency change in industrialized countries. *American Economic Review*, 84, 66-83.

[8] Fletcher, R. (1987). *Practical Methods of Optimization*. 2nd edn. Chichester: John Wiley & Sons.

[9] Grosskopf, S. (2003). Some remarks on productivity and its decomposition. *Journal of Productivity Analysis*, 20, 459-474.

[10] Kao, C. (2010). Malmquist productivity index based on common-weights DEA:

The case of Taiwan forests after reorganization. *Omega*, 38, 484-491.
[11] Pastor, J.T., & Lovell, C.A.K. (2005). A global Malmquist productivity index. *Economics Letters*, 88, 266-271.
[12] Seiford, L.M., & Zhu, J. (1999). Infeasibility of super-efficiency data envelopment analysis models. *INFOR*, 37, 174-187.
[13] Thrall, R.M. (1996). Duality, classification and slacks in DEA. *Annals of Operations Research*, 66, 109-138.
[14] Vicente, L.N., & Calamai, P.H. (1994). Bilevel and multilevel programming: A bibliography review. *Journal of Global Optimization*, 5, 291-306.

ON SUSTAINABLE PRINCIPLES IN MULTI OBJECTIVE PROGRAMMING PROBLEMS

Josip Matejaš, Tunjo Perić, Danijel Mlinarić
University of Zagreb, Faculty of Economics and Business
Kennedyjev trg 6, 10000 Zagreb, Croatia
E-mails: jmatejas@efzg.hr, tperic@efzg.hr, dmlinaric@efzg.hr.

Abstract: Usually in multi objective programming problems the set of solutions may be large and a question, which solution should be taken, is very important in practical situations. Therefore it is necessary to clearly state additional properties the chosen solution should have. In the paper we propose some sustainable principles which are naturally imposed by the practical problems which can be stated in the form of multi objective programming problems. We present a simple numerical method which is based on these principles and which yields the solution with desired properties. The method respects the priorities and aspirations of decision makers and enables iterations of the obtained solution.

Keywords: decision making; multi objective programming problem; priority, aspiration, iteration.

1 INTRODUCTION

A problem where several decision makers (*players*) try to optimize their utilities at the same time and on the same constraint set (*budget*) is well known as the multiple objective programming problem (MOPP) or multiple level programming problem (MLPP) if the players belong to different levels. Although there are many approaches and methods for solving MOPP and MLPP in the literature, there still remains unsatisfactory moments both from theoretical and practical aspects. Many methods involve complicated solving procedure which can be hardly understood and trusted by the players, the obtained solution (especially in MLPP) is often inefficient and/or unsatisfactory and useless (one player gets everything while the other gets nothing) etc. In the paper we present a simple numerical method which reflects the very nature of the problem.

Very often in practice players are not equal. They make decisions with different authorities. Some of them may be subordinate or superior to the other(s). Such situations are regularly encountered in everyday business life: employer and employee, seller and buyer, banker and client etc., as well as in everyday ordinary life: teacher and student, parents and children, young and old etc. Moreover, almost every institution is organized hierarchically. We see such level based structure everywhere: in education, politics, government and administration in general, business, companies, banks and economy in general, etc., up to sport, show business, art and ordinary social life. Mainly, such structure is necessary for normal functioning of institutions. According to the role in the hierarchy, each level has certain rights, importance, influence or limits which are regularly measured and quantified as stage, degree, weight, price, salary etc. We have to be able to recognize this differences from the mathematical formulation and computing process. So, in such cases, each player (or objective) is assigned with a particular *priority* (weight, right, importance, preference or significance) in the decision making process.

In everyday situations we enter with some expectations, hopes, wants, aims or aspirations which determine our behaviour, moves and activities. If a situation brings bad or undesirable effects and results, we try to get another chance to repair them. These are the basic principles in our everyday life. We need only to copy these principles in the solution process. So, numerical method that represents and apes real situation should respect the *aspirations* of the players and should ensure the possibility of *iterating* an unsatisfactory solution.

Finally, in the recent neoliberalism, maximization of the own profit at all costs is the main principle in economic and political life. All marketing, activities and efforts are focused in that direction without paying attention to the other players, ecology, natural and human resources. The obvious effects of this selfish principle is growing inequalities, economic crises, the poor become poorer, the richer become richer. Such cruel maximization principle is automatically involved in the existing numerical models for MOPP and MLPP. To prevent these undesirable effects, some *sustainable* principles must enter in the solution process. As we said, the priority of the players who belong to the higher levels or who are more important must be respected, but also, the care for players who belong to the lower levels or who are less important must be taken. It should not be allowed that, in the obtained solution, one player gets everything while the other gets nothing, as we shall see later in concrete examples.

This paper is an attempt to balance all these moments through a consistent numerical method which respects priorities and aspirations of the players, enables iterations (possible improvement of the solution) and which, by sustainable rules, respect the „strong'' and protect the „weak''. The paper is organized as follows. In Section 2, using a simple practical example, we reveal the main problems which become motivation for our work. We present the appropriate numerical method and its application in Section 3 and conclusions in Section 4.

2 PROBLEMS AND MOTIVATION

Let us consider a following basic problem which regularly occurs in everyday life. *Two players want to take a part of given budget (100%). What part could each of them take?* This problem can be stated as

$$\max_{(x.y)\in S}(x,y) \quad \text{where} \quad S=\{(x,y): x+y\le 100,\ x\ge 0,\ y\ge 0\}, \tag{1}$$

It can be considered as MOPP or MLPP if the objectives belong to different levels. We used (1) as a test problem. We have applied some of the most famous existing methods to solve it. Let (x^*, y^*) denotes the solution. We obtained the following results in Tab. 1 (for each method the corresponding reference is given).

Table 1: Overview of the results according to the literature.

Method for MOPP	**Solution**
Weight coefficients [3] ($w_1 \ge w_2$)	(100, 0)
Weight coefficients ($w_1 < w_2$)	(0, 100)
Goal programming [2], [6], [5] ($g_1 + g_2 \le 100$)	(g_1, g_2)
Goal programming ($g_1 \le 100$, $g_1 + g_2 > 100$)	(g_1, $100 - g_1$)
Goal programming ($g_1 > 100$)	(100, 0)
Lexicographic [4]	(100, 0)
MP [7] (d_1, d_2)	$100\cdot(d_1, d_2) / (d_1+d_2)$
STEM [1]	(50, 50) in the first step
Method for MLPP	**Solution**
Methods based on Stackelberg's model [8]	(100, 0)

Let us closely look at the obtained solutions. Are they satisfactory from the practical point of view, from the player's perspective?

The solutions (100,0) and (0,100), where one player gets everything and other one nothing, are completely unsatisfactory. Why? Because they implicitly exclude one of the players from decision process. Namely, these results are also the solutions of $\max_{(x.y)\in S} x$ and $\max_{(x.y)\in S} y$. As long as there exist feasible points $(x,y)\in S$, $x>0$, $y>0$, such solutions cannot be acceptable.

Besides, the bare formulation $\max(x, y)$ means that we try to achieve as much as possible for both, x and y, at the same time and not only for one of them. In goal programming method the obtained solution is satisfactory only for $g_1 + g_2 = 100$. If $g_1 + g_2 < 100$ then the solution is not efficient because both players could get more (see Fig. 1), while in all other cases the method respects only goal g_1 and ignores g_2. For example, if $g_2 = 3g_1$ then for $g_1 = 25, g_2 = 75$ the solution is $(25,75)$, for $g_1 = 50$, $g_2 = 150$ it is $(50,50)$ while for $g_1 = 100, g_2 = 300$ it is $(100,0)$. In MP method, both aspirations d_1, d_2 are fully respected in the frame of the given constraint set. The aspirations are also respected in the further steps in STEM method by additional cutting of the default constraint set.

Since the set of efficient solutions is large, which solution should be taken? Which properties it should have? Which principles for choosing a solution should be applied? These questions were the motivation for our work. Below we offer our answers to them.

3 NEW APPROACH

Practical problems which can be stated as MOPP or MLPP are, by their nature, very simple and so it should be the associated mathematical method. The high level and hardly understandable theoretical models, which are often met in the literature, are not suitable in such practical situations. The simple method which is easy to understand and trust by the players, is desirable. We consider the general MOPP,

$$\max_{x \in S} \left(F_1(x), F_2(x), ..., F_m(x)\right), \tag{2}$$

where $S \subset R^n$ is the given budget, *m* is the number of players and $F_i : S \to R$ is the objective function for player i, $i = 1,2,...,m$. Note that the number of objectives is not necessarily the number of players. Each player can have several objectives (goals) which serve him to make the right choice. To avoid confusion here we shall use the words player and objective as synonyms. It is enough to consider maximization problem only because any optimization can be stated in such form. Let p_i and a_i be the priority and aspiration of player i, respectively. Based on the MP method from [7], we propose the following method for solving problem (2),

$$\max \lambda \text{ where } x \in S, \ F_i(x) \geq \lambda p_i a_i, \ i = 1,2,...,m. \tag{3}$$

If the objectives are linear and S is defined by linear constraints then (3) is simple linear programming problem. Optimal solution λ^* shows to what maximal extent all the players could realize their aspirations, respecting their priorities on the given constraint set S. If the obtained solution is not satisfactory for some reasons, then the players can redefine their aspirations (priorities are fixed, imposed by original situation), and perform the next iteration by solving (3) again. We propose the following sustainable rules for definition of aspirations.

Rule 1. *Any aspiration has to be attainable.*

Rule 2. *An active player may require to increase his possible realization at most as much as he is ready to give up from his aspiration.*

Rule 3. *An active player has the right to retain at most as much of his possible realization as he is ready to give up.*

Rule 4. *Any player has the right to set his possible realization as his aspiration.*

Rule 1 means that $\min_{x \in S} F_i(x) \leq a_i \leq \max_{x \in S} F_i(x)$, although the method allows any aspiration (see Fig. 1). Rules 2 and 3 are related to the active players (players whose constraints in the optimal point are active, $F_i(x^*) = \lambda^* p_i a$) because they can change (unsatisfactory) solution by redefining their aspirations. (see also [7] for details). Rule 2 means that the new aspiration a_i' should satisfy $a_i' - F_i(x^*) \le a_i - a_i' \Rightarrow a_i' \le (a_i + F_i(x^*))/2$. Rule 3 enables any player to define the lower bound l_i for his realization, $F_i(x) \ge l_i$, which satisfies $l_i \le F_i(x^*) - l_i \Rightarrow l_i \le F_i(x^*)/2$. Thus, Rule 2 prevents megalomania of the active players especially those with the higher priority while Rule 3 prevents bankruptcy for the players with lower priority. Finally, Rule 4 ensures the lower bound for the next aspiration, $a_i' \ge F_i(x^*)$.

We apply and explain the proposed methodology on the initial problem (1) where method (3) reads

$$\max \lambda \ \text{ where } \ (x,y) \in S, \ \ x \ge \lambda p_1 a_1, \ \ y \ge \lambda p_2 a_2. \tag{4}$$

Problem (4) has a unique efficient solution

$$x^* = \frac{100}{p_1 a_1 + p_2 a_2} \cdot p_1 a_1, \quad y^* = \frac{100}{p_1 a_1 + p_2 a_2} \cdot p_2 a_2, \quad \lambda^* = \frac{100}{p_1 a_1 + p_2 a_2},$$

or (see Fig. 1)

$$T = (x^*, y^*) = 100 \cdot \left(\frac{1}{1+\mu\nu}, \frac{\mu\nu}{1+\mu\nu} \right) = 100 \cdot \left(\frac{1}{1+\mu\nu}, 1 - \frac{1}{1+\mu\nu} \right), \ \ \mu = \frac{p_2}{p_1}, \ \nu = \frac{a_2}{a_1}. \tag{5}$$

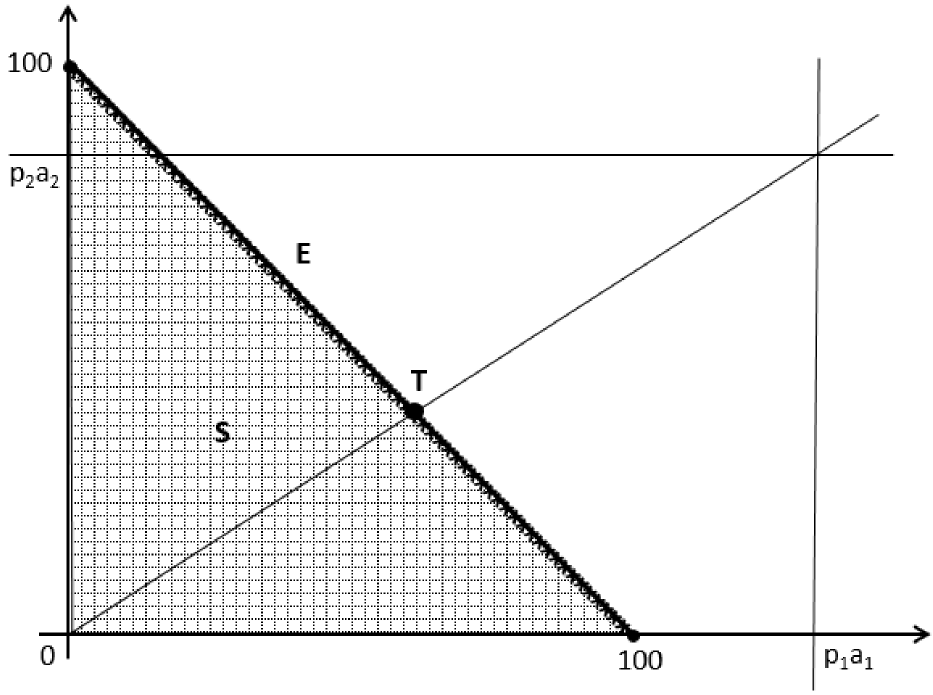


Figure 1: Aspirations and priorities are respected.

Here μ is the ratio that measures a level's relative weight of importance. For example, if 2 (3) is the priority for the first (the second) player, then the second player is $\mu = 1.5$ times more important (its priority is 50% higher) than the first one. The same is true whether the priorities are 4 and 6 or 0.6 and 0.9 and so on. Furthermore it means that the second player has 50% higher rights to realize his aspiration than the first one. And how could ν be interpreted? It shows how the aspiration of the second player is related to the aspiration of the first player. More precisely, for each unit which the first player aspires the second one aspires ν units.

For example if $a_1 = a_2$ then for $p_1 = p_2$ the solution (5) is $(50, 50)$, but for $p_1 = 2,\ p_2 = 3$ it is $(40, 60)$ which exactly shows that the second player has 50% higher rights to realize his aspiration than the first one. If, for the fixed $\mu \in \langle 0, +\infty \rangle$, we consider $T(x^*, y^*)$ from the relation (5) as a function of ν then $T(\nu)$ is bijection between $[0, +\infty]$ and the set of efficient points for problem (1), (see Fig. 1), $E = \{(x, y):\ 0 \le x \le 100,\ y = 100 - x\}$. This is a reason which justifies such treatment for aspirations and priorities. There are no preferred points in set E. Each one could be a solution depending on the aspirations and priorities (in fact on ratios μ and ν). Since the priority ratio μ is fixed for the given problem, the aspirations, and hence their ratio ν, can change during the solution process and so they can serve as a mechanism for improving the solution through successive iterations if it is unsatisfactory.

Note also that λ^* shows to what extent the aspirations according to the priorities, could be realized on set S: better $(\lambda^* > 1)$, equal $(\lambda^* = 1)$ or worse $(\lambda^* < 1)$ than the players expect it to be. For example, if $\mu = 1$ then for $d_1 = d_2 = 1000$ we have $x^* = y^* = 50$ and $\lambda^* = 0.05$ which means that the players could realize only 5% of their aspirations and for $d_1 = d_2 = 10$ we have again $x^* = y^* = 50$, but $\lambda^* = 5$ shows 500% realization.

If the obtained solution is not satisfactory for the players then they can redefine their aspirations by using Rules 1-4 and, through successive iterations, improve the solution in the desired direction. We illustrate the process for $\mu = 1$. Let $a_1^{(k)}, a_2^{(k)}$ and $(x^{(k)}, y^{(k)}, \lambda^{(k)})$ be the aspirations and the solution in k-th iteration, $k = 1, 2, 3, \ldots$, respectively. Suppose that the first player tries to increase his realization in each iteration (Rule 2) while the second one tries to retain as much as possible (Rule 4) that is an extreme situation. Generally, the choice of rules to be applied can be a matter of *cooperation* between the players. If the players define maximal initial aspirations, $a_1^{(1)} = a_2^{(1)} = 100$ (see Rule 1) the first solution is $(x^{(1)}, y^{(1)}) = (50, 50)$. For each subsequent iteration we have $a_1^{(k+1)} = (a_1^{(k)} + x^{(k)})/2$ (Rule 2) and $d_2^{(k+1)} = x_2^{(k)}$ (Rule 4), which yields the solutions in Tab. 2. Note that the second player has the right to restrict the set S, immediately in the first iteration, with additional constraint $y \ge y^{(1)}/2 = 25$ (Rule 3) which will be his protection in further iterations, but it is redundant here.

Table 2: Successive iterations of the solution.

k	$a_1^{(k+1)} = (a_1^{(k)} + x^{(k)})/2$	$x^{(k)}$	$y^{(k)} = a_2^{(k+1)}$	$\lambda^{(k)}$
1	100	50	50	0.5
2	75	60	40	0.8
3	67.5	62.790698	37.209302	0.930233
4	65.145349	63.646691	36.353309	0.976995
5	64.396020	63.917071	36.082929	0.992562
6	64.156546	64.003274	35.996726	0.997611
7	64.079910	64.030839	35.969161	0.999234
8	64.055375	64.039662	35.960338	0.999755
9	64.047518	64.042487	35.957513	0.999921
10	64.045003	64.043392	35.956608	0.999975

Thus, we obtained increasing sequences $x^{(k)} \to x^*$, $\lambda^{(k)} \to 1 = 100\%$ and decreasing one $y^{(k)} \to y^*$ where $64.043 < x^* < 64.044$ and $y^* = 100 - x^*$. We see how the stated rules gradually correct the aspirations to be realistic in the frame of given set S and thus enables realization to converge to 100%.

In the same way we can obtain the results for any $\mu \neq 1$ and any choice of initial aspirations. We can also see that the method cannot yield the solution (100,0) or (0,100) for any nontrivial priority and aspiration ($\mu, \nu \in \langle 0, +\infty \rangle$), because it fully respects them in the solving process.

4 CONCLUSIONS

Numerical method for solving multi objective problems, which is presented in the paper, respects the priorities and aspirations of the decision makers. Besides it allows iteration (improvement) of the obtained solution, according to the rules which respect (protect) decision makers with higher (lower) priorities. The method is summarized in the following algorithm.

Algorithm MOS (**M**ulti-**O**bjective-**S**olution)

1. Input: problem (2) with priorities p_i, $i = 1, 2, ..., m$.
2. Define the aspirations a_i, $i = 1, 2, ..., m$ according to Rule 1.
3. Solve (3).
4. If the solution is satisfactory then go to 7.
5. Using Rules 1-4 redefine the aspirations a_i, $i = 1, 2, ..., m$.
6. Go to 3.
7. End.

Using the proposed method decision makers are able to improve the obtained solution through successive iterations in the desired direction until they reach the state that satisfies everyone.

References

[1] Benayoun, R. J. de Montgolfier, Tergny, J., Larichev, O. 1971. Linear Programming with Multiple Objective Functions. Step Method (STEM), Mathematical Programming, Vol. 1 (3): 366-375.

[2] Charnes, A., Cooper, W. W. 1961. Management Models and Industrial Applications of Linear Programming, Vol. I, New York: Wiley.

[3] Fishburn, P. C. 1974. Lexicographic Orders, Utilities and Decision Rules: a Survey, Management Science, Vol. 20 (11): 1442-1471.

[4] Hwang, C. L., Masud, A. S. 1979. Multiple Objective Decision Making: Methods and Applications, New York: Springer Verlag.

[5] Ignizio, J. P. 1976. Goal Programming and Extensions, Massachusetts: Lexington Books.

[6] Lee, S. M. 1972. Goal Programming for Decision Analysis, Philadelphia: Auerbach Publishera.

[7] Matejaš J., Perić T. 2014. A new iterative method for solving multiobjective linear programming problem. Applied Mathematics and Computation, 243: 746–754.

[8] Sinha, A., Malo, P., Deb, K. 2018. A review on Bilevel Optimization: From Classical to Evolutionary Approaches and Applications. IEEE Transactions on Evolutionary Computation, *22(*2): 276–295.

DECISION MAKING IN COMPLEX DECENTRALIZED BUSINESS SYSTEMS BY MULTI-LEVEL MULTI-OBJECTIVE LINEAR PROGRAMMING METHODS

Tunjo Perić
University of Zagreb, Faculty of Economics and Business - Zagreb
Trg J.F. Kennedya 6, 10000 Zagreb, Croatia
E-mail: tperic@efzg.hr

Zoran Babić
University of Split, Faculty of Economics
Cvite Fiskovića 5, 21000 Split, Croatia
E-mail: babic@efst.hr

Slavko Matanović
High school of modern business Belgrade
Terazije 5, 11000 Beograd, Serbia
E-mail: slavko.matanovic@mbs.edu.rs

Abstract: This paper proposes a new methodology for multi-level multi-objective linear programming problem solving as an aid to decision making. The appliance of the suggested approach requires a certain degree of cooperation in preparing and making decisions between decision makers at all levels of decision-making in a complex decentralized business system. The efficiency of the procedure was tested on an example of optimization of production plan in a complex decentralized enterprise. The obtained results demonstrate the possibility of employing the recommended technique and confirm the high degree of its efficiency.

Keywords: Complex decentralized business systems, Multi-level decision making, Multi-objective programming methods

1 INTRODUCTION

A complex decentralized business system, such as a complex enterprise, is defined as a system consisting of a larger number of subsystems - enterprises, with the subsystems consisting of a plurality of sub subsystems, etc. Such a system has a hierarchical structure in which both the whole system and its subsystems have their goals that they try to maximize. Naturally, the objectives of the entire system are in conflict with the objectives of its subsystems while the sub-subsystem's objectives are in conflict with other subsystem's goals and with the objectives of the entire system, etc. The achievement of the goals of each subsystem depends on achieving the objectives of other subsystems and the entire system, so the consistency of their decisions is the prerequisite for the efficient functioning of the business system and its subsystems.

When deciding on a business system and its subsystems, it is necessary not only to consider maximizing its goals but also to take into account how these decisions will affect the degree of achievement of the goals of other subsystems and the entire business system. The methodology for multi-level multi-objective linear programming proposed as an aid to decision making, consists of the application of the method based on the cooperative theory of games ([2]).

The application of the suggested approach requires a certain degree of co-operation among decision makers in the decision-making process at all levels of decision-making in a complex decentralized business system. The efficiency of the proposed methodology was tested on the example of optimization of the production plan in a complex decentralized enterprise. The results obtained indicate the possibility of employing the novel approach as well as the high degree of efficiency of its utilization.

The rest of the paper is organized as follows. In Section 2 model of multi-level multi-objective linear programming is presented. Section 3 describes the suggested procedure to solve decision making problems. In Section 4 the new approach is tested on a practical production plan, and optimization of research investment, quantity of stocks and promotion investment in a supposed company. Conclusion and References are given at the end of the paper.

2 MODEL OF MULTI-LEVEL MULTI-OBJECTIVE LINEAR PROGRAMMING PROBLEM

Suppose that the decision model consists of $L + 1$ levels, upper levels (ULDM) and lower levels (LLDM), where we have p_0 decision makers at the level 0, p_1 decision makers at the level 1 and p_L decision makers at the lowest level ([1], [4]).

The model contains variables $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x}_L) \in \mathbf{R}^n$, $\mathbf{x}_0 \in \mathbf{R}^{n_{01}+n_{02}+\ldots+n_{0p0}}$, $\mathbf{x}_1 \in \mathbf{R}^{n_{11}+n_{12}+\ldots+n_{1p1}}$, …, $\mathbf{x}_L \in \mathbf{R}^{n_{L1}+n_{L2}+\ldots+n_{LpL}}$, where $n = n_{01} + \ldots + n_{0p_0} + n_{11} + \ldots + n_{1p_1} + \ldots + n_{L1} + \ldots + n_{Lp_L}$, $\mathbf{x}_0 = \left(\mathbf{x}_{01}, \mathbf{x}_{02}, \ldots, \mathbf{x}_{0p_0}\right)$, $\mathbf{x}_1 = \left(\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1p_1}\right)$, …, $\mathbf{x}_L = \left(\mathbf{x}_{L1}, \mathbf{x}_{L2}, \ldots, \mathbf{x}_{Lp_L}\right)$, $\mathbf{x}_{01} = \left(x_{01,1}, x_{01,2}, \ldots, x_{01,n_{01}}\right)$, …, $\mathbf{x}_{0p_0} = \left(x_{0p_0,1}, x_{0p_0,2}, \ldots, x_{0p_0,n_{0p0}}\right)$, $\mathbf{x}_{11} = \left(x_{11,1}, x_{11,2}, \ldots, x_{11,n_{11}}\right)$, …, $\mathbf{x}_{1p_1} = \left(x_{1p_1,1}, x_{1p_1,2}, \ldots, x_{1p_1,n_{1p1}}\right)$, …, $\mathbf{x}_{L1}^{li} = \left(x_{L1,1}^{li}, x_{L1,2}^{li}, \ldots, x_{L1,n_{L1}}^{li}\right)$, …, $\mathbf{x}_{Lp_L} = \left(x_{Lp_L,1}, x_{Lp_L,2}, \ldots, x_{Lp_L,n_{L,pL}}\right)$. The decision makers DM_{li} ($l = 0, 1, \ldots, L$; $i = 1, \ldots, p_l$) control variables $\boldsymbol{x}_{li}$.

The multi-level multi-objective linear programming model can be presented as:

$\left[\mathrm{DM}_{0,1}\right]$: $f_{0,1} = \max_{\mathbf{x}\in\mathbf{S}} f_{0,1}(\mathbf{x}) = \mathbf{c}_{01}^{01}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{01}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{01}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{01}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{01}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{01}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{01}$)

$\left[\mathrm{DM}_{0,2}\right]$: $f_{0,2} = \max_{\mathbf{x}\in\mathbf{S}} f_{0,2}(\mathbf{x}) = \mathbf{c}_{01}^{02}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{02}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{02}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{02}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{02}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{02}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{02}$)

……………………………………………………………………………………………

$\left[\mathrm{DM}_{0,p_0}\right]$: $f_{0,p_0} = \max_{\mathbf{x}\in\mathbf{S}} f_{0p_0}(\mathbf{x}) = \mathbf{c}_{01}^{0p_0}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{0p_0}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{0p_0}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{0p_0}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{0p_0}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{0p_0}\mathbf{x}_{Lp_L}$,
(controls variables $\mathbf{x}_{0p_0}$) (1)

where $\mathbf{x}_{11}$, $\mathbf{x}_{12}$, …, $\mathbf{x}_{1p_1}$ solve

$\left[\mathrm{DM}_{1,1}\right]$: $f_{1,1} = \max_{\mathbf{x}\in\mathbf{S}} f_{1,1}(\mathbf{x}) = \mathbf{c}_{01}^{11}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{11}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{11}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{11}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{11}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{11}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{11}$)

$\left[\mathrm{DM}_{1,2}\right]$: $f_{1,2} = \max_{\mathbf{x}\in\mathbf{S}} f_{1,2}(\mathbf{x}) = \mathbf{c}_{01}^{12}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{12}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{12}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{12}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{12}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{12}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{12}$)

………………………………………………………………………………………………..

$\left[\mathrm{DM}_{1,p_1}\right]$: $f_{1,p_1} = \max_{\mathbf{x}\in\mathbf{S}} f_{1p_1}(\mathbf{x}) = \mathbf{c}_{01}^{1p_1}\mathbf{x}_{01} + \ldots + \mathbf{c}_{0p_0}^{1p_1}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{1p_1}\mathbf{x}_{11} + \ldots + \mathbf{c}_{1p_1}^{1p_1}\mathbf{x}_{1p_1} + \ldots + \mathbf{c}_{L1}^{1p_1}\mathbf{x}_{L1} + \ldots + \mathbf{c}_{Lp_L}^{1p_1}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{1p_1}$)

where $\mathbf{x}_{L1}$, $\mathbf{x}_{L2}$, …, $\mathbf{x}_{Lp_L}$ solve

[DM$_{L,1}$]: $f_{L,1} = \max_{\mathbf{x}\in\mathbf{S}} f_{L,1}(\mathbf{x}) = \mathbf{c}_{01}^{L1}\mathbf{x}_{01} + ... + \mathbf{c}_{0p_0}^{L1}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{L1}\mathbf{x}_{11} + ... + \mathbf{c}_{1p_1}^{L1}\mathbf{x}_{1p_1} + ... + \mathbf{c}_{L1}^{L1}\mathbf{x}_{L1} + ... + \mathbf{c}_{Lp_L}^{L1}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{L1}$)
[DM$_{L,2}$]: $f_{L,2} = \max_{\mathbf{x}\in\mathbf{S}} f_{L,2}(\mathbf{x}) = \mathbf{c}_{01}^{L2}\mathbf{x}_{01} + ... + \mathbf{c}_{0p_0}^{L2}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{L2}\mathbf{x}_{11} + ... + \mathbf{c}_{1p_1}^{L2}\mathbf{x}_{1p_1} + ... + \mathbf{c}_{L1}^{L2}\mathbf{x}_{L1} + ... + \mathbf{c}_{Lp_L}^{L2}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{L2}$)
……………………………………………………………………………………..
[DM$_{L,pL}$]: $f_{L,p_L} = \max_{\mathbf{x}\in\mathbf{S}} f_{L,p_L}(\mathbf{x}) = \mathbf{c}_{01}^{Lp_L}\mathbf{x}_{01} + ... + \mathbf{c}_{0p_0}^{Lp_L}\mathbf{x}_{0p_0} + \mathbf{c}_{11}^{Lp_L}\mathbf{x}_{11} + ... + \mathbf{c}_{1p_1}^{Lp_L}\mathbf{x}_{1p_1} + ... + \mathbf{c}_{L1}^{Lp_L}\mathbf{x}_{L1} + ... + \mathbf{c}_{Lp_L}^{Lp_L}\mathbf{x}_{Lp_L}$, (controls variables $\mathbf{x}_{pl}$)
where
$\mathbf{S} = \left\{\mathbf{x} \in \mathbf{R}^n : \ \mathbf{A}_0\mathbf{x}_0 + \mathbf{A}_1\mathbf{x}_1 + ... + \mathbf{A}_L\mathbf{x}_L (\le, =, \ge)\mathbf{b}, \ \mathbf{x} \ge \mathbf{o}, \ \mathbf{b} \in \mathbf{R}^m, \mathbf{o} \in \mathbf{R}^n\right\} \ne \varnothing$,
$\mathbf{c}_0^{li} = \left(\mathbf{c}_{01}^{li}, \mathbf{c}_{02}^{li}, ..., \mathbf{c}_{0p_0}^{li}\right)$, $\mathbf{c}_1^{li} = \left(\mathbf{c}_{11}^{li}, \mathbf{c}_{12}^{li}, ..., \mathbf{c}_{1p_1}^{li}\right)$, …, $\mathbf{c}_L^{li} = \left(\mathbf{c}_{L1}^{li}, \mathbf{c}_{L2}^{li}, ..., \mathbf{c}_{Lp_L}^{li}\right)$, $\mathbf{c}_{01}^{li} = \left(c_{01,1}^{li}, c_{01,2}^{li}, ..., c_{01,n_0}^{li}\right)$, …,
$\mathbf{c}_{0p_0}^{li} = \left(c_{0p_0,1}^{li}, c_{0p_0,2}^{li}, ..., c_{0p_0,n_{0p_0}}^{li}\right)$, $\mathbf{c}_{11}^{li} = \left(c_{11,1}^{li}, c_{11,2}^{li}, ..., c_{11,n_{11}}^{li}\right)$, …, $\mathbf{c}_{1p_1}^{li} = \left(c_{1p_1,1}^{li}, c_{1p_1,2}^{li}, ..., c_{1p_1,n_{1p_1}}^{li}\right)$, …,
$\mathbf{c}_{L1}^{li} = \left(c_{L1,1}^{li}, c_{L1,2}^{li}, ..., c_{L1,n_{L1}}^{li}\right)$, …, $\mathbf{c}_{Lp_L}^{li} = \left(c_{Lp_L,1}^{li}, c_{Lp_L,2}^{li}, ..., c_{Lp_L,n_{Lp_L}}^{li}\right)$, $(l = 1, ..., L; \ i = 1, ..., p_i)$
are vectors in the objective functions, $\mathbf{A}_0$, $\mathbf{A}_l$, $l = 1, 2, ..., L$, are the matrices of the constraint coefficients, **b** is vector of the right hand of constraints and **o** is a null vector.

3 MULTI-LEVEL MULTI-OBJECTIVE LINEAR PROGRAMMING METHODOLOGY

To solve the ML-MOLP model we propose a methodology based on multi-objective linear programming method and cooperative game theory ([2]). Here we give a short overview of the MOLP method and the proposed algorithm to solve the ML-MOLP model ([3]). This method enables decision-makers to be significantly involved in the process of obtaining the preferred efficient solution.

The multi-objective programming problem represents the situation where (1) several (K) decision makers optimize their objectives, or (2) one decision maker optimizes several different objectives, at the same time and on the same constraint set. Here we analyse case (1). Each objective is given by the objective function $f_k(\mathbf{x}), \ k = 1, 2, ..., K$. If the analytic form of the constraint set and the objective functions is linear then we have a multi-objective linear programming (MOLP) problem.

There are two sets defined in this method,

$$\mathbf{D} = \left\{\mathbf{x} \in \mathbf{R}^n : \ \mathbf{x} \ge 0, \ f_k(\mathbf{x}) \ge d_k, \ k = 1, 2, ..., K\right\}, \qquad (2)$$

$$\mathbf{D}_\lambda = \left\{\mathbf{x} \in \mathbf{R}^n : \ \mathbf{x} \ge 0, \ f_k(\mathbf{x}) \ge \lambda d_k, \ k = 1, 2, ..., K\right\}, \ \lambda \ge 0, \qquad (3)$$

where d_k is the aspiration level which the decision maker DM_k wants to achieve $(f_k(\mathbf{x}) \ge d_k)$. At any point $\mathbf{x} \in \mathbf{D}$ all the decision makers achieve their aspirations fully, while at any point $\mathbf{x} \in \mathbf{D}_\lambda$ they achieve their aspirations to the relative extent of at least λ. The method is stated in the form: find the largest λ for which $\mathbf{D}_\lambda \cap \mathbf{S} \ne \varnothing$ or,

$$\max_{(\mathbf{x},\lambda)\in\mathbf{G}} \lambda, \qquad (4)$$

where

$$\mathbf{G} = \left\{(\mathbf{x}, \lambda) \in \mathbf{R}^{n+1} : \ \mathbf{x} \in \mathbf{S}, \ \lambda \ge 0, \ f_k(\mathbf{x}) \ge \lambda d_k, \ k = 1, 2, ..., K\right\},$$

which is a standard linear programming problem. The optimal solution λ^* shows to which (minimum) relative extent all the decision makers can realize their aspirations. For x^* being the optimal point, the indicator

$$\lambda_k = \frac{f_k(\mathrm{x}^*)}{d_k}, \quad k = 1, 2, ..., K \tag{5}$$

shows to which extent the decision maker DM_k can realize his own aspiration. Thus, the indicators measure the reality of decision makers' aspirations and can be used to improve the solution, if unsatisfactory, in the subsequent iterations (see [2] for details).

The algorithm of the ML-MOLP model solving can be presented as:

Step 1. Determine individual optimal values of all objective functions on the set of constraints **S** and form the pay-off table.
Step 2. Set the aspired lower and upper limit of all variables controlled by the decision makers of the level 0.
Step 3. Set an aspiration value $d_{l,i}$ $\left(l = 1, ..., L;\ i = 1, ..., p_L\right)$ for all objective functions at all levels.
Step 4. Formulate the model

$$\max_{(\mathbf{x},\lambda)\in\mathbf{G}} \lambda,$$

where

$$\mathbf{G} = \left\{(\mathbf{x},\lambda) \in \mathbf{R}^{n+1}:\ \mathbf{x} \in \mathbf{S},\ \lambda \geq 0,\ f_{l,i}(\mathbf{x}) \geq \lambda d_{l,i},\ l = 0, 1, ..., L,\ i = 1, ..., p_l\right\},$$

Step 5. Solve the model from Step 4 to get a solution of the ML-MOLP problem.
Step 6. If the obtained solution is not an empty set, go to Step 7, otherwise go to Step 8.
Step 7. Stop. The satisfactory solution of the ML-MOLP problem is obtained.
Step 8. Modify lower and/or upper aspiration limits of the variables controlled by the decision makers of the level 0 and/or the aspiration value $d_{l,i}$ $\left(l = 1, ..., L;\ i = 1, ..., p_L\right)$ of the objective functions (using relation (5)) and go back to Step 4.

4 PRACTICAL APPLICATION

Suppose a complex decentralized company with two decision makers on the level 0, two decision makers on the level 1 and two decision makers on the level 2. The decision makers on the level 0 maximize the total production, and the total investment in the development of the company, the decision makers on the level 1 maximize their net profits, while the decision makers on the level 2 maximize the total inventory, and the total investment in the promotion. Company produces 3 products (P_1, P_2, P_3) in the department 1, and 3 products (P_4, P_5, P_6) in the department 2. For the research into the development of the company there is an estimated 1000000 m.u. The annual cost amounts to 10% of the invested funds. The company requires that all the products should be produced in at least 500 units. In addition, the company should stock no less than 10% of its total production to ensure delivery safety. The management of the enterprise has set the inventory limit for each product to at least 50, and to at most 800 for each product. The enterprise management has determined that investment in promotion must not exceed 15% of the total gross profit of the company. Promotion investments may not exceed 15000 m.u. for each product. The total production must be at least 14000 units. Table 1 shows manufacturing data which are fixed by assumption.

Table 1: Data for the practical application

Data per unit	Products					
	P_1	P_2	P_3	P_4	P_5	P_6
Machines, department 1 (h) Capacity: 13000	2	1	3			
Machines, department 2 (h) Capacity: 12000				1	2	1
Sales price per unit (m.u.)	1400	1200	1300	900	1100	700
Gross profit per unit (m.u.)	100	120	80	150	200	180
Inventory cost per unit (m.u.)	8	6	10	10	6	8

Let $x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4}, x_{1,5}, x_{1,6}$ is the quantity of products P_1, P_2, P_3, P_4, P_5, P_6 respectively; $x_{1,7}, x_{1,8}$ is the capital invested in the departments 1, 2 respectively; $x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}, x_{2,5}, x_{2,6}$ is the quantity of stock products P_1, P_2, P_3, P_4, P_5, P_6 respectively, $x_{2,7}, x_{2,8}, x_{2,9}, x_{2,10}, x_{2,11}, x_{2,12}$ is the quantity of investment in promotion of products P_1, P_2, P_3, P_4, P_5, P_6 respectively.

The ML-MOLP model is presented as:

Level 0

$$\max_{\mathbf{x}\in\mathbf{S}} f_{0,1} = x_{1,1} + x_{1,2} + x_{1,3} + x_{1,4} + x_{1,5} + x_{1,6}\,; \quad \max_{\mathbf{x}\in\mathbf{S}} f_{0,2} = x_{1,7} + x_{1,8} \tag{6}$$

Level 1

$$\max_{\mathbf{x}\in\mathbf{S}} f_{1,1} = 100x_{1,1} + 120x_{1,2} + 80x_{1,3} - 0{,}1x_{1,7} - 8x_{2,1} - 6x_{2,2} - 10x_{2,3} - x_{2,7} - x_{2,8} - x_{2,9}$$

$$\max_{\mathbf{x}\in\mathbf{S}} f_{1,2} = 150x_{1,4} + 200x_{1,5} + 180x_{1,6} - 0{,}1x_{1,8} - 10x_{2,4} - 6x_{2,5} - 8x_{2,6} - x_{2,10} - x_{2,11} - x_{2,12}$$

Level 2

$$\max_{\mathbf{x}\in\mathbf{S}} f_{2,1} = x_{2,1} + x_{2,2} + x_{2,3} + x_{2,4} + x_{2,5} + x_{2,6}\,, \quad \max_{\mathbf{x}\in\mathbf{S}} f_{2,2} = x_{2,7} + x_{2,8} + x_{2,9} + x_{2,10} + x_{2,11} + x_{2,12}\,,$$

where

$$\mathbf{S} = \left\{ \begin{array}{l} \mathbf{x} : 2x_{1,1} + x_{1,2} + 3x_{1,3} \le 13000,\ x_{1,4} + 2x_{1,5} + x_{1,6} \le 12000,\ x_{1,7} + x_{1,8} \le 1000000,\ x_{1,1}, x_{1,2}, x_{1,3}, \\ x_{1,4}, x_{1,5}, x_{1,6} \ge 500,\ x_{2,1} \ge 0{,}1x_{1,1},\ x_{2,2} \ge 0{,}1x_{1,2},\ x_{2,3} \ge 0{,}1x_{1,3},\ x_{2,4} \ge 0{,}1x_{1,4},\ x_{2,5} \ge 0{,}1x_{1,5}, \\ x_{2,6} \ge 0{,}1x_{1,6},\ 50 \le x_{2,1}, x_{2,2}, x_{2,3}, x_{2,4}, x_{2,5}, x_{2,6} \le 800,\ x_{2,7} + x_{2,8} + x_{2,9} + x_{2,10} + x_{2,11} + x_{2,12} \le \\ 0{,}15(150x_{11} + 120x_{1,2} + 80x_{1,3} + 150x_{1,4} + 200x_{1,5} + 180x_{1,6}),\ x_{2,7}, x_{2,8}, x_{2,9}, x_{2,10}, x_{2,11}, x_{2,12} \le \\ 15000\ \ x_{1,1} + x_{1,2} + x_{1,3} + x_{1,4} + x_{1,5} + x_{1,6} \ge 14000,\ x_{1,7}, x_{1,8}, x_{2,7}, x_{2,8}, x_{2,9}, x_{2,10}, x_{2,11}, x_{2,12} \ge 0 \end{array} \right\}$$

Individual optimal values of all objective functions on the set of constraints **S** of the model (6) are presented in the following pay-off table:

Table 2: Pay-off table

	$f_{0,1}$	$f_{0,1}$	$f_{1,1}$	$f_{1,2}$	$f_{2,1}$	$f_{2,2}$
max $f_{0,1}$	**21750**	0	1168300	1840000	2175	0
max $f_{0,2}$	14000	**1000000**	544800	1157800	1400	0
max $f_{1,1}$	14000	0	**1168300**	599050	1400	0
max $f_{1,2}$	18000	0	642600	**1980300**	1800	0
max $f_{2,1}$	14000	0	630800	1145800	**4800**	0
max $f_{2,2}$	14000	0	554800	1067800	1400	**90000**

The aspired values of the variables controlled by decision makers on the Level 0 $(x_{1,7}, x_{1,8})$ are taken 500000 each. As the aspiration values of the objective functions in this step we have taken their optimal values.

To apply the proposed methodology the following linear programming model has been formed and solved:

$$\max_{\mathbf{x},\lambda\in\mathbf{G}} \lambda, \tag{7}$$

where

$$\mathbf{G} = \left\{ \begin{matrix} \mathbf{x},\lambda : \mathbf{x} \in \mathbf{S} \cap f_{0,1}(\mathbf{x}) \geq 21750\lambda,\ f_{0,2}(\mathbf{x}) \geq 1000000\lambda,\ f_{1,1}(\mathbf{x}) \geq 1167300\lambda,\ f_{1,2}(x) \geq \\ 1980300\lambda,\ f_{2,1}(\mathbf{x}) \geq 4800\lambda,\ f_{2,2}(x) \geq 90000\lambda,\ x_{1,7} \geq 500000\lambda,\ x_{1,8} \geq 500000\lambda \end{matrix} \right\}.$$

The following solution has been obtained:

$x_{1,1} = 1750,\ x_{1,2} = 8000,\ x_{1,3} = 500,\ x_{1,4} = 4862,\ x_{1,5} = 500,\ x_{1,6} = 6138,\ x_{2,1} = 800,\ x_{2,2} = 800,$ $x_{2,3} = 420,\ x_{2,4} = 800,\ x_{2,5} = 800,\ x_{2,6} = 800,\ x_{1,7} = 460370.10,\ x_{1,8} = 460370.10,\ x_{2,7} = 15000,$ $x_{2,8} = 7867,\ x_{2,9} = 15000,\ x_{2,10} = 15000,\ x_{2,11} = 15000,\ x_{2,12} = 15000,\ \lambda = 0.92074,\ f_{0,1} = 21750,$ $f_{0,2} = 920740.20,\ f_{1,1} = 1075701,\ f_{1,2} = 1823894,\ f_{2,1} = 4420,\ f_{2,2} = 82267.$

In the next steps decision makers cooperate to find a solution acceptable to all of them. If some of them are not satisfied with the achievement of the objective function value, it is necessary to reduce the level of aspiration of some other satisfied objective functions. In order to improve the value of functions f_{11} and f_{12} we have reduced the aspiration value of functions f_{01} to 18000, f_{21} to 3500 and f_{22} to 60000. Model (7) with the changed aspiration values is solved. The following solution has been obtained:

$x_{1,1} = 1750,\ x_{1,2} = 8000,\ x_{1,3} = 500,\ x_{1,4} = 2555,\ x_{1,5} = 722,\ x_{1,6} = 8000,\ x_{2,1} = 312,\ x_{2,2} = 800,$ $x_{2,3} = 50,\ x_{2,4} = 800,\ x_{2,5} = 800,\ x_{2,6} = 800,\ x_{1,7} = 468648.50,\ x_{1,8} = 468648.50,\ x_{2,7} = 15000,$ $x_{2,8} = 10297,\ x_{2,9} = 0,\ x_{2,10} = 15000,\ x_{2,11} = 15000,\ x_{2,12} = 15000,\ \lambda = 0.937297,\ f_{0,1} = 21527,$ $f_{0,2} = 937297,\ f_{1,1} = 1095044,\ f_{1,2} = 1856692,\ f_{2,1} = 3562,\ f_{2,2} = 70297.$

5 CONCLUSIONS

In this paper a new methodology to solve multi-level multi-objective linear programming problems has been proposed. The suggested procedure is based on the application of a multi objective programming method and cooperative game theory. The novel approach has been tested on a practical example of a supposed complex decentralized company planning.

The proffered technique is simple to use both for analysts and decision makers. Only the linear programming models are solved in a number of steps. From decision makers it requires information on the acceptable value of the objective functions and the decision variables they control (only from the top-level decision makers). For the future research we propose testing application efficiency of the proposed methodology on the real examples with a large number of variables and decision makers.

References

[1] Baky, I. A. (2010). Solving multi-level multi-objective linear programming problems through fuzzy goal programming approach, *Applied Mathematical Modelling,* Vol. 34, 2377–2387

[2] Matejaš, J., Perić, T. 2014. A new iterative method for solving multiobjective linear programming problem. Applied Mathematics and Computation, Vol. 243(9): 746-754

[3] Perić, T., Babić, Z., Matejaš, J. 2018. Comparative analysis of application efficiency of two iterative multi objective linear programming methods (MP method and STEM method). Central European Journal of Operations Research, Vol. 26(3): 565-583

[4] Perić, T., Babić, Z., Omerović, M. 2019. A fuzzy goal programming approach to solving decentralized bi-level multi-objective linear fractional programming problems. Croatian Operational Research Review, Vol. 10(1): 65-74.

A MULTI-CRITERIA, HIERARCHICAL MODEL FOR THE EVALUATION OF SCENARIOS THAT FACILITATE THE DEVELOPMENT OF DIGITAL COMPETENCES OF GYMNASIUM STUDENTS IN THE REPUBLIC OF SLOVENIA

Srečko Zakrajšek
IAM, College for Multimedia, Ljubljana
Leskoškova 12, SI-1000 Slovenia
E-mail: sreco.zakrajsek@iam.si

Eva Jereb, Uroš Rajkovič, Vladislav Rajkovič, Mojca Bernik
Faculty for Organizational Sciences, University of Maribor
Kidričeva cesta 55a, SI-4000 Kranj, Slovenia
E-mail: eva.jereb@fov.uni-mb.si; uros.rajkovic@fov.uni-mb.si;
vladislav.rajkovic@gmail.com; mojca.bernik@fov.uni-mb.si

Abstract: The development of digital competences of secondary-school students is one of the key tasks of contemporary education systems. The EU has been systematically dealing with this issue since 2005; in 2017, the digital competences for teachers (DidCompOrg) were published, and currently the framework of digital competences for secondary-school students, which will be finalised after 2021, is being prepared. Several EU states are trying to implement the attaining of digital competences of students according to different scenarios. We have decided to develop a multi-criteria hierarchical model according to the DEX methodology in order to evaluate various education scenarios.

Keywords: digital competences, secondary-school students, HMADM (Hierarchical Multi-Attribute Decision Making), DEX (Decision EXpert) methodology

1 INTRODUCTION

The field of digital competences is relatively new in education. The EU has been systematically developing it since 2015 in the framework of The Joint Research Centre - JRC. Already in 2006, the Norwegians have implemented digital competences as a key for achieving modern education into their national education programme, Italy began to measure the digital competences of secondary-school students after 2008, and in 2010, the OECD [1] began to pay attention to this field, along with other countries that began to address digital competences in education [5], [6].

On the EU level, activities in the field of digital competences in education have increased and began to be systematically developed in 2014, after the publication of the Horizon Report Europe [9] that was prepared by over 50 experts from 22 European countries. In 2015, the EU issued a publication [7], which presents the era of digital learning and the demands that the educational system is facing. In January 2018, the EU published a document [8] containing the recommendations for a modernisation of education in the EU. The document concludes that a modern education is key for the development of the EU, and each country has the autonomy and obligation to regulate their educational system.

The EU only sets frameworks and recommendations for digital competences in education systems [4], while the implementation remains in the domain of the state. Since no frameworks and goals have been set for the digital competences of secondary-school students yet, there are also no systematically defined and evaluated concrete recommendations for educational scenarios that could ensure the attaining of digital competences of secondary-school students in a certain environment.

There is no world-wide research that would systemically analyse different education scenarios that facilitate the acquisition of digital competences by secondary-school students and would become a model for the evaluation of individual scenarios. Our research, part of which is published in this article, presents a multi-criteria, hierarchical model that makes it possible to evaluate various education scenarios that facilitate the acquisition and development of digital competences of secondary-school students, while simultaneously providing institutions with a means of evaluating their potentials and needs in this area.

2 METHOD

The basic methodology on which the preparation of the multi-criteria model is based is MADM (Multi-Attribute Decision Making), which makes it possible to choose between various education scenarios, and also to evaluate the reasons for the differences and the current state, the potential and state of the individual institution or system, as well as the ideas on how to change that state. The model has been developed on the basis of Multi-Attribute Decision Making and primarily as a tool for evaluating various options (in our case various education scenarios). According to this model, the decision problem (field, option) is broken down into smaller problems that are easier to understand, evaluate in relation to each standard, and to master. The model consists of fields, basic and derived attributes that are situated into the criteria tree of value domains and the criteria they define and the utility functions F_i (or aggregation functions or convergence functions) of lower-level criteria. The parameters can be values that are described in words, whereas numerical parameters are described symbolically with value classes. According to the DEX (Decision EXpert) [12] method and with the programme DEXi, we built a model to evaluate education scenarios, whereby we also evaluated the requirements and conditions for the implementation and execution of a particular scenario and the anticipated results.

We constructed the entire tree on the basis of a review of literature and most of all our own research, in which experts from the ministry, professional institutions and schools also participated. The list of basic attributes was prepared according to the DEX methodology, based on a review of foreign research and the results of our own research with which we determined the key parameters that are significant for the acquisition of digital competences of secondary-school students in gymnasia and which we then verified with experts from the field of education. The cooperation of experts was very important in determining the importance (weight) of subattributes and attributes, as well as the correction of aggregated values according to the program's calculation. Our work leaned also on the results obtained by the researchers who prepared the model for the evaluation of the efficacy of the implementation of ICT in schools [3].

We implemented five phases of solving the decision problem: 1) identification of the problem; 2) identification of criteria (elements); 2.1) composing a list of criteria; 2.2) structuring the criteria (criteria tree); 2.3) determining measuring scales; 3) identification of utility functions; 4) description of variants; 5) evaluation and analysis of variants.

3 DESCRIPTION OF THE MODEL

We prepared a hierarchical model for the evaluation of various education scenarios that facilitate the acquisition of digital competences by secondary-school students. We identified the main attributes that define the process and proceeded to build the model according to the appropriate phases.

Figure 1 depicts three main areas and aggregate attributes that are important for the evaluation of digital competences in secondary-school students. The model contains 63 attributes, of which 40 are basic and 23 derived. The model contains 63 value domains and 20 functions. School organisation, education process and digital competences of secondary-school students are situated on the first level.

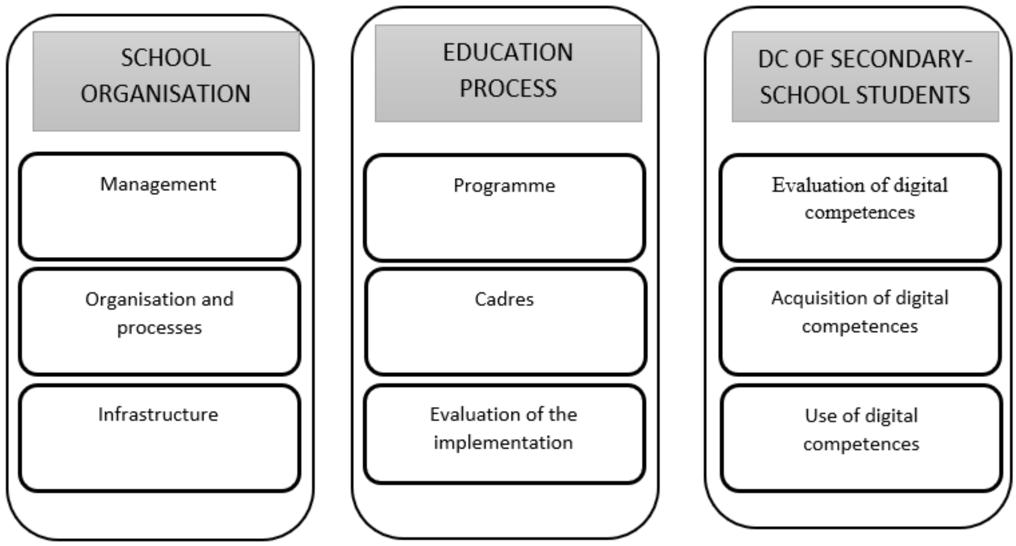


Figure 1: Depiction of three main areas and aggregate attributes of the model for evaluating education scenarios for the acquisition and development of digital competences in general gymnasia

To ensure clarity and due to the limitations of the DEXi programme, we split the basic elements into a maximum of three sub-areas and expanded the elements by assigning them a maximum of three successors. This ensures that we familiarise ourselves with each element in detail and can evaluate it as objectively as possible.

Figure 2 gives an example of the structure of attributes of the model for the field School organisation:

```
Criteria tree
Criterion -Description
Evaluation -Evaluation of the education scenario
├─School org. -School organisation
│ ├─Management
│ │ ├─Management -Management team
│ │ │ ├─Principal -Principal and assistants
│ │ │ ├─ICT coordinator -Information-communication technology coordinator
│ │ │ └─Teams -e-school development team
│ │ └─VSP -Vision, AWP and promotion
│ │   ├─Vision
│ │   ├─Implementation plan -AWP - Annual work plan
│ │   └─Promotion -Vision and strategy promotion
│ ├─O and P -Organisation and processes
│ │ ├─Organisation -Description of organisation
│ │ └─Processes -Description of processes
│ └─Infrastructure
│   ├─Rooms
│   │ ├─Classr., lab. -Classrooms, laboratories
│   │ ├─ICT R -ICT centre with library
│   │ └─A-S R -Above-standard rooms
│   ├─Equipment -Educational equipment, materials, accessories
│   │ ├─Hardware -Hardware
│   │ ├─Software -Software
│   │ └─Learn. materials -Learning materials and accessories
│   └─Connections -Connections, networks, bases,
│     ├─Networks -Links to networks
│     ├─E bases -Educational bases
│     └─O bases -Other bases
```

Figure 2: Example of the structure of attributes of the model for the field School organisation.

The scale on the sheets encompasses either four or three value domains that are determined by descriptive criteria.

Table 1 shows an example of a value domain with four criteria for the attribute Principal, whereas Table 2 presents an example of a value domain with three criteria for the attribute Classrooms and laboratories.

Table 1: Value domain and criteria of the attribute Principal (Principal, assistants)

No.	*Value*	*Criteria (description)*
1.	**D** **basic**	The principal and their assistants are not digitally competent and only include those digitalised activities into school activities that are necessary for the school's operation.
2.	**C** **better**	The principal and their assistants are not digitally competent but are aware of the importance of digital competences for education and enable teachers to implement the acquisition of digital competences of students in their subjects on their own initiative.
3.	**B** **good**	The principal and their assistants are quite digitally competent and encourage primarily teachers of voluntary subjects and younger co-workers to acquire digital competences.
4.	**A** **very good**	The principal and their assistants are digitally competent and actively involved into all activities that positively influence the development of digital competences of secondary-school students. They encourage all teachers to become digitally competent.

Table 2: Value domain and criteria of the attribute Classrooms, laboratories

No.	*Value*	*Criteria (description)*
1.	**BA** **Basic**	Classrooms and laboratories have basic equipment. The school has a computer classroom equipped with old computers and dedicated to the subject of Informatics.
2.	**GD** **Good**	Classrooms have classic equipment with a regular blackboard, canvass and AV means. Laboratories are equipped with certain didactic digital sets in addition to the classic equipment. The school has a computer classroom equipped with newer computers, dedicated primarily to the subject of Informatics.
3.	**VG** **Very good**	All classrooms are equipped with modern devices. Modern laboratories facilitate the digitalisation of experimental work. The gymnasium is equipped with several modern computer/multi-media classrooms for the purposes of teaching various subjects and the BYOD system.

Table 3 depicts the aggregated attributes on the first level, calculated from the utility functions of the subattributes. In evaluating individual areas, we applied the following evaluations and weights:

Table 3: Values of the aggregated attributes on the first level and weighing of areas.

SCHOOL ORGANISATION 30%

less suitable	suitable	good	very good	excellent
(1)	(2)	(3)	(4)	(5)

EDUCATION PROCESS 40 %

less suitable	suitable	good	very good	excellent
(1)	(2)	(3)	(4)	(5)

DIGITAL COMPETENCES OF SECONDARY-SCHOOL STUDENTS 30 %

insufficient	basic	good	very good	excellent
(1)	(2)	(3)	(4)	(5)

EVALUATION OF SCENARIO

unsuitable	less suitable	good	very good	excellent
(1)	(2)	(3)	(4)	(5)

4 RESULTS

We prepared a prototype model for the evaluation of various education scenarios that facilitate the acquisition of digital competences by secondary-school students. In this process, we utilised all key elements that affect the model and prepared the appropriate classification and structured questions and criteria. We applied the evaluation model to five education scenarios that were selected from a collection of several scenarios, as described in the article by Zakrajšek [10]; Table 5 shows the results of the evaluation of five education scenarios with our model.

Table 4: Selected education scenarios (ES) that facilitate the development of digital competences of secondary-school students in the Slovenian general gymnasium.

Reference of the education model	*Presentation of the scenario (upgrade or amendment of the current programme) – the programmes are sorted from 1 to 5 according to a growing efficacy and implementation difficulty.*
ES-1	**Various authentic programmes are implemented** into the gymnasium.
ES- 2	**Mandatory elective contents** that ensure the acquisition of **DC** are implemented into all four schoolyears.
ES-3	The acquisition of digital competences is implemented **only in two mandatory elective subjects**.
ES-4	The acquisition of digital competences **in a certain part of all Matura subjects**.
ES-5	The acquisition of digital competences **in all subjects**.

Table 5: The results of the evaluation of education scenarios.

The results of the evaluation of education scenarios.

Criterion	ES1	ES2	ES3	ES4	ES5
Evaluation	1	1	2	3	*5*
├─School org.	2	1	1	3	*5*
│ ├─Management	2	1	1	3	*5*
│ ├─O and P	3	3	3	*5*	*5*
│ └─Infrastructure	1	1	1	2	*5*
├─Ed. process	1	1	3	4	*5*
│ ├─Programme	1	1	*5*	*5*	*5*
│ ├─Cadres	1	1	2	3	4
│ └─Eval. of impl.	1	1	3	*5*	*5*
└─Students	2	3	3	3	4
├─Evaluation of DC	1	3	3	4	*5*
├─Acq. of DC	2	2	2	2	3
└─Use of DC	2	3	3	4	4

The evaluation of scenarios with the model shows that the best (excellent) scenario is ES-5, which reached 5 points. ES1 and ES2 scenarios are evaluated as unsuitable, as they do not ensure the acquisition of digital competences in subjects; ES3 and ES4 are less suitable and good, respectively. Each education scenario requires a certain modification of the vision and strategy of the school, the cadres, programme, infrastructure, implementation etc. By comparing the value domains of attributes in different scenarios according to individual criteria, we can also establish the requirements and necessary modifications that the implementation of an individual scenario demands.

5 CONCLUSIONS

This paper presents a multi-criteria hierarchical model for the evaluation of various education scenarios that facilitate the acquisition of digital competences by secondary-school students. At the same time, the model also facilitates the evaluation of the state and the potential of an individual school regarding the acquisition of digital competences by secondary-school students.

The MADM methodology DEX was applied, with the help of which we sort all significant attributes that influence the acquisition of digital competences into the appropriate tree, determine their value domains and criteria, as well as their weights. Because this methodology involves the use of pre-existing knowledge and the data is evaluated instead of precisely measured, the cooperation of experts, the harmonisation of their opinion and the verification of the solutions in practice are crucial.

The model facilitates the choice of the scenario most suitable for a specific environment, at the same time exposing the conditions that a certain environment must fulfil in order for a specific scenario to be implemented. As no such model has previously existed, this model is an original invention that together with the proposed scenarios facilitates the implementation and optimisation of modern education practices in secondary schools.

Acknowledgement
The authors thank the participating schools and their staff for their active role in the study.

References

[1] Are the new millennium learners making the grade? Technology use and educational performance in PISA. 2010. Vol. VI. Paris, France: CERI/OECD.

[2] Bohanec, M., Žnidaršič, M., Rajkovič, V., Bratko, I., Zupan, B. 2013. DEX methodology: three decades of qualitative multi-attribute modeling. Informatica: an international journal of computing and informatics, ISSN 0350-5596, 37(1): 49–54.

[3] Čampelj B., Karnet I., Brodnik A, Jereb E., Rajkovič U. 2018. A multi-attribute modelling approach to evaluate the efficient implementation of ICT in schools. Central European Journal of Operations Research. https://doi.org/10.1007/s10100-018-0595-y [Accessed 15/2/2018].

[4] Carretero, S.; Vuorikari, R. and Punie, Y. 2017. DigComp 2.1: The Digital Competence Framework for Citizens with eight proficiency levels and examples of use, EUR 28558 EN, doi:10.2760/38842. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC106281/web-digcomp2.1pdf_(online).pdf [Accessed 15/2/2018].

[5] Gill S., P. 2016. Promoting digital competence in secondary education: are schools there? Insights from a case study. New Approaches in Educational Research, 5(1): 57–63.

[6] Hatlevik, O. E., Christophersen, K.-A. 2013. Digital competence at the beginning of upper secondary school: identifying factors explaining digital inclusion. Computers & Education, 63: 240–247.

[7] Kampylis, P., Punie, Y., Devine, J. 2015. Promoting Effective Digital-Age Learning: A European Framework for Digitally-Competent Educational Organisations. Joint Research Centre, European Union.

[8] Łybacka K. 2018.REPORTon modernisation of education in the EU(2017/2224(INI)). European Parliament. http://www.europarl.europa.eu/doceo/document/A-8-2018-0173_EN.pdf [Accessed 15/4/2019].

[9] Johnson, L., Adams Becker, S., Estrada, V., Freeman, A., Kampylis, P., Vuorikari, R., and Punie, Y. 2014. Horizon Report Europe: 2014 Schools Edition. Luxembourg: Publications Office of the European Union.

[10] Zakrajšek, S. 2018. Possible education models for acquiring digital competence of students. Pedagoška obzorja (Didactica Slovenica), 33: 94-106.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Session 7:

Human Resources

SPATIAL INTERACTIONS AND THE REGIONAL EMPLOYMENT IN THE EU

Andrea Furková
University of Economics in Bratislava, Department of Operations Research and Econometrics
Dolnozemská cesta 1, 852 35 Bratislava, Slovakia
E-mail: andrea.furkova@euba.sk

Michaela Chocholatá
University of Economics in Bratislava, Department of Operations Research and Econometrics
Dolnozemská cesta 1, 852 35 Bratislava, Slovakia
E-mail: michaela.chocholata@euba.sk

Abstract: This paper deals with the estimation of spatial econometric model of employment rates across 259 NUTS 2 regions of the EU in 2018 regarding different region-specific factors as well as with quantification of their direct, indirect and total spatial impacts. The preliminary analysis confirming the spatial character of analysed data was followed by estimation of Spatial Durbin Model. The estimation of this model enables to calculate the global spillovers. The results confirmed our assumption of significant share of neighbouring regions as for *GDP* and compensation of employees in explaining regional employment rates. Significant influence of factors like educational attainment level and population density seems to be limited only to the particular region.

Keywords: regional employment rate, spatial interactions, Spatial Durbin Model (SDM)

1 INTRODUCTION

The issue of employment has been attracting both the politicians and the researchers for decades. One of the EU's (European Union) priorities declared in the strategic document Europa 2020, is to promote a high-employment economy and to achieve a target of 75 % employment rate of the population aged 20-64 by 2020. To face the problem of demographic change as well as to improve the position in the global competition, the EU needs to increase the labour force participation in order to deliver the economic, social and territorial cohesion (European commission, 2010). Plenty of studies have been published to analyse the regional (un)employment rates across various groups of EU regions considering different region-specific factors and using different methodological and empirical scope. The level of employment in a region is determined not only by individual region-specific factors, but its location in space plays a crucial role as well. Consideration of the spatial interactions is thus inevitable in creation of regional labour and employment policies supported by appropriate econometrical instruments in order to avoid the biased and incorrect results. From the studies investigating the issue of regional employment rate from the spatial econometric perspective can be mentioned, e.g., Perugini and Signorelli (2004), Franzese and Hays (2005), Pavlyuk (2011) and Chocholatá and Furková (2018).

The main aim of this paper is to estimate a spatial econometric model of the employment rates across 259 NUTS 2 (Nomenclature of Units for Territorial Statistics) regions of the EU in 2018 regarding different region-specific factors (GDP in PPS per inhabitant[1], population density, educational attainment level and compensation of employees per employee) as well as to quantify direct, indirect and total spatial impacts of individual factors. The rest of the paper is organized as follows: section 2 deals with the methodological aspects of analysis, section 3 is devoted to the data and empirical results and section 4 concludes.

[1] Gross domestic product in purchasing power standard per inhabitant

2 METHODOLOGICAL ASPECTS

This section provides brief overview of main methodological aspects related to empirical analysis of this paper. Instruments of spatial statistics and spatial econometrics seem to be appropriate in the situations when values of a variable under the consideration in nearby locations are more similar or related than values in locations that are far apart. Thus, if geographical location of units (e.g. states, regions, cities) matter or in other words if there are spatial interactions among units, this phenomenon is known as spatial autocorrelation.

Most of spatial econometric models explicitly allow for spatial dependence through spatially lagged variables. This group of spatial econometric models, models with spatially autoregressive process contains multiple specifications. Familiar SDM (Spatial Durbin Model)[2] model for cross–sectional data in matrix form can be written as follows:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{l}_N\alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \mathbf{u} \tag{1}$$

where $\mathbf{y}$ denotes $N \times 1$ vector of the observed dependent variable for all N observations (locations), $\mathbf{l}_N$ represents $N \times 1$ vector of ones associated with the constant term α, $\mathbf{X}$ represents a $N \times k$ matrix of exogenous explanatory variables (k denotes the number of explanatory variables), $\boldsymbol{\beta}$ is $k \times 1$ vector of unknown parameters to be estimated, $\mathbf{u}$ is $N \times 1$ vector of random errors, σ_v^2 is random error variance, $\mathbf{I}_N$ is N dimensional unit matrix and $\mathbf{W}$ is N dimensional spatial weighting matrix (the issues related to the spatial weighting matrix see e.g., Anselin and Rey, 2014). Vector $\boldsymbol{\gamma}$ has dimension of $k \times 1$ and parameter ρ represent important spatial autoregressive parameters.

On the one hand, complicated spatial structure of the spatial models causes estimation problems and there is a need for special estimation methods (review of models and estimation methods can be found in e.g. Anselin and Rey, 2014). At the same time, complicated spatial structure is a virtue of spatial econometrics, i.e. the ability to accommodate extended modelling strategies that describe multi-regional interactions. Unlike classical linear regression model, a change in a single region associated with any given explanatory variable will affect not only the region itself (called a direct impact), but potentially affect all other regions indirectly (called an indirect impact). Quantification of all impacts associated with r-th explanatory variable requires a construction of $\mathbf{S}_r(\mathbf{W})$ matrix which has $N \times N$ dimension and if we are assuming k explanatory variables, the total number of partial derivatives is $k \times N^2$. LeSage and Pace (2009) suggested summary impact measures: Average total impact (*ATI*), Average direct impact (*ADI*) and Average indirect impact (*AII*). These impacts can be calculated based on these formulas:

$$ATI = N^{-1}\mathbf{l}_N^{\mathrm{T}}\mathbf{S}_r(\mathbf{W})\mathbf{l}_N \tag{2}$$

$$ADI = N^{-1}\mathrm{tr}\left(\mathbf{S}_r(\mathbf{W})\right) \tag{3}$$

$$AII = ATI - ADI \tag{4}$$

where $\mathbf{S}_r(\mathbf{W})$ matrix for SDM model is defined as follows:

$$\begin{aligned} \mathbf{S}_r(\mathbf{W}) &= \mathbf{V}(\mathbf{W})\left(\mathbf{I}_N\beta_r + \mathbf{W}\gamma_r\right) \\ \mathbf{V}(\mathbf{W}) &= \left(\mathbf{I}_N - \rho\mathbf{W}\right)^{-1} = \mathbf{I}_N + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \rho^3\mathbf{W}^3 + \ldots \end{aligned} \tag{5}$$

[2] SDM model is presented in regard to our empirical analysis.

3 DATA AND EMPIRICAL RESULTS

The data set consists of the regional data for the employment rates (expressed in %) of population aged 25-64 across 259 NUTS 2 regions of the EU[3] over the period 2013-2018, GDP at current market prices in PPS per inhabitant in 2017, population density expressed as inhabitants per square kilometre in 2017, educational attainment level (expressed in %, 2017) of population aged 25-64 with upper secondary, post-secondary non-tertiary and tertiary education (levels 3-8) and compensation of employees in millions of euro per employee in 2016. The data were retrieved from the Eurostat web site (http://ec.europa.eu/eurostat/) and in the econometric models were considered in the form of natural logarithms. The employment rate of 2018 is supposed to depend on all the above-mentioned variables[4].

The regional averages of employment rate during the period 2013-2018 had the growing trend as illustrated by boxplot in Figure 1. Even though the interquartile range indicating the difference between the first and third quartiles as well as the standard deviation had the declining tendency, the employment rate disparities between well-developed regions and less-developed regions remained during the analysed period huge, reaching almost 40 percentage points. The boxplot identifies furthermore several lower outliers and no upper outliers.

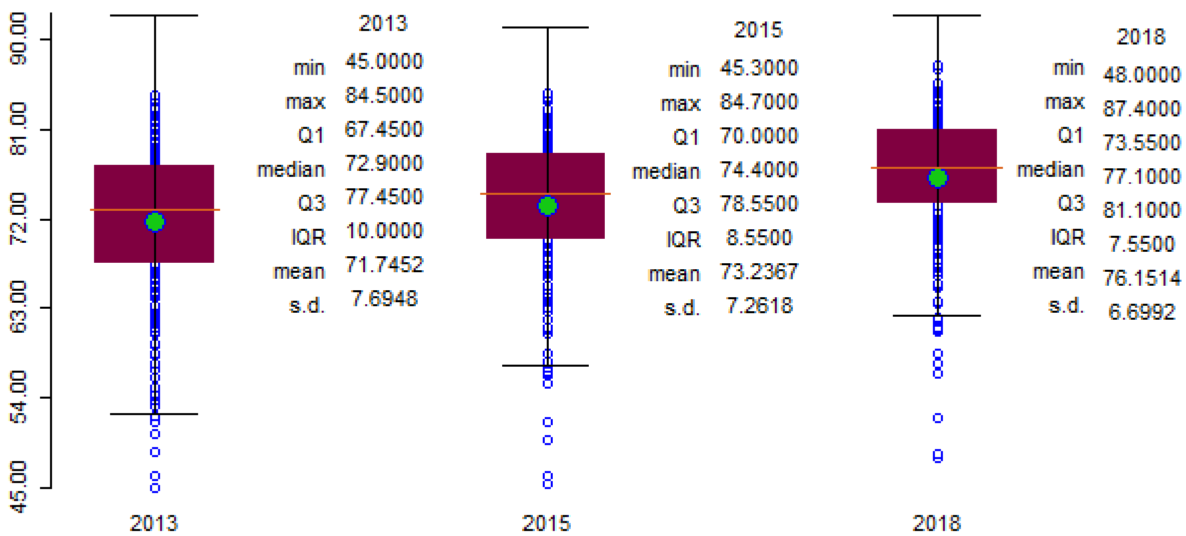


Figure 1: Boxplot of employment rate in 2013, 2015 and 2018
Source: authors' calculations in GeoDa

To have a precise look on the regional employment rate in 2018, the box map in Figure 2 clearly illustrates the unequal distribution of employment rate level over space. We can identify 14 regions with the employment rates between 48 – 62.225% located in Spain, southern part of Italy and Greece. The lowest employment rate of 48 % was detected for the Italian region Calabria. Employment rates of 62.225 – 73.55% were recorded in 51 regions located in Austria, Belgium, Bulgaria, Croatia, France, Greece, Hungary, Italy, Poland, Romania and Slovakia. Slightly higher employment rates of 73.55 – 77.1% were in the regions of different countries located mainly in eastern, south-eastern, western and south-western part

[3] After exclusion of isolated regions, the remaining data set consisted of 260 NUTS 2 regions, but due to the unavailability of some data, the region of Estonia was excluded, as well.

[4] These variables were denoted as follows: GDP in PPS per inhabitant – *GDP*, population density – *DEN*, educational attainment level – *EDU* and compensation of employees per employee – *COM*.

of the EU as well as regions located in Finland, UK and Ireland. The regions belonging to the next two categories are located mainly in central and northern part of the EU. The Swedish region Stockholm was the region with the highest employment rate of 87.4%. Extreme differences are visible also inside some analysed countries, e.g. among the Slovak regions, Italian regions and Austrian regions.

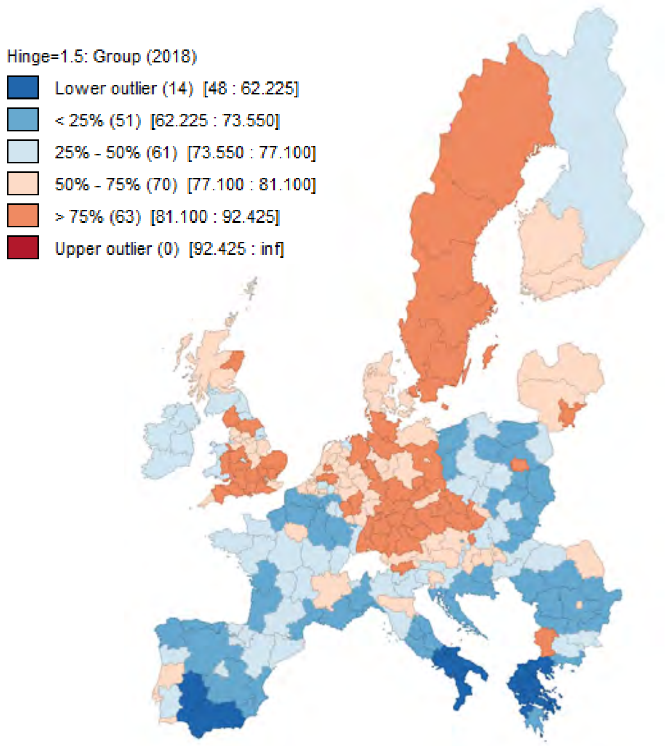


Figure 2: Box map of employment rate in 2018
Source: authors' calculations in GeoDa

The next step of our analysis was devoted to estimation of an econometric model of employment rates. The spatial dependence in analysed data was confirmed by calculation of Moran's *I* statistics[5] for all above mentioned variables.

We start our econometric analysis in accordance with "from general to specific" strategy, i.e., first we perform OLS (Ordinary Least Square) estimation (OLS model) and next we proceed with spatial modification of this model. As a part of the initial analysis, we have estimated several spatial versions of the OLS model (we do not present results due to insufficient space). We followed the *LM* test specifications as well as our assumption of the existence of global spillover effects in relation to modelling of regional employment. Finally, Table 1 provides estimation results of OLS model and its chosen spatial version – SDM model defined in (1). The estimation of SDM model was done by SML (Spatial Maximum Likelihood) estimator. Due to the confirmed spatial autocorrelation based on OLS estimation (LR test, LM tests), we will not pay further attention to OLS model. As for SDM model, we cannot ignore the fact that it is a model with global spillover effects and consequently statistical verification and interpretation of the parameters is more complicated (see section 2). Following LeSage and Pace (2009), we realized calculations and statistical verification of cumulative direct, indirect and total impacts (see Table 2).

[5] For the mathematical formula see e.g., Anselin and Rey (2014).

Table 1: Estimation results – OLS and SDM models

Estimation	OLS model	SDM model
	OLS	SML
α	2.0629***	0.3047***
$\beta_1 (\ln \mathbf{GDP})$	0.0865***	0.0586***
$\beta_2 (\ln \mathbf{DEN})$	-0.0040	-0.0118***
$\beta_3 (\ln \mathbf{EDU})$	0.3102***	0.3105***
$\beta_4 (\ln \mathbf{COM})$	0.0145	0.0161
$\gamma_1 (\mathbf{W} \ln \mathbf{GDP})$	–	0.0168
$\gamma_2 (\mathbf{W} \ln \mathbf{DEN})$	–	0.0122***
$\gamma_3 (\mathbf{W} \ln \mathbf{EDU})$	–	-0.2442***
$\gamma_4 (\mathbf{W} \ln \mathbf{COM})$	–	-0.0406**
ρ		0.7045***
R–squared	0.4560	–
Log likelihood	–	400.8311
Tests		
Moran's *I* (residuals)	12.02***	–
LM (lag)	140.6552***	–
Robust LM (lag)	6.0844**	–
LM (error)	146.6199***	–
Robust LM (error)	12.0491***	–
LR test	–	134.87***

Notes: Symbols ***, ** in both tables of the paper indicate the rejection of H_0 hypotheses at 1% and 5 % level of significance, respectively.
LR – Likelihood Ratio, *LM* – Lagrange Multiplier.
Source: authors' calculations in R

Table 2: Summary of direct, indirect and total impacts

	$\ln \mathbf{GDP}$	$\ln \mathbf{DEN}$	$\ln \mathbf{EDU}$	$\ln \mathbf{COM}$
Parameter estimate (β_1, β_2, β_3, β_4)	0.0586	-0.0118	0.3105	0.0161
Average direct impact (*ADI*)	0.0757***	-0.0106***	0.3030***	0.0074
Difference *ADI* and parameter estimate	0.0171	0.0012	-0.0075	-0.0086
Parameter estimate (γ_1, γ_2, γ_3, γ_4)	0.0168	0.0122	-0.2442	-0.0406
Average indirect impact (*AII*)	0.1794***	0.0121	-0.0786	-0.0906**
Difference *AII* and parameter estimate	0.1626	-0.0001	0.1656	-0.0499
Average total impact (*AT*I)	0.2551***	0.0015	0.2244***	-0.0831**
AII/*ATI*	0.7033	8.1980	-0.3505	1.0895
ADI/*ATI*	0.2967	-7.1980	1.3504	-0.0895

Source: authors' calculations in R

Let us focus on potential spillover effects, e.g. *GDP* variable. All impacts associated with this variable are statistically significant and have expected positive signs. The average direct impact does not match the estimate of the parameter β_1 and this difference (0.0171 – see Table 2) is the amount of feedback effects among the regions. Also, there is a difference (0.1626) between estimate of parameter γ_1 (spatial lag of *GDP*) and average indirect impact. If we perceived estimate of parameter γ_1 as an indirect impact, our conclusions regarding the *GDP* spillover effects would be wrong. Also average total impact equals 0.2551 and this impact if we just sum up the corresponding parameter values ($\beta_1 + \gamma_1$) would be equal to 0.0754, i.e. more than three times smaller and again our conclusions would be wrong. Another interesting fact is that up to 70% of total impact is attributed to indirect impact and only 30% to direct impact. We can conclude that neglecting the spatial interactions among regions can lead to truly misleading conclusions. Of course this is not only the case of *GDP* variable but also other variables[6] in our SDM model (see Table 2).

4 CONCLUSION

This paper was focused on spatial econometric analysis of regional employment rates in the EU, emphasizing the importance of spatial regional interactions among regions. Even, the initial spatial analysis based on the global spatial autocorrelation statistic and *LM* tests confirmed the assumption that the regional employment process is not a spatially isolated process. Consequently, we constructed several spatial econometric models but only SDM model is presented in paper. We prefer SDM specification for reasons of the ability of this model to capture global spillovers. Based on the SML estimates, we were able to quantify and statistically verify the summary measures of the direct, indirect and total impacts of the chosen employment determinants. The results clearly confirmed significant share of neighbouring regions as for *GDP* and compensation of employees in explaining regional employment rates. As for remaining factors, we found out that educational attainment level and population density at neighbouring regions do not have significant impact on employment rate performance of the particular region. The spatial decomposition of all impacts into marginal impacts corresponding to particular degrees of neighbourhoods would be a useful enrichment of our next research.

Acknowledgement
This work was supported by the Grant Agency of Slovak Republic – VEGA grant no. 1/0248/17 „Analysis of regional disparities in the EU based on spatial econometric approaches“.

References
[1] Anselin, L., Rey, S. J., 2014. Modern Spatial Econometrics in Practice. Chicago: GeoDa Press LLC.
[2] Chocholatá, M., Furková, A. 2018. The analysis of employment rates in the context of spatial connectivity of the EU regions. Equilibrium. Quarterly Journal of Economics and Economic Policy, 13(2): 181–213.
[3] European Commission 2010. Europe 2020. A European strategy for smart, sustainable and inclusive growth. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF [Accessed 05/02/2015].
[4] Franzese, R. J., Hays, J. C. 2005. Spatial econometric modeling, with application to employment spillovers and active-labor-market policies in the European Union. http://www-personal.umich.edu/~franzese/FranzeseHays.SpatialEcon.EmploymentSpillovers.ALP.pdf [Accessed 05/05/2018].
[5] LeSage, J. P., Pace, R. K. 2009. Introduction to Spatial Econometrics. Boca Raton, London & New York: Chapman & Hall/CRC.
[6] Pavlyuk, D. 2011. Spatial analysis of regional employment rates in Latvia. Scientific Journal of Riga Technical University, 2: 56-62.
[7] Perugini, C., Signorelli, J. 2004. Employment performance and convergence in the European countries and regions. European Journal of Comparative Economics, 1(2): 243-278.
[8] http://ec.europa.eu/eurostat/ [Accessed 05/05/2019].

[6] We do not interpret the results due to insufficient space.

PERCEPTIONS ON SOCIAL SUPERMARKETS' MANAGERS IN CROATIA, LITHUANIA, POLAND AND SERBIA

Blaženka Knežević, PhD
University of Zagreb, Faculty of Economics and Business, Department of Trade and International Business
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: bknezevic@efzg.hr

Petra Škrobot, PhD Candidate
University of Zagreb, Faculty of Economics and Business, Department of Trade and International Business
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: pskrobot1@efzg.hr

Berislav Žmuk, PhD
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: bzmuk@efzg.hr

Abstract: This paper deals with social supermarket as a new type of organizations emerged during economic crisis in European Union. Social supermarkets offer food, toiletries and other products to people in severe material deprivation. Their operation is based on donations of food, toiletries and other products. Therefore, reputation and recognition of social supermarkets' manager in the local community together with his/hers knowledge and skills determine their effectiveness and directly their success. In this paper, based on a primary research in Croatia, Serbia, Poland and Lithuania the general perception on social supermarkets' managers is analysed. The analysis has shown that the structure of respondents from the four observed countries who are agreeing that social supermarket managers should have a great reputation in the local community and that the reputation of a social supermarket manager greatly influences the success of collecting donations and fundraising activities can be considered to be at the same level.

Keywords: social supermarket, managers, Central and Eastern Europe, food waste reduction, poverty reduction

1 INTRODUCTION

Social supermarket is a new and specific form of social enterprises [7, 15] and a new retail format [1, 13]. As a new type of organizations, it fosters positive social change by fulfilling material needs of the socially disadvantaged groups and giving them an opportunity to preserve their dignity in an environment where they can choose various kinds of goods at extremely low prices [4, 14]. In some studies, authors emphasize that social supermarkets are nonprofit organizations that base their activity on volunteerism and charity and if they generate any profits, they use them for charitable activities [7].

There are two important operational characteristics in such form of organization: (1) authorized customers (people in material deprivation or at the risk of poverty) and (2) products donated free of charge [17]. Some social supermarkets do not sell products, but distribute products free of charge to people in need [12]. Moreover, based on conducted primary research [6, 12], social supermarkets dominantly collect donations of food and toiletries: (a) directly from producers, (b) from fast moving goods (and/or grocery) retailers and (c) from individuals. The structure of donation sources varies from county to country and the legal frameworks regarding food donations directly influences the structure of donation sources.

As a new type of organization developing since the recent economic crisis in European Union (since 2008), social supermarkets are subject to scientific research from various stand points: (1) theoretical where the aim is to define their mission, goals and position in food supply chains and in food waste prevention [4, 13, 17], (2) stakeholder's approach [4, 10], (3) retail mix structure analysis approach [1, 13], (4) comparative approach to operational characteristics in various markets [4, 6, 11].

However, the approach that addresses the perception of general public towards the role, position and knowledge of social supermarkets' managers is not analysed sufficiently. Especially, this topic is insufficiently researched in Central and Eastern European Countries (CEE). Therefore, in this paper we will focus to perceptions on social supermarkets' managers in four countries in CEE region: Croatia, Serbia, Poland and Lithuania. After the introduction, details about the conducted primary research are provided and the methodology used in the paper is listed. Descriptive statistics methods were used to describe respondents' main characteristics in chapter 3 and their perception of social supermarkets in chapter 4. In chapter 5 attitude of respondents towards social supermarket managers is inspected. Final chapter brings conclusions and suggestions for further research.

2 DATA AND METHODOLOGY

The social supermarkets topic is considered to be specific and sensitive topic. Therefore, web survey based on snowball sampling approach in selecting respondents was conducted [9]. At the beginning the invitation for participation in the survey was sent to overall 20 scientists who work at universities in Croatia, Lithuania, Poland and Serbia. The survey lasted from April to May 2018. In that period overall 419 fully completed questionnaires were received.

The questionnaire itself contained questions about demographic characteristics of respondents (gender, age, working status). After that respondents were asked about their general opinion about social supermarkets, about social supermarket managers and, finally, about the role of frameworks and institutions in the social supermarket area. Almost all questions about social supermarkets are presented in the Likert scale form [8]. Consequently, due to the non-probabilistic survey design and the used design of questionnaire questions, in the analysis focus will be given to descriptive statistical analysis. However, in addition, the non-parametric chi-square tests for equality of three or more proportions will be used as well [2]. The chi-square test will be used mainly to inspect questions given in the Likert scale form. It has to be emphasized that, the chi-square test requires that in each cell at least expected value of five is reached [3, 16]. The total sample size of 419 respondents includes all respondents from all four observed countries. Therefore, when those 419 respondents are split across the countries some problems with expected values of cells could appear. In such cases some categories at Likert scale questions are going to be merged to provent appearance of minimum expected cell values problems.

3 ANALYSIS OF RESPONDENTS' MAIN CHARACTERISTICS

In the web survey participated overall 419 respondents. The most respondents are from Poland (123) whereas the least number of respondents is achieved in Lithuania (71). In Croatia 117 respondents have taken part in the web survey. Overall 108 respondents are coming from Serbia. The structure of respondents according to their main demographic characteristics is given in Figure 1.

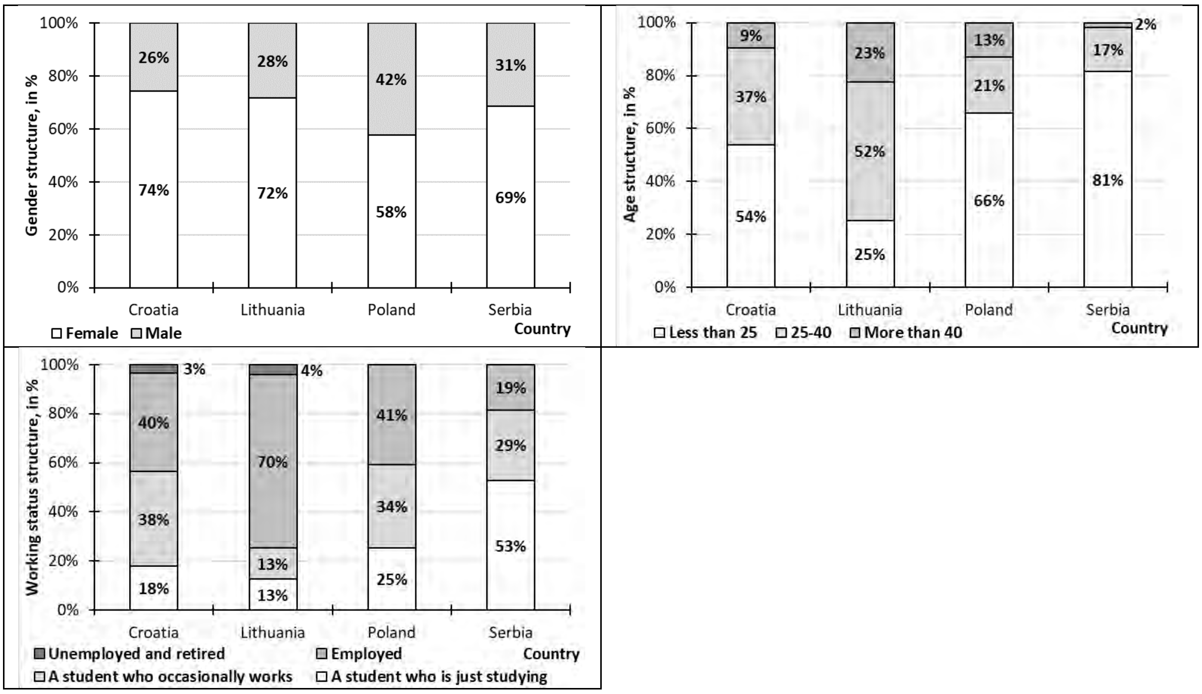


Figure 1 Main demographic characteristics of respondents – gender, age and working status structures

4 PERCEPTION TOWARDS SOCIAL SUPERMARKETS' PRESENCE AND MISSION

The respondents were asked about their opinions regarding social supermarkets presence in their city. The most respondents stated either that there is social supermarket in their city or that it would be very useful to have a social supermarket in the city. Detailed distribution of answers is provided in Figure 2 (left). In addition, respondents were asked to provide an answer whether they see reduction of food waste or reduction of poverty as the top priority of social supermarket mission. The distribution of answers is provided in Figure 2 (right).

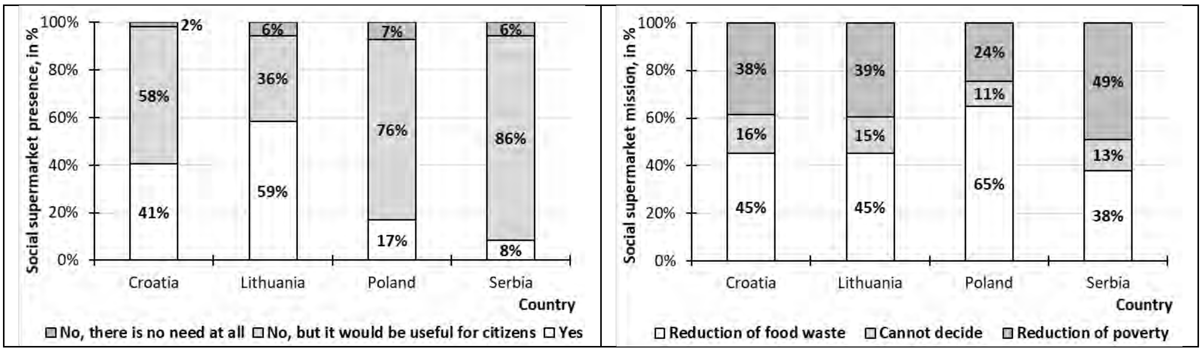


Figure 2 Social supermarkets presence (left) and Top priority of social supermarket mission (right)

5 SOCIAL SUPERMARKET MANAGERS

In the following part of the survey, respondents got a set of questions related to social supermarket managers. Overall five questions covered that area and all questions are presented in five-items Likert scale forms ranging from completely disagree (code 1) to completely agree (code 5).

Table 1 Main descriptive statistics results of social supermarkets managers variables

Variable	*Country*	*No of resp.*	*Mean*	*St. dev.*	*Coef. var.*	*Median*	*Mode*
Social supermarket managers must develop negotiating skills in order to establish the current flow of donations from companies, agricultural business and craftsmen, not just from individual citizens.	Croatia	117	4.41	0.66	15	5	5
	Lithuania	71	4.01	0.80	20	4	4
	Poland	123	4.04	0.81	20	4	4
	Serbia	108	4.45	0.72	16	5	5
	Overall	419	4.25	0.77	18	4	5
Social supermarket managers need more support in administrative procedures.	Croatia	117	4.30	0.75	17	4	5
	Lithuania	71	3.59	0.90	25	3	3
	Poland	123	3.73	0.92	25	4	3
	Serbia	108	4.18	0.86	21	4	5
	Overall	419	3.98	0.90	23	4	4
Social supermarket managers should have broader knowledge of management, marketing and/or finance.	Croatia	117	4.27	0.75	18	4	5
	Lithuania	71	3.79	0.94	25	4	4
	Poland	123	3.70	1.04	28	4	4
	Serbia	108	4.26	0.77	18	4	5
	Overall	419	4.02	0.92	23	4	4
Social supermarket managers should have a great reputation in the local community.	Croatia	117	4.21	0.81	19	4	5
	Lithuania	71	4.32	0.86	20	5	5
	Poland	123	4.14	0.85	21	4	5
	Serbia	108	4.15	0.95	23	4	5
	Overall	419	4.19	0.87	21	4	5
The reputation of a social supermarket manager greatly influences the success of collecting donations and fundraising activities.	Croatia	117	4.12	0.78	19	4	4
	Lithuania	71	4.31	0.86	20	5	5
	Poland	123	3.98	0.97	24	4	4
	Serbia	108	4.05	1.07	26	4	5
	Overall	419	4.09	0.93	23	4	5

In Table 1 main descriptive statistics results for the given social supermarkets managers' variables are presented. Again, those result are presented only to get insight about the distribution of answers and for comparison between other countries and overall. Generally speaking, the results pointed out that the respondents tend to agree or completely agree with the given statements in the observed variables.

The results of reliability analysis are shown in Table 2. Conducted reliability analysis resulted in Cronbach's alpha ranging from 0.6830 to 0.7927. On that way it can be concluded that the internal consistency here is questionable to acceptable [5, 18].

Table 2 Reliability analysis of social supermarkets managers variables, number of variables=5

Country	*No of responses*	*Cronbach's alpha*	*Standardized alpha*	*Average inter-item correlation*
Croatia	117	0.7927	0.7922	0.4465
Lithuania	71	0.6830	0.6843	0.3118
Poland	123	0.7254	0.7313	0.3579
Serbia	108	0.7838	0.7851	0.4273
Overall	419	0.7527	0.7548	0.3886

The results of conducted chi-square tests for equality of three or more proportions of social supermarkets managers variables are given in Table 3. According to the results, at significance level of 5%, the null hypothesis can be rejected in three, from five, cases. So, at three variables the structure of respondents who agreed or completely agreed with the given statements in a country is different than at other country.

Table 3 Chi-square tests for equality of three or more proportions of social supermarkets managers variables, responses agree and completely agree observed together

Variable	***No of responses***				***Com. prop.***	***Emp. Chi-square***	***p-value***
	Croatia	***Lithuania***	***Poland***	***Serbia***			
Social supermarket managers must develop negotiating skills in order to establish the current flow of donations from companies, agricultural business and craftsmen, not just from individual citizens.	106	51	91	96	0.8210	19.7452	0.0002
Social supermarket managers need more support in administrative procedures.	101	32	71	85	0.6897	47.4689	<0.0001
Social supermarket managers should have broader knowledge of management, marketing and/or finance.	96	43	76	89	0.7255	22.8500	<0.0001
Social supermarket managers should have a great reputation in the local community.	95	61	95	82	0.7947	3.2300	0.3575
The reputation of a social supermarket manager greatly influences the success of collecting donations and fundraising activities.	94	55	89	77	0.7518	3.2673	0.3522

However, at variables “social supermarket managers should have a great reputation in the local community” and “the reputation of a social supermarket manager greatly influences the success of collecting donations and fundraising activities” differences in proportions of respondents who are agreeing with the statements between the observed countries seems not to be statistically significant.

6 CONCLUSION

Primary research shows that perceptions on role, position and knowledge of social supermarkets’ managers in four countries in CEE region (Croatia, Serbia, Poland and Lithuania) is positive. In all countries, majority respondents completely agree or agree that social supermarkets’ managers should have a great reputation in the local because their reputation greatly influences the success of collecting donations and fundraising activities.

In addition, majority of respondents agree that they should develop their knowledge and skills in order to be able to manage current flow of donations from companies and agricultural businesses, not only from individual citizens.

However, conducted chi-square tests for equality of three or more proportions of social supermarkets managers variables shows that same structure of agreement is observed only for two out of five variables for all four countries, i.e. for statements “social supermarket managers should have a great reputation in the local community” and “the reputation of a social supermarket manager greatly influences the success of collecting donations and fundraising activities”. The largest difference in structure of answers is observed for statement “social supermarket managers need more support in administrative procedures” where in Croatia and Serbia majority of respondents strongly agree, while in Lithuania and Poland modal answer is neutral.

The main limitation of the research is use of the non-probabilistic survey design. Therefore, the research findings cannot be generalized. Except using a probabilistic survey design, for the future research improvements of the questionnaire are needed. On that way, use of more different statistical methods will be possible and more relevant conclusions could be brought.

In the further research it should be investigated how the business processes in social supermarkets can be further improved by joint action of all involved stakeholders.

References

[1] Bogetic, Z., Petkovic, G., Knezevic, B. 2018. Retail mix and its Specifics in Social Supermarkets. In Knezevic, B. (Ed.), Social supermarkets as entrepreneurial ventures in socially responsible economy (pp. 73-88). Zagreb: University of Zagreb, Faculty of Economics and Business.

[2] Bolboacă, S. D., Jäntschi, L., Sestraş, A. F., Sestraş, R. E., Pamfil, D. C. 2011. Pearson-Fisher Chi-Square Statistic Revisited. Information, 2(3):. 528-545.

[3] Cochran, W. G. 1952. The χ^2 test of goodness of fit. Annals of Mathematical Statistics, 23(3): 315-345.

[4] EU Fusions. 2015. Advancing social supermarkets across Europe. https://www.eu-fusions.org/phocadownload/feasibility-studies/Supermarkets/Advancing%20social%20supermarkets%20report.pdf [Accessed 25/05/2019].

[5] George, D., Mallery, P. 2003. SPSS for Windows step by step: A simple guide and reference. Boston: Allyn & Bacon.

[6] Holweg, C, Lienbacher, E. (Eds.) 2016. Social Supermarkets in Europe – Investigations from a retailing perspective in selected European countries, Institute for Retailing & Marketing, Vienna University of Economics and Business, Wien.

[7] Holweg, C., Lienbacher, E. 2011. Social Supermarkets – a New Challenge in Supply Chain Management and Sustainability. Supply Chain Forum, 23(4): 307-326.

[8] Joshi, A., Kale, S., Chandel, S., Pal, D. K. 2015. Likert Scale: Explored and Explained. British Journal of Applied Science & Technology, 7(4): 396-403.

[9] Kish, L. 1995. Survey Sampling. New York: John Wiley & Sons.

[10] Klindzic, M., Knezevic, B., Maric, I. 2016. Stakeholder Analysis of Social Supermarkets. Business Excellence, 10(1): 151-165.

[11] Knežević, B. 2018. Perception towards social supermarket mission and concept-primary research results. In Knezevic, B. (Ed.), Social supermarkets as entrepreneurial ventures in socially responsible economy (pp. 133-145). Zagreb: University of Zagreb, Faculty of Economics and Business.

[12] Knezevic, B., Škrobot, P. 2018. Social supermarkets' profiles in Croatia. In Knezevic, B. (Ed.), Social supermarkets as entrepreneurial ventures in socially responsible economy (pp. 147-158). Zagreb: University of Zagreb, Faculty of Economics and Business.

[13] Linebacher, E. 2012. Corporate Social Responsibility im Handel. Springer Gabler.

[14] Maric, I., Knezevic, B. 2014. Social Supermarkets as a New Retail Format Inspired by Social Needs and Philantrophy - Case of Croatia. In K., Demetri (Ed.), Global Business & Economics Anthology (pp. 278-286). Danvers, MA, USA: Business & Economics Society International.

[15] Maric, I., Knezevic, B., Dzambo, D. 2015. Social Supermarket Rijeka as a Social Innovation in Food Distribution. In Knego, N., Renko, S., Knežević, B. (Eds.), Trade Perspectives 2015: Innovations in Food Retailing (pp. 235-245). Zagreb: Faculty of Economics Zagreb.

[16] Roscoe, J. T., Byars, J. A. 1971. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. Journal of the American Statistical Association, 66(336): 755-759.

[17] Schnedlitz, P. (Ed.) 2011. Strukturanalyse Sozialmärkte in Österreich (Structural analysis of social supermarkets in Austria), Schriftenreihe Handel und Marketing, Vol. 74, Vienna.

[18] StatSoft. 2017. Example 1: Evaluating the Reliability of Items in a Questionnaire. http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Reliability/ReliabilityandItemAnalysis/Examples/Example1EvaluatingtheReliabilityofItemsinaQuestionnaire [Accessed 14/02/2019].

STRUCTURAL EQUATION MODELING IN THE CASE OF OLDER EMPLOYEES IN FINANCIAL SERVICE COMPANIES

Maja Rožman, Vesna Čančer
University of Maribor, Faculty of Economics and Business, Razlagova 14, 2000 Maribor, Slovenia
E-mails: maja.rozman1@um.si, vesna.cancer@um.si

Abstract: The paper presents and discusses the non-linear links between the individual constructs of the conceptual model of managing older employees in financial service companies in Slovenia. The analysis of the data set was based on exploratory and confirmatory factor analysis using structural equation modeling. The paper seeks to determine the effects of 1) leadership and employee relations on work satisfaction and 2) work satisfaction on work engagement in the case of older employees in financial service companies in Slovenia. The results show that both of the aforementioned effects are positive.

Keywords: WarpPLS, structural equation modeling, older employees, financial services companies

1 INTRODUCTION

Structural equation modeling (SEM) has been proven to be useful in exploring the links between constructs in various human resource management (HRM) multidimensional models [15]. Our research focuses on one of the important aspects of the sustainable profitability of a business: the satisfaction and engagement of older employees. Based on the HRM theoretical background, we built a conceptual multidimensional model of managing older employees in financial service companies in Slovenia. As we wanted to determine the impact of leadership and employee relationship—two important components of work satisfaction of older employees in financial service companies that further shape the level of work engagement of older employees—the conceptual model includes the following constructs: leadership, employee relations, employee satisfaction and employee work engagement. The main objective of this paper is to determine the impact of leadership and employee relations on work satisfaction of older employees, as well as to determine the impact of work satisfaction on the work engagement of older employees in financial service companies in Slovenia.

The findings of research [16] based on SEM analysis show that employee relations and leadership have a positive impact on employee satisfaction. The authors' findings indicate that older employees have good relationships with their colleagues and supervisors and therefore work in a more supportive work environment, have better health status and are more satisfied. Another study [13] suggests that work satisfaction is an important driver of work engagement. Additional findings [4] show that there is a positive relationship between work satisfaction, work motivation and work engagement. According to [10, 11], SEM is based on the linear or non-linear connections between constructs. The results obtained by WarpPLS show that the observed links in our model are non-linear. This paper aims to verify the following hypotheses:
H1: Leadership has a significant positive impact on the work satisfaction of older employees in financial service companies in Slovenia.
H2: Employee relations have a significant positive impact on the work satisfaction of older employees in financial service companies in Slovenia.
H3: Employee satisfaction has a significant positive impact on the work engagement of older employees in financial service companies in Slovenia.

2 DATA AND METHODOLOGY

The main survey involved 237 large- and medium-sized financial service companies, and from each company we selected up to three employees to participate in our research. Thus, 704 older employees responded to the questionnaire. In the literature, the definitions of older employees vary [2, 8]. In this paper, employees of ≥ 50 years of age were defined as older employees.

The respondents indicated on a 5-point Likert-type scale their agreement to the listed statements, where 1 = strongly disagree and 5 = completely agree (1 = I completely disagree, 2 = I do not agree, 3 = I partially agree, 4 = I agree, 5 = I completely agree). Items for the leadership construct were formed by [1], for the employee relations construct by [5], for the employee satisfaction construct by [7] and for the employee engagement construct by [13]. Within the empirical part, we established the justification to use the factor analysis on the basis of the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO ≥ 0.5) [9] and Bartlett's test of sphericity. With the purpose to improve the factors' interpretability and achieve a more even distribution of variance according to the factors, the rectangular rotation Varimax, which maximizes the variance of weight squares in every factor and simplifies the structure by columns, was used [6, 12]. Also, fulfillment of criteria regarding factor loadings ($\eta \geq 0.5$), communalities of variables ($h > 0.4$), and eigenvalues of factors ($\lambda \geq 1.0$) was analyzed [14]. The quality of the measurement model was measured by the variance explained for a particular construct. We checked the reliability of measurements within the scope of inner consistency with Cronbach's alpha coefficient [3]. As part of the convergent validity, we examined average variance extracted (AVE) and composite reliability coefficients (CR), keeping in mind the criteria AVE > 0.5 and CR > 0.7 and the criterion CR > AVE. In order to check for multicollinearity, we used variance inflation factors (VIF), considering the criterion VIF < 5.0 [6]. The quality of the structural model was measured by the R-squared and adjusted R-squared coefficients, reflecting the percentage of explained variance of latent variables in the structural model and the Stone-Geisser Q-squared coefficient. Thus, we examined the predictability value of the structural model. Acceptable predictive validity in connection with an endogenous latent variable is suggested by $Q^2 > 0$ [10]. To test the model, the following rules were also applied: average path coefficient (APC, $p < 0.05$), average R-squared (ARS, $p < 0.05$), average adjusted R-squared (AARS, $p < 0.05$), average block variance inflation factor (AVIF < 5.0), average full collinearity VIF (AFVIF < 5.0), goodness-of-fit (GoF ≥ 0.36), Sympson's paradox ratio (SPR ≥ 0.7), the R-squared contribution ratio (RSCR ≥ 0.9), statistical suppression ratio (SSR ≥ 0.7) and nonlinear causality direction ratio (NLBCD ≥ 0.7) [10, 11, 14]. To test the hypotheses, we used the path coefficient associated with a causal link in the model (γ) and indicator of Cohen's effect (f^2), with 0.02, 0.15, and 0.35 indicating the small, medium, and large effect sizes [10, 14]. The Statistical Package for the Social Sciences (SPSS) and WarpPLS software were used for data analysis.

3 RESEARCH ANALYSIS RESULTS AND DISCUSSION

The results in Table 1 show that the values of the measure of sampling adequacy and the results of Bartlett's test of sphericity for each construct (leadership, employee relations, work satisfaction, work engagement) suggest that the use of factor analysis is justified. The values of all communalities for all five construct are higher than 0.70; therefore, we have not eliminated any variable. Also, all factor loadings are higher than 0.70 and significant at the 0.001 level. For each construct, the one-dimensional factor solution was obtained. All measurement scales proved high reliability (all Cronbach's alpha > 0.80). In addition to the results in Table 1, the total variance explained for leadership is 81.8%, for employee relations is75.2 %, for employee satisfaction is 78.8% and for employee work engagement is 85.6%.

Table 1: Factor analysis results

Statement	***Factor label***	***Cronbach's alpha***	***Communalities***	***Factor loadings***
I have all necessary information to perform my work.	Leadership	0.967	0.749	0.870
I have everything I need to carry out my work tasks.			0.737	0.871
I have the possibility of independent thinking and decision-making in the workplace.			0.748	0.862
The company owner/manager fosters good relationships between employees			0.872	0.925
The company owner/manager of the company fosters good relationships between employees and superiors.			0.859	0.937
The company owner/manager emphasizes and encourages employee motivation in the workplace.			0.847	0.914
The company owner/manager ensures the work satisfaction and well-being of employees.			0.787	0.889
In the company, we have the possibility of training and education.			0.768	0.885
KMO = 0.919; Bartlett's Test of Sphericity: Approx. Chi-Square = 13416.452, df = 36, p < 0.01				
The presence of age discrimination is not felt among employees.	Employee relations	0.951	0.735	0.879
In the company, we do not feel the presence of age stereotypes.			0.757	0.886
There is no competition between older and younger employees in terms of who does better work.			0.714	0.779
We cooperate very well with colleagues in the performance of our tasks.			0.723	0.791
Employees appreciate the work of our colleagues.			0.855	0.903
In case of conflict, we solve the problem together and for the common benefit.			0.832	0.901
In the company, mutual trust and cooperation prevail.			0.842	0.922
KMO = 0.877; Bartlett's Test of Sphericity: Approx. Chi-Square = 5498.853, df = 15, p < 0.01				
At my workplace, I am satisfied with the working hours and distribution of work obligations.	Employee satisfaction	0.960	0.818	0.904
In this company, I am satisfied with the balance between my work and private life.			0.712	0.803
I am satisfied with the level of self-regulation of work speed that is enabled.			0.782	0.884
I am satisfied with the provision of job-sharing, which reduces the burden on the workplace.			0.746	0.843
I am satisfied with the interpersonal relationships in the company.			0.839	0.911
I am satisfied with the leadership in the company.			0.824	0.907
I am satisfied with the intergenerational cooperation and, thus, the distribution of work in the company.			0.827	0.910
KMO = 0.930; Bartlett's Test of Sphericity: Approx. Chi-Square = 7593.509, df = 21, p < 0.01				

Table 1 continued

Statement	***Factor label***	***Cronbach's alpha***	***Communalities***	***Factor loadings***
I do my work with passion.	Employee work engagement	0.964	0.862	0.948
I am engaged in the quality of my work.			0.847	0.936
I am engaged to achieve successful business results.			0.840	0.935
I feel connection with the company in which I work.			0.789	0.913
I am aware of the importance of innovation for our company and I am helping to develop the company.			0.790	0.927
I trust in my colleagues and the manager.			0.843	0.933
I feel that my work and job are important.			0.832	0.927
I am proud to be employed in this company.				0.911
I believe in the successful development and operation of the company.			0.796	0.934
KMO = 0.947; Bartlett's Test of Sphericity: Approx. Chi-Square = 13902.884, df = 36, p < 0.01				

Key quality assessment indicators of research model are presented in Table 2.

Table 2: Model fit and quality indicators

Quality indicators	***Criterion of quality indicators***	***Calculated values of indicators of model***
Average path coefficient (APC)	$p < 0.05$	0.638, $p < 0.001$
Average R-squared (ARS)	$p < 0.05$	0.897, $p < 0.001$
Average adjusted R-squared (AARS)	$p < 0.05$	0.897, $p < 0.001$
Average block variance inflation factor (AVIF)	AVIF < 5.0	2.096
Average full collinearity VIF (AFVIF)	AFVIF < 5.0	2.204
Goodness-of-fit (GoF)	GoF ≥ 0.1 (low) GoF ≥ 0.25 (medium) GoF ≥ 0.36 (high)	0.842
Sympson's paradox ratio (SPR)	SPR ≥ 0.7	1.000
R-squared contribution ratio (RSCR)	RSCR ≥ 0.9	1.000
Statistical suppression ratio (SSR)	SSR ≥ 0.7	1.000
Nonlinear causality direction ratio (NLBCD)	NLBCD ≥ 0.7	1.000

Table 2 shows that the indicators APC, ARS, AARS are statistically significant ($p < 0.001$), and the indicators AVIF and AFVIF are lower than 5.0 and are suitable. Indicator GoF shows the power of the underlying conceptual model [11], and the results of indicator GoF show that the model is highly appropriate. The values of indicators SPR, RSCR, SSR and NLBCD are higher than the minimal prescribed values and are suitable.

Table 3: Indicators of quality of structural model

Constructs	***Cronbach's α***	***CR***	***AVE***	***R^2***	***Adj. R^2***	***Q^2***	***VIF***
Leadership	0.976	0.979	0.808	(-)	(-)	(-)	1.473
Employee relations	0.948	0.957	0.736	(-)	(-)	(-)	1.687
Employee satisfaction	0.972	0.975	0.768	0.466	0.441	0.473	2.529
Employee work engagement	0.984	0.985	0.850	0.449	0.442	0.461	2.361

Note: (-) values cannot be calculated because the construct is a baseline

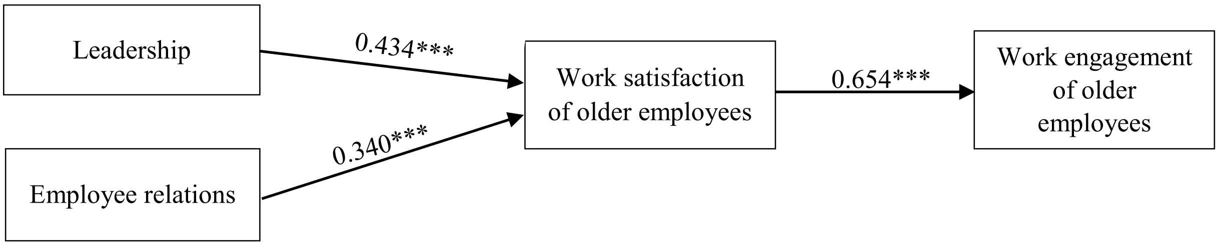
Table 3 indicates that the values of the latent variables' R^2, adjusted R^2 and Q^2 coefficients are greater than zero. Composite reliabilities (CR) for all five constructs are greater than 0.7. Also, values of AVE for all five constructs are greater than 0.5. As all CR values were higher than

AVE values, we confirmed the convergent validity for all the constructs studied. The VIF values ranged between 1.473 and 2.529 (VIF < 5.0), providing confidence that the structural model results were not affected by collinearity. The results of SEM and structural coefficients of links of the basic structural model are presented in Table 4. Also, Figure 1 presents the conceptual model with the values of path coefficients.

Table 4: Standardized Path Coefficients for Proposed Model

Hypothesized path	***Link direction***	***Shape of link***	***Path coefficient (γ)***	***Effect size (f^2)***	***Standard error***
LE→ES	Positive	Nonlinear	0.434***	0.206	0.032
ER→ES	Positive	Nonlinear	0.340***	0.213	0.034
ES→EWE	Positive	Nonlinear	0.654***	0.419	0.083

Note: ***p < 0.001; LE – leadership; ER – employee relations; ES – employee satisfaction; EWE – employee work engagement



Note: ***p < 0.001

Figure 1: Conceptual model of managing older employees with the values of path coefficients

The results in Table 4 show that leadership has a positive effect on the work satisfaction of older employees (LE→ES = 0.434, p < 0.001) in financial service companies. The value of Cohen's coefficient (f^2 = 0.206) is greater than 0.02 and shows that the effect of predictive latent variables is of medium strength. Also, employee relations have a positive effect on the work satisfaction of older employees (ER→ES = 0.340, p < 0.001). The value of Cohen's coefficient (f^2 = 0.213) shows that the effect of predictive latent variables is of medium strength. The results in Table 4 show that the work satisfaction of older employees has a positive effect on the work engagement of older employees (ES→EWE = 0.654, p < 0.001). The value of Cohen's coefficient (f^2 = 0.419) shows that the effect of predictive latent variables is of high strength. The results show that there is a non-linear connection between the individual constructs. We therefore verified and confirmed hypothesis 1 (leadership has a significant positive impact on the work satisfaction of older employees in financial service companies in Slovenia), hypothesis 2 (employee relations have a significant positive impact on the work satisfaction of older employees in financial service companies in Slovenia), and hypothesis 3 (employee satisfaction has a significant positive impact on the work engagement of older employees in financial service companies in Slovenia).

4 CONCLUSION

Based on the results, we found that leadership and employee relations have a positive effect on the work satisfaction of older employees in financial and insurance companies in Slovenia, as well as that work satisfaction has a positive effect on the work engagement of older employees. This is consistent with the findings of [1, 4, 13, 16], in which the authors found that leadership and employee relations have a positive impact on the work satisfaction of older employees and that the work satisfaction of older employees has a positive impact on the work engagement of older employees.

Our study is limited to the focus of older employees in Slovenia in medium-sized and large financial services companies. As an opportunity for future research, we recommend an upgrade of the measurement instrument with new constructs in the area of older employees. Also, our further research refers to analyzing different constructs (for example, stress, job burnout and work engagement) with structural equation modeling (SEM).

References

[1] Avery, D. R., McKay, P. F., Wilson, D. C. 2007. Engaging the aging workforce: The relationship between perceived age similarity, satisfaction with coworkers, and employee engagement. Journal of Applied Psychology, 92(6): 1542–1556.

[2] Brooke, L. 2003. Human resource costs and benefits of maintaining a mature-age workforce. International Journal of Manpower, 24(3): 260–283.

[3] Chronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. Psychometrika, 16(3): 297–334.

[4] Egan, T., Yang, B. and Bartlett, K. 2004. The effects of organizational learning culture and job satisfaction on motivation to transfer learning and turnover intention. Human Resource Development Quarterly, 15(3): 279-301.

[5] Gunnigle, P., Turner, T., Morley, M. 1998. Strategic integration and employee relations: the impact of managerial styles. Employee Relations, 20(2): 115–131.

[6] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. 2010. Multivariate data analysis. Upper Saddle River: Pearson Prentice Hall.

[7] Hayday, S. 2003. Questions to Measure Commitment and Job Satisfaction. https://www.employment-studies.co.uk/system/files/resources/files/mp19.pdf [Accessed 15/1/2019].

[8] Ilmarinen, J. 2001. Aging workers. Occupational & Enviromental Medicine, 55(8): 546–552.

[9] Kaiser, H. F. 1974. An index of factorial simplicity. Psychometrika, 39(1): 31–36.

[10] Kock, N. 2015. WarpPLS 5.0 User Manual. ScriptWarp Systems: Laredo, Texas. http://cits.tamiu.edu/WarpPLS/UserManual_v_5_0.pdf [Accessed 17/1/2019].

[11] Kock, N. 2016. Advantages of nonlinear over segmentation analyses in path models. International Journal of e-Collaboration, 12(4): 1-6.

[12] Manly, B. F. 2005. Multivariate statistical methods: A primer (3rd Edition). New York: Chapman & Hall/CRC.

[13] Robinson, D., Perryman, S., Hayday, S. 2004. The drivers of employee engagement. http://www.employmentstudies.co.uk/report-summary-drivers-employee-engagement [Accessed 17/1/2019].

[14] Tabachnick, B. G., Fidell, L. S. 2013. Using multivariate statistics (6th Edition). Boston: Pearson Education.

[15] Veingerl Čič, Ž., Šarotar Žižek, S., Čančer, V. 2017. Nonlinear connections in structural equation modeling: the case of service sector companies in Slovenia. In: Zadnik Stirn, L. (ed.), Zadnik Stirn, L., Kljajić Borštnar, M., Žerovnik, J., Drobne, S. (eds.). *SOR '17 proceedings*. Ljubljana: Slovenian Society Informatika, Section for Operational Research, pp. 350-355.

[16] Yang, T., Shen, Y. M., Zhu, M., Liu, Y. Deng, J., Chen Q., See, L. C. 2015. Effects of co-worker and supervisor support on job stress and presenteeism in an aging workforce: A structural equation modelling approach. International Journal of Environmental Research and Public Health, 13(1): 1-9.

NONRESPONSE IN BUSINESS WEB SURVEYS: SOURCES AND MEASURES

Berislav Žmuk
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: bzmuk@efzg.hr

Anita Čeh Časni
University of Zagreb, Faculty of Economics and Business, Department of Statistics
Trg J. F. Kennedyja 6, HR-10000 Zagreb, Croatia
E-mail: aceh@efzg.hr

Abstract: The nonresponse rates in web surveys have been often very high. Whereas the nonresponse in web surveys where the target population is individuals and/or households is quite good documented and investigated, there is a lack of research of nonresponse in business web surveys. Therefore, in the paper nonresponse in business web surveys is under the study. The main reasons for nonresponse in business web surveys are provided. Measurable main reasons for nonresponse and AAPOR standard definitions are used to suggest main elements for introducing new nonresponse measures.

Keywords: American Association for Public Opinion Research (AAPOR), business web survey, measuring nonresponse, nonresponse.

1 INTRODUCTION

Business surveys are a special kind of surveys in which target population is enterprises, firms or establishments, but the reporting units are employees of those enterprises. It has to be emphasized that there could be made some differences between term enterprise, firm and establishment. In [13] authors described the differences between those terms in detail. However, in order to keep appropriate simplicity level, in this paper all those three terms will be together covered by term business.

Despite the fact that business surveys have been conducted since the nineteenth century, only in recent period business surveys are being systematic and comprehensively treated [6]. In order to support the business surveys methodology International Conference on Establishment Statistics was initiated. However, since 1993 only five such conferences were held whereas the next one is scheduled to happen in 2020 [3]. If they are conducted on appropriate way, web surveys tend to have many advantages over other survey modes [11]. Consequently, it is not surprising that the popularity of web surveys is increasing. Couper [7] emphasizes importance of good quality of web surveys and provides a review of approaches and issues in web surveys. However, business web surveys are not represented well in a relevant literature and even if they are mentioned, only a couple of pages are devoted to them [5]. In general, the main problem of web surveys is low response rates [4, 16]. It has been shown that web survey rates are well below response rates achieved at other survey modes [12]. Consequently, low web survey response rates of 2% are not surprising any more. Similar response rates can be found at business web surveys as well [17].

The aims of the paper are to investigate and to list the main reasons of nonresponse in business web surveys and to single out main elements which can be used for making new nonresponse measures suited to business web surveys.

The paper is organized as follows. After the brief introduction to the research problem, in the second chapter main reasons of nonresponse in business web surveys are discussed. In the third chapter main elements needed for measuring nonresponse in business web surveys are investigated. The fourth chapter concludes the paper.

2 REASONS OF NONRESPONSE IN BUSINESS WEB SURVEYS

In order to get a response two steps should be done. Firstly, the units from target population should be successfully contacted and, secondly, they should be convinced enough to cooperate and participate in the survey [9]. The contact can be made by telephone, mail, telefax and personally. However, the most convenient way, for both researchers and respondents, is to send an e-mail invitation. That is a cheap and fast way of communication between those two sides.

No matter which mode of contact is preferred, it is suggested sending a prenotice to the respondents [15]. Prenotice has a role of making familiar respondents with a survey in which they are going to be invited to participate. However, prenotice if not written well it can have opposite effect of wanted. So, instead of decreasing nonresponse, prenotice can increase it. Therefore a researcher should pay attention how prenotice is written, which design is used, whether prenotice is personalized or not, and similar. In addition, a prenotice can be used to check respondents address or validity of e-mail address.

Generally speaking, a researcher should be very careful and think about every survey step which is under their control. But there are some survey parts which are out of researcher control. The list of main controllable and uncontrollable items regards nonresponse is provided in Table 1.

Table 1: Survey items under and out of researcher control (adopted from [16])

Under researcher control	***Out of researcher control***
• Respondent identification	• Business policy
• Contact strategies	• Environmental dependence
• Survey topic	• Availability of data
• Web survey questionnaire design	• Availability of resources
• Time schedules	• Respondents' authority
• Confidentiality	• Respondents' capacity
• Legal authority	• Respondents' motivation
• Survey sponsor	• Technological environment

It is noted earlier that prenotice can be used to check whether respondents contact data are correct or not. If a researcher is not sure in a sampling frame quality, prenotice can be used to estimate its validity. In case of business surveys the main problem with sampling frame is its datedness. Namely, the contact data of businesses can be changed and consequently no contact can be made. However, in business surveys data from official administrative sources are often used. Despite this fact, even in this case some problems with contact data are expected. The more wrong contact data are present in sampling frame, the higher nonresponse rate is.

Even if contact data is correct, it is questionable whether the right and competent person in business contacted is. That is another challenge for a researcher which becomes more emphasized in larger businesses. Contacting right person in business can be crucial in the business surveys. Namely, if wrong person in business is contacted, it is high probability that this person will not recognize importance of the survey and forward information about the survey to the right person who can provide answers to the questions in web survey.

If survey topic is of some interest of a business, it is more likely that it will participate in the survey. However, if the survey benefits are not recognized, businesses tend not to participate in the survey what leads to higher nonresponse rates. Therefore, a researcher should be careful about choosing survey title and how the survey will be described in the invitation letter.

Web survey questionnaire design should be good enough to keep interest of respondents from the beginning to the end of the questionnaire. On that way breakoffs, when respondent gives up from participating in the survey after certain number questions, and item nonresponse, when respondent does not provide answers on all questions, can be significantly avoided. It is hard to keep attention of employees for a long period. Because of that it is recommended that the questionnaire is as short and as simple as possible. Of course, an option to continue survey later on is highly recommended in business web surveys.

Researcher should be aware of time when the survey is conducted. It is not recommended to send any survey invoice to business in periods when annual and or tax reports should be made. In addition, dates of holidays should be also taken into account. Furthermore, the choice of day in a week and the time of day can play significant role as well [14]. Researcher should take optimal time for doing the survey on the field. In the other word the time span in which responses are collected should be neither too narrow nor too wide. In case of too narrow survey period maybe some business did not participate in the survey because they did not have enough time to spare but they would later do that. If the survey period is too long, there is high probability that the situation in the business, who participated in the surveys among the first ones, has changed and therefore they would choose other answers now.

Confidentiality of businesses and respondents is very important to them. If they do not trust the researcher that they will keep their identity secret, business will not participate in the survey. Therefore, the researcher must strictly follow the survey standards and obey privacy legislation. This should not be followed only during the process of collecting but in the case of publishing reports as well. In that case, researcher should take into account sensitive information and use disclosure control system. The system should be used especially when large enterprises or other enterprises with rare characteristics are observed.

There is great difference whether the survey is conducted by some institution like national bureau of statistics or some unknown person. Business will be more cooperative if they know who is doing the survey than if the survey is conducted by an unknown person. Therefore it is crucial to put names of survey sponsor, institutions who support the survey and other important persons who cast confidence.

Business policy is the first obstacle which is out of researcher control. A business can have simple policy of not participating in any survey. Unfortunately, not all enterprises share the information that they will not participate in the survey due to their business policy which makes difficult to find out the proportions of nonresponse due to this fact. In addition, such businesses are making additional costs for researchers who are trying to contact them.

If business climate is positive, it is more likely that business will participate in the survey. However, if economy is in recession and businesses have achieved negative business results, the rate of nonresponse should be higher.

Businesses differ in many things: size, main activity, ownership, and so on. Therefore the organisation of their data in business records can be different. Consequently, some of required information business could very easily provide whether some data are available only to very narrow number of people.

According to [16] availability of resources is the most important factor which determines businesses' ability to participate in the survey or not. As the first, a business should have some free and available employee who would take care about data collecting and participating in the survey. In addition, the question is for how long business can afford yourself that an employee does not work its main job but fills the survey up. Therefore businesses prefer to participate in surveys from which they could get some benefits and which are not too long, demanding and difficult to fill.

The great part of response burden is on respondents themselves. Respondents' authority refers to ability of taking responsibility for providing survey answers and releasing data about the business outside it [10]. Respondents' capacity refers to ability to understand survey questions, ability to collect needed information for survey and to ability of communicating the most appropriate answer in the survey for that business. Respondents' motivation refers to direct benefits which could they get and it is under high influence of their superiors.

Technological environment plays important role in web surveys. The main questions are whether respondents will: be able to open the web survey; see the questionnaire on the planned way; be able and know how to answer. Nowadays, web surveys can be filled in by using personal computer, tablet and smartphone. All that increases possibility that something will go wrong from technical point and which will result in nonresponse.

It has to be emphasized that the mentioned reasons of nonresponse in business web surveys are only the most important ones. Depending on the experience of a researcher some reasons of nonresponse in business web surveys can be set on negligible level.

3 MEASURING NONRESPONSE IN BUSINESS WEB SURVEYS

As stated in chapter 2, there are many different sources and reasons of nonresponse in business web surveys. The question is how to measure nonresponse on the most appropriate way. Unfortunately, the American Association for Public Opinion Research (AAPOR), the leading association of public opinion and survey research professionals [2], does not provide equations for calculating nonresponse but it is focused on measuring response rates (six measures), cooperation rates (four measures), refusal rates (three measures) and contact rates (three measures) [1]. In addition, those measures are written primary for face-to-face and phone surveys where sampling units are households. In Table 2 are given main elements of those measures.

Table 2: Main elements of response rate, cooperation rate, refusal rate and contact rate measures (adopted from [1])

Response rate	***Cooperation rate***	***Refusal rate***	***Contact rate***
Numerator: • Complete interview • Partial interview	Numerator: • Complete interview • Partial interview	Numerator: • Refusal and break-off	Numerator: • Complete interview • Partial interview • Refusal and break-off • Other
Denominator: • Complete interview • Partial interview • Refusal and break-off • Non-contact • Other • Unknown if household occupied • Unknown, other • Estimated proportion of cases of unknown eligibility that are eligible	Denominator: • Complete interview • Partial interview • Refusal and break-off • Other	Denominator: • Complete interview • Partial interview • Refusal and break-off • Non-contact • Other • Unknown if household occupied • Unknown, other • Estimated proportion of cases of unknown eligibility that are eligible	Denominator: • Complete interview • Partial interview • Refusal and break-off • Non-contact • Other • Unknown if household occupied • Unknown, other • Estimated proportion of cases of unknown eligibility that are eligible

AAPOR used the main elements listed in Table 2 and combined them to make different measures as certain ratios. Unfortunately, there is no corresponding measure which could be used as nonresponse measure. Namely, according to [8] "nonresponse occurs when a sampled unit does not respond to the request to be surveyed or to particular survey questions". However, very close to a nonresponse measures are refusal rate measures and the differences between 1 minus response rate measures. Still, there are not included items specific for business web surveys. Therefore in Table 3 the main elements which should be included in nonresponse measures are given.

Table 3: Main elements of nonresponse measures

Item	***Elements***
Sampling frame	• Contacted businesses (sent survey invitations by e-mail)
Business successfully received survey invitation e-mail	• Completed web questionnaires (employees have fully answered on all questions) • Partially completed web questionnaires (researcher must define what is mean by "partial") • Refusal (business received the e-mail with survey invitation and it has explicitly announced that it does not want to participate in the survey) • Break off (implicit refusal) • Attention problem (businesses excluding due to straight lining, speeding and similar) • Active looking (the business entered the survey but did not provide any answer) • Technical difficulties (business cannot run multimedia files and similar)
Business did not received survey invitation e-mail	• Wrong e-mail address (the e-mail is changed) • Mail-box quota exceeded (the e-mail inbox is full – the e-mail is not used anymore) • Out of office (business not available during the data collection period) • Other non-contact reasons (the survey invitation cannot be received due to different reasons)
Eligibility	• Ineligible businesses (businesses that turned out not to be eligible according to their answers) • Eligibility proportion (estimated proportion of businesses of unknown eligibility that are eligible)

The elements in Table 3 are listed by using appropriate elements from AAPOR measures and reasons of nonresponse in business web survey that can be measured. It is assumed that sampling frame from administrative sources is used and that it is additionally arranged for the survey purposes. Also, ability of collecting paradata is implied. Accordingly, following simple complete nonresponse measure of business web surveys can be written:

$$NR = \frac{CB - e \cdot (C + P) - IB}{CB - IB} \tag{1}$$

where NR is nonresponse ratio, CB is the number of contacted businesses, e is eligibility proportion, C is the number of completed web questionnaires, P is the number of partially completed web questionnaires, IB is the number of ineligible businesses. In dependence what a researcher wants to emphasize, from the equation 1 it could be taken or added some elements. Furthermore, instead of beginning from the number of contacted business, a researcher could begin by adding nonresponse elements in the numerator.

4 CONCLUSION

Due to their importance in modern society, it is of crucial importance to contact businesses and collect their thoughts, attitudes and other information about business related topics. The most convenient way, both for businesses and researcher, to do that is to conduct a web survey. However, due to different reasons businesses tend not to participate in such surveys and therefore nonresponse rates are high. Because of that it is very important to know the sources of nonresponse and pay additional attention to them. On that way, the nonresponse rate in a business web survey should be lowered.

In the paper main sources of nonresponses in business web surveys are listed and briefly explained. The impact strength of those sources is not the same in all business web surveys, but it depends on the target population, survey topic and similar. Still, there is a need to measure the nonresponse rate on an appropriate way. Thus, researchers could compare achieved nonresponse rates in their surveys with other survey. Moreover, researchers could estimate what their nonresponse rates should be by comparing with the achieved nonresponse rates in similar surveys. Therefore, in the paper are suggested main elements for constructing nonresponse rates measures in business web surveys.

In the future research, given elements should be used to make appropriate and easy to use nonresponse measure. Furthermore, it would be great if nonresponse would be separately observed for surveys conducted on personal computers, tablets and smartphones. The impact of contact and survey periods on nonresponse rates should be inspected in more detail.

References

[1] American Association for Public Opinion Research. 2016. Standard Definitions: Final Dispositions of Case Codes. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf [Accessed 21/01/2019].

[2] American Association for Public Opinion Research. 2019. Who We Are. https://www.aapor.org/About-Us/Who-We-Are.aspx [Accessed 21/01/2019].

[3] American Statistical Association. 2019. International Conference on Establishment Statistics. https://www.amstat.org/ASA/Meetings/ICES.aspx [Accessed 17/01/2019].

[4] Archer, T. M. 2008. Response rates to expect from Web-based surveys and what to do about it. Journal of Extension, 46(3). https://www.joe.org/joe/2008june/rb3.php [Accessed 17/01/2019].

[5] Callegaro, M., Lozar Manfreda, K., Vehovar, V. 2015. Web Survey Methodology. London: Sage.

[6] Copeland, K. R. 2018. How do Establishment Surveys Differ from Household Surveys? The Survey Statistician. International Association of Survey Statisticians. January 2018. http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_January_2018.pdf [Accessed 17/01/2019].

[7] Couper, M. P. 2000. Web Surveys: A Review of Issues and Approaches. Public Opinion Quarterly, 64(4): 464-494.

[8] Dillman, D. A., Eltinge, J. L., Groves, R. M., Little, R. J. A. 2002. Survey Nonresponse in Design, Data Collection, and Analysis. In Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (Eds.). Survey Nonresponse (pp. 3-26). New York: John Wiley & Sons.

[9] Dillman, D. A., Smyth, J. D., Christian, L. M. 2014. Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. Hoboken: John Wiley & Sons.

[10] Edwards, W. S., Cantor, D. 2004. Towards a Response Model in Establishment Surveys. In Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., Sudman, S. (Eds.) Measurement Error in Surveys (pp. 211-236). Hoboken: John Wiley & Sons.

[11] Evans, J. R., Mathur, A. 2005. The Value of Online Surveys. Internet Research, 15(2): 195-219.

[12] Petchenik, J., Watermolen, D. J. 2011. A cautionary note on using the Internet to survey recent hunter education graduates. Human Dimensions of Wildlife, 16(3): 216-218.

[13] Sadeghi, A., Talan, D. M., Clayton, R. L. 2016. Establishment, firm, or enterprise: does the unit of analysis matter? Monthly Labor Review. U.S. Bureau of Labor Statistics. November 2016. https://doi.org/10.21916/mlr.2016.51 [Accessed 17/01/2019].

[14] Saleh, A., Bista, K. 2017. Examining Factors Impacting Online Survey Response Rates in Educational Research: Perceptions of Graduate Students. Journal of MultiDisciplinary Evaluation, 13(29): 63-74.

[15] Tourangeau, R., Conrad, F., Couper, M. 2013. The Science of Web Surveys. New York: Oxford University Press.

[16] Willimack, D. K., Nichols, E., Sudman, S. 2002. Understanding Unit and Item Nonresponse In Business Surveys. In Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (Eds.). Survey Nonresponse (pp. 213-227). New York: John Wiley & Sons.

[17] Žmuk, B. 2018. Impact of Different Questionnaire Design Characteristics on Survey Response Rates: Evidence from Croatian Business Web Survey. Statistika: Statistics and Economy Journal, 98(1): 69-87.

The 15th International Symposium on
Operational Research in Slovenia

SOR ’19

Bled, SLOVENIA
September 25 - 27, 2019

Session 8:

Production and Management

FUZZY THREATNESS MATRICES IN PROJECT MANAGEMENT

Brožová Helena[1], Šubrt Tomáš[2], Rydval Jan[3], Pavlíčková Petra[4]
Czech University of Life Sciences Prague, Fac. of Economics and Management, Dept. of Systems Engineering
Kamýcká 129, 165 21 Praha 6 – Suchdol, Czech Republic
[1]brozova@pef.czu.cz, [2]subrt@pef.czu.cz, [3]rydval@pef.czu.cz, [4]pavlickovap@pef.czu.cz,

Abstract: The success of any project depends on many factors. The crucial question is to find out tasks causing delay of project deadline or failure of its objectives. We follow our recent research in the field of criticalness and failureness potential of tasks, we analyse project tasks criticalness and failureness and thus their threatness role in relation to project goals satisfaction. Our current research extends the criticalness and failureness evaluation by fuzzy numbers. Finally we introduce the fuzzy threatness matrix categorizing activities in a two-dimensional linguistic scale according to their potential threat to achieving project goals.

Keywords: Project management, task, criticalness potential, failureness potential, task threatness matrix, fuzzification, linguistic variable

1 INTRODUCTION

Modern project managers still have less and less time to quantitatively analyse activities in projects they manage. On the other hand, there is a growing number of project managers working in a multi-project environment and so-called program managers who manage project complexes within a coherent program over the longer term. With a greater number of managed projects, their experience, communication skills and soft skills are growing. However, quantitative approaches in project management are irreplaceable and so it is necessary to look for tools that will enable, through a qualified assessment, to estimate precisely the parameters of individual tasks, namely their riskiness, criticalness and the possibility of their potential failure, either absolute or only in a lower quality form or scope. Among these tools, the authors include fuzzification of criticalness and failureness of tasks and their subsequent multi-criteria evaluation and its graphical representation in the form of a threatness matrix, all with a maximum of informative power and a minimum of required information and interventions by the project manager.

2 MATERIAL AND METHODS

Project managers are often forced to make numerical estimates of parameters more or less based on their experience from already solved projects, or, on the contrary, based on standards or based on analysis of cost items, work effort, available time for task, etc. However, experts' experience and/or time needed to perform sophisticated analysis may not be sufficient. The experts are thus forced to make a quantified estimation without quantitative support.

The concept of task criticalness and failureness of a project or its part combines two types of views. From a purely mathematical view, task criticalness potential is based on aggregation of the task criticalness indicators expressed as crisp values based on task parameters and the task failureness potential can be expressed linguistically based on specific expert estimation of qualitative parameters considering the role of human factor. Nevertheless, the parameters of partial criticalness indicators can be estimated in the same way as failureness indicators.

The application of linguistic variables and their expression as fuzzy numbers into both criteria of evaluation, criticality and failure, can significantly increase the informative power and relevance of the information provided to the project's threats, as the expert is at least forced to quantitative inputs, because of using general linguistic variables ([4]). On the output side the linguistic variables also facilitate final decisions.

Combination of task criticalness and failureness potential will enable decision-makers to make a two-dimensional assessment of a task role to potentially risky impact on the project.

2.1 Concept of task criticalness

The task criticalness concept is continuously developing in time. The first idea was to delimit towards approaches based on the critical path using more sophisticated and more user-friendly defined indicators. Brozova et al. ([1], [2]) suggested to provide the overall evaluation of the task criticalness using quantitative crisp evaluation without soft knowledge about character of the tasks. The task criticalness potential is based on the multiple attribute decision making method using five indicators of the criticalness: task duration, slack, cost and work, and project structure.

The principle of calculating most of the indicators is similar in this approach, and so, for example, a procedure for cost criticalness indicator cc can be explain. This indicator is defined from the perspective of minimizing project cost. We can assume the task with low cost has a smaller impact on the total cost of the project than an expensive one. This indicator expresses the relative cost of each project activities and is defined as

$$cc_i = \frac{c_i - \min\limits_{k=1,2,..N} c_k}{\max\limits_{k=1,2,..N} c_k - \min\limits_{k=1,2,..N} c_k} \tag{1}$$

where cc_i is the cost criticalness of the task i, and c_i, c_k resp., are the cost of tasks i, k resp., N is the number of the tasks in the project. The cost criticalness indicator of the task transforms the task cost so that the higher value of this indicator shows higher criticalness. The value 1 corresponds to the highest criticalness and the value 0 to the lowest cost criticalness.

Other criticalness indicators are defined as follow.

$$ct_i = \frac{t_i - \min\limits_{k=1,2,..N} t_k}{\max\limits_{k=1,2,..N} t_k - \min\limits_{k=1,2,..N} t_k} \qquad cs_i = \frac{s_i - \max\limits_{k=1,2,..N} s_k}{\min\limits_{k=1,2,..k} s_K - \max\limits_{k=1,2,..N} s_k}$$

$$cp_i = \frac{p_i - \min\limits_{k=1,2,..N} p_k}{\max\limits_{k=1,2,..N} p_k - \min\limits_{k=1,2,..N} p_{Kk}} \qquad cw_i = \frac{w_i - \min\limits_{k=1,2,..N} w_k}{\max\limits_{k=1,2,..N} w_k - \min\limits_{k=1,2,..N} w_k} \tag{2}$$

where ct_i is the time criticalness of the task i, and t_i, t_k resp., are the duration of tasks i, k resp., cp_I is the structural criticalness of the task i, and p_i, p_k resp., are the probability the critical path will pass through the tasks i, k resp., cs_I is the slack criticalness of the task i, and s_i, s_k resp., are the slacks of tasks i, k resp., cw_i is the work criticalness of the task i, and w_i, w_k resp., are the work amount of tasks i, k resp., N is the number of the tasks in the project.

All these crisp values of criticalness indicators are then fuzzified using fuzzy linguistic variable CI which is defined using a quintuple (CI, T, U, M, G) where CI is the name of the variable, $T = \{T_1, T_2, \dots, T_6\}$ is the set of terms of CI, U is the universe - interval $\langle 0, 1 \rangle$, M is a semantic rule for associating each term with proper fuzzy set (number) defined on U, and G is a syntactic rule for generating the derived terms (Table 1). We suggest the following terms and their fuzzy interpretation using trapezoid fuzzy numbers and six step non-uniform fuzzy scale.

Table 1: Linguistic variable for criticalness indicators

T_j	**Linguistic term**	**Fuzzy number** $M(T_j) = (t_{1j}, t_{2j}, t_{3j}, t_{4j})$
T_1	Not at all critical	(0; 0; 0; 0.1)
T_2	Usually not critical	(0; 0.1; 0.2; 0.3)
T_3	Rather not critical	(0.2; 0.3; 0.4; 0.6)
T_4	Rather critical	(0.4; 0.6; 0.7; 0.8)
T_5	Usually critical	(0.7; 0.8; 0.9; 1)
T_6	Always critical	(0.9; 0.1; 1; 1)

The fuzzy value of each criticalness indicator is received as weighted sum of all values of linguistic variable where the weights are the membership function values of criticalness indicator. For instance, for cost criticalness indicator we receive the fuzzy value using formula

$$CC_I = (x_{I1}, x_{I2}, x_{I3}, x_{I4}) = \left(\left(\sum_{k=1}^{6} \mu_{T_k}(cc_I)t_{Ik1}\right), \left(\sum_{k=1}^{6} \mu_{T_k}(cc_I)t_{Ik2}\right), \left(\sum_{k=1}^{6} \mu_{T_k}(cc_I)t_{Ik3}\right), \left(\sum_{k=1}^{6} \mu_{T_k}(cc_I)t_{Ik4}\right)\right) \quad (3)$$

where cc_I is crisp value of cost criticalness indicator, $\mu_{T_k}(cc_I)$ is membership function and CC_I is its fuzzy value.

Criticalness potential C_I of task I is calculated as the weighted sum of individual fuzzy evaluation of indicators

$$C_I = (c_{I1}, c_{I2}, c_{I3}, c_{I4}) = v_t CT_I + v_s CS_I + v_p CP_I + v_c CC_I + v_w Cw_I \quad (4)$$

where CT_I is fuzzy time criticalness indicator of task I and v_t its weight, similarly CS_I is fuzzy slack criticalness indicator, CP_I is fuzzy structural criticalness indicator CC_I is fuzzy cost criticalness indicator, and CW_I is fuzzy work criticalness indicator.

The tasks are now split into five groups according to their criticalness potential C_I. This classification is made by mapping of fuzzy criticalness potential on the values of fuzzy linguistic variable - criticalness potential rate CR which is defined in Table 2 using a quintuple (CR, V, U, M, G) where CR is the name of the variable, $V = \{V_1, V_2, \dots, V_5\}$ is the set of terms of CR, U is the universe - interval $\langle 0, 1 \rangle$. We suggest the following terms and their fuzzy interpretation using triangular fuzzy numbers and five step non-uniform fuzzy scale.

Table 2: Linguistic variable for criticalness potential rate

V_j	**Linguistic term**	**Fuzzy number** $M(V_j) = (v_{1j}, v_{2j}, v_{3j}, v_{4j})$
V_1	Non-criticalness	(0; 0; 0.05; 0.15)
V_2	Weak criticalness	(0.05; 0.15; 0.25; 0.35)
V_3	Rather criticalness	(0.25; 0.35; 0.5; 0.6)
V_4	Strong criticalness	(0.5; 0.6; 0.75; 0.85)
V_5	Extreme criticalness	(0.75; 0.85; 1; 1)

The linguistic term expressing the classification of the task I criticalness potential is received using suitable method of linguistic approximation [3]. It can be based on simple measure of fuzzy numbers distance $d(C_I, V_j)$, which in case of fuzzy scale with ordered trapezoidal fuzzy numbers gives good results, simplicity and low computational complexity. The distance of fuzzy task potential of criticalness and its rate is calculated as

$$d(C_I, V_j) = \frac{\sum_{i=1}^{4} |c_{Ii} - v_{ij}|}{v_{4j} - v_{1j} + v_{3j} - v_{2j}} \quad (5)$$

where $C_I = (c_{I1}, c_{I2}, c_{I3}, c_{I4})$ is the criticalness potential of the task I and $V_j = (v_{1j}, v_{2j}, v_{3j}, v_{4j})$ is the term of linguistic variable.

The linguistic evaluation of the criticalness potential rate of the task I is received as linguistic term which is the nearest to the fuzzy criticalness potential

$$V_I\colon d(C_I, V_I) = \min_{j=1,\dots,5} d(C_I, V_j) \quad (6)$$

2.2 Concept of task failureness

Task failureness potential is based on the project magic triangle respecting three key parameters: project cost, duration and quality or resource work. Project task may fail in any of them. The evaluation of the project tasks failureness should be, similarly like of criticalness, done separately parameter by parameter, and complex evaluation is received using multiple attribute decision-making methods. Main difference between the task criticalness and task failureness is in mode of evaluation. Task failureness is seen as soft characteristic based on experience of experts. Any failureness can be judged by one or more experts (project manager, task manager, current expert) and his/her experiences and skills [1]. Each expert, his/her experiences respectively, can be weighted. Various approach to the complex evaluation of failureness should be used, ranking, scoring with 3 to 10 points scale, probability scale, pairwise comparisons or more soft approach as fuzzy evaluation.

In this paper we suggest vague evaluation using linguistic variable of failureness indicators FI which is defined using a quintuple (FI, T, U, M, G) where FI is the name of the variable, $T = \{T_1, T_2, \dots, T_6\}$ is the set of terms of FI, U is the universe - interval $\langle 0, 1 \rangle$ (Table 3). We suggest the following terms and their fuzzy interpretation using triangular fuzzy numbers and six step non-uniform fuzzy scale.

Table 3: Linguistic variable for failureness indicators

T_j	**Linguistic term**	**Fuzzy number** $M(T_j) = (t_{1j}, t_{2j}, t_{3j}, t_{4j})$
T_1	Not at all failing	(0; 0; 0; 0.1)
T_2	Usually not failing	(0; 0.1; 0.2; 0.3)
T_3	Rather not failing	(0.2; 0.3; 0.4; 0.6)
T_4	Rather failing	(0.4; 0.6; 0.7; 0.8)
T_5	Usually failing	(0.7; 0.8; 0.9; 1)
T_6	Always failing	(0.9; 0.1; 1; 1)

Each failureness indicator of the task I is expertly evaluated by fuzzy number and the failureness potential F_I of task I is than calculated as weighted sum of these fuzzy values:

$$F_I = (f_{I1}, f_{I2}, f_{I3}, f_{I4}) = u_t FT_I + u_Q FQ_I + v_c FC_I \qquad (7)$$

where FT_I is fuzzy time failureness indicator of task I and v_T its weight, similarly FQ_I is fuzzy quality failureness indicator, and FC_I is fuzzy cost failureness.

Again, the tasks are now split into five groups according to their values of failureness potential F_I. This classification is made by mapping of failureness potential on the values of fuzzy linguistic variable - failureness potential rate FR which is defined using a quintuple (FR, V, U, M, G) where FR is the name of the variable, $V = \{V_1, V_2, \dots, V_5\}$ is the set of terms of FR, U is the universe - interval $\langle 0, 1 \rangle$ (Table 4). We suggest the following terms and their fuzzy interpretation using triangular fuzzy numbers and five step non-uniform fuzzy scale.

Table 4: Linguistic variable for failureness potential rate

V_j	**Linguistic term**	**Fuzzy number** $M(V_j) = (v_{1j}, v_{2j}, v_{3j}, v_{4j})$
V_1	Non-failureness	(0; 0; 0.05; 0.15)
V_2	Weak failureness	(0.05; 0.15; 0.25; 0.35)
V_3	Rather failureness	(0.25; 0.35; 0.5; 0.6)
V_4	Strong failureness	(0.5; 0.6; 0.75; 0.85)
V_5	Extreme failureness	(0.75; 0.85; 1; 1)

The linguistic evaluation of the failureness potential rate of the task I is the nearest linguistic term receiving by the same way as criticalness potential rate.

2.3 Fuzzy threatness matrix

Using the linguistic terms and fuzzy scales to evaluating the task criticalness and failureness potentials is a suitable way how to express the individual and subjective assessment of the task ([4]). Furthermore, the suggested two-dimensional evaluation of task threatness should be adapted to the individual project environment and project needs.

The values of criticalness and failureness potential rate are used for placing of the tasks into cells of task threatness matrix (Figure 2). In the red area of the task threatness matrix, there are highly threatening tasks requiring great attention, and in the yellow area there are threatening tasks to be controlled, to ensure the successful completion of the project. The tasks in green area should not significantly influence the project.

It should be pointed out, that the results are affected by the type of fuzzy scale (i.e. uniform or non-uniform). It is also necessary to set up the proper scale size (for example three, five of seven points), and to divide the matrix into the areas, according to the selected scale size ([4]).

3 CASE STUDY

The tasks threatness matrix creation is described on the following small-scale project with 16 tasks (Figure 1). The Table 5 shows the initial quantitative values of criticalness elements, which are then transformed into criticalness indicators using formulas (1) and (2), and experts' evaluation of failureness indicators. The critical path of this project (Figure 1) is composed of tasks A, B, D, E, F, G, and O.

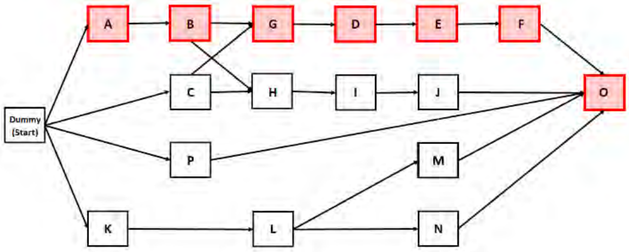


Figure 1: Small project

Table 5: Initial evaluation of the project tasks *(The grey-highlighted tasks lie on critical path.)*

Task	Days	Prob. of CP	Slack	Cost	Work	Time failureness	Quality failureness	Cost failureness
A	1	0.25	0	1120	1	Rather failing	Not at all failing	Always failing
B	4	0.25	0	3840	4	Usually not failing	Not at all failing	Not at all failing
C	1	0.25	4	1120	1	Rather failing	Rather failing	Rather not failing
D	1	0.25	0	5460	1	Usually not failing	Rather not failing	Not at all failing
E	1	0.25	0	960	1	Usually not failing	Usually not failing	Not at all failing
F	2	0.25	0	2400	2	Usually not failing	Usually not failing	Usually not failing
G	2	0.25	0	9180	3	Rather failing	Rather failing	Always failing
H	1	0.25	2	86200	1	Rather failing	Rather failing	Always failing
I	1	0.25	2	2320	2	Rather not failing	Rather not failing	Rather not failing
J	2	0.25	2	3440	3	Rather not failing	Rather not failing	Rather not failing
K	2	0.25	6	1920	2	Usually not failing	Usually not failing	Usually not failing
L	1	0.25	6	1200	1	Not at all failing	Usually not failing	Usually not failing
M	1	0.125	7	1200	1	Not at all failing	Usually not failing	Not at all failing
N	2	0.125	6	2400	2	Rather failing	Rather failing	Usually failing
O	5	1	0	15800	10	Rather not failing	Usually not failing	Usually not failing
P	2	0.25	9	2400	2	Rather failing	Rather failing	Usually failing

Using formulas (3), (4) or (7), (5) and (6) both for criticalness potential and failureness potential the complex evaluation of the project task is received (Table 6) and the project tasks can be split into the task threatness matrix (Figure 2).

For instance, it is seen the threatness of the task H, which is not critical (does not lie on critical path) but it requires attention due to its other characteristics (it is really costly and filing from the project triangle criteria), and contrary, the critical task B (lies on critical path) with low threatness.

Table 6: Task criticalness and failureness potential rates *(The grey-highlighted tasks lie on critical path.)*

Task	Fuzzy criticalness potential				Fuzzy criticalness rate	Fuzzy failureness potential				Fuzzy failureness rate
A	0.116	0.148	0.167	0.254	Weak criticalness	0.433	0.533	0.567	0.633	Rather failureness
B	0.252	0.332	0.39	0.5	Rather criticalness	0.000	0.033	0.067	0.167	Non-failureness
C	0.052	0.096	0.128	0.228	Weak criticalness	0.333	0.500	0.600	0.733	Rather failureness
D	0.116	0.177	0.225	0.312	Weak criticalness	0.067	0.133	0.200	0.333	Weak failureness
E	0.116	0.148	0.167	0.254	Weak criticalness	0.000	0.067	0.133	0.233	Weak failureness
F	0.133	0.204	0.262	0.357	Weak criticalness	0.000	0.100	0.200	0.300	Weak failureness
G	0.133	0.233	0.32	0.415	Weak criticalness	0.567	0.733	0.800	0.867	Strong failureness
H	0.497	0.574	0.606	0.661	Strong criticalness	0.567	0.733	0.800	0.867	Strong failureness
I	0.156	0.244	0.315	0.415	Weak criticalness	0.200	0.300	0.400	0.600	Rather failureness
J	0.107	0.178	0.249	0.357	Weak criticalness	0.200	0.300	0.400	0.600	Rather failureness
K	0.042	0.113	0.185	0.306	Weak criticalness	0.000	0.100	0.200	0.300	Weak failureness
L	0.026	0.058	0.089	0.202	Non-criticalness	0.000	0.067	0.133	0.233	Weak failureness
M	0	0.013	0.026	0.126	Non-criticalness	0.000	0.033	0.067	0.167	Non-failureness
N	0.042	0.095	0.147	0.268	Weak criticalness	0.500	0.667	0.767	0.867	Strong failureness
O	0.641	0.741	0.77	0.798	Strong criticalness	0.067	0.167	0.267	0.400	Weak failureness
P	0.016	0.075	0.133	0.241	Weak criticalness	0.500	0.667	0.767	0.867	Strong failureness

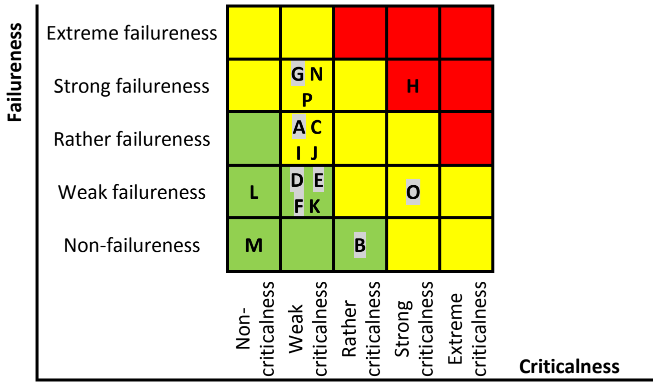


Figure 2: Task threatness matrix for case study project *(The grey-highlighted tasks lie on critical path.)*

4 CONCLUSION

This paper describes the project task threatness based on the criticalness potential evaluating the tasks criticalness and on the failureness potential evaluating the possibility of tasks failure. Both criteria are expressed by fuzzy terms and corresponding two-dimensional evaluation of task threatness is displayed in the task threatness matrix. This approach is useful with respect to the project management triangle and other indicators which have an impact on the project success. Important advantage of suggested threatness matrix is that it allows fuzzy assessments of the impact of individual tasks on project completion.

Acknowledgement

The research is supported by the Operational Programme Prague – Growth Pole of the Czech Republic - Implementation proof-of-concept activities CULS to promote technology transfer and knowledge into practice number: CZ.07.1.02/0.0/0.0/17_049/0000815 - KZ10.

References

[1] Brožová, H., Bartoška, J., Šubrt, T. 2014. Fuzzy Approach to Risk Appetite in Project Management. In Proceedings of the 32st International conference on Mathematical Methods in Economics (pp. 61-66). Olomouc: Palacký University.

[2] Brožová, H., Bartoška, J., Šubrt, T., Rydval, J. 2016. Task Criticalness Potential: A Multiple Criteria Approach to Project Management. Kybernetika, 52, 558-574.

[3] Johanyák, Zs. Cs., Kovács, Sz. 2005. Distance based similarity measures of fuzzy sets, In Proceedings of the 3rd Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence (pp. 265-276). Herl'any.

[4] Ross, T. J. 2010. Fuzzy Logic with Engineering Applications. Chichester: John Wiley & Sons Inc.

THE COMPARISON OF HOLT-WINTERS METHODS AND α-SUTTE INDICATOR IN FORECASTING THE FOREIGN VISITOR ARRIVALS IN INDONESIA, MALAYSIA, AND JAPAN

Liljana Ferbar Tratar
University of Ljubljana, Faculty of Economics
Kardeljeva pl 17, 1000 Ljubljana, Slovenia
E-mail: liljana.ferbar.tratar@ef.uni-lj.si

Ansari Saleh Ahmar
Universitas Negeri Makassar,
90222 Kampus UNM Gunungsari Makassar, Pare-Pare, Indonesia
E-MAIL: ansarisaleh@unm.ac.id

Abstract: Forecasting tourism is one of the important areas that need to be explored, as tourism is in direct contact with society. Tourism in the region is closely related to its economy, culture and the environment. As such, it affects the economic levels of the region, for example, increasing foreign exchange in the country and creating employment opportunities. Therefore, it is very important to know how tourism will develop in the future, which depends mainly on future demand (tourist arrivals). Many publications on tourism forecasting have appeared during the past years. Although different forecasting techniques can be used, the major conclusions are that time series models are simplest and the least expensive. The purpose of our research is to predict foreign visitor arrivals in Indonesia, Malaysia, and Japan by using Holt-Winters Methods (Additive, Multiplicative and Extended Holt-Winters Method) and α-Sutte Indicator. Data for our research is comprised of foreign visitor arrivals in Indonesia, Malaysia, and Japan from January 2008 to November 2017. The data is divided into 2 parts, namely fitting data and testing data. Based on the results of all four forecasting methods, we conclude that the Extended Holt-Winters method is most suitable. At the end of the analysis, using the Extended Holt-Winters method, we calculate monthly forecasts of tourist arrivals for all three countries in 2018.

Keywords: tourism, forecasting, foreign visitor arrivals, Holt-Winters method, α-Sutte Indicator.

1 INTRODUCTION

Things that people often dream about are traveling to foreign countries. Tourism is no longer just one of the forms of entertainment, but offers the possibility to increase knowledge about foreign countries, residents and different cultures. Tourism has become a way of life.

Tourism has many benefits for a region. For example, (1) the tourism of a region will generate large foreign exchange and will have an impact on improving the economy in an area [3,5], (2) cultural aspects – the development of tourism will provide an understanding of different cultures through the interaction of tourists with local communities (tourists will learn and appreciate the culture of the local community and understand the background of local culture [6]), (3) environmental aspects (benefits) – the local government will take care of and maintain the cleanliness of the tourist area [6].

From the above description, tourism is therefore very important for the area, so it needs to be carefully studied. In doing so, the forecast of the number of tourists in the region is of great help to us. Knowing the number of tourists, stakeholders in the region can adopt a policy related to the development of their territory.

Many publications on tourism forecasting have appeared during the past twenty years. The forecasting techniques can include time series models, the gravity model or expert-opinion techniques. The time series models are the simplest and least costly; the gravity model is best suited to handle international tourism flows; and expert-opinion methods are useful when data are unavailable [7].

The remainder of the paper is organized as follows. We begin with the description of the forecasting procedures (see Section 2). In Section 3, we present the data, methodology and results which allow us to compare different forecasting methods and to choose the most appropriate method. Finally, in Section 4, after the conclusions of our paper some further research steps are suggested.

2 FORECASTING METHODS

The Holt-Winters method of exponential smoothing involves trend and seasonality and is based on three smoothing equations: equation for level, for trend and for seasonality. The decision as to which method to use depends on time series characteristics: the additive method is used when the seasonal component is constant, the multiplicative method is used when the size of the seasonal component is proportional to the trend level [4].

α-Sutte Indicator was developed using the principle of the forecasting method of using the previous data [1]. It was developed using the adopted moving average method. The α-Sutte Indicator uses 4 previous data (Y_{t-1}, Y_{t-2}, Y_{t-3} and Y_{t-4}) as supporting data for forecasting and making the decision [2].

2.1 Holt-Winters' additive procedure (AHW)

The basic equations for the AHW method are:

$$L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (1)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (2)$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} \quad (3)$$

$$F_{t+m} = L_t + b_t m + S_{t-s+m} \quad (4)$$

where are L_t – estimation of variable in time *t*, Y_t – observed value, b_t – trend estimation of time series in time *t*, S_t – estimation of seasonality in time *t*, F_{t+m} – forecast in time *t* for *m* period ahead, *α, β, γ* – smoothing parameters in the interval [0, 1], *m* – number of forecasted periods, *s* – duration of seasonality (for example, number of months or quarters in a year).

For initialization of the additive method initial values of variable L_t, trend estimation b_t and seasonality estimation S_t are needed. To determine initial estimates we need at least one whole data season (that is, *s* data). Initialization of variable L_s is calculated with the formula:

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \cdots + Y_s) \quad (5)$$

For trend initialization it is more suitable if we use two whole seasons (that is, 2*s* data):

$$b_s = \frac{1}{s}\left(\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \cdots + \frac{Y_{s+s} - Y_s}{s}\right) \quad (6)$$

Seasonal indices are calculated as differences between observed value and variable estimation:

$$S_1 = Y_1 - L_s,\ S_2 = Y_2 - L_s, \ldots,\ S_s = Y_s - L_s \quad (7)$$

The method is proved to be (regarding costs and calculation itself) comparable with more complex methods (for example Box-Jenkins); in some cases the results gained with the Holt-Winters were even better than more complex methods [4].

2.2 Extended Holt-Winters' procedure (EHW)

The EHW method differs from AHW only in the equation for the level (1); all other equations remain the same as with the AHW (2 – 7). The equation for level now contains an additional smoothing parameter δ:

$$L_t = \alpha Y_t - \delta S_{t-s} + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (8)$$

This method allows to smooth the seasonal factors more or less than the AHW method, depending on the value of the parameter δ [4].

2.3 Holt-Winters' multiplicative procedure (MHW)

The basic equations for the MHW method are as follows:

$$L_t = \alpha(Y_t/S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{9}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \tag{10}$$

$$S_t = \gamma(Y_t/L_t) + (1-\gamma)S_{t-s} \tag{11}$$

$$F_{t+m} = (L_t + b_t m) \cdot S_{t-s+m} \tag{12}$$

The second of these equations (10) is identical to the second equation (2) of AHW. The only differences in the other equations are that the seasonal components are now in the form of products and ratios instead of being added and subtracted.

2.4 α-Sutte Indicator (SUTTE)

The equation of SUTTE method is ([2,8]):

$$F_t = \frac{1}{3}\left[\alpha\left(\frac{\alpha-\delta}{((\alpha+\delta)/2)}\right) + \beta\left(\frac{\beta-\alpha}{((\beta+\alpha)/2)}\right) + \gamma\left(\frac{\gamma-\beta}{((\gamma+\beta)/2)}\right)\right] \tag{13}$$

where $\alpha = Y_{t-3}$, $\beta = Y_{t-2}$, $\gamma = Y_{t-1}$, $\delta = Y_{t-4}$ and Y_{t-k} is observed data in period $(t-k)$.

3 CASE STUDY

3.1 Data

For research we used monthly data of foreign visitor arrivals in Indonesia, Malaysia, and Japan from January 2008 to November 2017. We acquired data from the website: (1) Badan Pusat Statistik (BPS-Statistics Indonesia); (2) Tourism Malaysia of Ministry of Tourism and Culture Malaysia; and (3) Statistics Bureau of Ministry of Internal Affairs and Communications, Japan.

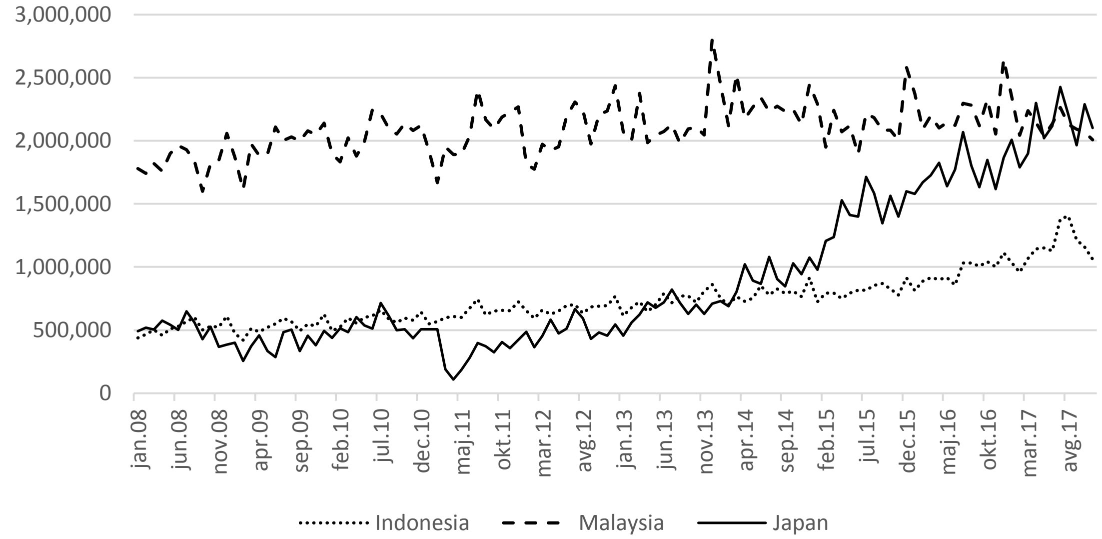


Figure 1: Foreign visitor arrivals in Indonesia, Malaysia, and Japan from January 2008 to November 2017.

Figure 1 shows the number of foreign visitor arrivals in Indonesia, Malaysia, and Japan between the years 2008 and 2017. It is evident that a growing trend and comprehensive (random) fluctuations are present in the data.

3.2 Methodology

The obtained data of each time series were split into initialization (the first two years, from January 2008 to December 2009), fitting (period from January 2008 to December 2016), and testing subset (period from January 2017 to November 2017). Fitting subset was used for method learning. With the testing subset we checked a time series learning ability. We calculated forecasting values for testing subset and then compare these values to independent-real data. To calculate forecasting values of testing subset we use a long-term forecasting approach for eleven (m=11) months ahead:

$$F_{t+m}, \ m = 1, 2, \dots, 11 \text{ (monthly forecasting)} \tag{14}$$

where t represents December 2016. Monthly long-term forecasting is important for the strategic planning decisions in the future.

For the evaluation of the forecasting methods we applied two forecasting accuracy measures, Mean Squared Error (MSE) and Mean Absolute Error (MAE):

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}(Y_t - F_t)^2 \tag{15}$$

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|Y_t - F_t| \tag{16}$$

where Y_t represents actual value, F_t forecasted value and N number of samples. MSE penalizes the errors proportional to their squares. Minimizing MSE therefore leads to smoothing parameters that produce fewer large errors at the expense of tolerating several small errors. The MAE penalty is proportional to the error itself. Minimizing MAE therefore leads to more small errors but also more frequent large errors. We use both objective functions as both aspects are important in practice. Of course, the lower values of MSE and MAE represent a better forecasting performance.

3.3 Forecasts for Indonesia

Table 1 shows MSE and MAE results for fitting set obtained with four different methods and the percentage of improvement of MSE and MAE, calculated by using the EHW method compared to the AHW, MHW and SUTTE method. EHW method is identified as the best method, not just in case of fitting set but also for testing set (see Table 2). It is obvious that with the EHW method a considerable reduction in the MSE and MAE can be reached. The results show that with the EHW method the MSE for testing set is reduced by more than 14% (25%, 30%) in comparison with AHW (MHW, SUTTE) method.

Table 1: MSE and MAE results of forecasting for Indonesia, fitting data.

INDONESIA			*Improvement*	
Fitting set (2010-2016)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	1,637,964,823.77	32,475.14	*EHW/method*	*EHW/method*
AHW	1,719,966,909.50	32,846.82	4.77%	1.13%
MHW	1,856,432,284.95	34,574.22	11.77%	6.07%
SUTTE	6,037,728,179.53	59,829.34	72.87%	45.72%

Table 2: MSE and MAE results of forecasting for Indonesia, testing data.

INDONESIA			*Improvement*	
Testing set (Jan-Nov 2017)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	12,522,672,017.26	97,694.25	*EHW/method*	*EHW/method*
AHW	14,667,585,431.30	100,606.20	14.62%	2.89%
MHW	16,847,254,534.97	110,122.94	25.67%	11.29%
SUTTE	18,032,018,775.72	111,326.00	30.55%	12.24%

3.4 Forecasts for Malaysia

The best fitting and testing result for MSE and MAE is reached by EHW method (Table 3 and Table 4). It is just a little bit lower than fitting performance, but significantly lower than testing performance of AHW and MHW method.

Table 3: MSE and MAE results of forecasting for Malaysia, fitting data.

MALAYSIA			*Improvement*	
Fitting set (2010-2016)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	21,624,253,750.05	109,285.24	*EHW/method*	*EHW/method*
AHW	21,882,965,131.56	109,458.78	1.18%	0.16%
MHW	22,528,196,047.93	109,529.26	4.01%	0.22%
SUTTE	70,618,631,313.33	199,981.72	69.38%	45.35%

Table 4: MSE and MAE results of forecasting for Malaysia, testing data.

MALAYSIA			*Improvement*	
Testing set (Jan-Nov 2017)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	25,234,146,629.73	83,289.63	*EHW/method*	*EHW/method*
AHW	43,942,291,730.75	103,486.76	42.57%	19.52%
MHW	46,409,847,320.42	142,211.72	45.63%	41.43%
SUTTE	43,615,238,632.09	146,375.90	42.14%	43.10%

3.5 Forecasts for Japan

Again, the excellent performance for fitting and testing set shows the EHW method. Although the results of the fitting set for the EHW and AHW method are very similar, the EHW method turns out to be significantly better for the testing set (Table 5 and Table 6).

Table 5: MSE and MAE results of forecasting for Japan, fitting data.

JAPAN			*Improvement*	
Fitting set (2010-2016)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	8,266,955,513.84	69,989.42	*EHW/method*	*EHW/method*
AHW	8,338,841,226.96	70,150.57	0.86%	0.23%
MHW	10,655,274,437.07	86,882.46	22.41%	19.44%
SUTTE	24,049,780,653.84	129,433.68	65.63%	45.93%

Table 6: MSE and MAE results of forecasting for Japan, testing data.

JAPAN			*Improvement*	
Testing set (Jan-Nov 2017)	***MSE***	***MAE***	***MSE***	***MAE***
EHW	15,319,208,751.29	129,427.98	*EHW/method*	*EHW/method*
AHW	22,753,518,944.41	139,454.62	32.67%	7.19%
MHW	80,530,197,423.19	1,116,039.96	80.98%	88.40%
SUTTE	69,466,709,126.94	225,004.39	77.95%	42.48%

3.6 Research findings

The best forecasting method is the method that has the smallest MSE and MAE value on the fitting and testing subset. Based on MSE and MAE as performance measures, the EHW method is identified as the best forecasting method for foreign visitor arrivals data in

Indonesia, Malaysia and Japan (see Tables 1-6). As the EHW method is the most appropriate method for all three time series, this method is used for forecasts of the foreign visitor arrivals in Indonesia, Malaysia and Japan from January to December 2018. Due to the dynamics of the time series, the authors propose a re-analysis when new data is available.

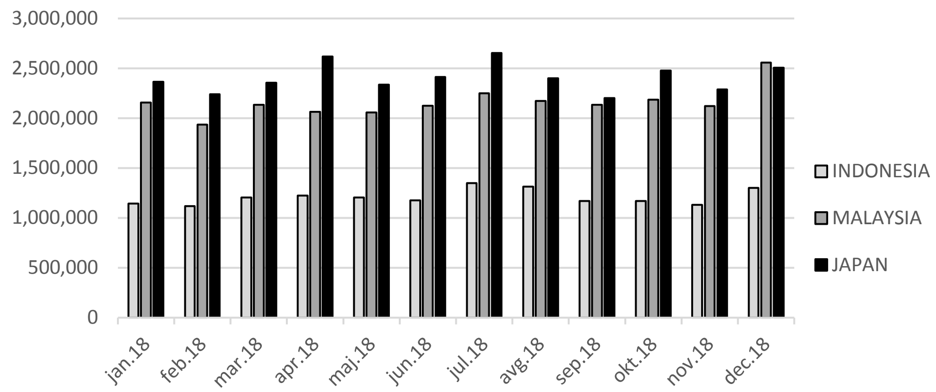


Figure 2: Forecast for foreign visitor arrivals in Indonesia, Malaysia, and Japan from January to December 2018.

4 CONCLUSION AND FURTHER RESEARCH

Tourism is a very important thing to be studied in various countries/regions because it is related to income and social economy in the region. The data on tourism that need to be explored are strongly connected to the number of tourists. Forecasts of the number of tourists in the future can affect tourism management so that stakeholders can adopt a policy that relates to tourists and tourism. In order to forecast the arrival of foreign visitors to Indonesia, Malaysia and Japan from January to December 2018, the EHW method proved to be more appropriate than other methods (AHW, MHW and SUTTE).

For further research, authors suggest that additional optimization using the initial parameters is used with Holt-Winters methods. Also, the possible way to improve SUTTE method would be to upgrade it by introducing seasonal variations.

References

[1] Ahmar, A.S., Rahman, A., Mulbar, U. 2018. α- Sutte Indicator: a new method for time series forecasting. IOP Conf. Ser. Mater. Sci. Eng., 366(1): 012018.

[2] Ahmar, A.S. 2018. A Comparison of α-Sutte Indicator and ARIMA Methods in Renewable Energy Forecasting in Indonesia. Int. J. Eng. Technol., 7(1.6): 20–22.

[3] Cárdenas-García, P.J., Sánchez-Rivero, M., Pulido-Fernández, J.I. 2013. Does Tourism Growth Influence Economic Development? J. Travel Res., 54(2): 206–221.

[4] Ferbar Tratar, L., Mojškerc, B., Toman, A. 2016. Demand forecasting with four-parameter exponential smoothing. Int. J. Prod. Econ., 181: 162–173.

[5] Lundberg, E. 2017. The importance of tourism impacts for different local resident groups: A case study of a Swedish seaside destination. J. Destin. Mark. Manag., 6(1): 46–55.

[6] Nikolla, I., Miko, D. 2013. Importance of tourism in community development. Mediterr. J. Soc. Sci., 4(9): 205.

[7] Sheldon, P.J., Var, T. 1985. Tourism forecasting: A review of empirical research. Journal of Forecasting, 4: 183-195.

[8] Sutiksno, D.U., Ahmar, A.S., Kurniasih, N., Susanto, E., Leiwakabessy, A. 2018. Forecasting Historical Data of Bitcoin using ARIMA and α-Sutte Indicator. J. Phys. Conf. Ser., 1028(1): 012194.

ON THE COST MINIMIZATION PROBLEM WITH CES TECHNOLOGY: REVERSE HÖLDER'S INEQUALITY APPROACH

Vedran Kojić, Zrinka Lukač
University of Zagreb, Faculty of Economics & Business, Department of Mathematics
Trg J. F. Kennedyja 6, HR–10000 Zagreb, Croatia
vkojic@efzg.hr, zlukac@efzg.hr

Abstract: A common application of mathematical programming in microeconomics is solving the firm's cost minimization problem. This is a constrained optimization problem that considers a firm minimizing its cost of producing a given level of output. A standard procedure for solving this problem is the use of differential calculus, i.e. Lagrange multiplier method. In order to find the solution by using calculus, a necessary and sufficient condition needs to be examined. If the technology is described by constant elasticity of substitution (CES) production function, the use of differential calculus is not trivial. Therefore, in this paper we provide a new complementary approach of obtaining the solution. Our methodology uses only the definition of a minimum and the reverse Hölder's inequality. We show that in case of CES production function our methodology provides an easier way of obtaining solution than the method based on calculus.

Keywords: microeconomics, cost minimization, CES production function, constrained optimization, reverse Hölder's inequality, without calculus

1 INTRODUCTION

One of the classical microeconomic problems in theory of the firm is the cost minimization problem with constant elasticity of substitution (CES) production function. This is a typical example of a constrained optimization problem, where the cost is to be minimized over the positive amounts of two inputs, while the constraint in equality form represents a firm's desired level of produced output. In many microeconomic textbooks the case of two inputs is solved and discussed by means of differential calculus, i.e. Lagrange multiplier method (see for instance [2]). The use of calculus is justified by imposing assumptions that input quantities are continuous (real) variables and that CES production function is a real continuous function. Although in theory this assumption of continuity is very useful, it does not correspond to the real world practice. In practice, input amounts are very often discrete variables and therefore the production function is not a continuous function but a sequence. Thus, if we look at the problem of the cost minimization as the problem of discrete optimization, the use of calculus becomes questionable and it gives rise to some new questions: can we solve this problem with some other tool that does not use calculus? If so, will the solution of the problem change or remain the same? Would that another method be easier to apply than calculus? The purpose of this paper is to answer these questions. We claim that the cost minimization problem with the CES technology can be solved without calculus, using only the definition of a minimum and a certain mathematical inequality, i.e. the reverse Hölder's inequality (the use of mathematical inequalities as alternative methods for solving other microeconomic problems can be found for example in [3], [4] and [5]). Although the cost minimization problem with CES production function can be formulated as the problem of several variables (inputs), here we consider only n=2 inputs (the analysis of the case of n>2 inputs is beyond the scope of this paper and it is left for further research).

The structure of this paper is as follows. After the introduction, the notation that we use is given in the second section. The third section presents the Hölder's and the reverse Hölder's inequality. In the fourth section, first we treat the problem by using the Lagrange multiplier method and then we solve the problem by using the reverse Hölder's inequality. We discuss the arguments in favour of the reverse Hölder's inequality approach in comparison to the calculus approach. The fifth section concludes the paper.

2 NOTATION

In this paper, we use the same notation as in [2], which is as follows:

$\mathbb{R}_+^2$	$\mathbb{R}_+^2 \equiv \{(x_1, x_2): x_i \geq 0,\ i = 1, 2\} \subset \mathbb{R}^2$
$x = (x_1, x_2) \in \mathbb{R}_+^2$	vector of inputs (amount of input i is x_i, i=1,2)
$w = (w_1, w_2) \in \mathbb{R}_+^2$	vector of prevailing market prices at which the firm can buy inputs $x = (x_1, x_2)$
$y > 0$	given amount of firm's output
$\alpha_1 > 0, \alpha_2 > 0$	positive constants such that $\alpha_1 + \alpha_2 = 1$
$\rho \in \langle -\infty, 1\rangle \setminus \{0\}$	substitution coefficient

3 HÖLDER'S AND REVERSE HÖLDER'S INEQUALITY

Here we state the Hölder's and the reverse Hölder's inequality, which are essential for the method that we propose in this paper.

Theorem 3.1. (Hölder's inequality) If $a_k \geq 0,\ b_k \geq 0$ for k = 1, 2,…, n, and $\frac{1}{p} + \frac{1}{q} = 1$ with p>1, then

$$\left(\sum_{k=1}^{n} a_k^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} b_k^q\right)^{\frac{1}{q}} \geq \sum_{k=1}^{n} a_k b_k\ , \tag{1}$$

with equality holding if and only if $\alpha a_k^p = \beta b_k^q$ for k = 1, 2,…, n, where α and β are real nonnegative constants with $\alpha^2 + \beta^2 > 0$.
Proof. See [6].

Theorem 3.2. (Reverse Hölder's inequality) If $a_k > 0, b_k > 0$ for k = 1, 2,…, n, and $\frac{1}{p} + \frac{1}{q} = 1$ with p<0 or q<0, then

$$\left(\sum_{k=1}^{n} a_k^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} b_k^q\right)^{\frac{1}{q}} \leq \sum_{k=1}^{n} a_k b_k\ , \tag{2}$$

with equality holding if and only if $\alpha a_k^p = \beta b_k^q$ for k = 1, 2,…, n, where α and β are real nonnegative constants with $\alpha^2 + \beta^2 > 0$.
Proof. The reverse Hölder's inequality follows from Theorem 3.1. Here is the sketch of the proof (for details see [6]). Assume that p<0 and let $P = -p/q$, $Q = 1/q$. Then $1/P + 1/Q = 1$ with P>0 and Q>0. Therefore, according to (1) we have

$$\left(\sum_{k=1}^{n} A_k^P\right)^{\frac{1}{P}} \left(\sum_{k=1}^{n} B_k^Q\right)^{\frac{1}{Q}} \geq \sum_{k=1}^{n} A_k B_k\ ,$$

where $A_k > 0$ and $B_k > 0$ for k = 1, 2, …, n. The last inequality for $A_k = a_k^{-q}$ and $B_k = a_k^q b_k^q$ becomes (2). Q.E.D.

Corollary 3.3. (Reverse Hölder's inequality for *n*=2) If $a_k > 0, b_k > 0$ for k = 1, 2, and $\frac{1}{p} + \frac{1}{q} = 1$ with p<0 or q<0, then

$$\left(a_1^p + a_2^p\right)^{\frac{1}{p}} \left(b_1^q + b_2^q\right)^{\frac{1}{q}} \le a_1 b_1 + a_2 b_2 \,, \tag{3}$$

with equality holding if and only if $\frac{a_1^p}{b_1^q} = \frac{a_2^p}{b_2^q}$.

4 PROBLEM FORMULATION, ANALYSIS AND MAIN RESULTS

Let us first formulate the problem as it is presented in [2]. Suppose the firm's technology is represented by the two-input CES production function. Its cost-minimization problem for a given level of output y can be formulated as

$$\min_{x_1 \ge 0, x_2 \ge 0} \; w_1 x_1 + w_2 x_2 \tag{4}$$

$$\text{s.t.} \; \left(\alpha_1 x_1^\rho + \alpha_2 x_2^\rho\right)^{\frac{1}{\rho}} = y \,, \tag{5}$$

where the function $f\left(x_1, x_2\right) = \left(\alpha_1 x_1^\rho + \alpha_2 x_2^\rho\right)^{\frac{1}{\rho}}$ from (5) is called the CES production function.

4.1 Lagrange multiplier method approach

Let us first give an overview of the standard approach to solving this problem, which is the differential calculus approach, specifically the Lagrange multiplier method (see for example [2]). Here it has been implicitly assumed that all the variables as well as the CES production function are continuous. The corresponding Lagrangian function for the problem (4)-(5) is

$$L\left(x_1, x_2, \lambda\right) = w_1 x_1 + w_2 x_2 - \lambda \left[\left(\alpha_1 x_1^\rho + \alpha_2 x_2^\rho\right)^{\frac{1}{\rho}} - y\right]. \tag{6}$$

The first order conditions lead to the following two conditions:

$$\frac{w_1}{w_2} = \frac{\alpha_1}{\alpha_2} \cdot \left(\frac{x_1}{x_2}\right)^{\rho - 1} \,, \tag{7}$$

$$y = \left(\alpha_1 x_1^\rho + \alpha_2 x_2^\rho\right)^{\frac{1}{\rho}}. \tag{8}$$

Solving (7) for x_1, substituting in (8), and rearranging gives

$$y = x_2 \left(\alpha_1 \cdot \left(\frac{w_1}{w_2} \cdot \frac{\alpha_2}{\alpha_1}\right)^{\frac{\rho}{\rho - 1}} + \alpha_2\right)^{\frac{1}{\rho}}. \tag{9}$$

From (9) and (7) we get solution for x_1 and x_2:

$$x_1^* = y \cdot \left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}} \cdot \left(\alpha_1 \left(\frac{w_1}{\alpha_1}\right)^{\frac{\rho}{\rho-1}} + \alpha_2 \left(\frac{w_1}{\alpha_2}\right)^{\frac{\rho}{\rho-1}}\right)^{-\frac{1}{\rho}}, \tag{10a}$$

$$x_2^* = y \cdot \left(\frac{w_2}{\alpha_2}\right)^{\frac{1}{\rho-1}} \cdot \left(\alpha_1 \left(\frac{w_1}{\alpha_1}\right)^{\frac{\rho}{\rho-1}} + \alpha_2 \left(\frac{w_1}{\alpha_2}\right)^{\frac{\rho}{\rho-1}}\right)^{-\frac{1}{\rho}}. \tag{10b}$$

To prove that (x_1^*, x_2^*) from (10a) and (10b) is the unique minimizer for the problem (4)-(5), the sign of the bordered Hessian determinant of the Lagrangian function (6) must be examined, which is not trivial (see for instance [1], where authors examined the properties of the CES production function with multiple inputs using differential calculus). However, it can be proved that (x_1^*, x_2^*), given by (10a) and (10b), is the unique minimizer of the problem (4)-(5), with corresponding minimum cost given by

$$c_{\min} = w_1 x_1^* + w_2 x_2^* = y \cdot \left(w_1 \left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}} + w_2 \left(\frac{w_2}{\alpha_2}\right)^{\frac{1}{\rho-1}}\right)^{\frac{\rho-1}{\rho}}. \tag{11}$$

4.2 Reverse Hölder's inequality approach

Let us present the method for solving problem (4)-(5) by using the reverse Hölder's inequality approach. Note that the method does not assume the continuity of variables and the CES production function and therefore is valid for discrete case as well.

From (5) we have

$$x_2 = \left(\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2} x_1^\rho\right)^{\frac{1}{\rho}}. \tag{12}$$

Using (12), we transform the problem (4)-(5) into the following unconstrained problem:

$$\min_{x_1 \ge 0} g(x_1) = w_1 x_1 + w_2 \left(\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2} x_1^\rho\right)^{\frac{1}{\rho}}, \tag{13}$$

where g is a real function of a real variable. Let $a_1 = w_1 \left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{1}{\rho}}$, $a_2 = w_2$, $b_1 = \left(\frac{\alpha_1}{\alpha_2}\right)^{\frac{1}{\rho}} \cdot x_1$, $b_2 = \left(\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2} x_1^\rho\right)^{\frac{1}{\rho}}$, $p = \frac{\rho}{\rho - 1}$, $q = \rho$. Note that $\frac{1}{p} + \frac{1}{q} = \frac{\rho - 1}{\rho} + \frac{1}{\rho} = 1$. Furthermore, note that for all $\rho \in \langle -\infty, 0 \rangle$ the inequalities $0<p<1$ and $q<0$ hold, while for all $\rho \in \langle 0, 1 \rangle$ we have $p<0$ and $0<q<1$. Thus, in any of these two cases we can apply Corollary 3.3. on (13):

$$g(x_1) = \underbrace{w_1\left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{1}{\rho}}}_{a_1}\underbrace{\left(\frac{\alpha_1}{\alpha_2}\right)^{\frac{1}{\rho}} x_1}_{b_1} + \underset{a_2}{w_2}\underbrace{\left(\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2}x_1^\rho\right)^{\frac{1}{\rho}}}_{b_2}$$

$$\geq \left[w_1^p\left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{p}{\rho}} + w_2^p\right]^{\frac{1}{p}} \cdot \left[\left(\left(\frac{\alpha_1}{\alpha_2}\right)^{\frac{1}{\rho}} x_1\right)^q + \left(\left(\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2}x_1^\rho\right)^{\frac{1}{\rho}}\right)^q\right]^{\frac{1}{q}} \tag{14}$$

$$= y\left(w_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}} + w_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}}\right)^{\frac{\rho-1}{\rho}}.$$

Equality in (14) holds if and only if $\frac{a_1^p}{b_1^q} = \frac{a_2^p}{b_2^q}$, which is equivalent to

$$\frac{\left(w_1\left(\frac{\alpha_2}{\alpha_1}\right)^{\frac{1}{q}}\right)^{\frac{\rho}{\rho-1}}}{\left(\left(\frac{\alpha_1}{\alpha_2}\right)^{\frac{1}{q}} x_1\right)^{\rho}} = \frac{w_2^{\frac{\rho}{\rho-1}}}{\frac{y^\rho}{\alpha_2} - \frac{\alpha_1}{\alpha_2}x_1^\rho} . \tag{15}$$

From (15) we get

$$\overline{x_1} = y \cdot \left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}} \cdot \left(\alpha_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{\rho}{\rho-1}} + \alpha_2\left(\frac{w_1}{\alpha_2}\right)^{\frac{\rho}{\rho-1}}\right)^{-\frac{1}{\rho}}, \tag{16}$$

and by combining (16) and (12) we get

$$\overline{x_2} = y \cdot \left(\frac{w_2}{\alpha_2}\right)^{\frac{1}{\rho-1}} \cdot \left(\alpha_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{\rho}{\rho-1}} + \alpha_2\left(\frac{w_1}{\alpha_2}\right)^{\frac{\rho}{\rho-1}}\right)^{-\frac{1}{\rho}} . \tag{17}$$

Thus, from (14) it follows that

$$g(x_1) \geq y\left(w_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}} + w_1\left(\frac{w_1}{\alpha_1}\right)^{\frac{1}{\rho-1}}\right)^{\frac{\rho-1}{\rho}} \tag{18}$$

for all $x_1 \geq 0$, where equality in (18) holds if and only if x_1 is given by (16). By definition, this means that the function g has a global minimum which is achieved at the unique minimizer given by (16). Thus, we observe that the results (14), (16) and (17) obtained by the reverse Hölder's inequality are equivalent to (11), (10a) and (10b). Therefore, this approach gives the same results as Lagrange multiplier method, but unlike the differential calculus, our method is appropriate for the discrete case as well. Moreover, in comparison to differential calculus, the reverse Hölder's inequality does not require examining the first and the second order conditions, which are quite complicated in case of the CES production function. Note that, unlike the Lagrange multiplier method, the method based on the reverse Hölder's inequality immediately obtains the value of the global minimum for the problem (4)-(5).

5 CONCLUSION

In this paper, we treat a standard and a very important microeconomic problem of firm's cost minimization problem with the CES production function in case of two inputs. We present and compare two approaches to solve this problem: calculus approach, i.e. Lagrange multiplier method, and reverse Hölder's inequality approach. The use of differential calculus demands non-trivial examination of the first and the second order condition in order to prove that the global minimum and unique minimizer exist. On the other hand, by using the reverse Hölder's inequality, the problem can be solved in two steps only, where the values of the global minimum and the unique minimizer are obtained immediately (there is no need for complicated calculation required by the calculus method). Moreover, if the firm's inputs are given as discrete variables, the cost minimization problem becomes a problem of discrete optimization, so formally the calculus cannot be applied. However, the reverse Hölder's inequality approach does not require the continuity of the inputs as variables, making it an appropriate method for the discrete case.

References

[1] Avvakumov, S. N., Kiselev, Yu. N., Orlov, M. V., Taras'ev, A. M. (2010). Profit maximization problem for Cobb-Douglas and CES production function. Computational Mathematics and Modeling, 21 (3): 336-378.

[2] Jehle, G. A., Reny, P. J. (2011). Advanced Microeconomic Theory. Third Edition. Pearson.

[3] Kojić, V. (2015). A non-calculus approach to solving the utility maximization problem using the Cobb-Douglas and CES utility function. Croatian Operational Research Review, 6: 269-277.

[4] Kojić, V. (2017). Solving the consumer's utility-maximization problem with CES and Cobb-Douglas utility function via mathematical inequalities. Optimization Letters, 11 (4): 875-884.

[5] Kojić, V., Lukač. Z. (2018). An alternative approach to solving cost minimization problem with Cobb-Douglas technology. Central European Journal of Operations Research, 26 (3): 629-643.

[6] Mitrinović, D. S. (1970). Analytic Inequalities. Berlin: Springer-Verlag.

ON THE PROPERTIES
OF THE SATO PRODUCTION FUNCTION

Vedran Kojić, Zrinka Lukač and Krunoslav Puljić
University of Zagreb, Faculty of Economics & Business, Department of Mathematics
Trg J. F. Kennedyja 6, HR–10000 Zagreb, Croatia
vkojic@efzg.hr, zlukac@efzg.hr, kpuljic@efzg.hr

Abstract: Production function with only one output plays one of the key roles in the theory of the firm. Some of the common examples of production functions in the main microeconomic literature are, for example, CES, Cobb-Douglas or Leontief production function. In order to be a production function, a given mathematical function must satisfy certain properties. Various properties of the CES, Cobb-Douglas and Leontief production functions are proved and very well known in general. However, one of the not so common production functions is the "Sato function". In this paper, we prove that Sato function satisfies the necessary assumptions for a function to be a production function. These requirements are continuity, strict monotonicity, strict quasiconcavity and that a positive amount of output requires positive amounts of some of the inputs. To the best of our knowledge, our results are new and unknown in the microeconomic literature.

Keywords: Sato production function, continuity, strict monotonicity, strict quasiconcavity

1 INTRODUCTION

According to [7], Ryuzo Sato proposed a new production function with one output given by the following expression:

$$Y = F(K,L) = \frac{K^2L^2}{aK^3 + bL^3}, \quad a > 0, b > 0, \tag{1}$$

where Y is the total output, K is the capital, and L is the labor. In his paper "On the Stability of Growth Equilibrium" presented at the 1963 meeting of the Econometric Society, Sato used the function (1) to illustrate the stability conditions of the growth equilibrium. The production function (1) can also be found in some more recent research, such as in [6]. However, in order to be a production function, mathematical function must satisfy four fundamental assumptions. Referring to [3], it is known that the production function $f:\mathbb{R}_+^n \to \mathbb{R}_+$ has to be continuous, strictly increasing, strictly quasiconcave and $f(0)=0$. However, to the best of our knowledge, there are no rigorous mathematical proofs that the Sato function satisfies the four assumptions mentioned previously. Hence, the purpose of this paper is to give the strict and rigorous proofs of the Sato function properties.

The structure of this paper is as follows. After the introduction, in the second section we present the notation. The third section contains preliminary mathematical facts that we use in the rest of the paper. The fourth section presents the main results, while the fifth section presents the conclusion and outlines the further research.

2 NOTATION

We use the following notation:

x	the amount of the first input (capital, for example)
y	the amount of the second input (labor, for example)
a, b	positive real numbers, a>0, b>0
$\mathbb{R}_+$	the set of all non-negative real numbers

$\mathbb{R}_+^n$	$\mathbb{R}_+^n \equiv \{(x_1, x_2, \ldots, x_n) : x_i \geq 0,\ i = 1, 2, \ldots, n\} \subset \mathbb{R}^n,\ n \in \mathbb{N}$
$x \leq y$	$x_i \leq y_i,\ i = 1, 2, \ldots, n$, where $x = (x_1, x_2, \ldots, x_n),\ y = (y_1, y_2, \ldots, y_n)$
$x \ll y$	$x_i < y_i,\ i = 1, 2, \ldots, n$, where $x = (x_1, x_2, \ldots, x_n),\ y = (y_1, y_2, \ldots, y_n)$
int K	interior of the set K

3 PRELIMINARY FACTS

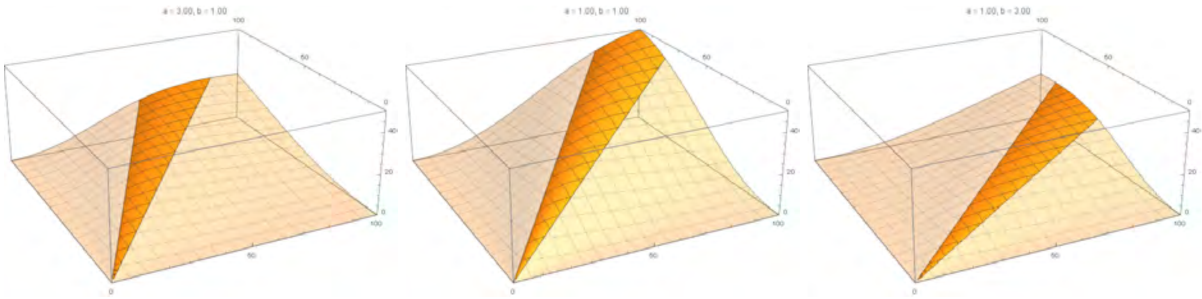
Instead of the function (1), we treat the function $f : S \cup \{(0,0)\} \subset \mathbb{R}_+^2 \to \mathbb{R}_+$,

$$f(x,y) = \begin{cases} \dfrac{x^2 y^2}{ax^3 + by^3}, & \text{for } (x,y) \in S, \\ 0 & , \text{for } x = y = 0, \end{cases} \tag{2}$$

where

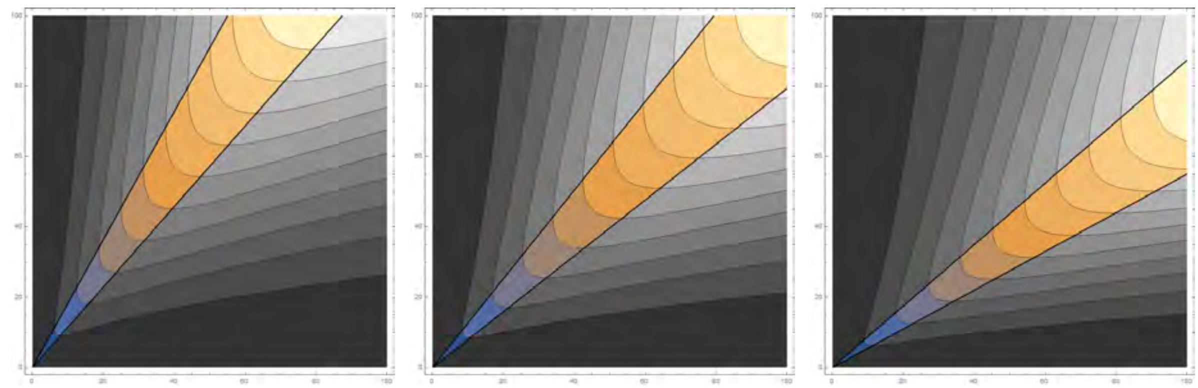
$$S = \left\{(x,y) \in \mathbb{R}_+^2 : \left(2bxy^5 - ax^4y^2\right)\left(ax^3 + by^3\right)^{-2} \geq 0, \left(2ax^5y - bx^2y^4\right)\left(ax^3 + by^3\right)^{-2} \geq 0, x^2 + y^2 > 0\right\} \tag{3}$$

Notice that here we use the notation x instead of the capital K and the notation y instead of the labor L. Furthermore, in economics, the set S is called the "economic region of production". Economic region of production is the set of all inputs (x,y) where the isoquants are downward-sloping. According to [1], uneconomic region of production is "the region of upward-sloping or backward-bending isoquants. In the uneconomic region, at least one input has a negative marginal product, i.e. $MP_K<0$ or $MP_L<0$ (in our notation, MP_K is f_x and MP_L is f_y).



Source: Authors.

Figure 1: Sato function with a=3, b=1; a=b=1 and a=1, b=3. Economic region is dark colored.



Source: Authors.

Figure 2: Sato function isoquants with a=3, b=1; a=b=1 and a=1, b=3. Economic region is colored.

A firm that wants to minimize its production costs should never operate in a region of upward-sloping or backward-bending isoquants.“ This is the main reason why the domain of the Sato function needs to be $S \cup \{(0,0)\}$, instead of $\mathbb{R}_+^2$. The graph of the Sato function with *a*=3, *b*=1; *a*=*b*=1 and *a*=1, *b*=3 can be seen in Figure 1. Figure 2 illustrates the corresponding economic regions in these three cases. Since

$$f_x(x,y) = \frac{2bxy^5 - ax^4y^2}{\left(ax^3+by^3\right)^2} = \underbrace{\frac{xy^2}{\left(ax^3+by^3\right)^2}}_{\geq 0} \cdot \left(2by^3 - ax^3\right) \text{, and} \tag{4}$$

$$f_y(x,y) = \frac{2ax^5y - bx^2y^4}{\left(ax^3+by^3\right)^2} = \underbrace{\frac{x^2y}{\left(ax^3+by^3\right)^2}}_{\geq 0} \cdot \left(2ax^3 - by^3\right), \tag{5}$$

the set S becomes

$$S = \left\{(x,y) \in \mathbb{R}_+^2 \setminus \{(0,0)\} : 2by^3 \geq ax^3, 2ax^3 \geq by^3\right\} = \left\{(x,y) \in \mathbb{R}_+^2 \setminus \{(0,0)\} : \sqrt[3]{\frac{a}{2b}}x \leq y \leq \sqrt[3]{\frac{2a}{b}}x\right\}. \tag{6}$$

It is trivial to prove that $S \cup \{(0,0)\}$ is a convex cone. Now we present several well-known mathematical facts that we use in the rest of the paper.

Theorem 3.1. (Inequality between arithmetic and geometric mean – AG-inequality)
For all real numbers $x \geq 0, y \geq 0$, the following inequality holds:

$$\sqrt{xy} \leq \frac{x+y}{2}. \tag{7}$$

Equality in (7) holds if and only if *x*=*y*.
Proof. Inequality (7) is equivalent to $\left(\sqrt{x}-\sqrt{y}\right)^2 \geq 0$, which is obviously true. Equality holds if and only if $\sqrt{x} = \sqrt{y}$, which is equivalent to *x*=*y*. Q.E.D.

Theorem 3.2. Let $f: D_f \subseteq \mathbb{R}^2 \to \mathbb{R}$ be a differentiable function on the open set D_f. If $f_x(x,y) > 0$ for all $(x,y) \in D_f$, then the restriction $f|_{\{\mathbb{R} \times \{\tilde{y}\}\} \cap D_f}$ of the function f is strictly increasing on $\{\mathbb{R} \times \{\tilde{y}\}\} \cap D_f$ for all fixed $\tilde{y}$ such that $(x,\tilde{y}) \in D_f$.
Proof. This is the consequence of Lagrange's Mean Value Theorem. The proof is analogous to the proof of the similar statement in case of the one-variable function, which can be seen, for instance, in [4]. Note that analogous statement holds for partial derivative f_y. Q.E.D.

Theorem 3.3. Any convex subset of $\mathbb{R}^n$ is path connected and hence connected.
Proof. See [5].

Assumption 3.4. (Properties of the production function) The production function, $f: \mathbb{R}_+^n \to \mathbb{R}_+$, is continuous, strictly increasing, strictly quasiconcave on $\mathbb{R}_+^n$, and $f(0) = 0$. (See [3].)

4 MAIN RESULTS

In this section we prove that Sato function f given by (2) satisfies all properties from Assumption 3.4.
Proposition 4.1. Sato function is continuous and $f(0) = 0$.

Proof. Sato function *f* satisfies $f(0)=0$ by definition (see relation (2)). Furthermore, since *f* is a rational function of two variables, it is of the class C^∞ on S (in fact, it is of the class C^∞ on $\mathbb{R}^2 \setminus \{(0,0)\}$), and therefore it is continuous on S. It remains to prove that Sato function is continuous in the point (0,0). Hence, we have to prove that relation (8) holds:

$$\text{for all } \varepsilon > 0 \text{ there exists } \delta = \delta(\varepsilon) > 0 \text{ such that for all } (x,y) \in S,$$
$$\sqrt{(x-0)^2 + (y-0)^2} < \delta \Rightarrow \left| \frac{x^2 y^2}{ax^3 + by^3} - 0 \right| < \varepsilon. \tag{8}$$

For any arbitrary but fixed $\varepsilon > 0$, let $\delta = 2\sqrt{2}\sqrt{ab} \cdot \varepsilon$. Since $\sqrt{x^2 + y^2} < 2\sqrt{2}\sqrt{ab} \cdot \varepsilon$, we have

$$\frac{x^2 y^2}{ax^3 + by^3} \overset{AG}{\leq} \frac{x^2 y^2}{2\sqrt{ab} \cdot x^{3/2} y^{3/2}} = \frac{\sqrt{xy}}{2\sqrt{ab}} \overset{AG}{\leq} \frac{1}{2\sqrt{ab}} \sqrt{\frac{x^2 + y^2}{2}} = \frac{\sqrt{x^2 + y^2}}{2\sqrt{2}\sqrt{ab}} < \frac{2\sqrt{2}\sqrt{ab} \cdot \varepsilon}{2\sqrt{2}\sqrt{ab}} = \varepsilon . \tag{9}$$

This completes the proof. Q.E.D.

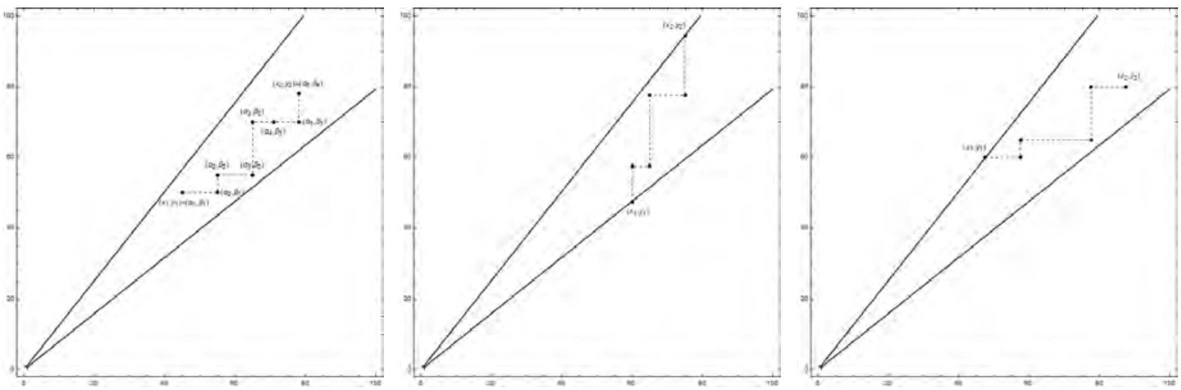
Proposition 4.2. Sato function is strictly increasing.
Proof. We prove that for any two points $(x_1, y_1), (x_2, y_2) \in S \cup \{(0,0)\}$ the following implication holds:

$$(x_1, y_1) \prec (x_2, y_2) \Rightarrow f(x_1, y_1) < f(x_2, y_2). \tag{10}$$

Since $S \cup \{(0,0)\}$ is convex cone in $\mathbb{R}^2$, according to Theorem 3.3., $S \cup \{(0,0)\}$ is connected. Hence, there exist finite number of points $(\alpha_i, \beta_j) \in S \cup \{(0,0)\}$, $i \in \{1,2,...,p\}$, $j \in \{1,2,...,q\}$, $p, q \in \mathbb{N}$, $x_1 = \alpha_1 < \alpha_2 < \cdots < \alpha_p = x_2$ and $y_1 = \beta_1 < \beta_2 < \cdots < \beta_q = y_2$, such that

$$(x_1, y_1) = (\alpha_1, \beta_1) \leq \cdots \leq (\alpha_s, \beta_t) \leq (\alpha_u, \beta_v) \leq \cdots \leq (\alpha_p, \beta_q) = (x_2, y_2), \tag{11}$$

where for any two adjacent points (α_s, β_t) and (α_u, β_v) in (11), inequality $(\alpha_s, \beta_t) \leq (\alpha_u, \beta_v)$ implies that $\alpha_s = \alpha_u$, $\beta_t < \beta_v$, or $\alpha_s < \alpha_u$, $\beta_t = \beta_v$ (see Figure 3).



Source: Authors.

Figure 3: Examples of three paths between points (x_1, y_1) and (x_2, y_2).

Therefore, without the loss of generality, it is sufficient to prove the case where $(x_1, y_1), (x_1, y_2), (x_2, y_2) \in S$ (see Figure 4) implies

$$(x_1, y_1) \prec (x_2, y_2) \Rightarrow f(x_1, y_1) < f(x_2, y_2). \tag{12}$$

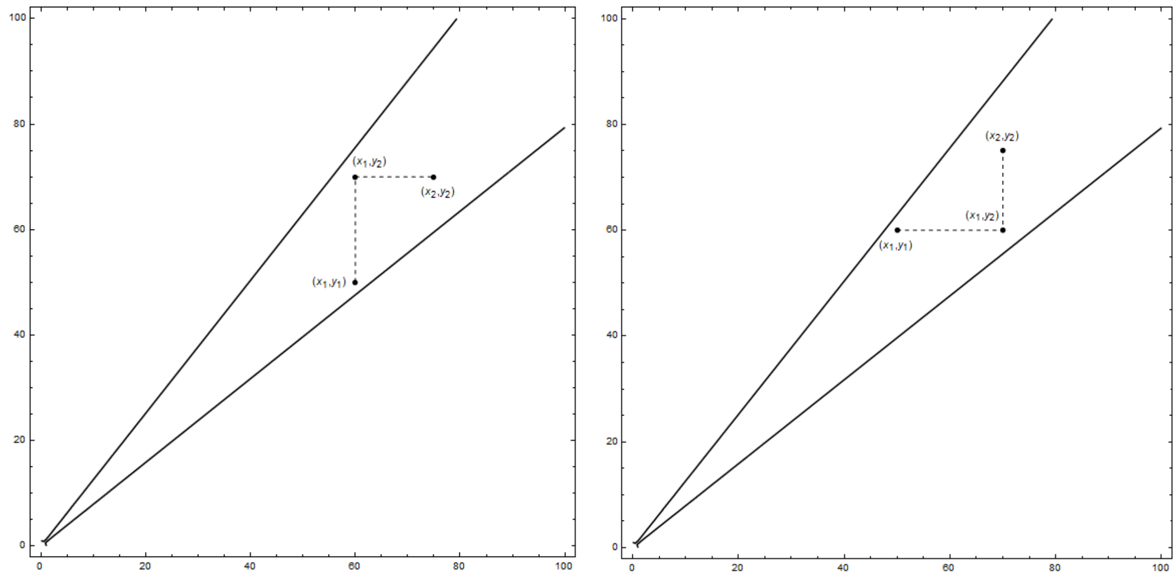
Since $f_x(x,y)>0,\ f_y(x,y)>0$ for all $(x,y)\in \operatorname{int} S$, by Theorem 3.2., we have

$$(x_1,y_1)\le(x_1,y_2)\Rightarrow f(x_1,y_1)<f(x_1,y_2), \tag{13}$$

and

$$(x_1,y_2)\le(x_2,y_2)\Rightarrow f(x_1,y_2)<f(x_2,y_2). \tag{14}$$

Combining (13) and (14), we get (12), which completes the proof. Q.E.D.



Source: Authors.

Figure 4: Two paths between points (x_1, y_1) and (x_2, y_2).

Proposition 4.3. Sato function is strictly quasiconcave.

Proof. Let us first note that in 1969 Geithman and Stinson in [2] commented on diminishing returns of Sato function. However, they did not show that Sato function is strictly quasiconcave. Now, let us prove this proposition. Since $S\cup\{(0,0)\}$ is a convex set, it is sufficient to prove that the determinant of the bordered Hessian of function *f* in any point $(x,y)\in \operatorname{int} S$ is positive, i.e.

$$\det H_f(x,y)=\begin{vmatrix}0 & f_x & f_y\\ f_x & f_{xx} & f_{yx}\\ f_y & f_{xy} & f_{yy}\end{vmatrix}=2f_xf_yf_{xy}-f_x^2f_{yy}-f_y^2f_{xx}>0. \tag{15}$$

For the Sato function we have

$$f_x=\frac{2bxy^5-ax^4y^2}{\left(ax^3+by^3\right)^2},\ f_y=\frac{2ax^5y-bx^2y^4}{\left(ax^3+by^3\right)^2}, \tag{16}$$

$$f_{xx}=\frac{2y^2\left(a^2x^6+b^2y^6-7abx^3y^3\right)}{\left(ax^3+by^3\right)^3},\ f_{yy}=\frac{2x^2\left(a^2x^6+b^2y^6-7abx^3y^3\right)}{\left(ax^3+by^3\right)^3},\ \text{and} \tag{17}$$

$$f_{xy}=\frac{2xy\left(-a^2x^6-b^2y^6+7abx^3y^3\right)}{\left(ax^3+by^3\right)^3}. \tag{18}$$

After long calculation we find

$$\det H_f(x,y) = \frac{2x^4y^4\left(-b^2y^6+7abx^3y^3-a^2x^6\right)}{\left(ax^3+by^3\right)^5} = \underbrace{\frac{2a^2x^{10}y^4}{\left(ax^3+by^3\right)^5}}_{>0}\left(-t^2+7t-1\right), \tag{19}$$

where $t=\dfrac{by^3}{ax^3}$. From (6) we get

$$\frac{1}{2} \le \frac{by^3}{ax^3} \le 2 \Leftrightarrow t \in \left[\frac{1}{2}, 2\right]. \tag{17}$$

Since $-t^2+7t-1>0$ for all $t \in \left\langle \dfrac{7-3\sqrt{5}}{2}, \dfrac{7+3\sqrt{5}}{2} \right\rangle \supset \left[\dfrac{1}{2}, 2\right]$, (19) implies that $\det H_f(x,y)>0$, which completes the proof. Q.E.D.

5 CONCLUSION AND FURTHER RESEARCH

In microeconomic theory, every production function must satisfy four assumptions: (i) continuity, (ii) strict monotonicity, (iii) strict quasiconcavity and that (iv) a positive amount of output requires positive amounts of some of the inputs. In this paper, we proved that the Sato function satisfies these assumptions indeed. To the best our knowledge, our results are new and unknown in the microeconomic theory literature. We emphasize that only after the analysis provided by our results Sato function can be called "Sato production function". For further research, we will investigate the same properties of the generalized Sato production function with at least two inputs. Moreover, one of the future goals is to solve a very important microeconomic problem which is the cost minimization problem with Sato production function.

References

[1] Besanko, D., Braeutigam, R. R. (2014). Microeconomics. Fifthy Edition. Wiley.

[2] Geithman, D. T., Stinson, B. S. (1969). A note on Diminishing Returns and Linear Homogeneity. The American Economist, 13 (1), 77–79.

[3] Jehle, G. A., Reny, P. J. (2011). Advanced Microeconomic Theory. Third Edition. Pearson.

[4] Lukač, Z. (2014). Matematika za ekonomske analize. Zagreb: Element.

[5] Manetti, M. (2015). Topology, Springer-Verlag.

[6] Roubalová, L., Hampel, D., Viskotová, L. (2018). Technological Progress at the Sectoral Level: the Sato Production Function Approach. Proceedings of the 36th International Conference Mathematical Methods in Economics, Jindřichův Hradec, Czech Republic, 470–475.

[7] Sato, R. (1964). Diminishing Returns and Linear Homogenity: Comment. The American Economic Review, 54 (5), 744–745.

LEVEL OF IMPLEMENTATION OF LEAN MANUFACTURING TOOLS: A CASE STUDY IN THE NORTH OF PORTUGAL

Ângela Silva
Business School, Viana do Castelo Polytechnic Institute, and
Centro de Investigação em Organizações,
Mercados e Gestão Industrial (COMEGI), Lusíada University, Portugal
E-mail: angela.a@esce.ipvc.pt

Wellington Alves
Business School, Viana do Castelo Polytechnic Institute, and
ALGORITMI Research Centre, University of Minho,
Escola de Engenharia, Depto Produção e Sistemas, Campus de Azurém, Portugal
E-mail: wellingtonb@esce.ipvc.pt

Helena Sofia Rodrigues
Business School, Viana do Castelo Polytechnic Institute, and
Center for Research and Development in Mathematics and Applications (CIDMA),
University of Aveiro, Portugal
E-mail: sofiarodrigues@esce.ipvc.pt

Abstract: Nowadays business environment is very unstable, complex and requires a quick response from the companies, with a better allocation of their scarce resources, and a clearer strategic focus. The Lean Manufacturing requires that companies make the best use of their resources eliminating wastes. In this research, in order to evaluate the level of implementation of Lean tools in the companies located in the North of Portugal, an online survey was conducted. Results show that the 5S and TPM methods are the ones that have higher progress of implementation, and the Kanban tool has a lower level.

Keywords: Lean Management, 5S, Kanban, Kaizen, Just in time

1 INTRODUCTION

The instability of the business market and the growth of companies supply chain has been improved the organization's performance to become more efficient, flexible and faster to answer first to the changes in the business environments.

Lean Manufacturing plays an important role in supporting companies to overcome environmental, social and economic impacts attributed to the production processes, which has been a major concern for the industrial sector lately.

The Lean philosophy promotes efficiency and elimination of waste, focusing on a high customer service level. Based on that, Lean tools have been adopted by many companies to best improve their operations. Implementation of Lean manufacturing tools in any type of organizations can bring many benefits, such as waste reduction and improving operating efficiency [2]–[6]. This shows that Lean is not limited to one type or size of the company, but rather all types, sizes and industries that attempt to increase their competitive advantages, operations and profits in the regional and global markets [2]. However, in the literature, some studies were made suggesting that the implementation of the Lean Manufacturing concepts in industries are different in function of their dimension.

In this research, to evaluate the level of implementation of Lean tools in different types and dimension companies, an online survey was conducted with a set of companies from the North of Portugal. Then, Lean issues and practices such as Kaizen philosophy, 5S (Sort/Set in order/Shine/Standardize/Sustain), Total Productive Maintenance (TPM), Kanban, Just in time (JIT), stock reduction, Kaizen circles and collaboration with suppliers were analysed.

2 LITERATURE REVIEW

Lean Manufacturing is focused on the reduction of waste and improvement of operational efficiency using a set of different tools to get these objectives. Many of these tools can be successfully used in isolation, which makes it much easier to get started, but on the other hand, the benefits will propagate as more tools are used, as they do support and reinforce each other.

In the literature, some studies demonstrate the influence of the application of the Lean methods and tools in different performance indicators: Belekoukias *et al.*[9] have analysed the impact of Lean methods and tools on the operational performance of manufacturing organisations and the results indicate that JIT and Autonomation have the strongest significance on operational performance while Kaizen, TPM and value stream mapping (VSM) seem to have a lesser, or even negative, effect on it. More recently Garza-Reyes *et al.* [10] investigate the effect of the same five essential Lean methods, i.e., JIT, Autonomation, Kaizen/continuous improvement, TPM and VSM, on four commonly used indicators for the compliance of environmental performance, i.e., material use, energy consumption, non-product output, and pollutant releases.

Regarding the implementation of the Lean tools based on the companies' dimension, the large amount of research was focused on large scale organizations. More recently, the research on Lean concepts applications in Small, Medium Enterprises (SME) is increasing ([2], [12], [13]), due to the existence of a large number of these organization in the global territory [8].

There are many Lean methods and tools that can be used to improve the organization's performance. One of these consists in the JIT method [14]; the authors suggest that JIT is playing a significant role to achieve a high service level at a minimum cost. As mentioned, the TPM and Kaizen/continuous improvement methods also have a huge impact on the organization's performance. Different tools are used to implement these methods. The 5S, for example, is a simple tool which develops discipline and cleanliness at the workplace, maximizing efficiency and productivity.

Another important aspect related to Lean implementation is related to the close relationship between human resources and all the supply chain elements (suppliers, partners, and clients). The involvement of the top managers and the engagement of the workers in the implementation process is very important to get the performance objectives intended [15].

To evaluate the level of implementation of Lean tools in different companies dimension, in the North of Portugal, a survey was developed. It was implemented to a sample of 120 organizations, from micro to large scale dimension, focusing in a specific group of 9 methods and tools identified as Engagement of workers, Continuous Improvement, 5S, TPM, Kanban, JIT, Stock Reduction, Kaizen Circles, and Suppliers Relationship.

3 METHODOLOGY

To study the level of implementation of lean procedures in a set of Portuguese companies, a survey was conducted. The questionnaire was designed based on the work developed by Jabbour *et al.* [16]. The questionnaire consisted of two parts; the first contains general questions about the characterization of the companies, such as dimension, number of employees related to logistics, and turnover. The second part, the main one, consists of nine Lean attributes (Table 1). Each company was asked to rate their level of implementation of lean practices, with each item on a five-level Likert scale, from 1 (Not implemented) to 5 (Completely implemented).

The sample taken is a convenient one due to time and budget constraints. Companies were asked their willingness to fill out the questionnaire, published online through Google Docs forms, and 102 answers were obtained from multisector companies.

Table 1: Level of "Lean Management" practices

Question	***Description***
LM1	Engagement of workers
LM2	Continuous improvement
LM3	5S (Sort/Set in order/Shine/Standardize/Sustain)
LM4	Total productive maintenance (TPM)
LM5	Kanban (pull system)
LM6	Just in Time (JIT)
LM7	Stock reduction
LM8	Kaizen Circle (discussion groups to improve processes)
LM9	Collaboration with suppliers

4 ANALYSIS OF THE RESULTS

An analysis and discussion of the results were made, using a statistical approach through the software IBM SPSS version 24.

4.1 Sample characterization

The selected companies had a large spectre of characteristics, as summarized in Table 2. The results showed that the dimension of the companies related to the number of workers is very heterogeneous. More than 50% of the companies had a micro or small dimension, meeting the Portuguese business fabric.

Regarding the number of employees associated with the logistics area, it is possible to observe that a large number of companies had up to three workers associated with this field. It should be noted that the companies inquired are multisector, so this value is within the expected. Besides, a great number of companies had a turnover, by year, more than five million euros (36.3%).

Table 2: Technical record of participating companies

Dimension on the company (number of employees)	***Percent***		***Number of employees associated with logistics***	***Percent***		***Turnover (in euros)***	***Percent***
Micro (up to 10)	32.4		[0;3[	34.3		[0;100k[	13.7
Small (10-50)	25.5		[3;6[	20.6		[100k; 250k[	10.8
Medium (50-250)	16.7		[6;9[	5.9		[250k 500k[	9.8
Large (more than 250)	25.5		[9;12[	9.8		[500k; 1M[	10.8
			[12;15[	2.0		[1M; 5M[	18.6
			15 or more	27.5		5M or more	36.3

4.2 Lean management practices

Lean management practices should be implemented by managers, who are trained in lean concepts, and they are passed on throughout the organization. To understand the level of Lean procedures in Portuguese companies, there were pointed out nine questions. Table 3 compiles some descriptive statistics related to these procedures. For all items, the five points Likert scale achieve the highest score (five), meaning that the level of implementation of the environmental practices is unequal between companies.

For the case of practices LM3 and LM4 are the ones that have higher means values, meaning more progress of implementation. On the other hand, the lowest means values are related to the procedures LM5 and LM1. This could be explained by the fact that the 5S tool and the TPM method are considered hard lean practices which are more extensively used than soft Lean practices (Engagement)[15]. Also, the 5S is a simple implementation tool which allows rapid results with high visual impacts, consisting on the first tool to use for clean and straighten the workplace.

Table 2: Descriptive statistics for Lean management practices

Environment managment practices	*Min*	*Max*	*Mean*	*St. Dev.*
LM1	1	5	2.49	1.481
LM2	1	5	3.12	1.381
LM3	1	5	3.20	1.328
LM4	1	5	3.42	1.238
LM5	1	5	2.25	1.369
LM6	1	5	2.65	1.426
LM7	1	5	3.16	1.241
LM8	1	5	2.63	1.400
LM9	1	5	3.06	1.296

Figure 1 shows the boxplots for all questions. This graphic is according to Table 3, giving information about the use of the entire scale. Besides, 50% of inquired companies, partially implemented (level 3) the practices LM6 to LM9. All these practices are related to Production Pull System which requires the collaborations with the suppliers and the implementation of JIT method, getting stocks reduction. The enterprises that are trying to implement the Pull System should implement all these methods and tools at the same time, on the same level. In contrast, practice LM5 has a lower level of implementation, where 50% of companies selected levels 1 (not implemented) or 2 (starting to implement). The methods like Kanban, JIT are used by large companies and international groups generates a certain fear and barrier [7] Another curiosity is that only 25% of the companies have selected the full achievement/implementation of these practices.

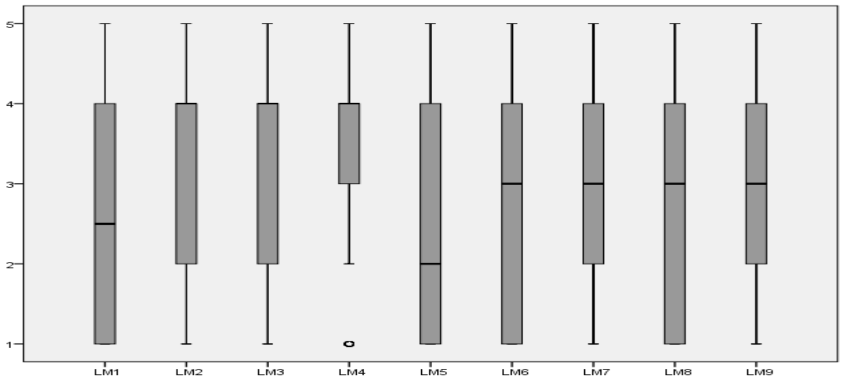


Figure 1: Boxplots of Lean management practices

In order to understand how these practices are correlated, Table 4 is presented. The correlations between items are not very high. However, the item LM4 stands out to the correlation with LM2 and LM3. TPM supports the predictive, preventive and corrective maintenance activities to achieve efficient production equipment and relies on tools such as 5S, single minute exchange of die (SMED), overall equipment effectiveness (OEE), planned, autonomous and quality maintenance and initial control before starting production [10].

Another important question related to this theme, is the level of implementation of Lean practices, according to the numbers of workers. According to Matt and Rauch [7], the Lean production methods and instruments are not equally applicable to large and small companies. Consequently, the level of accomplishment of Lean procedures was also analyzed, taking into consideration the dimension of the company.

Table 3: Matrix correlation between environment management practices [For all values, correlation is significant at the 0.01 level (2-tailed)].

Item	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9
LM1	1								
LM2	0.591	1							
LM3	0.625	0.576	1						
LM4	0.572	0.671	0.696	1					
LM5	0.558	0.518	0.555	0.491	1				
LM6	0.660	0.464	0.675	0.489	0.655	1			
LM7	0.572	0.532	0.600	0.607	0.454	0.586	1		
LM8	0.528	0.633	0.530	0.537	0.587	0.385	0.581	1	
LM9	0.398	0.560	0.505	0.435	0.388	0.338	0.407	0.695	1

The results presented in Figure 2, show from the consulted companies, micro and small companies have the lowest levels of implementation of Lean procedures, which can be justified by the fact that these companies do not feel the need to implement these systems to be productive. Another explanation is related to the challenge of the implementation of some integrated Lean production systems due to specific knowledge and money spent.

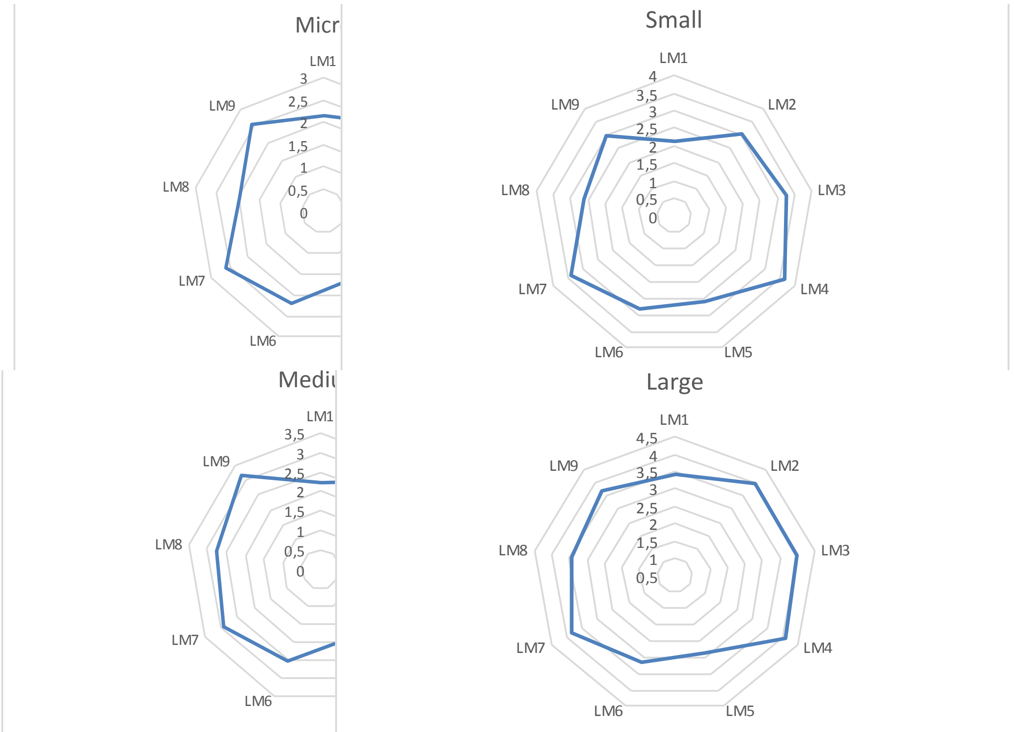


Figure 2: Average level of environment management practices, by companies' dimension

5 CONCLUSIONS

In this work, it was possible to analyze the implementation level of Lean practices in the North of Portugal. Despite being an initial analysis, the results showed the 5S and TPM methods are the ones that have higher progress of implementation, and the kanban tool has a lower level of implementation, in general. It is also possible to conclude that micro and small companies have the lowest levels of implementation of Lean procedures.

Acknowledgement
This research was supported by the FCT – Fundação para a Ciência e Tecnologia, through the project UID/EMS/04005/2019 (Silva); the project UID/CEC/00319/2019 (Alves); and the project UID/MAT/04106/2019 (Rodrigues).

References

[1] M. A. Lewis, "Lean production and sustainable competitive advantage," *Int. J. Oper. Prod. Manag.*, vol. 20, no. 8, pp. 959–978, Aug. 2000.

[2] A. Alkhoraif, H. Rashid, and P. McLaughlin, "Lean implementation in small and medium enterprises: Literature review," *Oper. Res. Perspect.*, no. December, p. 100089, 2018.

[3] N. Cardoso, A. C. Alves, M. Figueiredo, and A. Silva, "Improving workflows in a hospital through the application of lean thinking principles and simulation," *"Proceedings Int. Conf. Comput. Ind. Eng. CIE,"* no. October, pp. 11–13, 2017.

[4] V. Resende, A. C. Alves, A. Batista, and Â. Silva, "Financial and human benefits of lean production in the plastic injection industry: An action research study," *Int. J. Ind. Eng. Manag.*, vol. 5, no. 2, pp. 61–75, 2014.

[5] J. A. Garza-Reyes, V. Kumar, S. Chaikittisilp, and K. H. Tan, "The effect of lean methods and tools on the environmental performance of manufacturing organisations," *Int. J. Prod. Econ.*, vol. 200, 2018.

[6] I. Belekoukias, J. A. Garza-Reyes, and V. Kumar, "The impact of lean methods and tools on the operational performance of manufacturing organisations," *Int. J. Prod. Res.*, vol. 52, no. 18, pp. 5346–5366, Sep. 2014.

[7] D. T. Matt and E. Rauch, "Implementation of Lean Production in small sized Enterprises," *Procedia CIRP*, vol. 12, pp. 420–425, 2013.

[8] A. D. Jewalikar and A. Shelke, "Lean Integrated Management Systems in MSME Reasons, Advantages and Barriers on Implementation," in *Materials Today: Proceedings*, 2017, vol. 4, no. 2, pp. 1037–1044.

[9] I. Belekoukias, J. A. Garza-Reyes, and V. Kumar, "The impact of lean methods and tools on the operational performance of manufacturing organisations," *Int. J. Prod. Res.*, vol. 52, no. 18, pp. 5346–5366, 2014.

[10] J. A. Garza-Reyes, V. Kumar, S. Chaikittisilp, and K. H. Tan, "The effect of lean methods and tools on the environmental performance of manufacturing organisations," *Int. J. Prod. Econ.*, 2018.

[11] A. K. Möldner, J. A. Garza-Reyes, and V. Kumar, "Exploring lean manufacturing practices' influence on process innovation performance," *Journal of Business Research*, 2018.

[12] A. Pearce, D. Pons, and T. Neitzert, "Implementing lean—Outcomes from SME case studies," *Oper. Res. Perspect.*, vol. 5, pp. 94–104, 2018.

[13] M. Almanei, K. Salonitis, and Y. Xu, "Lean Implementation Frameworks: The Challenges for SMEs," *Procedia CIRP*, vol. 63, pp. 750–755, 2017.

[14] S. Abdul, R. Khan, D. Qianli, and Y. Zhang, "A Survey Study : Important Factors in Just-in-Time Implementation," *Traffic Transp. Eng.*, vol. 2, no. 5, pp. 74–80, 2017.

[15] Y. Larteb, A. Haddout, and M. Benhadou, "Successful Lean Implementation: The Systematic and Simultaneous Consideration of Soft and Hard Lean Practices," *Int. J. Eng. Gen. Sci.*, vol. 3, no. 2, pp. 1258–1270, 2014.

[16] A. B. L. de S. Jabbour, C. J. C. Jabbour, W. R. de S. Freitas, and A. A. Teixeira, "Lean and green?: evidências empíricas do setor automotivo brasileiro," *Gestão & Produção*, vol. 20, no. 3, pp. 653–665, 2013.

COOPERATIVENESS IN DUOPOLY FROM AN EVOLUTIONARY GAME THEORY PERSPECTIVE

Ilko Vrankić, PhD
University of Zagreb, Faculty of Economics and Business, Trg J. F. Kennedyja 6, Zagreb, Croatia
E-mail: ivrankic@efzg.hr

Mirjana Pejić Bach, PhD
University of Zagreb, Faculty of Economics and Business, Trg J. F. Kennedyja 6, Zagreb, Croatia
E-mail: mpejic@efzg.hr

Tomislav Herceg, PhD
University of Zagreb, Faculty of Economics and Business, Trg J. F. Kennedyja 6, Zagreb, Croatia
E-mail: therceg@efzg.hr

Abstract: In a duopoly analysis companies can partake in different mutual relations. In case when the quantity of production is a decision variable, the equilibria can be of either Cournot, Stackelberg or cartel type. In these market structures, companies pick different behavioural patterns, thus making different profits. In cooperative equilibrium, both companies have an incentive to cheat since a unanimous move can improve their own profit, but since it lowers competitor's profits in the same time, this equilibrium collapses to Cournot equilibrium where profits are lower. In Stackelberg equilibrium, a dominant company yield greater profits at the expense of the lower follower's profit and the total industry profit in Cournot equilibrium. Therefore this paper is looking for the best long-term behaviour pattern of a company, using a new perspective unifying microeconomic analysis of market structures and evolutionary game theory; it is assumed that companies pick, with certain probabilities, either cooperative, reactive or dominant strategies. Companies in time adapt their choices by increasing the choice probability of those strategies which fitness is greater than the average profit, applying the replicatory dynamics. The paper also demonstrates that the decision-making of companies strongly depends in the long run on the starting set of strategies.

Keywords: Evolutionary game theory, duopoly, cooperative strategy, dominant strategy, reactive strategy, the tit-for-tat strategy

1 INTRODUCTION

Duopolistic market structures in the case when quantity is decision variable can be based on several strategies which lead to Cournot (reactive), Stackelberg (dominant) or cartel equilibrium (cooperative). In most literature, Cournot players are considered to be naïve since they keep behaving as if their peer would never change their mind. However, a different insight is offered by [4] who shows that best survivors are not always those who maximize their profit, at least in the short run.

Definition of the long run is a time in which no variable is fixed. The other possibility to define long run is a time at which partakers on the market have enough time to adapt. In that period of time, players can adapt their decisions and evolve to a new way of behaviour. This kind of thinking is, among others, described in detail in [5] under the replicator dynamics concept, developed under evolutionary game theory concept. In that analysis, a probability of choices is analysed and brought into relation with the payoffs for each strategy. When a probability growth rate of each strategy is observed, one can identify to which strategy would players evolve (in other words, they learn which strategy is more lucrative to be picked) in the long run, reaching stability.

Stability of the static points of the system of expected payoffs is tested using a Stability of nonlinear systems theorem [6]. This method uses eigenvalues of Jacobi matrices of probability growth rates and gives a precise result where the system tends to in the long run. The other similar analyses (like [3]) do not offer a full-scale stability check, which is, therefore, a contribution of this paper.

This paper first analyses cooperative and reactive strategies only, yielding a result where reactive strategies are more likely to be picked in the long run. After that, a dominant strategy is introduced, and a completely different result is obtained. Even though it might seem like a deficiency of the model, it is actually its power – it would show that when players realize a wider spectrum of options, their expected decisions change. These analyses then open a wide spectrum of combinations and different strategies that might arise in heads of decision makers, making a model very dynamic, but far reaching in its thinking. These findings would then confront two contradictory views of Cournot-Nash equilibrium since if an evolution brought companies to Cournot point, who ca be brave enough to call evolution naïve?

2 COMPANY STRATEGIES AND PAYOFF MATRICES

A linear demand is assumed, $p = a - bY = a - b(y_1 + y_2)$, where p is a price of a good, Y total quantity of production, y_1 and y_2 produced quantities of the 1st and the 2nd company respectively, *a* is a reservation price and *b* is a price reaction to a unit quantity increase, *a* and *b* being real number parameters. It is also assumed that marginal costs are constant and equal for both companies, *c*, where reservation price, *a*, is greater than marginal cost.

In the perfectly competitive equilibrium in the long run, companies have only normal profits where p = c and equilibrium quantity is $Y^C = \frac{a-c}{b}$. In Cournot duopoly each, company sets their production quantity by which it maximizes their profit taking the competitors quantity as given. Therefore, firstthe company's residual demand is $p = (a - by_2) - by_1$ and its profit, $\Pi_1 = py_1 - cy_1 = (a - c - by_2)y_1 - by_1^2$. From necessary first order conditions a first company's rea,ction function is deducted, $y_1 = \frac{1}{2}Y^C - \frac{1}{2}y_2$. By analogy, reaction function of the sthe econd company is $y_2 = \frac{1}{2}Y^C - \frac{1}{2}y_1$. Cournot or Nash equilibrium is obtained at the intercept of reaction functions, where each company supplies 1/3 of the perfectly competitive market is $y_1 = y_2 = \frac{1}{3}Y^C$, or Y= $\frac{2}{3}Y^C$. Then their profits are $\Pi_1 = \Pi_2 = \frac{b}{9}(Y^C)^2$.

In the Stackelberg model of quantitative leadership, a dominant company sets its production by which it maximizes its own profit, knowing the reaction curve of the reactionary company (follower) which behaves under Cournot assumption. Residual demand of a dominant company is $\boldsymbol{p = a - b\left(y_1 + \frac{1}{2}Y^C - \frac{1}{2}y_1\right) = \frac{a+c}{2} - \frac{b}{2}y_1}$ and its profit is $\boldsymbol{\Pi_1 = py_1 - cy_1 = \frac{a-c}{2}y_1 - \frac{b}{2}y_1^2}$. First order necessary conditions for dominant company profit maximization it follows that a dominant company supplies half of the perfectly competitive market, $\boldsymbol{y_1 = \frac{1}{2}Y^C}$. Stubbing it in a reaction curve of a reactionary company (follower) it is obtained that the follower supplies half-of-the-half, or a quarter of the market, $\boldsymbol{y_2 = \frac{1}{4}Y^C}$. Therefore, their profits are (or vice versa, if they switch leader – follower roles) $\boldsymbol{\Pi_1 = \frac{b}{8}(Y^C)^2}$; $\boldsymbol{\Pi_2 = \frac{b}{16}(Y^C)^2}$.

In cooperative equilibrium companies maximize total instead of individual profit, $\boldsymbol{\Pi_1 = pY - cY = (a - c)Y - bY^2}$. From necessary first order conditions it follo,ws that they jointly cover half of the perfectly competitive market, Y= $\boldsymbol{\frac{1}{2}Y^C}$, providing, in a symmetric cooperative equilibrium: $\boldsymbol{y_1 = y_2 = \frac{1}{4}Y^C}$. In that case their profits ,are $\boldsymbol{\Pi_1 = \Pi_2 = \frac{b}{8}(Y^C)^2}$.

Unanimous change in production in the cooperative equilibrium a company can, at least in the short run, increase their own profit at the expense of a large fall in price, followed by a large fall in profits of a cheated co-operator. In case when the first company remains cooperative, but the second becomes reactive, then its production is $\boldsymbol{y_2 = \frac{1}{2}(Y^C - y_1) = \frac{3}{8}Y^C}$. It rises overall production, thus lowering the price and the profit of the cooperative company, but rising the profit of a reactive company to $\boldsymbol{\Pi_1 = \frac{6b}{64}(Y^C)^2 \text{ and } \Pi_2 = \frac{9b}{64}(Y^C)^2}$.

In case when both companies behave dominantly, they both produce for ½ of the competitive market which lowers price to p = c causing normal profits for both companies. These data provide the following pay-off matrix:

		Company 2 cooperative	reactive	dominant
Company 1	cooperative	$\left(\frac{1}{8},\frac{1}{8}\right)$	$\left(\frac{3}{32},\frac{9}{64}\right)$	$\left(\frac{1}{16},\frac{1}{8}\right)$
	reactive	$\left(\frac{9}{64},\frac{3}{32}\right)$	$\left(\frac{1}{9},\frac{1}{9}\right)$	$\left(\frac{1}{16},\frac{1}{8}\right)$
	dominant	$\left(\frac{1}{8},\frac{1}{16}\right)$	$\left(\frac{1}{8},\frac{1}{16}\right)$	$(\mathbf{0},\mathbf{0})$

Figure 1: Pay-off matrix for cooperative, reactive and dominant strategies

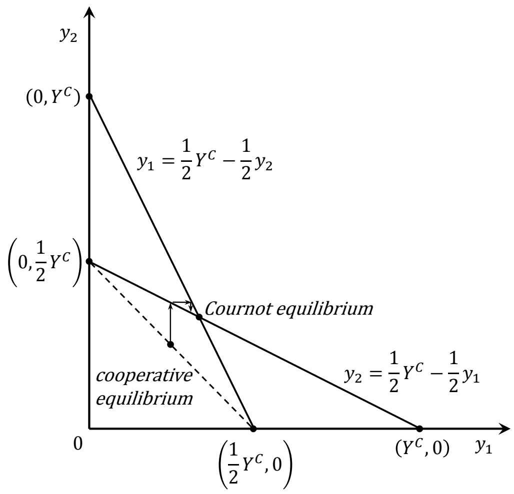


Figure 2: Cooperative and Cournot equilibria

Figure 1 reveals that in cooperative equilibrium both companies can improve their profits switching from cooperative to a reactive strategy $\left(\boldsymbol{\frac{9}{64} > \frac{1}{8}}\right)$, but in time, companies would adapt and shift to Cournot equilibrium (Figure 2). It can be deduced that cheating increases profit in the short run, but after sequential adjustments, both companies earn less in Cournot equilibrium.

3 REPLICATOR DYNAMICS

Under Cournot assumption, each company announces its own profit maximizing production level knowing their peers' production level. In Cournot equilibrium, no company has an incentive for a unanimous change in their production level, which is the basic property of the Nash equilibrium. In some literature, Cournot companies are described as naïve since they believe that competitors would stick to their own choices, but there is no such view of the Nash equilibrium, which makes literature incoherent. This paper will try to show, using an evolutionary or replicator dynamics, that behaviour of Cournot players is coherent with their long-term goals.

In order to support this proposition, two first rows and columns of the pay-off matrix (Figure 1) are taken. Let *P* be a probability of the cooperative strategy of a duopolistic company (like in [1]). Then the fitness of both companies and the market average are:

$$C = \frac{3}{32} + \frac{1}{32}P;\ R = \frac{1}{9} + \frac{17}{576}P\ ;\ M = \frac{P^2}{576} + \frac{7P}{576} + \frac{1}{9} \qquad (1)$$

Figure 3 shows the fitness of a cooperative and a reactive company.

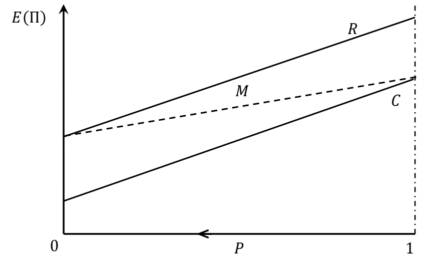


Figure 3: Fitness of a cooperative and a reactive company

Figure 3 shows that for each probability of choice of a cooperative strategy, the profit of the reactive company is larger than the profit of a cooperative company. The basic idea of the natural selection is that survivors are those with greater payoffs, while those with smaller payoffs adapt or die out. Therefore it means that companies would more frequently pick reactive as compared to the cooperative strategies until the latter disappears completely. As a result of evolutionary dynamics, companies would end up in Cournot equilibrium. This, in turn, means that cheating is an evolutionary response of economic subjects although it makes the worse off in the long run.

Even more interesting results could be obtained when the entire pay-off matrix (Figure 1) is analysed. Let P_1, P_2, and P_3 be probabilities of choice of cooperative, reactive, and dominant strategies responsively. Then the fitness of the cooperative, reactive, and dominant companies, as well as the total fitness, are as follows:

$$C = \frac{1}{8}P_1 + \frac{3}{32}P_2 + \frac{1}{16}P_3 = \frac{1}{8}P_1 + \frac{3}{32}P_2 + \frac{1}{16}(1 - P_1 - P_2) = \frac{1}{16}P_1 + \frac{1}{32}P_2 + \frac{1}{16}$$
$$R = \frac{9}{64}P_1 + \frac{1}{9}P_2 + \frac{1}{16}P_3 = \frac{9}{64}P_1 + \frac{1}{9}P_2 + \frac{1}{16}(1 - P_1 - P_2) = \frac{5}{64}P_1 + \frac{7}{144}P_2 + \frac{1}{16}$$
$$D = \frac{1}{8}P_1 + \frac{1}{8}P_2\ ;\ M = P_1C + P_2R + P_3D = \frac{3}{16}P_1 + \frac{3}{16}P_2 - \frac{1}{16}P_1^2 - \frac{11}{144}P_2^2 + \frac{9}{64}P_1P_2 \qquad (2)$$

Rule of a thumb is that those payoffs which are greater than the average payoff would be picked more frequently. According to replicator dynamics, a strategy probability growth rate is measured as a difference between the fitness of an observed payoff and total fitness (see [2] for similar analysis) :

$$\dot{P}_1(t) = P_1(t)(C - M) = P_1\left(-\frac{1}{8}P_1 - \frac{5}{32}P_2 + \frac{1}{16} + \frac{1}{16}P_1^2 + \frac{11}{144}P_2^2 + \frac{9}{64}P_1P_2\right)$$
$$\dot{P}_2(t) = P_2(t)(R - M) = P_2\left(-\frac{7}{64}P_1 - \frac{5}{32}P_2 + \frac{1}{16} + \frac{1}{16}P_1^2 + \frac{11}{144}P_2^2 + \frac{9}{64}P_1P_2\right)$$
$$\dot{P}_3(t) = P_3(t)(D - M) = P_3\left(-\frac{1}{16}P_1 - \frac{1}{16}P_2 + \frac{1}{16} + \frac{1}{16}P_1^2 + \frac{11}{144}P_2^2 + \frac{9}{64}P_1P_2\right) \qquad (3)$$

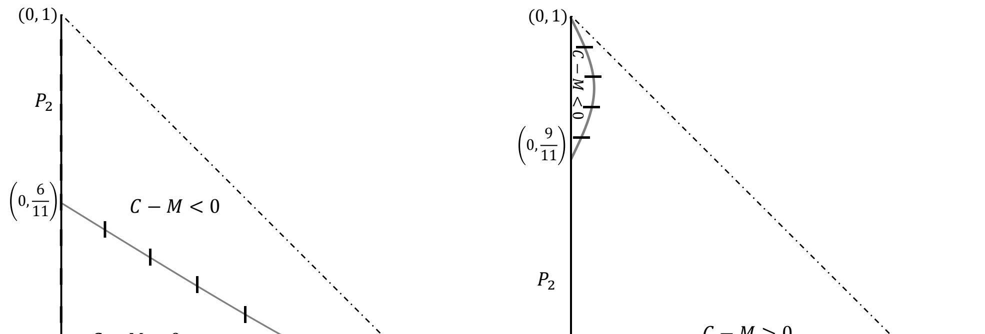


Figure 4: Isoclines for cooperative and reactive strategies when a third strategy is a dominant strategy
a) Isoclines when $\dot{P}_1 = 0$; b) Isoclines when $\dot{P}_2 = 0$

Since $\boldsymbol{P_1 + P_2 + P_3 = 1}$ the last row of (24) can be omitted. In this system of autonomous differential equations for each pair $(\boldsymbol{P_1, P_2})$, a vector $(\boldsymbol{\dot{P}_1, \dot{P}_2})$ indicated a direction in which probabilities change. Figure 4 shows isoclines when $\boldsymbol{\dot{P}_1 = 0}$ & $\boldsymbol{\dot{P}_2 = 0}$.

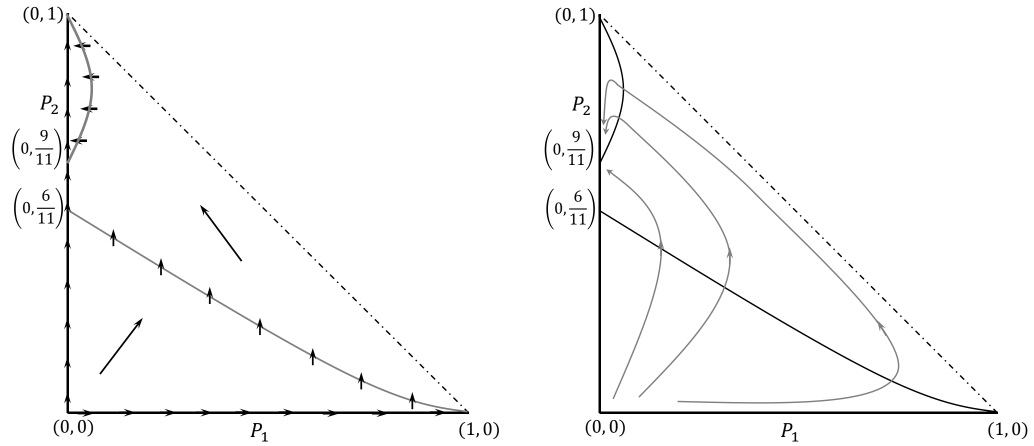


Figure 5: Vector field and probability orbits of cooperative and reactive strategies when the third strategy is dominant : a) Vector field; b) Orbits

Based on the analysis of a $\boldsymbol{\dot{P}_1}$ & $\boldsymbol{\dot{P}_2}$ sign in the areas which jointly bound the isoclines one can determine the direction of the change in the choice probability of cooperative and reactive strategies and their orbits (Figure 5).

4 STABILITY OF STATIONERY POINTS

This system has four stationery points: $(\boldsymbol{0,0}), (\boldsymbol{1,0}), (\boldsymbol{0,1})$ & $\left(\boldsymbol{0, \frac{9}{11}}\right)$. This paper introduces stability test, applying Stability of nonlinear systems theorem [6]. In order to make such a test, a calculation of Jacobian and eigenvalues is in order. Jacobi matrix is:

$$\boldsymbol{J(P_1, P_2)} = \begin{bmatrix} -\frac{1}{4}P_1 - \frac{5}{32}P_2 + \frac{1}{16} + \frac{3}{16}P_1^2 + \frac{11}{144}P_2^2 + \frac{9}{32}P_1P_2 & -\frac{5}{32}P_1 + \frac{11}{72}P_1P_2 + \frac{9}{64}P_1^2 \\ -\frac{7}{64}P_2 + \frac{1}{8}P_1P_2 + \frac{9}{64}P_2^2 & -\frac{7}{64}P_1 - \frac{5}{18}P_2 + \frac{1}{16} + \frac{1}{16}P_1^2 + \frac{11}{48}P_2^2 + \frac{9}{32}P_1P_2 \end{bmatrix} \quad (5)$$

Jacobian and eigenvalues for points $(\boldsymbol{0,0}), (\boldsymbol{1,0}), (\boldsymbol{0,1})$ & $\left(\boldsymbol{0, \frac{9}{11}}\right)$ are respectively:

$$\boldsymbol{J(P_1, P_2)} = \begin{bmatrix} \frac{1}{16} & 0 \\ 0 & \frac{1}{16} \end{bmatrix}; \lambda_1 = \lambda_2 = \frac{1}{16} \quad ; \boldsymbol{J(P_1, P_2)} = \begin{bmatrix} 0 & -\frac{1}{64} \\ 0 & -\frac{3}{64} \end{bmatrix}; \lambda_1 = 0; \lambda_2 = -\frac{3}{64} \quad (6)$$

$$\boldsymbol{J(P_1, P_2)} = \begin{bmatrix} -\frac{5}{288} & 0 \\ \frac{1}{32} & \frac{1}{72} \end{bmatrix}; \lambda_1 = \frac{1}{72}; \lambda_2 = -\frac{5}{288}; \boldsymbol{J}\left(0, \frac{9}{11}\right) = \begin{bmatrix} -\frac{5}{352} & 0 \\ \frac{9}{1936} & -\frac{1}{88} \end{bmatrix}; \lambda_1 = -\frac{1}{88}; \lambda_2 = -\frac{5}{352}$$

According to the above mentioned theorem, due to positive eigenvalues, point (0, 0) is unstable.

Since for point (1, 0) one of the eigenvalues is neither positive nor negative (0), one cannot determine stability at this point. Intuitively, it is stable if one does not move away from it. However, if one does move, there is no way to come back. For point (0, 1) the first eigenvalue is positive, hence stationery point is unstable. Finally, for point $\left(\boldsymbol{0, \frac{9}{11}}\right)$ eigenvalues are negative, therefore stationery point $\left(\boldsymbol{0, \frac{9}{11}}\right)$ is asymptotically stable.

Figure 5 orbits show the evolution of a company in a duopoly which, under certain probabilities, can pick cooperative, reactive, or dominant strategies. It can be shown that in

82% of the cases, companies choose reactive strategies, while in only 18% of the cases they behave dominantly.

5 CONCLUSION

Evolutionary game theory offers a different view of the long run. It shows in detail, using replicator dynamics, which strategies would be more frequently used after a period of adaptation (evolution) in which players learn how the system works. It is also very important due to the fact it merges crude mathematic apparatus and the uncertainty of human behaviour, which is what the entire economics is all about.

The system itself is framed by the set of strategies among which players can pick. As this set changes, the expected payoffs alter the probability of choices of certain strategies, while probability growth rates of these strategies show what players are inclined to do as the time goes by. Vector field of this probability growth rates direct system to a steady state, further analysed by stability test using eigenvalues, which strengthens the analysis of the replicator dynamics in evolutionary game theory. That steady state shows which strategies are more likely to be picked in the long run.

This paper has shown how a choice among several strategies brings different long-run states depending on the total set of strategies. At first, only two strategies are introduced – cooperative and reactive. Replicator dynamics analysis has shown that in 100% of the cases long run equilibrium would lead to an equilibrium where payoffs are lower (Cournot equilibrium) due to the fact that it is more lucrative to cheat in the short run.

After a third strategy is introduced, a dominant strategy, it was found out that in 18% of the cases players would choose dominant strategies, while in 82% of the cases their response would be a reactive strategy. This finding is very important – it shows that as the set of strategies gets wider, the probability of choice changes due to the fact that players in time recognize the power of the chosen strategy. It also shows that Cournot players, which are reactive, are not naïve, but evolutionary adaptive. Most of the literature shows Cournot player as a naïve one, even though they come to Nash equilibrium which is never referred to as a result of naïve players. This finding is crucial since it denounces naïve as a description of Cournot players since their choices are evolutionary choices, which nobody dares to call naïve.

This paper has only made a first step in the analyses of a much wider set of choices. Further analyses will introduce tit-for-tat strategy and increase the total number of strategies, showing how the system gets more complex, but more interesting when the ingenuity of players introduces new strategies they choose from. However, even though this paper analyses only two and three strategies, it does not render these analyses less realistic since in real life in many cases players short-list their choices to only two or three to make their decisions easier.

References

[1] Barron, E. N. 2013. Game Theory: an Introduction, 2nd Ed. Hoboken, NJ, USA: Wiley

[2] Kopal, R., Korkut, D. 2014. Uvod u teoriju igara. 2. Izdanje. Zagreb: Effectus

[3] Sánchez Carrera E. J., Pavlinović, S. 2013. Evolution of the Place Attachment: An Economic Approach. CRORR, Vol. 4

[4] Schaffer, M. E. 1989. Are profit Maximizers the Best Survivors?. Journal of Economic Behaviour and Organization 12, 29-45. North-Holland.

[5] Schecter, S., Gintis, H. 2016. An Introduction to Classical and Evolutionary Models. Princeton, NJ, USA: Princeton University Press

[6] Simon, C. P., Blume, L. 1994. Mathematics for Economists. New York, USA: W. W. Norton&Co.

The 15th International Symposium on
Operational Research in Slovenia

SOR '19

Bled, SLOVENIA
September 25 - 27, 2019

Appendix

Authors' addresses

Addresses of SOR'19 Authors

(The 15th International Symposium on OR in Slovenia, Bled, SLOVENIA, September 25 – 27, 2019)

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
1.	Kolos Csaba	Ágoston	Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest Institute of Economics, Research Centre for Economic and Regional Studies, Hungarian Academy of Sciences	Fővám tér 8. Tóth Kálmń u. 4.	1093 1097	Budapest	Hungary	kolos.agoston@ uni-corvinus.hu
2.	Ansari Saleh	Ahmar	Universitas Negeri Makassar	90222 Kampus UNM Gunungsari Makassar		Pare-Pare	Indonesia	ansarisaleh@unm.ac.id
3.	Wellington	Alves	Business School, Viana do Castelo Polytechnic Institute, and ALGORITMI Research Centre, University of Minho, Escola de Engenharia, Depto Produção e Sistemas, Campus de Azurém				Portugal	wellingtonb@ esce.ipvc.pt
4.	Agnès	Ansari	CNRS/IDRIS				France	agnes.ansari@idris.fr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
5.	Evin	Aslan Oğuz	Faculty of Electrical Engineering University of Ljubljana, Nielsen	Tržaška cesta 25 Obrtniška ulica 15	1000 6000	Ljubljana Koper	Slovenia	evin.aslanoguz@ nielsen.com
6.	Zoran	Babić	University of Split, Faculty of Economics	Cvite Fiskovića 5	21000	Split	Croatia	babic@efst.hr
7.	Tatiana	Baltesová	L. Pasteur University Hospital	Trieda SNP 1		Košice	Slovak Republic	tatiana.baltesova@ unlp.sk
8.	Dariusz	Banas	University of Economics	1 Maja 50	40-287	Katowice	Poland	
9.	Anton Florijan	Barišić	University for applied sciences VERN, Zagreb	Trg Drage Iblera 10	10000	Zagreb	Croatia	afbarisic@chronos.hr
10.	János	Baumgartner	University of Pannonia, Faculty of Information Technology, Department of Computer Science and Systems Technology	Egyetem str. 10.	8200	Veszprém	Hungary	baumgartner@ dcs.uni-pannon.hu
11.	Nina	Begičević Ređep	University of Zagreb, Faculty of organization and informatics	Pavlinska 2	42000	Varaždin	Croatia	nbegicev@foi.hr
12.	Tomaž	Berlec	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	tomaz.berlec@fs.uni-lj.si
13.	Mojca	Bernik	University of Maribor, Faculty of Organisational Sciences	Kidričeva 55	4000	Kranj	Slovenia	mojca.bernik@ fov.uni-mb.si
14.	David	Bogataj	University of Padova, Department of Management and Engineering	Stradella San Nicola 3	36100	Vicenza	Italy	david.bogataj@unipd.it

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
15.	Marija	Bogataj	University of Ljubljana, SEB & INRISK				Slovenia	marija.bogataj@ ef.uni-lj.si
16.	Drago	Bokal	University of Maribor, Faculty of Natural Sciences and Mathematics Institute of Mathematics, Physics, and Mechanics	Koroška cesta 160 Jadranska 19	2000 1000	Maribor Ljubljana	Slovenia	drago.bokal@um.si
17.	Immanuel	Bomze	University of Vienna, Department of Statistics and Operations Research (ISOR)	Oskar-Morgenstern-Platz 1	1090	Vienna	Austria	immanuel.bomze@ univie.ac.at
18.	Aua-aree	Boonperm	Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University Rangsit Center		12120	Pathumthani	Thailand	aua-aree@ mathstat.sci.tu.ac.th
19.	Ivan	Bratko	Slovenian Academy of Sciences and Arts University of Ljubljana, Faculty of Computer and Information Science	Novi trg 2 Večna pot 113	1000	Ljubljana	Slovenia	bratko@fri.uni-lj.si
20.	Andrej	Bregar	Informatika d.d.	Vetrinjska ulica 2	2000	Maribor	Slovenia	andrej.bregar@ informatika.si
21.	Alenka	Brezavšček	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	alenka.brezavscek@ um.si
22.	Helena	Brožová	Czech University of Life Sciences Prague, Fac. of Economics and Management, Dept. of Systems Engineering	Kamýcká 129	165 21	Praha 6 – Suchdol	Czech Republic	brozova@pef.czu.cz

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
23.	Sergio	Cabello	University of Ljubljana and IMFM		1000	Ljubljana	Slovenia	
24.	Francisco	Campuzano-Bolarín	Universidad Politécnica de Cartagena	Campus Muralla del Mar	30202	Cartagena (Murcia)	Spain	francisco.cmpuzano@upct.es
25.	Katarína	Cechlárová	Institute of Mathematics, Faculty of Science, P.J. Šafárik University	Jesenná 5	040 01	Košice	Slovak Republic	katarina.cechlarova@upjs.sk
26.	Michaela	Chocholatá	University of Economics in Bratislava, Department of Operations Research and Econometrics	Dolnozemská cesta 1	852 35	Bratislava	Slovak Republic	michaela.chocholata@euba.sk
27.	Éric	Colin de Verdière	Iniversité Paris-Est, LIGM, CNRS, ENPC, ESIEE Paris, UPEM			Marne-la-Vallé	France	
28.	Boris	Cota	Faculty of Economics and Business, University of Zagreb, Croatia, Department of Macroeconomics and Economic Development	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	bcota@net.efzg.hr
29.	Ivana	Cunjak Mataković	Centar revizija d.o.o,	V. Mačeka 28	47000	Karlovac	Croatia	ivana.cunjak@gmail.com
30.	Radoslaw	Cymer	Universität Augsburg				Germany	r.cymer@web.de
31.	Vesna	Čančer	University of Maribor, Faculty of Economics and Business	Razlagova 14	2000	Maribor	Slovenia	vesna.cancer@um.si
32.	Anita	Čeh Časni	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedy 6	10000	Zagreb	Croatia	aceh@efzg.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
33.	Mirjana	Čižmešija	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	mcizmesija@efzg.hr
34.	Zsolt	Darvay	Babes-Bolyai University, Faculty of Mathematics and Computer Science			Cluj-Napoca	Romania	
1.	Balázs	Dávid	InnoRenew CoE; University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies	Livade 6 Glagoljaška ulica 8	6310 6000	Izola Koper	Slovenia	balazs.david@ innorenew.eu balazs.david@ famnit.upr.si
2.	Marko	Debeljak	Jozef Stefan Institute, Department of Knowledge Technologies	Jamova cesta 39	1000	Ljubljana	Slovenia	marko.debeljak@ijs.si
3.	Mihael	Debevec	University of Ljubljana, Faculty of Mechanical Engineering, Chair of Manufacturing Technologies and Systems, Laboratory for Handling, Assembly and Pneumatics	Aškerčeva 6	1000	Ljubljana	Slovenia	mihael.debevec@ fs.uni-lj.si
4.	Antonio	Delgado	University of Sevilla, Departamento de Ciencias Agroforestales			Sevilla	Spain	adelgado@us.es
5.	Blaženka	Divjak	University of Zagreb, Faculty of organization and informatics	Pavlinska 2	42000	Varaždin	Croatia	bdivjak@foi.hr
6.	Gregor	Dolinar	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	gregor.dolinar@ bf.uni-lj.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
7.	Rok	Drnovšek	University Medical Centre Ljubljana	Zaloška cesta 2	1000	Ljubljana	Slovenia	rok.drnovsek@kclj.si
8.	Samo	Drobne	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova cesta 2	1000	Ljubljana	Slovenia	samo.drobne@ fgg.uni-lj.si
9.	Ksenija	Dumičić	University of Zagreb, Faculty of Economics and Business, Department of Statistics	Trg J.F. Kennedy 6	10000	Zagreb	Croatia	kdumicic@net.efzg.hr
10.	Marianna	E.-Nagy	Budapest University of Technology and Economics	Muegyetem rkpt. 3.	1111	Budapest	Hungary	enagym@math.bme.hu
11.	Ayşegül	Engin	Department of Business Decisions and Analytics, University of Vienna	Oskar Morgenstern Platz 1	1090	Viena	Austria	ayseguel.engin@ univie.ac.at
12.	Nataša	Erjavec	Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	nerjavec@efzg.hr
13.	Liljana	Ferbar Tratar	University of Ljubljana, Faculty of Economics	Kardeljeva pl 17	1000	Ljubljana	Slovenia	liljana.ferbar.tratar@ ef.uni-lj.si
14.	Aljaž	Ferencek	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	aljaz.ferencek@ student.um.si
15.	Petra	Fic	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	petra.ficc@gmail.com
16.	Marzena	Filipowicz-Chomko	Bialystok University of Technology	ul. Wiejska 45A	15-351	Białystok	Poland	m.filipowicz@pb.edu.pl

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
17.	Andrea	Furková	University of Economics in Bratislava, Department of Operations Research and Econometrics	Dolnozemská cesta 1	852 35	Bratislava	Slovak Republic	andrea.furkova@ euba.sk
18.	Boštjan	Gabrovšek	FS, University of Ljubljana; Institute of Mathematics, Physics and Mechanics; FMF, University of Ljubljana	Aškerčeva 6; Jadranska 19	1000	Ljubljana	Slovenia	bostjan.gabrovsek@ fs.uni-lj.si
19.	Elif	Garajová	Charles University, Faculty of Mathematics And Physics, Department of Applied Mathematics	Malostranské nám. 25	11800	Prague	Czech Republic	elif@kam.mff.cuni.cz
20.	Alberto	Garcia Fernandez	CNRS/IDRIS				France	alberto.garciafernandez @idris.fr
21.	Margareta	Gardijan Kedžo	University of Zagreb Faculty of Economics and Business, Department of Mathematics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	mgardijan@efzg.hr
22.	Alberto	Garre	Wageningen University & Research, Laboratory of Food Microbiology	P.O. Box 17	6700 AA	Wageningen	Netherlands	alberto.garreperez @wur.nl
23.	Helena	Gaspars-Wieloch	Poznan Universtity of Economics and Business, Department of Operations Research	Al. Niepodleglosci 10	61-875	Poznan	Poland	helena.gaspars@ ue.poznan.pl
24.	Blaž	Gašperlin	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	blaz.gasperlin@ student.um.si
25.	Petra	Grošelj	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	petra.groselj@bf.uni-lj.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
26.	Dobroslav	Grygar	University of Žilina, Department of Mathematical Methods and Operations Research	Univerzitna 8215/1	01026	Žilina	Slovak Republic	dobroslav.grygar@ fri.uniza.sk
27.	Nicolò	Gusmeroli	Alpen-Adria-Universität			Klagenfurt	Austria	
28.	Márton	Gyetvai	Institute of Economics, Research Centre for Economic and Regional Studies, Hungarian Academy of Sciences Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest	Tóth Kálmń u. 4. Fővám tér 8.	1093 1097	Budapest	Hungary	gyetvai.marton@ krtk.mta.hu
29.	László	Hajdu	University of Szeged, Institute of Informatics University of Primorska, Faculty of mathematics, Natural Sciences and information technologies InnoRenew CoE	Árpád tér 2 Muzejski trg. 2 Livade 6	6720 6000 6310	Szeged Koper Izola	Hungary Slovenia	laszlo.hajdu@ innorenew.eu
30.	Anita	Harmina	VERN' University of Applied Sciences	Trg Drage Iblera 10	10000	Zagreb	Croatia	anita.harmina@vern.hr
31.	Niko	Herakovič	University of Ljubljana, Faculty of Mechanical Engineering, Chair of Manufacturing Technologies and Systems, Laboratory for Handling, Assembly and Pneumatics	Aškerčeva 6	1000	Ljubljana	Slovenia	niko.herakovic@ fs.uni-lj.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
32.	Tomislav	Herceg	University of Zagreb, Faculty of Economics and Business	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	therceg@efzg.hr
33.	Milan	Hladík	Charles University, Faculty of Mathematics and Physics, Department of Applied Mathematics	Malostranské nám. 25	11800	Prague	Czech Republic	hladik@kam.mff.cuni.cz
34.	Vladimír	Holý	University of Economics, Prague	Winston Churchill Square 1938/4	130 67	Prague 3	Czech Republic	vladimir.holy@vse.cz
35.	Eloy	Hontoria	Technical University of Cartagena / Business Management		30202	Cartagena	Spain	eloy.hontoria@upct.es
36.	Jaroslav	Horáček	Charles University, Faculty of Mathematics And Physics, Department of Applied Mathematics	Malostranské nám. 25	11800	Prague	Czech Republic	horacek@ kam.mff.cuni.cz
37.	Timotej	Hrga	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	timotej.hrga@ lecad.fs.uni-lj.si
38.	Tomi	Ilijaš	Arctur, d.o.o.	Industrijska cesta 15	5000	Nova Gorica	Slovenia	tomi.ilijas@arctur.si
39.	Tibor	Illés	Budapest University of Technology and Economics, Institute of Mathematics	Egry József u. 1	1111	Budapest	Hungary	illes@math.bme.hu
40.	Mario	Jadrić	University of Split, Faculty of Economics, Business and Tourism, Department of Business Informatics	Cvite Fiskovića 5	21000	Split	Croatia	mario.jadric@efst.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
41.	Timotej	Jagrič	Faculty of Economics and Business (University of Maribor)	Razlagova 14	2000	Maribor	Slovenia	timotej.jagric@um.si
42.	Marko	Jakšič	School of Economics and Business, University of Ljubljana	Kardeljeva ploščad 17	1000	Ljubljana	Slovenia	marko.jaksic@ef.uni-lj.si
43.	Saša	Jakšić	Faculty of Economics and Business, University of Zagreb, Croatia, Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	sjaksic@efzg.hr
44.	Jaroslav	Janáček	University of Žilina, Faculty of Management and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	jaroslav.janacek@ fri.uniza.sk
45.	Peter	Jankovič	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	peter.jankovič@ fri.uniza.sk
46.	Ľudmila	Jánošíková	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	ludmila.janosikova@ fri.uniza.sk
47.	Sławomir	Jarek	University of Economics in Katowice, Department of Operations Research	ul. 1 Maja 50	40-287	Katowice	Poland	slawomir.jarek@ ue.katowice.pl
48.	Slobodan	Jelić	J. J. Strossmayer University of Osijek - Department of Mathematics	Trg Ljudevita Gaja 6	31000	Osijek	Croatia	sjelic@mathos.hr
49.	Eva	Jereb	University of Maribor, Faculty of Organisational Sciences	Kidričeva 55	4000	Kranj	Slovenia	eva.jereb@fov.uni-mb.si
50.	Joanna	Józefowska	Poznan University of Technology, Faculty of Computing	Piotrowo 3	60-965	Poznań	Poland	joanna.jozefowska@ cs.put.poznan.pl

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
51.	Dragan	Jukić	Department of Mathematics, J. J. Strossmayer University of Osijek	Trg Ljudevita Gaja 6	31000	Osijek	Croatia	jukicd@mathos.hr
52.	Elza	Jurun	University of Split, Faculty of Economics, Business and Tourism	Cvite Fiskovića 5	21000	Split	Croatia	elza@efst.hr
53.	Nikola	Kadoić	University of Zagreb, Faculty of organization and informatics	Pavlinska 2	42000	Varaždin	Croatia	nkadoic@foi.hr
54.	Michael	Kahr	University of Vienna, Department of Statistics and Operations Research (ISOR)	Oskar-Morgenstern-Platz 1	1090	Vienna	Austria	m.kahr@univie.ac.at
55.	Tina	Kep	University of Maribor, Faculty of agriculture and life science	Pivola 11	2311	Hoče	Slovenia	tina.kep@student.um.si
56.	Sandi	Klavžar	University of Ljubljana, Faculty of Mathematics and Physics University of Maribor, Faculty of Natural Sciences and Mathematics Institute of Mathematics, Physics and Mechanics			Ljubljana Maribor Ljubljana	Slovenia	sandi.klavzar@fmf.uni-lj.si
57.	Jernej	Klemenc	FS, University of Ljubljana	Aškerčeva 6	1000	Ljubljana	Slovenia	jernej.klemenc@fs.uni-lj.si
58.	Mirjana	Kljajić Borštnar	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	mirjana.kljajic@fov.uni-mb.si
59.	Davorin	Kofjač	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	davorin.kofjac@fov.uni-mb.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
60.	Michal	Koháni	University of Žilina, Department of Mathematical Methods and Operations Research	Univerzitna 8215/1	01026	Žilina	Slovak Republic	michal.kohani@ fri.uniza.sk
61.	Vedran	Kojić	University of Zagreb, Faculty of Economics & Business, Department of Mathematics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	vkojic@efzg.hr
62.	Miha	Konjar	University of Ljubljana / Faculty of Architecture	Zoisova cesta 12	1000	Ljubljana	Slovenia	miha.konjar@fa.uni-lj.si
63.	Andrej	Košir	Faculty of Electrical Engineering University of Ljubljana,	Tržaška cesta 25	1000	Ljubljana	Slovenia	andrej.kosir@fe.uni-lj.si
64.	Erzsébet	Kovács	Head, Department of Operational Research and Actuary Sciences Corvinus University of Budapest	Fővám tér 8	1093	Budapest	Hungary	erzsebet.kovacs@ uni-corvinus.hu
65.	László	Kovács	Department of Computer Science, Corvinus University of Budapest	Fővám tér 8.	1093	Budapest	Hungary	laszlo.kovacs2@ uni-corvinus.hu
66.	Blaženka	Knežević	University of Zagreb, Faculty Economics and Business, Department of Trade and International Business	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	bknezevic@efzg.hr
67.	Manja	Krajnčič	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	manja.krajncic@ student.um.si
68.	Aleš	Kresta	VŠB – Technical University of Ostrava, Department of Finance	Sokolská tř. 33	702 00	Ostrava	Czech Republic	ales.kresta@vsb.cz

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
69.	Miklós	Krész	InnoRenew CoE	Livade 6	6310	Izola	Slovenia	miklos.kresz@ innorenew.eu
70.	Janez	Kušar	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	janez.kusar@fs.uni-lj.si
71.	Marek	Kvet	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	marek.kvet@ fri.uniza.sk
72.	Markus	Leitner	University of Vienna, Department of Statistics and Operations Research (ISOR)	Oskar-Morgenstern-Platz 1	1090	Vienna	Austria	markus.leitner@ univie.ac.at
73.	Dean	Lipovac	InnoRenew CoE; University of Primorska, Andrej Marušič Institute	Livade 6; Muzejski trg. 2	6310; 6000	Izola; Koper	Slovenia	dean.lipovac@ innorenew.eu
74.	Shiang-Tai	Liu	Vanung University, Graduate School of Business and Management	No.1, Wanneng Rd, Zhongli District	320	Taoyuan	Taiwan	stliu@vnu.edu.tw
75.	Zrinka	Lukač	University of Zagreb, Faculty of Economics & Business, Department of Mathematics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	zlukac@efzg.hr
76.	Snježana	Majstorović	Department of Mathematics, Josip Juraj Strossmayer University of Osijek	Trg Ljudevita Gaja 6	31000	Osijek	Croatia	smajstor@mathos.hr
77.	Vili	Malnarič	TPV d.o.o.	Kandijska cesta 60	8000	Novo mesto	Slovenia	v.malnaric@tpv.si
78.	Fulgencio	Marín-García	Universidad Politécnica de Cartagena	Campus Muralla del Mar	30202	Cartagena (Murcia)	Spain	pentxo.marin@upct.es

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
79.	Aleš	Marjetič	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova cesta 2	1000	Ljubljana	Slovenia	ales.marjetic@ fgg.uni-lj.si
80.	Slavko	Matanović	High school of modern business Belgrade	Terazije 5	11000	Beograd	Serbia	slavko.matanovic@ mbs.edu.rs
81.	Josip	Matejaš	University of Zagreb, Faculty of Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	jmatejas@efzg.hr
82.	Metka	Mesojedec		Dolenje Mokro Polje 22	8310	Šentjernej	Slovenia	metka.mesojedec@ gmail.com
83.	Jerzy	Michnik	University of Economics in Katowice	ul. 1 Maja 50	40-287	Katowice	Poland	jerzy.michnik@ ue.katowice.pl
84.	Lorena	Mihelač	ŠC Novo mesto, IT and Music Department & International Postgraduate School Jožef Stefan		8000	Novo mesto	Slovenia	lorena.mihelac@ sc-nm.si
85.	Stanislav	Mikolajčík	University of Žilina, Faculty of Management Science and Informatics	Univerzitná 1	010 26	Žilina	Slovak Republic	Stanislav.Mikolajcik@ stud.uniza.sk
86.	Ljubica	Milanović Glavan	Faculty of Economics and Business, University of Zagreb, Department of Informatics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	ljmilanovic@efzg.hr
87.	Marija	Milavec Kapun	University of Ljubljana, Faculty of Health Sciences	Zdravstvena pot 5	1000	Ljubljana	Slovenia	marija.milavec@ zf.uni-lj.si
88.	Danijel	Mlinarić	University of Zagreb, Faculty Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	dmlinaric@efzg.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
89.	José Andrés	Moreno-Nicolás	Universidad Politécnica de Cartagena	Campus Muralla del Mar	30202	Cartagena (Murcia)	Spain	josea.moreno@upct.es
90.	Matic	Muc	TPV d.o.o.	Kandijska cesta 60	8000	Novo mesto	Slovenia	m.muc@tpv.si
91.	Tina	Novak	University of Ljubljana, Faculty of mechanical engineering	Aškerčeva cesta 6	1000	Ljubljana	Slovenia	tina.novak@fs.uni-lj.si
92.	Anders Qvale	Nyrud	NMBU, Faculty of Environmental Sciences and Natural Resource Management,	Universitetstunet 3	1432	Ås	Norway	anders.qvale.nyrud@nmbu.no
93.	Petra	Pavlíčková	Czech University of Life Sciences Prague, Fac. of Economics and Management, Dept. of Systems Engineering	Kamýcká 129	165 21	Praha 6 – Suchdol	Czech Republic	pavlickovap@pef.czu.cz
94.	Polona	Pavlovčič Prešeren	University of Ljubljana, Faculty of Civil and Geodetic Engineering	Jamova cesta 2	1000	Ljubljana	Slovenia	polona.pavlovcic-preseren@fgg.uni-lj.si
95.	Karmen	Pažek	University of Maribor, Faculty of agriculture and life science	Pivola 10	2311	Hoče	Slovenia	karmen.pazek@um.si
96.	Mirjana	Pejić-Bach	University of Zagreb, Faculty Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	mpejic@efzg.hr
97.	Aljoša	Peperko	FS, University of Ljubljana; Institute of Mathematics, Physics and Mechanics	Aškerčeva 6; Jadranska 19	1000	Ljubljana	Slovenia	aljosa.peperko@fs.uni-lj.si
98.	Tunjo	Perić	University of Zagreb, Faculty Economics and Business	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	tperic@efzg.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
99.	Miha	Pipan	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva cesta 6	1000	Ljubljana	Slovenia	miha.pipan@fs.uni-lj.si
100.	Snježana	Pivac	University of Split, Faculty of Economics, Business and Tourism, Department of Quantitative Methods	Cvite Fiskovića 5	21000	Split	Croatia	spivac@efst.hr
101.	Diana	Plačková	Institute of Mathematics, Faculty of Science, P.J. Šafárik University:	Jesenná 5	040 01	Košice	Slovak Republic	diana.plackova@ student.upjs.sk
102.	Janez	Povh	University of Ljubljana, Faculty of mechanical engineering; Institute of mathematics, physics and mechanics Ljubljana	Aškerčeva cesta 6; Jadranska ulica 19	1000	Ljubljana	Slovenia	janez.povh@ fs.uni-lj.si
103.	Matej	Praprotnik	Laboratory for Molecular Modeling, National Institute of Chemistry	Hajdrihova 19	1001	Ljubljana	Slovenia	praprot@cmm.ki.si
104.	Boris	Prevolšek	University of Maribor, Faculty of tourism	Cesta prvih borcev 36	8250	Brežice	Slovenia	boris.prevolsek@um.si
105.	Jernej	Prišenk	University of Maribor, Faculty of agriculture and life science	Pivola 10	2311	Hoče	Slovenia	jernej.prisenk@um.si
106.	Jernej	Protner	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva cesta 6	1000	Ljubljana	Slovenia	jernej.protner@fs.uni-lj.si
107.	Krunoslav	Puljić	University of Zagreb, Faculty of Economics & Business, Department of Mathematics	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	kpuljic@efzg.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
108.	Miroslav	Rada	University of Economics, Prague, Department of Financal Accounting and Auditing	Nám. W. Churchilla 4	13067	Prague	Czech Republic	miroslav.rada@vse.cz
109.	Uroš	Rajkovič	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55 a	4000	Kranj	Slovenia	uros.rajkovic@ fov.uni-mb.si
110.	Vladislav	Rajkovič	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55 a	4000	Kranj	Slovenia	vladislav.rajkovic@ fov.uni-mb.si
111.	Nada	Ratković	University of Split, Faculty of Economics, Business and Tourism	Cvite Fiskovića 5	21000	Split	Croatia	nada.ratkovic@efst.hr
112.	Robert	Repnik	Faculty of Natural Sciences and Mathematics, Physics Department, University of Maribor Association for Technical Culture of Slovenia	Koroška 160 Zaloška c. 65	2000 1000	Maribor Ljubljana	Slovenia	robert.repnik@um.si
113.	Bertrand	Rigaud	CNRS/CC-IN2P3				France	bertrand.rigaud@ cc.in2p3.fr
114.	Petra Renáta	Rigó	Budapest University of Technology and Economics, Department of Differential Equations Babes-Bolyai University, Faculty of Mathematics and Computer Science			Budapest Cluj-Napoca	Hungary Romania	
115.	Lidija	Rihar	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	lidija.rihar@fs.uni-lj.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
116.	Helena Sofia	Rodrigues	Business School, Viana do Castelo Polytechnic Institute, and Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro,				Portugal	sofiarodrigues@ esce.ipvc.pt
117.	Marco	Rorro	CINECA				Italy	m.rorro@cineca.it
118.	Ewa	Roszkowska	University of Białystok, Faculty of Economy and Management	ul. Warszawska 63	15-062	Białystok	Poland	e.roszkowska@ uwb.edu.pl
119.	Črtomir	Rozman	University of Maribor, Faculty of agriculture and life science	Pivola 10	2311	Hoče	Slovenia	crt.rozman@um.si
120.	Maja	Rožman	University of Maribor, Faculty of Economics and Business	Razlagova 14	2000	Maribor	Slovenia	maja.rozman1@um.si
121.	Darja	Rupnik Poklukar	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva ulica 6	1000	Ljubljana	Slovenia	darja.rupnik@fs.uni.lj.si
122.	Gregor	Rus	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55a	4000	Kranj	Slovenia	gregor.rus4@um.si
123.	Jan	Rydval	Czech University of Life Sciences Prague, Fac. of Economics and Management, Dept. of Systems Engineering	Kamýcká 129	165 21	Praha 6 – Suchdol	Czech Republic	rydval@pef.czu.cz
124.	Kristian	Sabo	Department of Mathematics, J. J. Strossmayer University of Osijek	Trg Ljudevita Gaja 6	31000	Osijek	Croatia	ksabo@mathos.hr
125.	Blaž	Sašek	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	blaz.sasek@ student.um.si

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
126.	Jaap	Schröder	Wageningen University and Reasearch, Plant Science Group			Wageningen	Netherlands	jaap.schroder@wur.nl
127.	Rogier	Schulte	Wageningen University and Reasearch, Department of Plant Sciences			Wageningen	Netherlands	rogier.schulte@wur.nl
128.	Ângela	Silva	Business School, Viana do Castelo Polytechnic Institute and Centro de Investigação em Organizações,Mercados e Gestão Industrial (COMEGI), Lusíada University				Portugal	angela.a@esce.ipvc.pt
129.	Wutiphol	Sintunavarat	Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University Rangsit Center		12120	Pathumthani	Thailand	wutiphol@ mathstat.sci.tu.ac.th
130.	Andreja	Smole	Cosylab, Control System Laboratory		1000	Ljubljana	Slovenia	andreja.smole@ cosylab.com
131.	Sølvi Therese	Strømmen Wie	NMBU, Faculty of Environmental Sciences and Natural Resource Management	Universitetstunet 3	1432	Ås	Norway	solvi.wie@nmbu.no
132.	Zoltán	Süle	University of Pannonia, Faculty of Information Technology, Department of Computer Science and Systems Technology	Egyetem str. 10	8200	Veszprém	Hungary	sule@dcs.uni-pannon.hu
133.	Eszter	Szénási	Budapest University of Technology and Economics, Department of Differential Equations				Hungary	

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
134.	Boško	Šego	University of Zagreb, Department of Mathematics and Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	bsego@net.efzg.hr
135.	Marijana	Šemanović	University of Split, Faculty of Economics, Business and Tourism, Postgraduate Specialist Study	Cvite Fiskovića 5	21000	Split	Croatia	marijana.semanovic@ gmail.com
136.	Andrej	Škraba	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	andrej.skraba@um.si
137.	Tihana	Škrinjarić	University of Zagreb, Department of Mathematics and Department of Statistics	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	tskrinjar@net.efzg.hr
138.	Ana	Škrlec	J&T Banka d.d.	Aleja kralja Zvonimira 1	42000	Varaždin	Croatia	ana.skrlec@jtbanka.hr
139.	Petra	Škrobot	University of Zagreb, Faculty Economics and Business, Department of Trade and International Business	Trg J.F. Kennedya 6	10000	Zagreb	Croatia	pskrobot1@efzg.hr
140.	Tomáš	Šubrt	Czech University of Life Sciences Prague, Fac. of Economics and Management, Dept. of Systems Engineering	Kamýcká 129	165 21	Praha 6 – Suchdol	Czech Republic	subrt@pef.czu.cz
141.	Špela	Tertinek	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	spela.tertinek@ student.um.si
142.	Jaka	Toman	University of Maribor, Faculty of Organizational Sciences	Kidričeva cesta 55 a	4000	Kranj	Slovenia	jaka.toman@gmail.com

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
143.	Petra	Tomanová	University of Economics, Prague, Department of Econometrics	W. Churchill Sq. 1938/4	130 67	Prague 3	Czech Republic	petra.tomanova@vse.cz
144.	Aneta	Trajanov	Jozef Stefan Institute, Department of Knowledge Technologies	Jamova cesta 39	1000	Ljubljana	Slovenia	aneta.trajanov@ijs.si
145.	Tadeusz	Trzaskalik	Department of Operations Research, University of Economics in Katowice	ul. 1 Maja 50	40-287	Katowice	Poland	tadeusz.trzaskalik@ ue.katowice.pl
146.	Anita	Varga	Budapest University of Technology and Economics	Muegyetem rkpt. 3.	1111	Budapest	Hungary	vanita@math.bme.hu
147.	Ágnes	Vaskövi	Department of Finance, Corvinus University of Budapest	Fővám tér 8.	1093	Budapest	Hungary	agnes.vaskovi@ uni-corvinus.hu
148.	Alen	Vegi Kalamar	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	alen.vegi.kalamar1994@ gmail.com
149.	Rudolf	Vetschera	Department of Business Decisions and Analytics, University of Vienna	Oskar Morgenstern Platz 1	1090	Viena	Austria	rudolf.vetschera@ univie.ac.at
150.	Jožef	Vinčec	University of Maribor, Faculty of agriculture and life science	Pivola 10	2311	Hoče	Slovenia	jozef.vincec@ student.um.si
151.	Josipa	Višić	Faculty of Economics, Business and Tourism/Department of Economics	Cvite Fiskovića 5	21 000	Split	Croatia	josipa.visic@efst.hr
152.	Petr	Volf	Institute of Information Theory and Automation, AS CR	Pod vodarenskou vezi 4		Prague 8	Czech Republic	volf@utia.cas.cz

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
153.	Ilko	Vrankić	University of Zagreb, Faculty of Economics and Business	Trg J. F. Kennedyja 6	10000	Zagreb	Croatia	ivrankic@efzg.hr
154.	Andreas	Vroutsis	EPCC				UK	A.Vroutsis@ epcc.ed.ac.uk
155.	Marija	Vuković	University of Split, Faculty of Economics, Business and Tourism, Department of Quantitative Methods	Cvite Fiskovića 5	21000	Split	Croatia	marija.vukovic@efst.hr
156.	Bože	Vuleta	Franciscan Institute for the Culture of Peace,	Poljudsko šetalište 24	21000	Split	Croatia	fbvuleta@gmail.com
157.	David	Wall	Teagasc - Crops, Environment and Land Use Programme, Johnstown Castle			Wexford	Ireland	david.wall@teagasc.ie
158.	Anlan	Wang	VŠB – Technical University of Ostrava, Department of Finance	Sokolská tř. 33	702 00	Ostrava	Czech Republic	anlan.wang.st@vsb.cz
159.	Tomasz	Wachowicz	University of Economics in Katowice, Department of Operations Research	ul. 1 Maja 50	40-287	Katowice	Poland	tomasz.wachowicz@ uekat.pl
160.	Angelika	Wiegele	Alpen-Adria-Universität			Klagenfurt	Austria	
161.	Lidija	Zadnik Stirn	University of Ljubljana, Biotechnical Faculty	Jamnikarjeva 101	1000	Ljubljana	Slovenia	lidija.zadnik@bf.uni-lj.si
162.	Srečko	Zakrajšek	IAM, College for Multimedia	Leskoškova 12	1000	Ljubljana	Slovenia	sreco.zakrajsek@iam.si
163.	Jovana	Zoroja	University of Zagreb, Faculty Economics and Business, Department of Informatics	Trg J.F. Kennedyja 6	10000	Zagreb	Croatia	jzoroja@efzg.hr

	First name	Surname	Institution	Street and Number	Post code	Town	Country	E-mail
164.	Tadej	Žerak	University of Maribor, Faculty of Natural Sciences and Mathematics	Koroška cesta 160	2000	Maribor	Slovenia	tadej.zerak@ student.um.si
165.	Janez	Žerovnik	Faculty of Mechanical Engineering Institute of Mathematics, Physics and Mechanics	Aškerčeva 6 Jadranska 19	1000	Ljubljana	Slovenia	janez.zerovnik@ fs.uni-lj.si
166.	Maja	Žibert	University of Maribor, Faculty of agriculture and life science	Pivola 10	2311	Hoče	Slovenia	maja.zibert@ student.um.si
167.	Berislav	Žmuk	University of Zagreb, Faculty of Economics and Business	Trg J.F.Kennedyja 6	10000	Zagreb	Croatia	bzmuk@efzg.hr
168.	Anja	Žnidaršič	University of Maribor, Faculty of Organizational Sciences	Kidričeva 55a	4000	Kranj	Slovenia	anja.znidarsic@ fov.uni-mb.si
169.	Tena	Žužek	University of Ljubljana, Faculty of Mechanical Engineering	Aškerčeva 6	1000	Ljubljana	Slovenia	tena.zuzek@fs.uni-lj.si

The 15th International Symposium on Operational Research in Slovenia - SOR '19
Bled, SLOVENIA, September 25 - 27, 2019

Slovenian Society INFORMATIKA
Section for Operational Research
- organizer –

University of Maribor
Faculty of Organizational Sciences
- co-organizer –

University of Ljubljana
Faculty of Mechanical Engineering
- co-organizer –

Partnership for
Advance Computing in Europe
- sponsor –

Association of European Operational Research Societies
- co-sponsor -